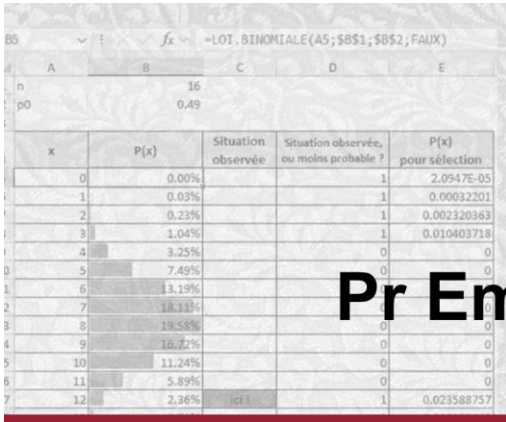
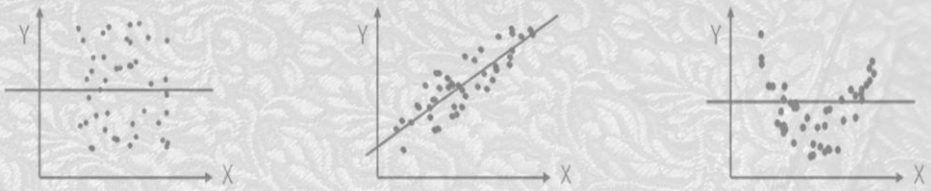




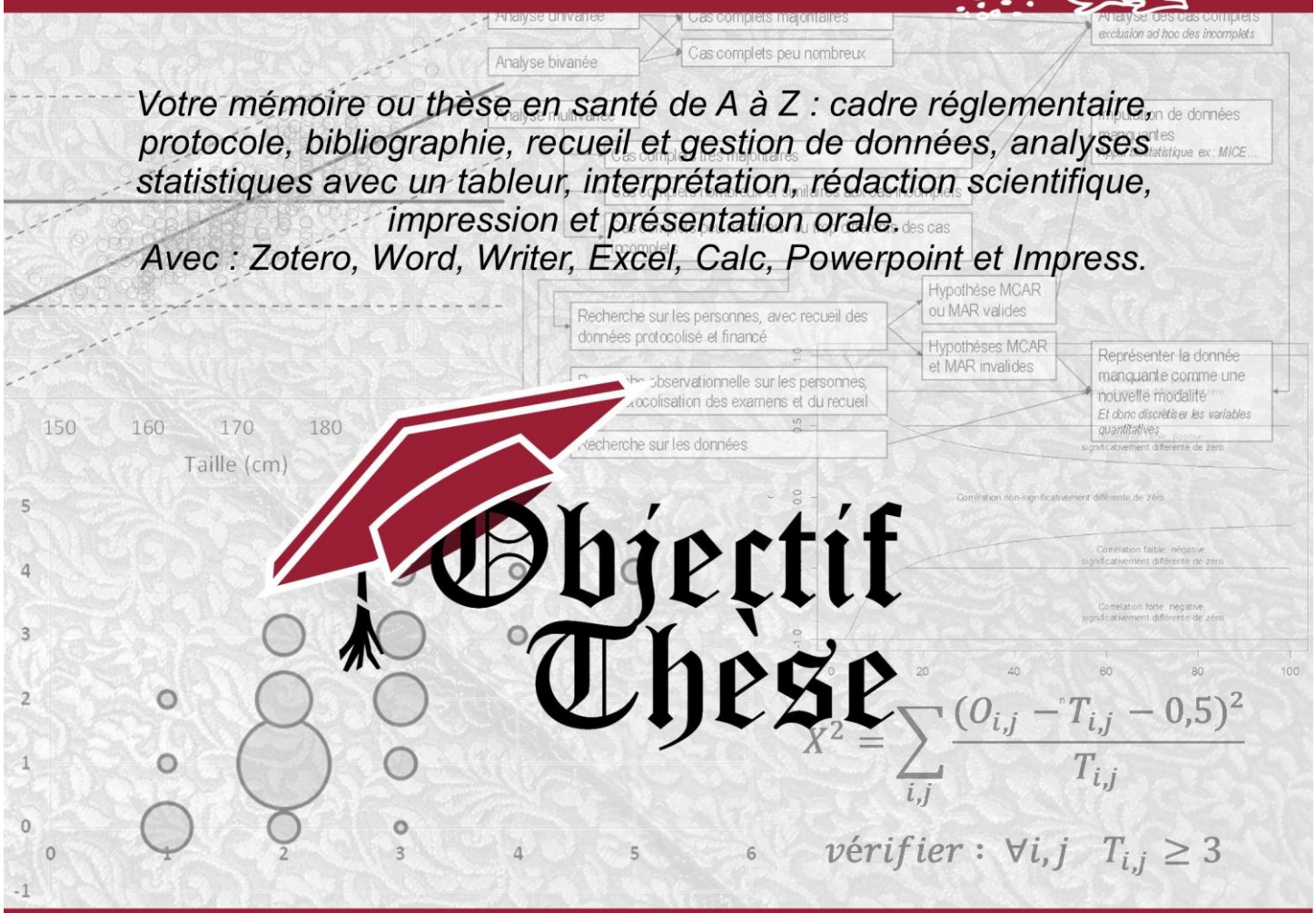
Pr Emmanuel Chazard

Objectif Thèse niveau 2 :

Poulet consciencieux

Votre mémoire ou thèse en santé de A à Z : cadre réglementaire, protocole, bibliographie, recueil et gestion de données, analyses statistiques avec un tableur, interprétation, rédaction scientifique, impression et présentation orale.

Avec : Zotero, Word, Writer, Excel, Calc, Powerpoint et Impress.



Objectif Thèse

http://editions.chazard.org

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Objectif thèse niveau 2 : « Poulet consciencieux »

Vous devez mener une étude quantitative en santé sans logiciel de statistique ? Nous suivrons ensemble toutes les étapes pour mener un mémoire académique (M1, M2, thèse d'exercice, thèse d'université) : cadre réglementaire, protocole, bibliographie, recueil et gestion de données, analyses statistiques univariées et bivariées avec un tableur, interprétation des résultats, rédaction scientifique, impression et présentation orale. Nous aborderons les logiciels suivants : Zotero, Word, Writer, Excel, Calc, Powerpoint et Impress.

Pr Emmanuel Chazard

http://editions.chazard.org

Texte, illustrations et couverture : Emmanuel Chazard

Relecture : Clémence Duriez

Dessin en couverture : Frédérique Chazard

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Emmanuel Chazard est Professeur de Médecine à l'Université de Lille et au CHU de Lille. Il est titulaire de trois masters, une thèse de Médecine en Santé Publique, une thèse d'Université en biostatistique et informatique médicale, une Habilitation à Diriger des Recherches, et tout le tralala. Il occupe de nombreuses responsabilités (directions d'équipes, Conseil de l'Ordre des Médecins, collège d'enseignants CIMES, congrès EMOIS, etc.). Il a encadré une centaine de mémoires académiques. Il adore ses étudiants et souhaite soulager leur angoisse relative à la thèse ou aux mémoires de master. Il se décrit lui-même comme « un mec cool », ce qui est sa seule blague drôle. Il parle de lui à la troisième personne, et est vraisemblablement l'auteur de ces quelques lignes. Il n'a pas d'humour.

Trèves de plaisanteries, ce cours (0% IA, 100% expertise et expérience) fait directement suite au programme Objectif Thèse, qui accompagne des internes de médecine vers la préparation de leur thèse d'exercice depuis de nombreuses années. Les cours initialement diffusés ont été améliorés et colligés sous forme d'ouvrage, destiné à tous les étudiants de filières de santé, de deuxième et troisième cycle. Cet ouvrage est édité par http://editions.chazard.org. Il existe en une version PDF gratuite et une version papier qui peut être commandée en ligne.

© 2025 Emmanuel Chazard

Tous droits de traduction, d'adaptation et de reproduction par tous procédés réservés pour tout pays. En application de la loi du 1er juillet 1992, il est interdit de reproduire, même partiellement, la présente publication sans l'autorisation de l'auteur.

Dépôt légal auprès de la BnF avril 2025, sous le numéro 10000001128414.

ISBN 978-2-9579934-1-3

9 782957 993413

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Sommaire

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Sigles, acronymes et abréviations

AOMI	Artériopathie Oblitérante des Membres Inférieurs
API	*Application Programming Interface*
ARC	Attaché de Recherche Clinique
ATC	*Anatomic Therapeutic and Chemical classification*
ATIH	Agence Technique de l'Information sur l'Hospitalisation
AUC	*Area under the Curve*
BDD	Base De Données
BnF	Bibliothèque nationale de France
CCAM	Classification Commune des Actes Médicaux
CIl	Correspondant Informatique et Liberté
Cim10	Classification Internationale des Maladies, 10ème version
CNAMTS	Caisse Nationale d'Assurance Maladie des Travailleurs Salariés
Cnil	Commission Nationale de l'Informatique et des Libertés
Consort	*Consolidated Standards of Reporting Trials*
CP	Capacité Prédictive
CPP	Comité de Protection des Personnes
CSP	Catégorie Socio-Professionnelle
DFG	Débit de Filtration Glomérulaire
DPD	Délégué à la Protection des Données (synonyme de DPO)
DPO	*Data Protection Officer* (synonyme de DPD)
DS	Déviation Standard
e-CRF	*Electronic clinical research form*
EMA	*European Medicines Agency*
EVA	Echelle Visuelle Analogique
FDA	*Food and Drug Administration* (USA)
GS	*Gold Standard*
HIPAA	*Health Insurance Portability and Accountability Act* (loi USA 1996)
HTML	*HyperText Markup Language*
IA	Intelligence Artificielle
IC95	Intervalle de Confiance à 95%
IF	*Impact Factor*
IMMRaD	*Introduction Material Methods Results and Discussion*
Insee	Institut National de la Statistique et des Etudes Economiques
IRPP	Impôt sur le Revenu des Personnes Physiques
JCR	Journal Citation Report
LPP	Liste des Produits et Prestations
MAR	*Missing At Random*
MCAR	*Missing Completely At Random*
Mice	*Multiple Imputation by Chained Equations*
MNAR	*Missing Not At Random*
MR*xxx*	Ixième méthodologie de référence de la CNIL (ex : MR005)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Nipals	*Nonlinear Iterate PArtial Least Squares*
NLM	National Library of Medicine
NSN	Nombre de sujets nécessaires
OMS	Organisation Mondiale de la Santé
OR	*Odds ratio* (rapport des cotes)
PMSI	Programme de Médicalisation des Systèmes d'Information
Prisma	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
RCT	*Randomized Controlled Trial*
RGPD	Règlement Général de Protection des Données
RIPH	Recherche Impliquant la Personne Humaine
RNIPH	Recherche N'Impliquant pas la Personne Humaine (sur les données)
ROC	*Receiver Operating Characteristic*
RR	Risque Relatif
Sampra	*Software for Analysis and Management of Publications & Research Assessment*
SD	*Standard deviation* (déviation standard)
Se	Sensibilité
SIDA	Syndrome d'ImmunoDéficience Acquise
Sigaps	Système d'Interrogation, de Gestion, d'Analyse des publications Scientifiques
Sigrec	Système d'Information et de Gestion de la Recherche et des Essais Cliniques
SNDS	Système National des Données de Santé
Sniiram	Système National d'Identification Inter-Régimes de l'Assurance Maladie
Sp	Spécificité
Strobe	*STrengthening the Reporting of OBservational Studies*
VIH	Virus de l'Immunodéficience Humaine
VPN	Valeur Prédictive Négative
VPP	Valeur Prédictive Positive
WoS	Web of Science

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Préambule

Ce message s'adresse à toi, le Jeune. Si tu veux dominer le monde et asservir l'humanité, tu dois d'abord rédiger et soutenir ton mémoire académique. C'est là que j'interviens : lis bien mon cours, ensuite seulement tu accompliras ton destin !

Le présent ouvrage s'adresse aux étudiants désireux de comprendre et réaliser des études quantitatives, notamment pour mener à bien leur projet de mémoire académique en santé (master, mémoire de fin d'étude, thèse d'exercice, thèse d'université).

Etudiants en fin d'études en santé, en Master santé, en Thèse d'exercice ou en Thèse d'université en santé : cet ouvrage correspond exactement à ce dont vous avez besoin pour réaliser votre mémoire académique, de la conception de l'étude à l'écriture des résultats d'analyse statistique. Pour ce qui est des analyses statistiques, nous verrons ici comment les réaliser avec un tableur (Excel ou Calc), et proposerons des conduites à tenir opérationnelles adaptées aux étudiants qui ne souhaitent pas utiliser un logiciel de programmation en statistique.

Internes en santé publique, en Master Statistique ou Thèse d'université orientée statistique : cet ouvrage vous sera utile pour tout ce qui n'est pas lié aux analyses statistiques. Il ne sera pas suffisant s'il est attendu de vous que vous preniez en main un logiciel de statistique, et que vous réalisiez vous-même des analyses multivariées.

N'hésitez pas à visiter le site http://editions.chazard.org : vous y trouverez les différentes versions de cet ouvrage (**PDF gratuit**, version **papier broché**, version **eBook**). Si vous trouvez ce cours trop simple ou trop long, vous trouverez sans doute également d'autres ouvrages plus adaptés à votre projet.

Si vous appréciez cet ouvrage, n'hésitez pas à en parler autour de vous et à le diffuser !

A Clémence, Frédérique et Louis

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

	Objectif Thèse niveau 1 : *Poussin pressé*	Objectif Thèse niveau 2 : *Poulet consciencieux*	Objectif Thèse niveau 3 : *Coq méthodique*
Conception, formalités, bibliographie	☑ abrégé	☑ détaillé	☑ détaillé
Recueil, correction et transformation de données	☑ abrégé, avec un tableur	☑ détaillé, avec un tableur	☑ avancé, avec R
Analyse statistique univariée et bivariée	☑ abrégée, avec un tableur	☑ détaillée, avec un tableur	☑ détaillée, avec R
Analyse statistique multivariée, rapport automatisé	-	-	☑ détaillée, avec R
Rédaction, traitement de texte, bibliographie, impression, diaporama	☑ abrégée	☑ détaillée	☑ détaillée

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Concevoir l'étude

La conception de l'étude passe par plusieurs étapes que nous évoquerons successivement.

Il vous faudra d'abord mener une **recherche bibliographique**, qui est un passage obligé avant même de concevoir l'étude à proprement parler (voir chapitre 1 Mener une recherche bibliographique en page 12)

Il vous faudra ensuite identifier le **cadre réglementaire de l'étude**, et ce, très précocement : il est interdit de mener certaines recherches sans autorisations, et les délais d'obtention de ces autorisations peuvent disqualifier d'emblée certains types d'études, selon le calendrier de votre propre mémoire académique (voir chapitre 2 Déterminer le type réglementaire d'étude en page 30 puis chapitre 3 Identifier les autorisations en page 33)

Ensuite, vous devrez déterminer si votre recherche est **qualitative ou quantitative** (si elle est purement qualitative, la plupart des chapitres de cet ouvrage vous seront inutiles 😉) (voir chapitre 4 Enquête quantitative ou qualitative ? en page 39 puis chapitre 5 Déroulement d'une étude quantitative en page 41).

Ensuite, il vous faudra déterminer le **design méthodologique** de votre étude (voir chapitre 6 Designs les plus fréquents en étude quantitative en page 42). Si votre étude est un questionnaire, il vous faudra comprendre les notions d'**échantillonnage** (voir chapitre 7 Questionnaires : taux de sondage, taux de réponse en page 51).

Enfin, la dernière étape de la conception vous amènera peut-être à calculer le **nombre de sujets nécessaire**, ou à savoir expliquer pourquoi vous ne réaliserez pas ce calcul (voir chapitre 8 Calculer le nombre de sujets nécessaires en page 54).

Vous pourrez également trouver de l'aide sur l'excellent blog du Dr Michaël Rochoy, qui a préparé une version reliée de ses billets : https://www.mimiryudo.com/blog/

1 Mener une recherche bibliographique

1.1 Définitions, environnement de publication

1.1.1 Littérature blanche ou grise

On définit traditionnellement comme relevant de la « **littérature blanche** », les documents publiés par des éditeurs dont c'est l'activité principale, et soumis à une certaine traçabilité. Cela inclut :

- Les articles de journaux périodiques scientifiques (ex : JAMA, BMJ, etc.)
- Les articles de journaux périodiques non-scientifiques (ex : le Figaro, le Monde, etc.)
- Les livres édités, que leur contenu soit académique, polémique, littéraire, etc.

Il existe d'autres documents qui n'ont pas été publiés par ce procédé, et n'ont pas toujours la même traçabilité : ils relèvent alors de la « **littérature grise** ». Ceci inclut notamment :

- les rapports techniques, les documents issus des autorités nationales ou internationales (ex : HAS, FDA)
- les mémoires académiques (masters, thèses d'exercice, thèses d'université, etc.)
- les sites web, les posts sur les réseaux sociaux
- les prospectus, tracts, annonces, brochures, catalogues
- etc.

Parmi tous ces documents, ceux qui ont une distribution suffisamment large sont soumis, depuis François Premier, à une obligation d'enregistrement à la Bibliothèque nationale de France, la BnF[1]. A travers cette dichotomie blanche/grise, les éditeurs essaient de vous faire croire qu'eux seuls sont garants de la qualité des contenus publiés. Bien évidemment il n'en est rien, comme l'illustre l'important nombre de livres ou journaux de désinformation publiés par des éditeurs.

1.1.2 Littérature fiable ou non

Cette dichotomie blanche/grise n'a aucun intérêt au regard de la recherche scientifique, car elle n'apporte pas de garantie de fiabilité.

On pourra plutôt distinguer trois grandes catégories de documents.

La première contient des **sources de qualité**, que vous pourrez citer directement (tout en restant critique) :

- **les articles scientifiques revus par les pairs** [1] publiés dans des journaux scientifiques (revues scientifiques). Nous reparlerons de ces articles scientifiques un peu plus bas. Pour ceux qui nous intéresse, leur qualité est garantie par le journal qui les publie, et cette qualité repose sur un processus contradictoire de publication.
- **les rapports scientifiques** publiés par certaines **agences étatiques** considérées comme fiables, comme la HAS, l'EMA, la FDA, etc. Ces rapports citent généralement des articles scientifiques. Même s'ils ne sont pas soumis au même processus de sélection que les articles, d'autres éléments de leur processus d'édification les rendent généralement fiables (conférence de consensus, comité d'experts, rédaction collégiale, etc.).
- **certains ouvrages académiques (livres)** publiés par des **auteurs reconnus**. Ces ouvrages ne doivent être considérés ni comme infaillibles (du fait de l'absence de contrôle par un tiers), ni comme nécessairement à jour. Cependant, ils peuvent être utilisés pour documenter des connaissances largement partagées et consensuelles (ex : méthode statistique ancienne, anatomie, éléments bien établis de physio-pathologie, etc.). Ils ne devraient pas être considérés comme aussi fiables pour des informations plus récentes ou discutées.

La deuxième contient des **sources de qualité intermédiaire**, qui énoncent des connaissances et citent des sources de meilleure qualité, présentées précédemment. Ces sources ne devraient pas être citées. Elles peuvent cependant être utilisées pour **retrouver les articles originaux**, qu'il faudra toujours relire avant de les citer. Ces sources de qualité intermédiaire sont :

- des **mémoires académiques** (masters, thèses, habilitations à diriger des recherches)
- des **rapports techniques** adressés à des autorités publiques
- les pages de **Wikipedia**
- tout contenu produit par une **société savante** ou un collège d'enseignants
- etc.

La dernière catégorie contient des sources autres, qui devront être considérées comme **non-fiables**, jusqu'à preuve du contraire (elles pourront être citées pour faire référence à l'actualité ou aux opinions infondées par exemple) :

- des pages didactiques produites par des sociétés commerciales (ex : tutoriels de société réalisant des statistiques, éditeurs de contenus grand public, etc.)

[1] Certains utilisent l'anglicisme « papier » (scientific peer-reviewed paper) pour désigner ces articles. Le terme « scientifique » fait référence aux méthodologies de mise à jour de connaissances, et non au champ disciplinaire tel qu'il est entendu au collège ou au lycée. On peut faire de la recherche scientifique en littérature, si les méthodes employées sont scientifiques. Inversement, un ingénieur utilise des outils issus de disciplines scientifiques, mais n'est pas à proprement parler un scientifique s'il ne participe pas à des travaux de recherche.

- des sites spécialisés destinés aux patients, femmes enceintes et parents, non-édités par des sociétés savantes (ex : éditeurs de contenus, laboratoires pharmaceutiques, associations de patients, forums vivant de la publicité, etc.)
- la presse scientifique grand public
- la presse non-scientifique (même sans mauvaise intention, il est fréquent que les journalistes comprennent mal les informations scientifiques qu'ils relayent)

1.1.3 Les articles scientifiques revus par les pairs

Au premier rang des sources considérées comme fiables (mais toujours à lire avec un sens critique) se trouvent les articles scientifiques revus par les pairs. Un article scientifique revu par les pairs est référencé dans une **base de données bibliographique comme Pubmed** (http://pubmed.gov)[2]. Il l'est parce que l'ensemble du journal dans lequel il a été publié l'est. Le journal lui-même est référencé parce qu'il s'est engagé sur plusieurs points :

- un processus éditorial formalisé, fondé sur la **peer review**, ou revue par les pairs (cf. plus bas)
- une garantie de qualité scientifique des articles publiés, l'engagement à permettre et tracer le débat sur les articles publiés et, si nécessaire, la rétractation des articles discrédités
- une garantie déontologique et d'indépendance des éditeurs, des reviewers et des auteurs

1.1.4 Processus de revue par les pairs, ou *peer review*

Les journaux scientifiques disposent d'un comité éditorial qui reçoit les propositions d'articles. Il est généralement composé de scientifiques académiques, qui sont indemnisés ou non par l'éditeur.

Un premier filtre rapide est réalisé, en fonction de la qualité évidente de l'article mais surtout du champ disciplinaire du journal (ex : un journal dédié à l'obstétrique ne s'intéresse pas aux prothèses de hanche, à moins qu'il s'agisse spécifiquement d'une étude sur les femmes enceintes porteuses de telles prothèses).

Ensuite, le comité éditorial recherche et sollicite des **pairs** : ce sont des **scientifiques indépendants** qui connaissent bien le sujet abordé par l'article candidat. Généralement, ce sont des **chercheurs concurrents** des auteurs. Ces pairs interviennent généralement bénévolement. Ils donnent un **avis** (favorable ou non) et, surtout, exigent des **améliorations** dans l'article soumis.

Le comité éditorial fait la synthèse des remarques, et les renvoie à l'auteur. S'ensuivent plusieurs tours. A l'issue de chaque tour, les réponses peuvent être les suivantes :

- **Rejet** de l'article
- Acceptation possible sous réserve de **révision majeure** : d'importantes modifications sont demandées, y compris si besoin des modifications de l'étude, et presque toujours des modifications de présentation des résultats
- Acceptation possible sous réserve de **révision mineure** : il est demandé aux auteurs de modifier la présentation et d'apporter des précisions
- **Acceptation, et révision d'édition** : ce dernier tour consiste en des échanges avec les employés du journal, pour parfaire la mise en forme (ex : typographie, orthographe, unités de mesure, résolution des images, précision des affiliations, etc.)
- **Publication**, généralement rapidement en ligne, puis en version papier le cas échéant, selon l'agenda de publication du journal (en tenant compte de numéros thématiques, etc.)

Ce processus est schématisé en Figure 1.

Un article de qualité qui est apprécié par un journal suit en général trois tours successifs : révision majeure, puis mineure, puis édition. Certains articles sont rejetés immédiatement,

d'autres subissent 2 ou 3 tours avant d'être finalement rejetés, d'autres sont acceptés après 5 tours, etc. Tout est possible.

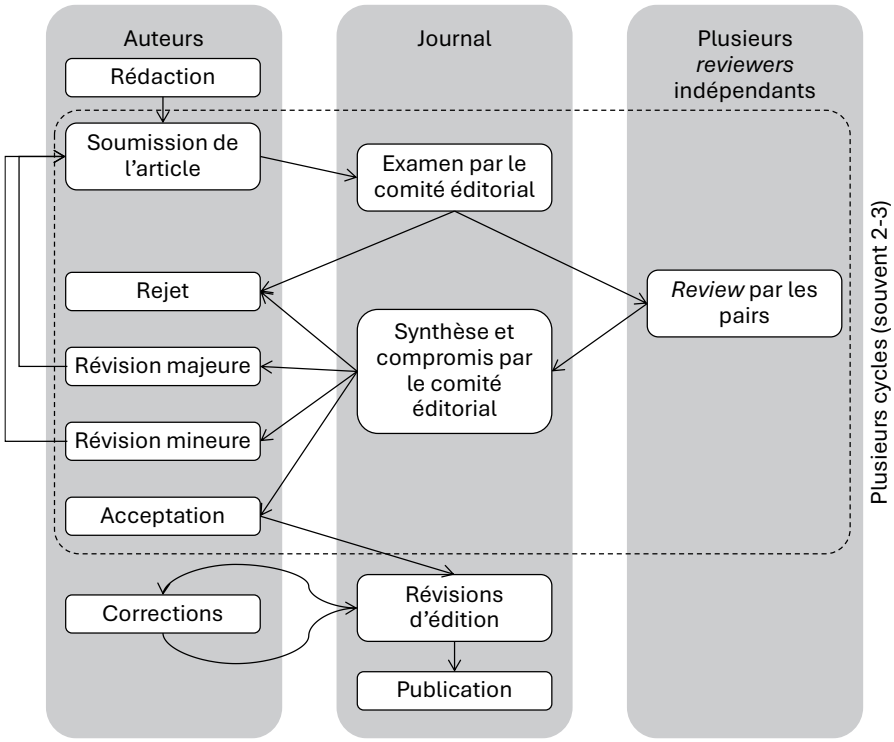


Figure 1. Schématisation du processus de revue par les pairs d'un journal scientifique

Finalement, ce processus garantit :

- Un **filtrage** des articles en fonction de la **thématique**, pour assurer qu'un journal s'adresse à un lectorat spécifique (ex : discipline, chercheurs ou cliniciens, etc.)
- Un **filtrage** des articles en fonction de leur **qualité**
- Une **amélioration itérative** de l'article : sous la pression des *reviewers*, les articles sont généralement de bien meilleure qualité à leur publication que lors de leur première soumission. Ce processus est parfois douloureux pour les auteurs, mais améliore considérablement la fiabilité des résultats, la neutralité des conclusions et la qualité globale des articles

1.1.5 Bases de données bibliographiques

Certaines bases de données bibliographiques intègrent des journaux scientifiques en fonction de critères qui leur sont propres. La base de données **Medline**, de la **NLM** (*National Library of Medicine*), est la base bibliographique de référence en recherche en santé. **Pubmed**[2] **http://pubmed.gov** est le moteur de recherche assis sur Medline. Nous vous conseillons de rechercher les articles scientifiques **uniquement sur Pubmed**. Les critères d'inclusion d'un journal dans cette base de données sont essentiellement fondés sur :

- La qualité et l'indépendance du processus de revue par les pairs du journal
- L'indépendance du journal vis-à-vis de diverses influences, dont les laboratoires pharmaceutiques
- Le fait que les articles du journal candidat à l'indexation sont déjà cités par des journaux eux-mêmes déjà présents dans la base de données (nous reviendrons sur le concept de **bibliométrie** plus tard)

Web Of Science[3] est une base de données plus large, tant en termes de champ disciplinaire couvert, que de tolérance sur la qualité des journaux inclus. Il peut être nécessaire d'y recourir dans les circonstances suivantes :

- Vous vous intéressez à un champ disciplinaire non-couvert par Pubmed (ex : ingénierie non-appliquée à la santé)
- Vous vous intéressez à des articles de moindre qualité, parce que le sujet est très rarement couvert
- Vous vous intéressez à un sujet extrêmement récent, discuté dans certaines conférences scientifiques mais ne faisant pas encore de publications de qualité

Il existe de nombreuses bases de données **affiliées aux éditeurs**. Par exemple, *Science-direct* est la base de données de l'éditeur Elsevier : elle contient tous les journaux d'Elsevier, quelle que soit leur qualité, et ignore tous les journaux des autres éditeurs, quelle que soit leur qualité également.

1.1.6 Droits d'auteur et copyright, pour l'auteur du contenu

Pour comprendre la section suivante relative aux modèles économiques, il faut comprendre une notion réglementaire importante. En France, les articles scientifiques (notamment) sont protégés par le droit d'auteur. Dans d'autres pays, comme les USA ou le Royaume Uni, ces créations sont protégées par le copyright, qui n'a ni la même définition, ni la même portée, et est aliénable. Les articles scientifiques rédigés par un chercheur français, et publiés par un journal des Etats-Unis ou du Royaume Uni par exemple, sont soumis en France au droit d'auteur, et possiblement au copyright dans le pays dont le journal est issu.

Nous vous proposons une lecture très simplifiée et opérationnelle ci-dessous. Cette lecture ne satisfera pas un juriste spécialiste, mais vous permettra de vous comporter d'une manière simple et non-répréhensible.

Le **droit d'auteur** est l'ensemble des droits inaliénables qu'ont les auteurs sur leur œuvre. Les auteurs restent détenteurs à jamais de ce droit d'auteur. Le droit d'auteur porte sur « le fond » de l'article. Ce droit comprend notamment les droits moraux : revendiquer et assumer la paternité. En France, les idées et concepts ne peuvent pas être protégés : rien n'interdit de propager une idée, tant que ce n'est pas un plagiat (recopie stricte d'un passage trop long, ou sans citer la source).

Le **copyright** est un droit moral restreint sur une œuvre. Considérez qu'il concerne en particulier le droit de reproduction de l'article mis en forme. Il correspond au travail d'édition mené par les employés de l'éditeur (corrections typographiques, présentation, mise en forme des tableaux, des références, etc.). En quelque sorte, le copyright porte sur « la forme » de l'article plus que sur le fond.

Il faut comprendre que, dans les modèles classiques d'édition, lorsqu'un auteur français propose un article à un éditeur des USA ou du Royaume Uni et que ce dernier le publie :

- L'auteur reste détenteur du droit d'auteur, qui porte sur le fond
- L'auteur cède le « copyright », qui porte sur le contenu mis en page, à l'éditeur

Si un auteur recopie le fond de son article, avec une mise en page différente, il le peut car il détient le droit d'auteur. En revanche, il n'est pas systématiquement autorisé à diffuser l'article mis en page, car ce travail de mise en page est couvert par le copyright, que l'auteur a cédé à l'éditeur.

1.1.7 Droits d'auteur et copyright, pour celui qui cite un contenu

En tant qu'étudiant, vous pouvez être autorisé à citer ou réutiliser un contenu déjà publié dont vous n'êtes pas l'auteur, sous conditions. Vous devez cependant respecter le droit d'auteur et le copyright.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Pour ce qui est des **idées et concepts** : vous êtes autorisés à réutiliser sans aucune limitation toutes les idées et tous les concepts. Vous devez, chaque fois que possible, indiquer d'où viennent ces idées et concepts, à l'aide de références bibliographiques.

Pour ce qui est du **texte** : vous pouvez bien évidemment réutiliser des mots en reformulant les phrases, sans limitation. Vous pouvez également citer des passages un peu plus longs (phrases entières, paragraphes) à condition de les citer entre guillemets et d'en préciser la source (sous forme de référence bibliographique notamment). Dans le cas contraire, ce serait un **plagiat** (atteinte au droit d'auteur). La plupart des universités sont dotées de systèmes de **contrôle anti-plagiat**, auxquels sont systématiquement soumises toutes les thèses avant leur soutenance (exemple en Figure 2). Pour ce qui est des autres mémoires ou travaux académiques, les enseignants des universités ont un accès facilité à de tels systèmes. Ces systèmes recherchent sur le web des passages de texte similaires à ceux de votre mémoire, et émettent des alertes si ces passages sont trop longs et trop nombreux. Il n'existe pas de seuil à ne pas dépasser : ces logiciels émettent des alertes, qui sont contrôlées de manière ciblée par un humain.

Pour ce qui est des **images (schémas, photos)**, la plupart d'entre elles sont couvertes par le droit d'auteur. Vous pouvez réutiliser librement les images sous **licence Creative Commons**, ou sous d'autres licences ouvertes. Théoriquement, vous devriez obtenir l'autorisation de l'auteur pour toutes les autres images. En pratique, on ne vous en voudra pas si vous recopiez quelques images, à condition de clairement citer la source. S'il s'agit d'un schéma, vous avez toujours le droit de vous inspirer d'un schéma existant, de le refaire vous-même, et de citer la source (non pas pour l'œuvre, mais pour l'idée et les concepts).

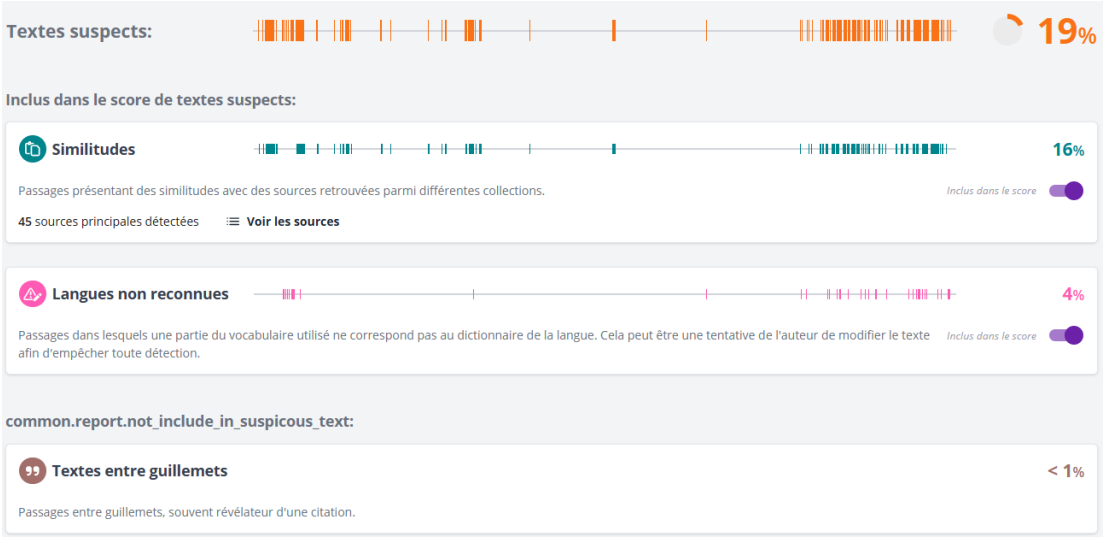


Figure 2. Exemple de rapport de détection de plagiat par Compilatio Magister®
Chaque passage est ensuite détaillé, avec la source trouvée sur internet

1.1.8 Modèles économiques de publication d'articles

Les modèles économiques de publication des articles scientifiques ont beaucoup évolué ces vingt dernières années.

Modèle traditionnel par abonnement : les auteurs publient bénévolement et gratuitement, les *reviewers* sont bénévoles, et les lecteurs paient un abonnement leur permettant d'accéder aux numéros du journal. Cet accès peut par exemple comporter les exemplaires en papier du journal pour la période couverte, et un accès électronique aux articles plus anciens. Il est à noter que les auteurs conservent leur droit d'auteur, mais renoncent au copyright : ils ne sont généralement pas autorisés à diffuser eux-mêmes leurs articles mis en forme.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Modèle *open access* : les auteurs paient pour publier chaque article (généralement entre 1500€ et 3000€ ; certains journaux exigent même 500€ pour soumettre, sans garantie de publication !), les *reviewers* sont bénévoles ou rémunérés, mais en contrepartie les lecteurs peuvent accéder librement et sans limitation à tous les articles du journal. Selon les journaux, les articles publiés sont placés sous un copyright permettant un accès libre (ex : licence *Creative Commons*). Ce modèle a permis d'augmenter les indices bibliométriques des journaux concernés, qui étaient plus cités. En contrepartie, les institutions publiques protestent car, si elles publient beaucoup, elles paient plus que par le système des abonnements.

Modèles mixtes : de nombreux journaux traditionnels ont dû s'adapter pour faire face aux journaux *open access*. Leur modèle reste fondamentalement celui de l'accès par abonnement avec publication gratuite, mais des adaptations se voient, comme par exemple :

- La proposition aux auteurs d'une option *gold open access* : ils paient, mais en contrepartie leur article sera visible de tous immédiatement
- L'autorisation de la diffusion par les auteurs de leurs articles sans aucune mise en forme (plus précisément, dans la version acceptée par les *reviewers* mais avant intervention des salariés du journal)
- L'autorisation de la diffusion par les auteurs de leurs articles mis en forme, passée une période de 12 ou 24 mois d'embargo
- La diffusion systématique en *open access* de tous les articles, même soumis gratuitement, passée une période de 12 ou 24 mois d'embargo

Conférences scientifiques avec actes (articles ou abstracts publiés) : certaines conférences scientifiques proposent depuis fort longtemps un modèle assez différent. Il s'agit de congrès, auxquels les participants paient tous une inscription, qu'ils soient ou non auteurs. Ces inscriptions financent le congrès (location de salle, alimentation) et la publication des « actes du congrès ». Les auteurs soumettent gratuitement un article (mais s'engagent à venir en payant leur inscription, en cas d'acceptation). Les *reviewers* sont bénévoles. Les articles sont présentés oralement en séance à tous les participants, et sont débattus. Ils sont ensuite publiés en *open access* à l'issue de la conférence.

Le Tableau 1 présente une comparaison des trois modèles économiques classiques de publication.

Tableau 1. Comparaison des principaux modèles économiques de publication

Type	Journal classique	Journal open access	Conférence avec actes
Support	Article papier et électronique	Article électronique uniquement	Présentation orale puis article papier ou électronique
Financement	Abonnements des lecteurs	Paiement par les auteurs	Inscription des participants
Lecteurs : accès aux articles complets	Au prix d'un abonnement payant (ou passé un certain délai)	Ouvert et gratuit	Ouvert et gratuit
Auteurs : publication d'un article	Gratuite et bénévole	Payante	Gratuite et bénévole, mais inscription au congrès obligatoire
Reviewers : sélection et amélioration des soumissions	Bénévole	Bénévole (parfois rémunérée)	Bénévole
Impact factor	Selon le journal	Selon le journal, généralement plus élevé	Non-calculé car non-périodique

Les revues prédatrices sont des journaux qui détournent le modèle ***open access*** à des fins purement mercantiles. La définition de ces revues n'est pas toujours évidente. On retrouve souvent (attention, ces caractéristiques ne sont ni nécessaires, ni suffisantes) :

- Un modèle exclusivement *open access*, avec des prix aisément négociables à la baisse sous réserve de paiement rapide en ligne
- Des options *fast track*, qui permettent notamment aux doctorants de publier très vite leurs articles en payant plus, ce qui leur permettra de soutenir leur thèse d'université avant la fin de leur inscription administrative
- Des noms de journaux ronflants, éventuellement très proches de noms de prestigieux journaux
- Une acceptation systématique, avec un processus de *review* très rapide et complètement bâclé
- L'absence de référencement par Pubmed
- Un siège social souvent (de moins en moins) localisé dans un pays en voie de développement

La question des revues prédatrices est bien suivie est expliquée par des auteurs de renom tel Hervé Maisonneuve[4]. La Conférence des Doyens des facultés de Médecine publie une liste des revues considérées comme non-prédatrices[5].

1.1.9 Bibliométrie

La « bibliométrie » désigne l'ensemble des opérations de quantification réalisables sur les articles et les journaux scientifiques. Ces opérations visent généralement à mesurer la quantité de travail de publication, et la qualité de ces publications, de manière automatisée.

Lorsque les journaux, et donc les articles scientifiques, sont référencés Pubmed ou WoS, la base de données structure également les références bibliographiques de ces articles, et tente de les raccrocher aux publications déjà référencées dans cette base de données. Cela permet de calculer des **indices bibliométriques**, qui permettent par exemple de :

- Quantifier la quantité et la qualité des citations d'un seul article par les autres articles, et donc :
- Quantifier la quantité et la qualité des publications d'un auteur ou de la structure dans laquelle il intervient (son affiliation)
- Quantifier la quantité et la qualité des journaux qui publient ces articles

Ces indices bibliométriques s'appuient sur des API (*Application Programming Interface*) qui permettent d'interroger automatiquement les données structurées de Pubmed ou WoS. Certains indices bibliométriques sont plus connus :

- L'***impact factor*** **(IF)**, ou facteur d'impact, est un indice quantitatif rapportant combien les articles d'un journal sont cités par les autres journaux. Cet *impact factor* est très élevé pour les grands journaux, mais favorise les journaux généralistes, qui peuvent être cités par toutes les disciplines (ex : NEJM, BMJ, Nature, JAMA, etc.). Il pénalise en comparaison les disciplines comportant peu de chercheurs (ex : aide médicale à la procréation, informatique de santé, etc.). L'IF est un indice calculé par une société privée (*Clarivate*®, anciennement *Thomson-Reuters*®) dans son *Journal Citation Report* (JCR), selon ses propres critères. Cela peut également prendre en compte des critères arbitraires (ex : exclusion des journaux non-périodiques) voire disciplinaires (ex : exclusion temporaire du journal *Methods of Medicine*, pour fraude sur les citations). Cet indicateur est cependant relativement neutre, de qualité, et clairement documenté, au point d'être largement utilisé.
- Le **rang d'impact factor**, calculé en A+/A/B/C/D/E par Sampra ou SIGAPS, ou en Q1/Q2/Q3/Q4 par WoS, tient compte de l'impact factor mais uniquement au sein de chaque discipline, afin de gommer le biais précédent. Il repose sur la classification des journaux dans des disciplines, comme le fait la NLM.

- Le nombre de points **Sigaps** peut être utilisé pour évaluer la production scientifique d'un hôpital (avec un impact sur son financement), d'un service ou d'un membre du personnel hospitalier (ex : médecin chercheur). Il s'appuie sur les rangs d'IF calculés sur Pubmed, donc principalement en santé.
- Le nombre de points **Sampra** peut être utilisé pour évaluer la production scientifique d'une université (sans impact sur son financement), d'une équipe de recherche ou d'un membre du personnel universitaire. Il s'appuie sur les rangs d'IF calculés sur WoS.
- De nombreux index, référencés par Anne-Wil Harzing (projet *Publish or Perish*)[6], visent à calculer un indice individuel par chercheur
- Etc.

Tous ces indices sont globalement cumulatifs, au sens où les chercheurs âgés ont généralement cumulé plus de points que les jeunes chercheurs. Cependant, il se peut que les coefficients utilisés soient mis à jour rétrospectivement, et que le score d'un chercheur diminue brutalement alors qu'il continue à publier.

1.2 Pourquoi mener une recherche bibliographique ?

Si votre travail consiste en une revue de la littérature ou une méta-analyse, la recherche bibliographique est, littéralement, votre travail. Dans tous les autres cas, vous aurez également besoin absolument de mener une recherche bibliographique à plusieurs étapes.

Premièrement, **en amont du travail**.

Vous devez **savoir si le travail a déjà été réalisé**, tout simplement. S'il a déjà été réalisé, il vous faudra prendre connaissance du procédé exact et des résultats. Cela vous permettra peut-être de détecter des failles dans le travail réalisé, et de proposer une méthode pour **combler ces failles**. Si le travail réalisé est déjà excellent, il vous faudra argumenter sur le besoin de le **répliquer** dans un contexte différent. Vous serez alors dans l'obligation de mettre en place une **méthodologie très proche et comparable** au travail déjà réalisé.

Si le travail n'a pas été réalisé, il vous faudra néanmoins bien connaître le domaine pour **éviter les pièges** que les autres articles ont déjà identifiés et traités (ex : exclure certaines sous-populations, vérifier la validité d'une mesure au préalable, etc.). Il vous faudra également envisager dès le début le calcul d'indicateurs qui sont les mêmes que les autres articles, afin de permettre une c**omparabilité de vos résultats**.

La recherche bibliographique est également indispensable pour la rédaction de votre document final. Nous reviendrons dessus plus tard, dans le chapitre Rédiger et présenter le document en page 236).

Ensuite, pour l'**introduction de votre mémoire**, la bibliographie vous permettra de :

- Montrer que le travail que vous allez réaliser est **nécessaire**, du point de vue de la connaissance collective
- Montrer que le travail **n'a pas été réalisé**, ou doit être **complété**, ou doit être **répliqué** sur une sous-population différente
- Montrer, par un faisceau d'éléments externes, que ce travail est **réalisable** et **peut répondre** à la question posée
- Montrer, plus généralement, que vous êtes **crédible** car vous connaissez bien le domaine étudié et que vous vous appuyez sur les connaissances collectives

Enfin, pour la **discussion de votre mémoire**, la bibliographie réalisée (y compris celle déjà citée dans l'introduction) vous permettra de :

- **Comparer** vos résultats à ceux des autres
- **Comparer** vos méthodes à celles des autres
- Identifier et discuter les **biais** éventuels au regard de connaissances externes (y compris pour conclure que ces biais ne suffisent pas à décrédibiliser l'étude)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- Montrer l'**apport** de votre étude dans la connaissance collective
- Envisager les **suites** à donner à votre étude

A travers ce rapide exposé, nous voyons que **la recherche bibliographique est indispensable** à la réalisation d'un mémoire académique ou d'un travail scientifique (voir Figure 3). Cette recherche doit être débutée **avant de commencer le travail**.

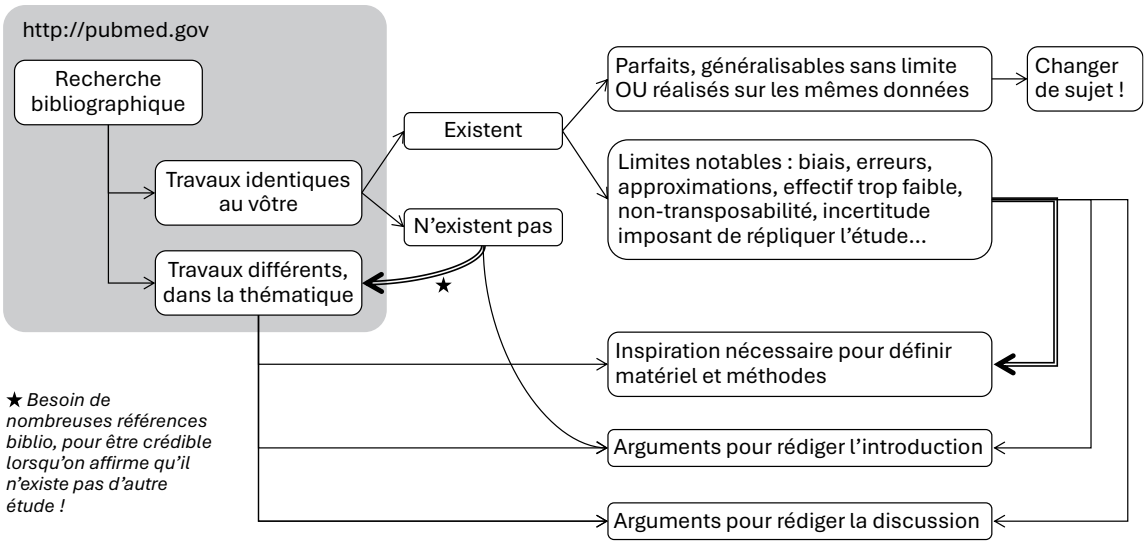


Figure 3. Arbre décisionnel : place de la recherche bibliographique

1.3 Mener une recherche bibliographique sur Pubmed

1.3.1 Fonctionnement de l'interface Pubmed, en bref

Pour les raisons expliquées dans le chapitre 1.1.5 Bases de données bibliographiques en page 15, votre recherche bibliographique pourra la plupart du temps être faite uniquement sur **Pubmed**[2].

Le site web http://pubmed.gov propose une fenêtre de recherche. Lorsque vous saisissez une requête, le moteur affiche les résultats sous la forme d'une liste d'articles (Figure 4). La fenêtre est intuitive et ne mérite pas un tutoriel en soi. La liste de réponses est affichée, ainsi que de nombreux filtres qui peuvent être activés a posteriori, laissant ainsi voir la réduction itérative des réponses possibles.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

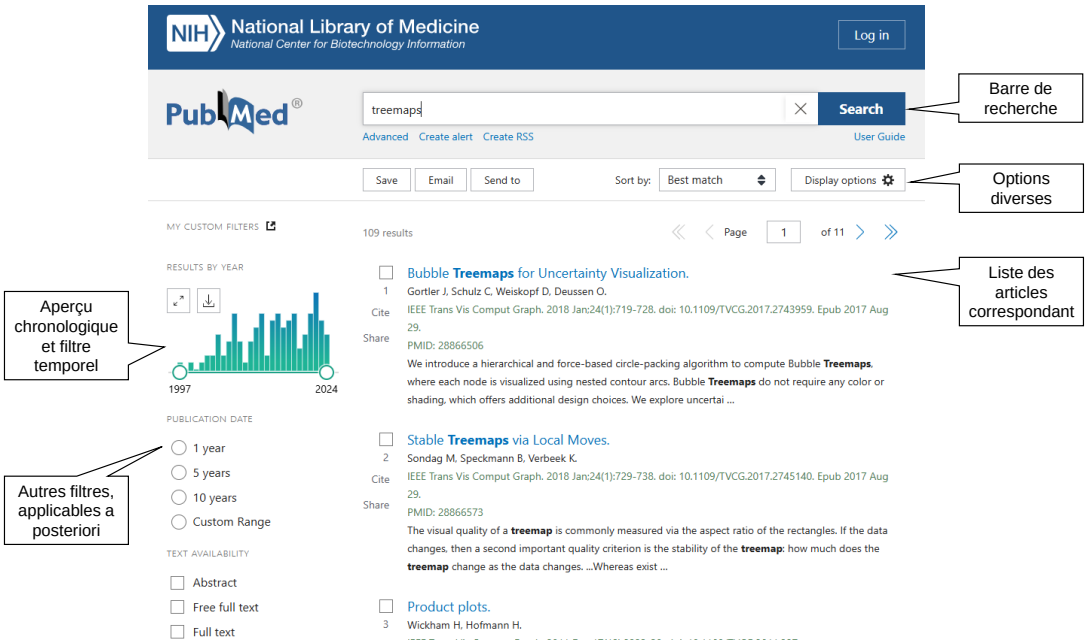


Figure 4. Fenêtre de recherche de http://pubmed.gov

Il est très important de comprendre que, sans l'afficher explicitement, Pubmed fonctionne selon deux modes : le **mode clinicien**, et le **mode chercheur** (ces appellations sont officieuses).

Le **mode clinicien** est hélas le **mode par défaut**. Il se traduit par :

- Une recherche des mots entrés par l'utilisateur en position « titre ou abstract » par défaut
- Une recherche des mots entrés par l'utilisateur en position de « mot clef MeSH », lorsque c'est possible, et leur substitution automatique et transparente[2] par certains synonymes
- Une interprétation des mots clefs utilisés lors de la requête : certains mots sont automatiquement étendus à leur forme plurielle, certaines fautes d'orthographe sont automatiquement corrigées
- L'affichage des articles dans l'ordre « best match », qui n'est pas explicitement documenté et peut évoluer sans informer les utilisateurs

Ce comportement est utile pour le clinicien qui cherche rapidement une réponse à une question, mais ne souhaite pas lire plus d'un ou deux articles. Dans ce cas d'usage, il sera très profitable de cocher la case correspondant au filtre « systematic review », qui apportera une réponse au clinicien à un problème très précis.

Pour le chercheur, **aucune de ces options n'est acceptable**. Non seulement le moteur cherche autre chose que ce que le chercheur lui demande, mais en plus l'ensemble des transformations ou tris appliqué sont susceptibles de changer dans le temps : une même requête peut très bien, le lendemain, afficher des résultats différents et dans un ordre différent.

Le moteur passe en **mode chercheur** lorsqu'il est interrogé sous la forme d'une **chaîne de requête** (sans demande explicite de l'utilisateur). Alors, le moteur exécute strictement la chaîne de requête sans la modifier. L'ordre de tri, depuis peu, reste en « best match », mais peut aisément être corrigé si nécessaire. Une chaîne de requête comprend :

- des termes, éventuellement délimités par des guillemets
- systématiquement assortis d'un indicateur de portée (titre, auteur, abstract, date…), entre crochets

[2] Pour rappel, ce qui est transparent est invisible

- le tout agencé à l'aide de parenthèses, de AND et de OR

Dans l'exemple reproduit en Équation 1, on recherche tous les articles qui, à la fois :

- contiennent dans le titre le terme *ECG*, ou *EKG*, ou *electrocardiogram*, ou *electrocardiograms* (notez que c'est à nous de préciser les différentes orthographes et la forme plurielle)
- contiennent dans le titre le terme *automated* ou, *automatic* ou *computer* ou *computerized* (notez que c'est à nous d'imaginer et préciser les différentes manières de nommer l'interprétation automatisée)
- ont été publiés entre 2010 et 2024 (soit 15 années complètes, pas 14)

Pour bien préparer une chaîne de requête pour Pubmed, il faut suivre plusieurs étapes (elles seront exposées par la suite, puis synthétisées en Figure 5 en page 26) :

- identifier les concepts, souvent au nombre de 1, 2 ou 3
- identifier, pour chaque concept, les synonymes et les termes équivalents (sans être forcément des synonymes), avec pour chacun les différentes typographies, et les formes plurielles
- agencer le tout sans erreur avec des parenthèses, guillemets, AND et OR
- faire tourner cette chaîne de requête, parcourir rapidement quelques titres et abstracts, et l'améliorer itérativement

```
(       ("ECG"[Title]) OR ("EKG"[Title]) OR
("electrocardiogram"[Title]) OR ("electrocardiograms"[Title])
) AND (
("automated"[Title]) OR ("automatic"[Title]) OR
("computer"[Title]) OR ("computerized"[Title])
) AND (
("2010"[Date - Publication] : "2024"[Date - Publication])     )
```

Équation 1. Exemple de chaîne de requête (simplifiée) pour Pubmed.gov

1.3.2 Etape 1 : identifier les concepts

Généralement, une requête comportera 1, 2, 3, au maximum 5 concepts. Ces concepts, de notre expérience, s'entendent en termes d'intersection. Des exemples seront plus parlants :

Exemple 1 : On s'intéresse aux effets indésirables liés à l'interruption d'un traitement par statine. On identifie 3 concepts, en relation logique :

Interruption_de(statine) → effet_indésirable

En l'état actuel des choses, il n'est ni nécessaire, ni possible, de rendre compte du lien logique dans une requête. Il suffit donc de lister les 3 concepts d'intérêt :

- Statine
- Interruption (d'un traitement)
- Effet indésirable

Comme, de toute manière, il adviendra un moment ou un humain contrôlera et filtrera les résultats, le plus simple est de chercher tous les articles qui contiennent à la fois ces trois concepts, autrement le premier ET le deuxième ET le troisième. On notera que cet opérateur logique ET (AND) est commutatif, l'ordre des concepts n'a donc pas d'importance.

Exemple 2 : On s'intéresse aux résultats d'évaluation des interpréteurs automatisés d'ECG. On identifie 3 concepts :

- *ECG*
- *Interprétation automatisée (ici le terme automatisé est plus spécifique)*
- *Evaluation*

1.3.3 Etape 2 : identifier les termes dans chaque concept

Pour chaque concept, il faut à présent identifier une liste de termes correspondants. Ces termes sont bien entendu le nom que vous donnez au concept, mais aussi :

- ses variants typographiques
- ses synonymes
- les termes reliés mais pas synonymes (ex : sous-types, termes proches, etc.)
- les formes plurielles de tous les termes précédents

Il faut faire au mieux à cette étape, mais ce sont surtout les tests itératifs qui nous permettront de stabiliser ces listes. Notez qu'ajouter un terme que personne n'utilise, au pire, sera inutile mais pas gênant.

Exemple 1 : Voici par exemple les termes qu'on pourra utiliser pour identifier le concept « statine » :

- Le terme « statine » lui-même, avec différents variants (notez que le terme « statine » est erroné, cela n'a pas d'importance) :
  - statin, statins
  - statine, statines
- Son synonyme (dans ce cas ce n'est pas très utile) :
  - HMG Co-A reductase inhibitor, HMG Co-A reductase inhibitors
- Différentes statines, partant du principe qu'un article qui ne parle que de la fluvastatine nous intéressera, même si notre étude est plus large :
  - fluvastatin
  - lovastatin
  - mevastatin
  - pitavastatin
  - pravastatin
  - rosuvastatin
  - simvastatin

Exemple 1, suite : Voici par exemple les termes qu'on pourra utiliser pour identifier le concept « interruption du traitement » :

- stop, stops
- discontinuation, discontinuations
- interruption, interruptions
- deprescription, deprescriptions

Exemple 1, fin : Pour identifier le concept « effet indésirable », on pourrait utiliser les termes suivants et leurs pluriels :

- adverse event, adverse events
- adverse effect, adverse effects
- adverse reaction, adverse reactions

On voit qu'il suffit alors de rechercher le terme « adverse », qui inclut nécessaire tous les précédents.

1.3.4 Etape 3 : agencer la requête

On rappellera que l'opérateur **OR** correspond à l'union : il augmente le nombre de réponses. Si on cherche des chaussettes et des chaussures, on cherche en réalité des objets qui soient une chaussette ou une chaussure (OR dans ce cas).

L'opérateur **AND** correspond à l'intersection : il diminue le nombre de réponses. Si on recherche des chaussettes rouges, on cherche en réalité des objets qui soient à la fois une chaussette, et de couleur rouge (AND dans ce cas).

Le principe général est le suivant.

Premièrement, la requête est articulée autour de conditions (chacune obligatoire) portant sur les concepts :

(concept1) ***AND*** *(concept2)* ***AND*** *(concept3)* ***AND*** *(limites_de_dates)*

Deuxièmement, chaque concept est à remplacer par la liste des termes qui le composent : il en faut au moins un, mais aucune d'entre eux n'est obligatoire.

Chaque terme est associé à un indicateur de portée : [Title] ou [Title/Abstract] qu'on peut noter [TiAb]. Les autres indicateurs de portée ne vous seront pas utiles.

Enfin, un terme composé de plusieurs mots doit obligatoirement être mis entre guillemets (les guillemets sont facultatifs pour les autres).

Les passages à la ligne n'ont pas d'impact, et les parenthèses peuvent être utilisées *a volo* pour lever toute ambiguïté.

Ainsi, dans notre exemple, le concept de statine est à remplacer par :

("statin"[Title/Abstract]) ***OR*** *("statins"[Title/Abstract])* ***OR***
("statine"[Title/Abstract]) ***OR*** *("statines"[Title/Abstract])* ***OR***
("fluvastatin"[Title/Abstract]) ***OR*** *("lovastatin"[Title/Abstract])* ***OR***
("mevastatin"[Title/Abstract]) ***OR*** *("pitavastatin"[Title/Abstract])* ***OR***
("pravastatin"[Title/Abstract]) ***OR*** *("rosuvastatin"[Title/Abstract])* ***OR***
("simvastatin"[Title/Abstract]) ***OR***
("HMG Co-A reductase inhibitor"[Title/Abstract]) ***OR***
("HMG Co-A reductase inhibitors"[Title/Abstract])

Ne vous embêtez pas à faire cela à la main ! Sur le site ObjectifThèse[7] (http://objectifthese.org) je vous propose un fichier tableur gratuit, très simple d'utilisation, qui permet de générer automatiquement une requête simplement en remplissant les cases du tableur.

1.3.5 Etape 4 : itérations

Personne n'est parfait : utilisez une première fois le fichier que je vous fournis pour construire une requête, puis testez-la, afin de détecter les **anomalies**, mais surtout trouver des **termes équivalents** aux concepts que vous recherchez. Pour ce faire, lisez en diagonale une vingtaine d'abstracts à chaque tour, ainsi que les titres des articles cités par ces articles, qu'on peut souvent lire sur Pubmed, . Ne prenez le temps de lire en détail les articles que lorsque votre requête sera stabilisée. La définition de la requête est donc un travail **itératif**, qui sera d'autant plus aisé à réaliser que vous vous aiderez d'un constructeur de requête comme celui que je vous propose.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

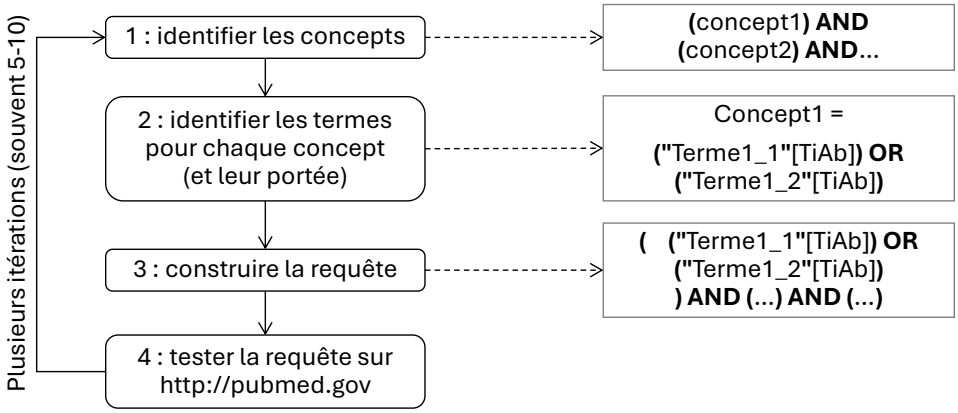


Figure 5. Construction itérative d'une requête pour Pubmed

1.3.6 Exemple 1

Dans cet exemple exposé précédemment, on s'intéresse aux effets indésirables liés à l'interruption d'un traitement par statine.

Du fait du grand nombre de réponses, nous avons souhaité limiter la recherche aux termes présents dans le titre de l'article.

Année minimale : **2010**
Année maximale : **2024**

concept 1	concept 2	concept 3	concept 4
Title	**Title**	**Title**	**Title/Abstract**
statin	stop	adverse event	
statins	discontinuation	adverse effect	
statine	interruption	adverse events	
statines	deprescription	adverse effects	
fluvastatin		adverse reaction	
lovastatin		adverse reactions	
mevastatin			
pitavastatin			
pravastatin			
rosuvastatin			
simvastatin			
HMG Co-A reductase inhibitor			
HMG Co-A reductase inhibitors			

Figure 6. Utilisation du fichier d'Objectif Thèse :
articles portant sur les effets indésirables liés à l'arrêt d'une statine

Le fichier produit automatiquement la requête suivante, qu'il suffit de copier dans la boîte de recherche Pubmed :

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

(("statin"[Title]) OR ("statins"[Title]) OR ("statine"[Title]) OR ("statines"[Title]) OR ("fluvastatin"[Title]) OR ("lovastatin"[Title]) OR ("mevastatin"[Title]) OR ("pitavastatin"[Title]) OR ("pravastatin"[Title]) OR ("rosuvastatin"[Title]) OR ("simvastatin"[Title]) OR ("HMG Co-A reductase inhibitor"[Title]) OR ("HMG Co-A reductase inhibitors"[Title])) AND (("stop"[Title]) OR ("discontinuation"[Title]) OR ("interruption"[Title]) OR ("deprescription"[Title])) AND (("adverse event"[Title]) OR ("adverse effect"[Title]) OR ("adverse events"[Title]) OR ("adverse effects"[Title]) OR ("adverse reaction"[Title]) OR ("adverse reactions"[Title])) AND (("2010"[Date - Publication] : "2024"[Date - Publication]))

Équation 2. Equation de recherche Pubmed : articles portant sur les effets indésirables liés à l'arrêt d'une statine

La construction pas-à-pas de cette requête fait l'objet d'une vidéo qui pourra vous être utile. Elle est référencée sur le site : http://www.objectifthese.org

1.3.7 Exemple 2

Dans cet autre exemple exposé précédemment, on s'intéresse à l'évaluation des interpréteurs automatisés d'ECG.

Nous avons modifié la requête au fil des itérations. Nous avions tout d'abord recherché les termes en titre ou abstract, mais constaté que les résultats n'étaient pas assez spécifiques. Inversement, avant de restreindre aux titres des articles, nous avons constaté que les titres étaient plus littéraires et pouvaient parfois se limiter à des termes très spécifiques. Voici quelques commentaires sur certains concepts :

- Le concept d'ECG a dû être complété par certaines anomalies électrocardiographiques qui, seules, persistaient dans le titre (fibrillation atriale, etc.).
- Le concept d'évaluation a dû être complété par le terme « outperforms » (A fait mieux que B)
- Assez rapidement, nous avons été frustrés par les abstracts annonçant qu'il faudrait ensuite évaluer le dispositif qu'ils avaient conçu. Nous avons ajouté un quatrième concept, stipulant qu'un résultat d'évaluation soit cité dans le titre ou l'abstract.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Année minimale :	**2010**
Année maximale :	**2019**

concept 1	concept 2	concept 3	concept 4	concept 5
Title	**Title**	**Title**	**Title/Abstract**	**Title**
automated atrial fibrillation	accuracy	algorithm	F1	
bundle branch block	accurate	algorithms	F-score	
ECG	analyses	artificial intelligence	kappa	
ECGs	analysis	atrial fibrillation detection	predictive value	
EKG	analyzed	automated	recall	
EKGs	analyzing	automatic	sensitivity	
electrocardiogram	assess	computer	specificity	
electrocardiograms	assessing	computer-interpreted		
electrocardiographic	assessment	computerized		
electrocardiography	detection	device incorporated		
ST-elevation	evaluate	machine learning		
	evaluated	neural network		
	evaluation	program		
	identification	programs		
	interpretation	software-based		
	outperform			
	outperforms			
	performance			
	prediction			
	validation			

Figure 7. Utilisation du fichier d'Objectif Thèse :
articles portant sur l'évaluation des interpréteurs automatisés d'ECG

Comme l'illustre cet exemple, c'est le caractère itératif de notre démarche qui a permis de détecter les failles de la version précédente, et de faire progresser notre requête Pubmed. Ces itérations sont plus aisées sur un fichier aisé à maintenir, avec recalcul automatique de la requête de recherche.

Le fichier produit automatiquement la requête suivante, qu'il suffit de copier dans la boîte de recherche Pubmed :

(("automated atrial fibrillation"[Title]) OR ("bundle branch block"[Title]) OR ("ECG"[Title]) OR ("ECGs"[Title]) OR ("EKG"[Title]) OR ("EKGs"[Title]) OR ("electrocardiogram"[Title]) OR ("electrocardiograms"[Title]) OR ("electrocardiographic"[Title]) OR ("electrocardiography"[Title]) OR ("ST-elevation"[Title])
) AND (
("accuracy"[Title]) OR ("accurate"[Title]) OR ("analyses"[Title]) OR ("analysis"[Title]) OR ("analyzed"[Title]) OR ("analyzing"[Title]) OR ("assess"[Title]) OR ("assessing"[Title]) OR ("assessment"[Title]) OR ("detection"[Title]) OR ("evaluate"[Title]) OR ("evaluated"[Title]) OR ("evaluation"[Title]) OR ("identification"[Title]) OR ("interpretation"[Title]) OR ("outperform"[Title]) OR ("outperforms"[Title]) OR ("performance"[Title]) OR ("prediction"[Title]) OR ("validation"[Title])
) AND (
("algorithm"[Title]) OR ("algorithms"[Title]) OR ("artificial intelligence"[Title]) OR ("atrial fibrillation detection"[Title]) OR ("automated"[Title]) OR ("automatic"[Title]) OR ("computer"[Title]) OR ("computer-interpreted"[Title]) OR ("computerized"[Title]) OR ("device incorporated"[Title]) OR ("machine learning"[Title]) OR ("neural network"[Title]) OR ("program"[Title]) OR ("programs"[Title]) OR ("software-based"[Title])
) AND (

("F1"[Title/Abstract]) OR ("F-score"[Title/Abstract]) OR ("kappa"[Title/Abstract]) OR ("predictive value"[Title/Abstract]) OR ("recall"[Title/Abstract]) OR ("sensitivity"[Title/Abstract]) OR ("specificity"[Title/Abstract])) AND (("2010"[Date - Publication] : "2019"[Date - Publication]))

Équation 3. Equation de recherche Pubmed : articles portant sur l'évaluation des interpréteurs automatisés d'ECG

1.4 Utiliser un logiciel de gestion de la bibliographie

Zotero[8] est un logiciel open source et gratuit de gestion de la bibliographie. Ce logiciel est très simple et très performant. Nous vous recommandons très vivement de le télécharger à l'adresse https://www.zotero.org/download/

Zotero est tellement bien conçu, qu'il vous fera gagner du temps dès la première journée d'utilisation, ce qui est très rare. Si vous réalisez votre mémoire académique avec Zotero, même avec seulement une trentaine de références bibliographiques, vous gagnerez **plusieurs heures de travail** ! Les références du présent ouvrage ont été gérées avec Zotero.

Zotero est présenté plus tard dans cet ouvrage, dans la section 2 Installer et utiliser Zotero, logiciel de bibliographie en page 243.

Lorsque vous aurez gratuitement installé Zotero, il vous sera très facile de :

- Importer des items dans votre bibliothèque Zotero :
  - Importer une ressource avec toutes ses métadonnées simplement en cliquant sur un bouton de votre navigateur
  - Gérer simplement les ressources (indexation, classement, dédoublonnage, rapatriement du PDF)
  - Partager vos ressources entre vos différentes machines (stockage cloud) ou entre collègues (bibliothèques partagées) ou sur un site web
- Citer des ressources dans votre document de thèse
  - Citation immédiate dans le texte par un bouton
  - Mise en forme et numérotation automatique des citations
  - Mise en forme et mise à jour automatique de la bibliographie en fin de document
  - Choix du style de bibliographie selon l'aspect souhaité ou le journal visé

La vidéo d'Objectif Thèse explique très simplement l'utilisation de Zotero :
http://www.objectifthese.org

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2 Déterminer le type réglementaire d'étude

Après avoir réalisé la bibliographie, il vous faudra déterminer le type réglementaire d'étude. Une manière simple de classifier ces études est de s'intéresser aux objets que vous souhaitez étudier (voir Figure 8) : il peut s'agir de données bibliographiques, de données préexistantes, de personnes humaines (patients notamment), ou de professionnels de santé. L'objet de cette section n'est pas de dresser une typologie complète des études scientifiques, mais simplement de vous aider à identifier le cadre réglementaire qui en découle. Cette typologie est volontairement simplifiée, et présente l'état connu en 2025. N'hésitez pas à vérifier ces informations au moment où vous lisez cet ouvrage.

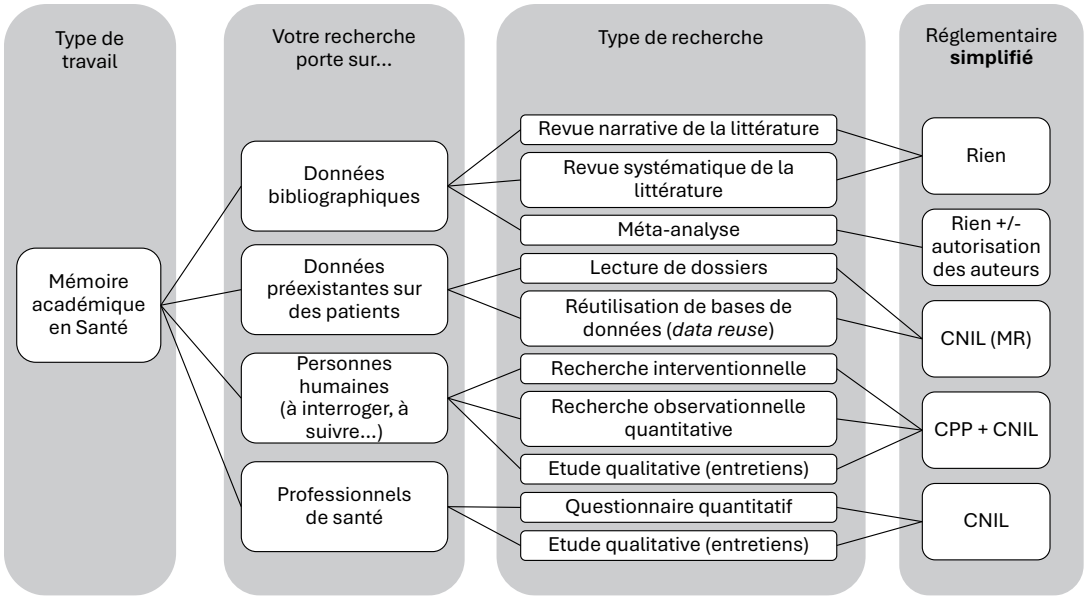


Figure 8. Typologie réglementaire des études en 2025

2.1 Etude portant sur des données bibliographiques

Certaines études portent sur des données bibliographiques, le plus souvent des articles scientifiques, ou plus généralement des documents. Ces études ne nécessitent le plus souvent **aucune autorisation** car elles ne portent ni sur des personnes, ni sur des données de santé. On distingue plusieurs sous-types (voir Figure 8).

Les **revues narratives de la littérature** consistent à lire des documents choisis par l'auteur, et à en proposer une synthèse. Cette synthèse est souvent à la fois qualitative (ex : commenter le résultat ou les méthodes) et quantitative (ex : calculer la proportion d'articles qui appliquent telle ou telle méthode). La sélection des documents inclus est réalisée arbitrairement par l'auteur en fonction de son intérêt.

Les **revues systématiques de la littérature** consistent là aussi à lire des documents et à en proposer une synthèse quantitative et qualitative. La différence avec les revues narratives est que les auteurs s'engagent à ce que le processus de sélection des documents soit systématique et reproductible. Il peut s'agir au mieux d'une chaîne de requête pour un moteur de recherche tel Pubmed[2]. Cette chaîne doit pouvoir être rejouée par le lecteur et conduire au même nombre de résultats. Par la suite, les auteurs explicitent le processus de sélection des ressources, et détaillent les effectifs évincés à chaque étape (ex : lecture du titre, lecture de l'abstract, recherche du document complet, lecture du document complet, etc.). L'ensemble est généralement matérialisé par un diagramme de flux, ou ***flowchart***. Les auteurs de revues systématiques de la littérature sont encouragés à suivre les recommandations (ou *guidelines*)

Prisma[9]. Le caractère exhaustif de la recherche donne un poids supplémentaire aux résultats, en particulier les résultats quantitatifs (ex : « dans 25% des articles, l'analyse est ajustée sur tel facteur »).

Les **méta-analyses** sont des analyses d'analyses. Si le préfixe grec meta- (du grec ancien μετά, metá) signifie « au-delà, après », il est fréquemment utilisé pour indiquer la récursivité. Les méta-analyses sont des analyses d'analyses. En d'autres termes, il s'agit de revues de la littérature dans lesquelles les résultats des documents cités sont eux-mêmes exploités pour recalculer des indicateurs, comme si on avait réalisé une étude en incluant tous les sujets des différentes études. Prosaïquement, si plusieurs analyses proposent des estimations de la prévalence d'une maladie, les auteurs de la méta-analyse pourront aisément recalculer une prévalence de la maladie comme étant la moyenne pondérée des prévalences précédemment publiées. Cet exemple est simpliste mais parlant. Les méta-analyses visent très souvent à réestimer une quantité précisant l'association entre un facteur de risque et un état pathologique par exemple. Ce recalcul peut parfois se faire en récupérant les chiffres publiés dans l'article originale, ou parfois en sollicitant les auteurs pour obtenir des données complémentaires, tel un tableau d'individus.

Si une méta-analyse implique un partage de données par les auteurs d'une autre étude, ce partage doit avoir été autorisé par la loi dont dépend le pays qui a hébergé la première recherche. Hormis ce cas précis, les études portant sur des données bibliographiques ne nécessitent aucune autorisation, dans la mesure où les données analysées sont des données agrégées rendues publiques par leurs auteurs.

Les études précédemment citées peuvent inclure une phase d'analyse de données. Le contenu de cet ouvrage, bien qu'il ne soit pas consacré aux études bibliographiques, pourra aider lorsque viendra le temps de l'analyse statistique. Le cas particulier des méta-analyses ne sera cependant pas traité.

2.2 Etude portant sur des données préexistantes de patients

Certaines analyses portent sur des données de patients, mais les investigateurs ne rencontrent jamais lesdits patients. On parle alors de **recherche sur les données**, ou **recherches n'impliquant pas la personne humaine, RNIPH**. On parle également de **réutilisation de données**, ou utilisation secondaire de données (*data reuse* ou *secondary use of data*). Il peut s'agir de données structurées et contenues dans des bases de données, qui sont directement analysées. Il peut également s'agir de données non-structurées ou semi-structurées, tel le dossier médical papier d'un patient, qui seront relues par un humain et ressaisies via un questionnaire (voir Figure 8).

Dans tous les cas, les données ont au préalable été recueillies pour une finalité première, qui est généralement le soin du patient. La recherche qui réutilise ces données tire opportunément profit des données existantes, pour une finalité secondaire différente de la finalité première. Ce **changement de finalité** permet de caractériser une recherche sur les données. C'est également lui qui engendre des difficultés techniques et méthodologiques.

Ces études sont **généralement rétrospectives**, ce qui leur permet de tirer profit de gros volumes de données. Cependant, rien n'interdit que ces études soient prospectives si elles sont mises en place simultanément à un nouveau recueil de données pour le soin. Ces études sont **toujours observationnelles** : elles ne modifient pas le cours de la prise en charge des patients. Dans le cas contraire, il s'agirait d'études sur les personnes, comme nous le verrons plus bas (ex : les *registry-based controlled trials*).

Toutes les études portant sur les données nécessitent une autorisation de la **CNIL** (commission nationale de l'informatique et des libertés), ou peuvent s'inscrire dans une des **MR** (méthodologies de référence) de la CNIL.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.3 Etude portant sur des personnes humaines

Les recherches impliquant la personne humaine (**RIPH**)[10,11] sont celles durant lesquelles les investigateurs rencontrent leurs sujets (humains vivants) d'étude, et que ces sujets ne sont pas eux-mêmes des professionnels de santé. Il peut par exemple s'agir de patients, de leurs proches, ou de sujets sains.

Il convient alors de savoir si l'étude modifie la prise en charge normale de ces personnes.

Cas n°1 : si la prise en charge de ces personnes est notablement modifiée, on parle de **recherche interventionnelle**. Il peut s'agir de recherches au cours desquelles on réalise un acte qu'on n'aurait pas réalisé en temps normal (ex : réaliser une chirurgie, administrer un traitement), ou au cours desquelles le choix de l'attitude thérapeutique est modifié par l'étude (ex : on avait réellement le choix entre deux médicaments équivalents, mais le fait qu'une étude soit en place modifie la manière dont ce choix est réalisé).

Cas n°2 : si la prise en charge n'est en rien modifiée par l'étude, on parle de **recherche observationnelle**. Ainsi, si on souhaite étudier l'impact du tabagisme sur le risque de cancer du poumon, dans la mesure où ce sont les patients eux-mêmes qui ont décidé de fumer ou non, l'étude est bien observationnelle. Le caractère observationnel de l'étude ne suffit pas à garantir son innocuité pour le patient : même si le cours du soin n'est pas modifié, il se peut que le patient soit soumis à des examens cliniques ou paracliniques potentiellement dangereux, comme une épreuve d'effort (rarement mais potentiellement mortelle), une imagerie exposant aux rayons X, une prise de sang, ou même un questionnaire. Les **questionnaires eux-mêmes peuvent être dangereux** d'une certaine manière : si on pose trop de questions sur le suicide à un patient ayant une personnalité paranoïaque, il est possible que cela augmente sa propension à se suicider.

Il existe des situations intermédiaires qu'on trouve notamment dans les **designs quasi-expérimentaux**, qui sont les designs **ici-ailleurs** et **avant-après**. Ces designs permettent d'obtenir des conclusions comparables aux études interventionnelles (avec moins de certitude), tout en restant dans un cadre observationnel.

Design **ici-ailleurs** : on peut comparer un service de soins dans lequel un traitement est systématiquement proposé, à un autre service de soins qui propose un autre traitement.

Design **avant-après** : on peut également étudier les prises en charge dans un service, puis ces mêmes prises en charge après mise en place d'un nouveau protocole.

Dans tous les cas, si les traitements proposés (études interventionnelles et designs quasi-expérimentaux) et les examens complémentaires proposés (études observationnelles) correspondent aux bonnes pratiques actuelles, on parlera de **recherche portant sur le soin courant**. Les résultats obtenus seront intéressants, mais n'atteindront pas le même niveau de preuve que les essais randomisés contrôlés, car les patients n'auront pas été affectés purement aléatoirement à un bras ou l'autre.

Les distinctions que nous avons tenté de faire dans ce paragraphe n'affectent pas l'information principale à retenir : dès qu'il s'agit d'une recherche impliquant la personne humaine (même un simple questionnaire), une autorisation du **CPP** (comité de protection des personnes) est indispensable. Une autorisation de la **CNIL** est également requise.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.4 Etude portant sur des professionnels de santé

Comme souvent dans la loi française, les professionnels de santé sont moins protégés que les autres personnes[3]. Il en est de même lorsque les professionnels de santé sont le sujet de l'étude.

Ainsi, lorsque les professionnels de santé sont l'objet de l'étude, l'étude entre dans le cadre des **évaluations de pratiques professionnelles**, et nécessite une autorisation de la **CNIL** mais pas d'autorisation du CPP. Les conditions suivantes sont requises :

- Les professionnels sont le sujet exclusif de l'étude
- Aucune donnée recueillie ne concerne leur propre état de santé
- Aucune donnée recueillie n'implique des données individuelles de patients

3 Identifier les autorisations nécessaires

3.1 Quelle autorisation pour quelle étude ?

La Figure 9 reprend en la simplifiant la Figure 8. Voici les autorisations nécessaires en fonction du type réglementaire d'étude :

- Études portant sur des données bibliographiques : aucune autorisation n'est requise (hormis en cas de partage de données pour certaines méta-analyses)
- Etudes portant sur des données : autorisation CNIL
- Etudes portant sur des personnes (hors professionnels de santé) : autorisation du CPP et de la CNIL
- Etude portant sur des professionnels de santé : autorisation CNIL

Nous détaillerons par la suite ces organismes et les principales motivations.

L'objectif est simplement de vous présenter ici les grands cadres CNIL et CPP. Il existe en réalité des sous-types d'études, et des procédures plus ou moins complexes associant d'autres organismes. La CNIL et le CPP restent cependant les principaux points d'entrée.

[3] Par exemple, les coordonnées des professionnels de santé sont largement diffusées par l'Assurance Maladie et réutilisées par des annuaires à but lucratif, sans que ces professionnels puissent s'y opposer.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Figure 9. Autorisation nécessaire en fonction du type réglementaire

3.2 Protection des données et CNIL

3.2.1 Enjeu de la protection des données

Un premier enjeu réglementaire est de protéger les données de santé des patients, car ces données sont sensibles. Voici des exemples de données codées qu'on peut trouver sur des patients dans des bases de données nationales telle la base nationale du PMSI (programme de médicalisation des systèmes d'information), dans une thématique grivoise :

- Des codes diagnostiques CIM10 :
  - F522 Échec de la réponse génitale, insuffisance érectile[4]
  - F524 Éjaculation précoce
- Des codes thérapeutiques CCAM :
  - JHLA002 Pose d'une prothèse pénienne hydraulique sans composant extra-caverneux
  - JHAA004 Allongement du pénis par section du ligament suspenseur
- Des codes de dispositifs médicaux LPP :
  - 3174580 Implant pénien expansible par mécanisme hydraulique
- Des codes de médicaments ATC :
  - G04BE08 Tadalafil[5]

On notera que des codes de natures diverses permettent de déduire les diagnostics (actes, médicaments, prothèses, etc.) : l'absence de code diagnostique ne suffit pas à empêcher de découvrir les pathologies des patients.

Naturellement, on peut trouver de nombreux autres codes, dont la divulgation serait plus grave et pourrait avoir de nombreux impacts :

- Nuire à une personnalité publique (ragots, chantage, extorsion de fonds)
- Nuire à un proche (dénigrement, etc.)
- Empêcher l'accès au crédit bancaire
- Augmenter le coût des assurances
- Barrer l'accès à un recrutement, une promotion ou l'attribution de nouvelles responsabilités
- Léser les intérêts immobiliers et financiers (viager, etc.)
- Etc.

[4] Kiki tout mou
[5] Un cousin du Viagra®

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Les bases de données (BDD) qui décrivent des individus, qu'elles contiennent ou non des données de santé, peuvent être classées en trois catégories : anonymes, nominatives et indirectement nominatives.

Les **bases de données nominatives** comprennent directement tout ou partie du nom et du prénom, ou certaines informations personnelles identifiables[12]. Une seule des informations suivantes est suffisante pour que la base de données soit nominative : un numéro de sécurité sociale, un numéro de titre d'identité, un numéro de compte bancaire, un numéro de téléphone, un courriel, ou une image du visage. Certaines combinaisons d'informations moins précises sont également considérées comme nominatives.

Les **bases de données anonymes** sont celles qui ne permettent pas la réidentification des individus. Or, la plupart des bases de données que l'on croit être anonymes sont en réalité **indirectement nominatives** : une combinaison de variables de notoriété publique permet de réidentifier les individus. Nous l'illustrons dans le chapitre suivant. En France, pour la Loi, dès qu'une base de données comporte un enregistrement par individu, elle est au moins indirectement nominative : elle ne doit jamais être considérée comme anonyme.

3.2.2 Réidentification dans les BDD indirectement nominatives

La réidentification d'individus suit principalement le scénario suivant. Il ne s'agit pas de partir de certaines informations très précises (ex : taux de potassium dans le sang, pathologie) pour déduire des informations publiques (ex : âge). Il s'agit au contraire de partir d'informations de notoriété publique (ex : âge, sexe, code postal, date d'hospitalisation) pour découvrir des informations médicales confidentielles (ex : maladie, actes, médicaments ; voir Figure 10).

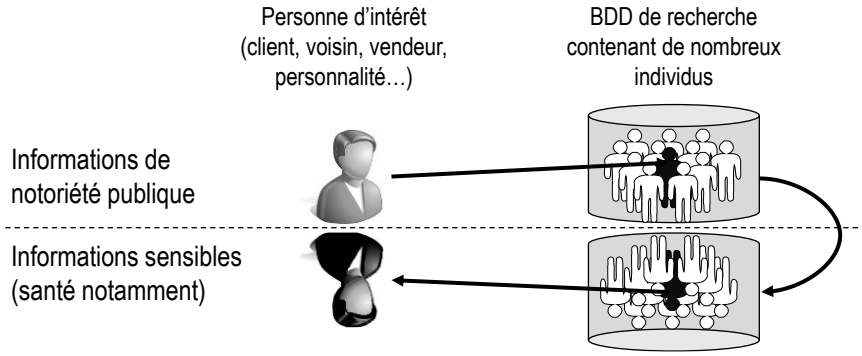


Figure 10. Scénario d'identification dans une BDD indirectement nominative

Cette réidentification est d'autant plus aisée que la base de données d'intérêt est de grande portée, et que son périmètre est précisément connu (ex : tous les patients hospitalisés dans un hôpital précisé).

L'exemple fictif suivant (Tableau 2) s'intéresse à la réidentification d'une personne que nous connaîtrions, dans le SNDS, base nationale de l'Assurance Maladie ouverte à certains chercheurs. On observe que 5 informations très simples sont suffisantes pour identifier cet individu dans une grande base de données.

Tableau 2. Exemple fictif de réidentification d'un individu dans une base nationale

Informations de départ (cumulées)	Nombre de correspondances
Aucune information	67 000 000
Il a 22 ans	760 000
C'est un homme	320 000
Il vit en Indre-et-Loire (37)	3200
Il est né le 22 janvier	9
Il a été hospitalisé cette année	1

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

La Figure 11 illustre la rapidité de décroissance du nombre de correspondance (une échelle logarithmique est nécessaire). On observe qu'il est très facile d'identifier des individus à partir d'informations de notoriété publique. Ces informations sont connues de leurs proches, mais peuvent désormais aisément être retrouvées sur Internet, qu'il s'agisse d'informations diffusées par l'individu (ex : réseaux sociaux, *curriculum vitae*, etc.) ou par des tiers (ex : Legifrance pour les concours ou nominations de la fonction publique, Assurance Maladie pour les professionnels de santé, etc.).

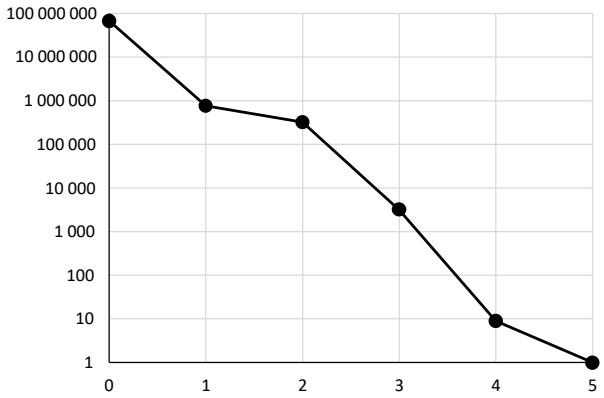


Figure 11. Courbe de réidentification
(x=nombre d'informations ; y=nombre de correspondances, échelle logarithmique)

3.2.3 La CNIL

La **CNIL, Commission Nationale de l'Informatique et des Libertés**[13–15], est l'organisme chargé de veiller sur la confidentialité des données. Elle a été créée en 1978. La Loi Informatique et Liberté a ensuite inspiré le **RGPD, règlement général de protection des données**, à la suite de quoi la Loi française a été mise en conformité avec le RGPD[16].

Une autorisation explicite de la CNIL est obligatoire pour les recherches **sur les personnes, sur les données, et sur les professionnels de santé** (en-dehors des méthodologies de référence, évoquées plus bas).

Une déclaration CNIL comportera les éléments permettant à la CNIL d'accorder ou non l'autorisation de recueil, traitement, conservation et partage des données.

Les objectifs et la portée de l'autorisation concernent en particulier :

- Le recueil de l'**autorisation** des sujets
  - Le plus souvent par ***opt in*** (le sujet est expressément d'accord pour participer)
  - Parfois par ***opt out*** (par défaut, le sujet est d'accord, mais peut s'y opposer expressément) : ces cas restrictifs sont définis par la Loi
- L'interdiction de collecter certaines données **sensibles** (ethniques, religieuses, politiques…)
- La pseudonymisation ou la gestion des **identités** en fonction des besoins de l'analyse
- La **parcimonie** : limitation au minimum nécessaire de la collecte des données utiles à la recherche
- La sécurité du **stockage**
- L'absence d'**échange** de données, ou la justification et la sécurisation de ces échanges
- L'absence de recoupement ou **appariement** de données, ou l'encadrement strict de cela
- L'absence de **diffusion** des données individuelles en-dehors du traitement
- L'intensité de l'**agrégation** des données anonymes destinées à être publiées
- La **destruction planifiée** des données individuelles une fois le traitement terminé

La CNIL délivre des autorisations *ad hoc*, ou permet dans certains cas d'adhérer à des méthodologies de référence pour simplifier et accélérer les traitements.

3.2.4 Les méthodologies de référence

Les **MR, méthodologies de référence**[17,18], proposent des cadres standardisés pour certains traitements de données. Si un organisme adhère à une MR, alors chaque recherche entrant dans une MR est autorisée par défaut, sous réserve d'être validée et enregistrée par le Correspondant Informatique et Liberté (**CIL**), qui est généralement le *Data Protection Officer* (**DPO**) ou délégué à la protection des données (**DPD**) de l'organisme. Ces méthodologies évoluent, et nous citerons seulement deux exemples.

La 4ème méthodologie de référence, ou **MR004**[17], s'intéresse à la réutilisation de données des dossiers de patients d'un établissement de soins (ex : cabinet de groupe, hôpital). Si l'étude poursuit une finalité de santé publique, l'accord des participants est obtenu par *opt out*.

La 5ème méthodologie de référence, ou **MR005**[18], s'intéresse à la réutilisation de données du SNDS (système national des données de santé, incluant le SNIIRAM, géré par l'Assurance Maladie) et de la base nationale du PMSI (programme de médicalisation des systèmes d'information, gérée par l'ATIH). Les traitements sont réalisés à des fins de recherche uniquement, à distance sur des plateformes informatiques mises à disposition par l'ATIH et l'Assurance Maladie.

3.3 Protection des personnes et CPP

3.3.1 L'enjeu de la protection des personnes, en bref

Il est nécessaire de protéger les sujets contre les **effets individuels néfastes** que pourrait avoir un protocole de recherche sur eux, tout en gardant à l'esprit les **effets bénéfiques collectifs** que pourrait avoir ce même protocole pour les autres individus, qui bénéficieront des connaissances ainsi acquises, que ces connaissances soient favorables ou défavorables au traitement évalué. Rappelons que les recherches menées sur les sujets peuvent leur nuire individuellement, que ce soit par le traitement qui leur est assigné, ou par les examens complémentaires qui seront réalisés, y compris un questionnaire (cf. supra), ou même simplement le fait d'être inclus dans un protocole et d'entendre régulièrement parler d'un sujet.

Pour ces mêmes raisons, si une recherche est légitime, le nombre de sujets à inclure dans cette recherche, appelé « **nombre de sujets nécessaires (NSN)** », nécessite là aussi d'être pris en considération.

Si un certain nombre est suffisant pour atteindre l'objectif fixé, il n'est pas nécessaire d'en inclure plus, et donc il n'est pas éthique de soumettre inutilement un nombre supplémentaire de sujets aux risques inhérents à la recherche.

Inversement, si un certain nombre est nécessaire pour atteindre l'objectif fixé, il n'est pas souhaitable d'en inclure moins, car on n'atteindrait pas l'objectif fixé, et alors tous les patients inclus auraient été soumis à un risque inutilement.

Cette notion sera détaillée dans la section 8 Calculer le nombre de sujets nécessaires en page 54.

Dans le même ordre d'idée, toute **faiblesse méthodologique** de l'étude menacerait la recevabilité scientifique de ses résultats, auquel cas les sujets auraient inutilement été exposés à un risque. La faiblesse méthodologique d'une étude est donc indirectement un risque éthique, au même titre qu'avoir inclus trop peu de sujets dans une étude.

On voit ainsi que les aspects éthiques liés à la recherche sur les personnes impliquent en particulier :

- De savoir si oui ou non une recherche peut être conduite et sous quelles conditions

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- De savoir combien de sujets doivent participer à cette recherche
- De s'assurer de la rigueur méthodologique de la recherche planifiée

3.3.2 Les CPP

L'autorisation d'un **Comité de Protection des Personnes** (CPP)[19] est indispensable avant toute recherche impliquant la personne humaine (hors professionnels de santé) : que cette recherche soit observationnelle ou interventionnelle, qu'il s'agisse d'une recherche à haut risque ou simplement sur le soin courant, et même s'il s'agit d'un simple questionnaire.

Le CPP se prononcera ainsi sur 3 points notamment :

- Le protocole est-il recevable du point de vue de **l'éthique et de la sécurité des patients** ?
- Le **nombre de sujets nécessaires** (NSN) est-il correctement évalué ?
- La **rigueur méthodologique** de la recherche est-elle au niveau de l'état de l'art ?

Les documents soumis au CPP éclaireront son jugement de manière très large, jusqu'aux ajustements prévus en cas de difficultés de recrutement, aux conflits d'intérêts des investigateurs, etc.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

4 Enquête quantitative ou qualitative ?

Les **études quantitatives** visent à **calculer des quantités** (ex : proportions, moyennes, etc.). Toutes les études de réutilisation de données, ou utilisant un questionnaire, sont quantitatives. Seules les études quantitatives (et les études qui en tirent parti) apportent une preuve généralisable à la population. Parmi les études quantitatives, les **enquêtes quantitatives** utilisent des **questionnaires** pour interroger des personnes.

Les **enquêtes qualitatives** visent à **lister des options**, sans calculer de quantité.

Le caractère qualitatif ou quantitatif d'une étude **ne tient pas à l'objet étudié**, mais bien à la famille de **méthodes employées**. Une étude peut évaluer un phénomène subjectif à l'aide de questionnaires comportant des questions binaires ou des échelles de Likert. Exemple : « indiquez, de 1 à 5, votre accord avec l'affirmation suivante : "je trouve mon métier utile à la société" ». Cette étude est alors quantitative, bien que le phénomène étudié soit hautement subjectif. Inversement, il serait possible d'étudier un phénomène quantifiable par une étude qualitative. Exemple : « Comment percevez-vous votre taille corporelle par rapport aux sièges conducteur des véhicules du marché ? ».

/!\ Le texte qui suit vise uniquement à comparer les **enquêtes quantitatives** *et les* **enquêtes qualitatives***, qui toutes deux interrogent des individus. Nous en excluons les autres études quantitatives et des études bibliographiques.*

Les **enquêtes quantitatives** s'appuient généralement sur des questionnaires structurés, avec des questions majoritairement fermées. Elles impliquent la saisie des réponses dans un tableau qui sera ensuite analysé statistiquement, avec au minimum des calculs de moyennes ou de proportions. Grâce à l'inférence statistique, elles seules permettent d'extrapoler les observations réalisées à toute la population étudiée.

Les **enquêtes qualitatives** suivent divers protocoles, un exemple typique étant l'**entretien semi-dirigé**. Ces entretiens sont volontairement ouverts, et visent à recueillir un panel large de réponses. Ils sont généralement réalisés sur un petit nombre d'individus, en recherche de variance maximale, c'est-à-dire en espérant recueillir des avis très variés, très différents, et sans aucune attente d'obtenir un avis moyen et central. L'objectif est généralement de lister toutes les options possibles. Le faible nombre d'individus, et surtout le caractère non-aléatoire de leur sélection, interdit tout calcul de moyenne ou de fréquence.

Une enquête qualitative seule a rarement un intérêt. Inversement, une enquête quantitative basée sur un questionnaire qui aurait été construit sans enquête qualitative prendrait le risque d'omettre de poser les bonnes questions. Le **design mixte**[20,21] est idéal. Sa forme la plus naturelle consiste à mener une enquête qualitative peu formelle pour lister les options possibles, mieux comprendre un sujet et éviter les oublis, puis construire une enquête quantitative pour apporter une réponse scientifique à une question. Dans ce cas, **l'enquête qualitative préalable permet d'améliorer la pertinence de l'enquête quantitative**, mais c'est bien l'enquête quantitative qui, seule, peut répondre à la question scientifique. Dans des processus itératifs, il est également possible de réaliser les enquêtes qualitatives pour expliquer des résultats intermédiaires et suggérer des améliorations de l'enquête quantitative.

Le Tableau 3 compare deux types d'enquête typiques : les enquêtes quantitatives par questionnaire, et les enquêtes qualitatives par entretien semi-dirigé.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 3. Comparaison d'enquêtes qualitatives et quantitatives typiques

	Enquête quantitative par questionnaire	**Enquête qualitative par entretien semi-dirigé**
Effectif	Des centaines d'individus	6-12 personnes
Support	Questionnaire fermé (cases à cocher…)	Guide d'entretien, ouvert
Qui mène ?	Auto-administré : rempli par le sujet Hétéro-administré : rempli par l'enquêteur, face au sujet	Entretien mené par enquêteur de manière à « faire parler » le sujet
Durée	Idéalement max. 5 minutes par sujet	Souvent une demi-heure ou une heure par sujet
Objectif stratégique	Répondre à une question scientifique	Préparer ou interpréter une enquête quantitative
Objectif opérationnel	Estimer des moyennes, des proportions…	Lister toutes les options possibles, comprendre

A partir de ce point, cet ouvrage s'intéressera uniquement aux études quantitatives.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

5 Déroulement d'une étude quantitative

Les études quantitatives suivent toutes un déroulement similaire (voir Figure 12).

La plupart des études utilisent un questionnaire (partie gauche de la Figure 12). Dans ce cas, l'analyse de la bibliographie complétée par des réunions d'experts ou une enquête qualitative permet de concevoir un questionnaire. Ce questionnaire est utilisé pour interroger des personnes (qu'il soit auto-administré ou hétéro-administré), ou réutiliser des données en demandant à un opérateur de relire des dossiers et compléter lui-même le questionnaire.

D'autres études tirent parti de bases de données existantes (partie droite de la Figure 12). Ces bases de données ne sont pas directement analysables : les données sont trop complexes (plusieurs tables) et trop abondantes au regard de la question posée. Dans ce cas, l'analyse de la bibliographie complétée par des réunions d'experts ou une enquête qualitative permet de concevoir le procédé d'extraction de caractéristiques (transformation et simplification de données).

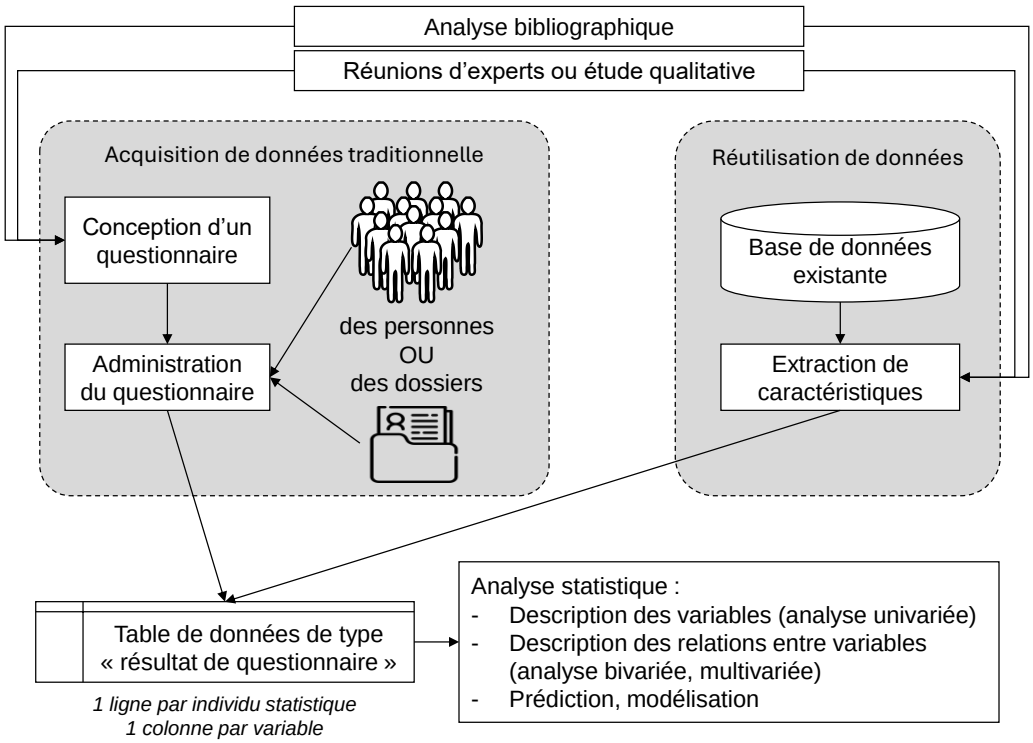


Figure 12. Déroulement typique d'une étude quantitative

Dans tous les cas (partie basse de la Figure 12), l'étude produit un tableau de données comportant essentiellement **une ligne par individu** statistique (une personne, une consultation, un séjour hospitalier…) et **une colonne par variable**.

Ce tableau est ensuite soumis à l'analyse statistique. De manière systématique, une **analyse univariée** décrit chaque variable, une par une. Dans de nombreux cas, cette analyse est déjà suffisante pour répondre à la question scientifique. Viennent ensuite des **analyses bivariées**, qui analysent le lien entre variables deux à deux, et des **analyses multivariées**, qui utilisent plusieurs variables en même temps. Ces analyses bivariées et multivariées peuvent suivre un objectif descriptif ou prédictif, selon l'objectif de l'étude.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6 *Designs* les plus fréquents en étude quantitative

Nous présenterons ici les designs les plus fréquents, pour les études quantitatives portant sur des personnes ou des professionnels de santé.

En pratique, cette classification peut s'avérer manichéenne, dans la mesure où la plupart des études sont en réalité composites. Cette classification a principalement une vocation pédagogique, et il est utile de chercher à identifier de quel cadre votre étude se rapproche le plus. Cette classification est fortement marquée par un historique « papier & crayon » des études épidémiologiques historiques, dans lesquelles peu de variables étaient recueillies, et peu d'analyses numériques étaient réalisées. En pratique de nos jours, l'ensemble de données recueillies permettent de nombreuses analyses opportunistes, rendant plus floue la classification proposée.

Cette présentation est simpliste, et ne liste que les designs les plus fréquents. Elle porte la marque du passé. Cette section ne doit pas être considérée comme un cours spécialisé sur les designs.

Nous suivrons la hiérarchie suivante :

- Etudes observationnelles
    - Etudes descriptives
        - Etude transversale
        - Etude longitudinale
    - Etudes analytiques
        - Cohorte
            - Cohorte prospective
            - Cohorte historique
            - Etude pronostique
        - Exposé-non-exposé
        - Cas-témoin
- Etudes interventionnelles
    - Etudes non-comparatives
    - Etudes comparatives
    - Etudes quasi-expérimentales

6.1 Etudes observationnelles

Les études observationnelles, rappelons-le, sont celles dans lesquelles le cours du soin n'est pas modifié par l'étude, en termes d'**actes thérapeutiques** (entretien à visée thérapeutique, manipulation à visée thérapeutique, médicaments, chirurgie, etc.). Cela n'empêche pas que des **actes diagnostiques** supplémentaires soient réalisés pour l'étude (questionnaire, examen clinique, examen paraclinique, etc.). Ces actes diagnostiques peuvent parfois être dangereux (ex : épreuve d'effort chez certains patients), c'est pourquoi il ne faut pas penser que les études observationnelles sont toujours moins risquées que les études interventionnelles. On distingue (artificiellement) les études descriptives, qui ne s'intéressent qu'à des variables indépendamment les unes des autres (ex : quelle est la prévalence du tabagisme ? quelle est la survie après un diagnostic de cancer du poumon ?), des études analytiques, qui recherchent le lien entre deux variables, généralement une exposition et une maladie (ex : le tabagisme augmente-t-il le risque de cancer du poumon ?).

Cette distinction est en réalité artificielle, dans la mesure où dès qu'on dispose de plusieurs variables, on peut aisément réaliser des analyses bivariées. Inversement, toutes les analyses statistiques commencent par une phase d'analyse descriptive univariée.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6.1.1 Etudes descriptives

Une étude descriptive vise prosaïquement à mesurer l'importance en population d'un phénomène. Exemples :

- distribution d'une variable (ex : taille des enfants en fonction de l'âge)
- fréquence d'un caractère (ex : prévalence d'une maladie)
- risque instantanée de survenue d'un événement (ex : incidence d'une maladie, risque de décès chez les personnes présentant une maladie)

Contrairement à ce qu'on pourrait penser, les études descriptives sont **difficiles et coûteuses** à mettre en œuvre, notamment car tout biais de sélection des individus entache fortement la faculté à généraliser les résultats de l'échantillon vers la population.

6.1.1.1 Etude transversale

Une étude transversale tente principalement d'établir **une mesure à un instant donné** ou, faute de mieux, sur une période contrôlée. Exemples :

- Quel est votre poids ? => étude de la distribution d'une variable
- A telle date précise, êtes-vous fumeur ? => estimation d'une prévalence
- Actuellement (en espérant contrôler les dates de passage du questionnaire), êtes-vous fumeur ? => estimation d'une prévalence
- Sur telle période précisée, avez-vous été hospitalisé ? => estimation d'une incidence. Il faut noter dans ce cas que la durée de la période est la même pour toutes les personnes interrogées, et que la réponse peut être entachée d'un biais de mémorisation.

6.1.1.2 Etude longitudinale

Une étude longitudinale tente principalement de mesurer l'**évolution d'un phénomène durant une période de suivi**. Cela suppose d'inclure les individus (à une date variable selon les individus, pour des raisons organisationnelles) et de les suivre un certain temps. Certaines méthodes d'analyse utilisées (ex : analyses de survie) supporteront très bien d'avoir des **durées de suivi variables** selon les individus : il faudra simplement consigner cette information.

Exemples :

- Durant le suivi, quel est l'évolution de votre hémoglobine glyquée ? => analyse d'une donnée fonctionnelle
- Durant le suivi, observe-t-on une première hospitalisation ? => estimation d'une incidence (risque instantané, événement absent ou considéré comme unique)
- Durant le suivi, observe-t-on des épisodes d'hypoglycémie ? => estimation d'une fréquence d'événements dans le temps (risque instantané, événements répétés)

6.1.2 Etudes analytiques

Une étude analytique vise prosaïquement à évaluer le lien entre une exposition et un phénomène de santé (ex : tabac → cancer du poumon). En réalité, cela se terminera presque toujours par étudier le lien entre de nombreux facteurs et un phénomène d'intérêt (ex : tabac + âge + sexe + amiante → cancer du poumon), c'est pourquoi cette classification est un peu manichéenne à l'épreuve de la réalité.

Les principaux types d'études analytiques sont représentés en Figure 13.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

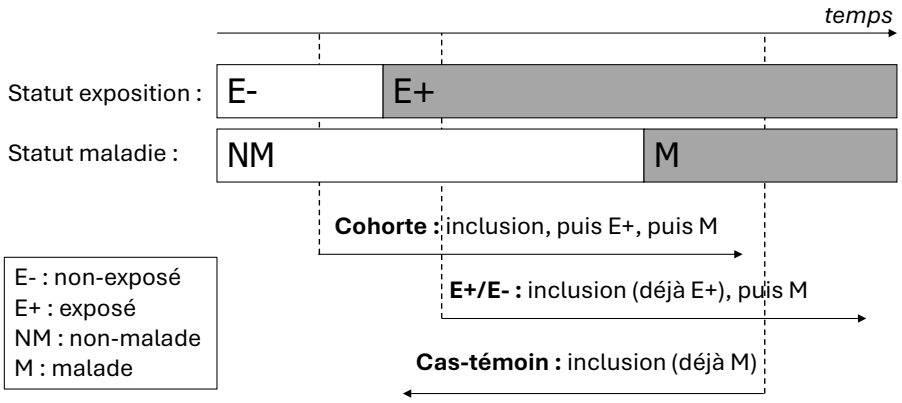


Figure 13. Principaux types d'études analytiques (exemple d'une personne exposée puis malade)

6.1.2.1 Cohorte

6.1.2.1.1 Cohorte prospective

Dans les études de cohorte prospective, les individus sont inclus et suivis dans le temps (Figure 13). Leur inclusion se fait sans tenir compte de l'exposition qu'on étudiera. Dans le cas typique, l'individu n'est pas encore exposé, ou on ne connait pas encore son statut vis-à-vis de l'exposition.

Ex : on inclut 300 adolescents de 12 ans (non-fumeurs, non-cancéreux). On les suit durant 30 ans. Certains d'entre eux commenceront à fumer. Puis certains d'entre eux développeront un cancer du poumon.

Les informations relatives à l'exposition et à la maladie sont toutes découvertes au fil du suivi.

Ces études sont prospectives, très longues et très coûteuses, car il faut financer les salaires des personnes qui prendront régulièrement des nouvelles des patients inclus.

Ces études permettent d'estimer le risque d'être exposé, puis le risque de devenir malade, pour toute la population dont l'échantillon est issu. Elles incluent par essence une connaissance précise des décès et des perdus de vue. Ces connaissances seront prises en compte dans l'estimation de tout risque dépendant du temps.

Ces études permettent donc de calculer des taux de prévalence et d'incidence, un **risque relatif** de l'exposition pour la maladie, et un **odds ratio** entre l'exposition et la maladie (des indicateurs seront explicités par la suite).

Ces caractéristiques sont représentées dans le Tableau 4.

6.1.2.1.2 Cohorte historique

Les études de cohorte historique correspondent à l'émulation rétrospective d'une cohorte (Figure 13). Ce sont en quelque sorte des **cohortes virtuelles sur dossier**. Elles nécessitent de disposer d'**un recueil préexistant, qui avait été prospectif**. Il peut d'agir de dossiers de patients hospitalisés, de registre d'infirmerie scolaire, de dossier militaire, etc. Si ce dossier est la source exclusive d'information, il s'agit alors d'une **recherche sur les données**.

Leur inclusion se fait sans tenir compte de l'exposition qu'on étudiera. Dans le cas typique, l'individu n'est pas encore exposé, ou on ne connait pas encore son statut vis-à-vis de l'exposition.

Les études de cohorte historique partagent les avantages des études de cohorte prospective, sans les inconvénients : elles sont peu coûteuses, très rapides à réaliser (il faut simplement le temps de retrouver et colliger les informations) et donc peu onéreuses, en rapport au nombre d'individus.

Leur principale faiblesse est celle de la réutilisation de données : la finalité du recueil n'étant généralement pas la même que celle de la recherche, il se peut que les informations recueillies à l'époque ne soient pas suffisamment précises pour la recherche menée aujourd'hui.

Ces caractéristiques sont représentées dans le Tableau 4.

6.1.2.1.3 Etude pronostique

Les études pronostiques sont un **cas particulier de cohorte** (prospective ou historique). Cette distinction n'est pas méthodologique mais plutôt liée à l'intérêt médical et à la portée clinique des preuves qui en découlent. Certains trouvent cette distinction un peu *old school*. Pour cette raison, les études pronostiques ne sont pas séparées des cohortes sur la Figure 13 et dans le Tableau 4.

L'étude pronostique s'intéresse généralement à une catégorie précise de personnes (ex : les patients qui ont eu telle chirurgie ; les patients chez lesquels on vient de diagnostiquer un cancer du poumon, etc.). Elle les suit dans le temps, s'intéresse à la survenue d'un événement secondaire (ex : décès, complication, etc.), comme dans une étude descriptive longitudinale. Ensuite, sans avoir précisément fixé un seul critère d'exposition, elle recherche des facteurs de risque de l'événement en question (comme les études de cohorte). Il existe donc plusieurs différences (subtiles) avec une cohorte traditionnelle :

- La « maladie » devient un critère d'inclusion et non un critère d'intérêt, et c'est une « deuxième maladie » qui devient un critère d'intérêt.
- L'exposition n'est pas partie intégrante du design de l'étude : cette étude recherchera opportunément tous les facteurs de risque ou protecteurs parmi les variables disponibles
- Dans la mesure où les expositions ne sont pas expérimentales, la portée décisionnelle de ces études est moindre que celle des études interventionnelles (ce que certains journaux scientifiques étiquettent, de manière simpliste, comme « design non-comparatif »)

6.1.2.2 Exposé-non-exposé

L'étude exposé-non-exposé (notée E+/E-) consiste à recruter côte-à-côte deux cohortes : une cohorte d'individus dont on sait déjà qu'ils présentent l'exposition à étudier, et une cohorte d'individus dont on sait déjà qu'ils ne présentent pas l'exposition à étudier (Figure 13). C'est donc une cohorte en deux sous-groupes. Le suivi est ensuite prospectif, et on découvrira quels individus développent la maladie.

Ex : on inclut 300 adultes de 30 ans (non-cancéreux) : 150 fumeurs et 150 non-fumeurs. On les suit durant 15 ans. Certains d'entre eux développeront un cancer du poumon.

Ces études commencent plus tard dans le cours de vie des sujets, et donc atteignent le même objectif que les études de cohorte, mais plus rapidement et pour un coût moindre.

Ces études ne permettent pas de connaître l'incidence de l'exposition, car la prévalence de l'exposition a été fixée par le protocole (ex : 1/2 ou 1/3). Ceci est aussi un avantage, car si l'exposition est très rare, cela permet de « booster » la proportion d'individus exposés, par rapport à une cohorte. Cela peut permettre de recruter moins d'individus au total. L'incidence de la maladie est toujours étudiable séparément chez les exposés et les non-exposés, mais il ne faut pas considérer l'incidence de la maladie dans l'ensemble de l'étude. Ces études incluent les décès et les perdus de vue, mais elles ignorent sans doute les formes très précoces et très agressives de la maladie étudiée.

Comme les études de cohorte, ces études permettent de calculer un **risque relatif** de l'exposition pour la maladie, et un **odds ratio** entre l'exposition et la maladie (des indicateurs seront explicités par la suite).

Ces caractéristiques sont représentées dans le Tableau 4.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6.1.2.3 Cas-témoin

Les études cas-témoin proposent une approche rétrospective très différente des premières (Figure 13). Le principe est de recruter séparément un groupe de malades et un groupe de non-malades. Ces personnes sont ensuite interrogées pour retrouver dans le passé leur éventuelle exposition. Un fort biais de ces études est le **biais de mémorisation**. Elles sont par exemple déconseillées pour retrouver les facteurs de risque des démences, ou pour retrouver des facteurs de risque inconnus du sujet (ex : composition précise des aliments consommés 10 ans plus tôt).

Ces études sont rétrospectives, réalisées par interrogatoire, et donnent donc des résultats très rapidement et pour un coût généralement modéré.

Lorsque la maladie est rare, cela permet de « booster » la proportion d'individus malades (on choisira généralement arbitrairement 1/2 ou 1/3). La prévalence de la maladie n'est donc plus étudiable. La prévalence de l'exposition, retrouvée par interrogatoire, est donc interprétable respectivement chez les malades et chez les non-malades, mais ne l'est plus dans l'étude entière.

Une grande faiblesse de ces études est qu'elles ignorent totalement les personnes mortes de la maladie et plus généralement les personnes dont le suivi aurait été interrompu dans une cohorte (perdus de vue). Elles sont souvent entachées d'un fort **biais de sélection** : seuls les survivants sont inclus.

La seule quantité calculable est l'**odds ratio**, qui sera défini plus tard. Il est strictement interdit d'estimer le risque relatif dans de telles études.

6.1.2.4 Synthèse

Le Tableau 4 propose une synthèse comparative des études observationnelles analytiques précédemment citées.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 4. Comparaison des principaux types d'études observationnelles analytiques (sous réserve de biais)

		Cohorte prospective	Cohorte historique	Exposé / non exposé	Cas-témoin
Informations sur exposition et maladie	**Inclusion**	Ø	Ø	E+/E-	M/NM
	Interrogatoire ou enquête	Ø	E+/E- et M/NM	Ø	E+/E-
	Au fil du suivi	E+/E- puis M/NM	Ø	M/NM	Ø
Etude	**Temporalité**	Prospective	Rétrospective	Prospective	Rétrospective
	Durée	Très longue	Courte	Longue	Courte
	Coût	Très élevé	Faible	Elevé	Faible
Population	**% exposés**	Conforme	Conforme	Fixé par protocole	Distordu
	% malades	Conforme	Conforme	Distordu	Fixé par protocole.
	Inclut les morts et perdus de vue	OUI	OUI	OUI	NON
Statistiques calculables	**Taux prévalence & incidence exposition**	OUI	OUI	NON	Séparément chez M ou NM
	Taux prévalence & incidence maladie	OUI	OUI	Séparément chez E+ ou E-	NON
	Risque relatif	OUI	OUI	OUI	NON
	Odds ratio	OUI	OUI	OUI	OUI

6.2 Etudes interventionnelles

Les études interventionnelles sont des études durant lesquelles le cours normal du soin est modifié intentionnellement par le protocole d'étude, en termes d'interventions thérapeutiques. En clair, lorsque les patients sont inclus, cette inclusion est susceptible de modifier la manière dont on les soigne, et cette décision est imposée par l'étude (ex : tirage au sort).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Compte tenu de leur coût, du parcours réglementaire et du temps requis, il n'est généralement pas conseillé de mener seul une de ces études dans le cadre d'un mémoire académique en santé.

6.2.1 Etudes non-comparatives

Les études interventionnelles non-comparatives sont généralement des études descriptives ou pronostiques (voir chapitre 6.1.2.1.3 Etude pronostique page 45) dans lesquelles l'inclusion implique la réalisation d'une intervention pour tous.

Ex : on réalise une nouvelle intervention chirurgicale sur 20 patients. On observe les suites chirurgicales.

Ces études n'étant pas comparatives, leur portée en termes de connaissances est limitée, et elles ont de fortes chances de ne pas être autorisées, sauf cas particulier.

D'une certaine manière, les phases 1 des essais thérapeutiques entrent dans ce cadre. Il en est de même des études pilotes, qui visent à acquérir de nouvelles connaissances comme préalable à des recherches plus poussées et structurées.

6.2.2 Etudes comparatives

6.2.2.1 Principe général

Ces études visent à comparer deux groupes (ou plus), qui sont généralement :

- Un bras ayant subi l'intervention thérapeutique qu'on souhaite évaluer
- Un bras « comparateur » ou « contrôle » qui n'a pas subi l'intervention à évaluer. Nous verrons que différentes options sont alors possibles.

On parle alors d'**essai contrôlé**, au sens où il existe un bras contrôle.

6.2.2.2 Classification selon le type d'intervention

Si l'intervention relève des actes thérapeutiques réalisables par certaines professions de santé (ou proches) réglementées, on parle d'**essai thérapeutique**. Ces actes peuvent être par exemple :

- L'administration d'un ou plusieurs médicaments
- La réalisation d'un vaccin
- La réalisation d'un acte de chirurgie précisé
- La réalisation d'un soin infirmier protocolisé
- Une manipulation spécifique réalisée par un kinésithérapeute
- Une technique d'entretien mise en œuvre par un psychologue
- Etc.

Si l'intervention peut être réalisée en-dehors de toute profession réglementée, on parlera d'**intervention de santé publique** :

- Séances de sport
- Modification du régime alimentaire pour une personne non-malade
- Autosurveillance à l'aide d'un procédé quelconque non-médical
- Entretiens de soutien avec des bénévoles associatifs
- Etc.

Cette classification a une importance en termes **réglementaires**, mais pas en termes méthodologique ou statistique.

6.2.2.3 Classification selon le bras contrôle

Le bras contrôle peut être de différentes natures.

Absence d'intervention :

Cette attitude est **déconseillée** car le bras intervention sera alors le seul à bénéficier de l'**effet placebo**. L'effet placebo est lié au conditionnement : la plupart des sujets ont vécu de nombreuses fois une amélioration de leurs symptômes après la prise d'un médicament. Cette association répétée conditionne leur cerveau au point que les administrations d'un produit inactif ont également un effet thérapeutique partiel[6]. De même, ce bras bénéficiera de l'**effet cocooning** : le simple fait d'être écouté et pris en charge améliore déjà réellement certains symptômes (ex : douleur, troubles du sommeil, anxiété, constipation, etc.). De plus, inversement, le fait de ne pas subir d'examen complémentaire diminue les chances de détecter des complications.

Prise en charge similaire, sans l'intervention spécifique :

Il peut être proposé au groupe contrôle de subir la même prise en charge, exception faite de l'acte à évaluer. Cette prise en charge comprendra par exemple soins infirmiers, hospitalisation factice de même durée, protocole de suivi identique, etc. Cette attitude est valable, mais permet de mesurer l'efficience de l'intervention, sans mesurer sa supériorité. En outre, cette attitude peut amener à croire que cette intervention en particulier est efficace, alors qu'il n'en est rien. Ainsi par exemple, si la sophrologie améliore la qualité de vie des patients cancéreux, cela ne valide en rien les fondements pseudo-scientifiques de la sophrologie. Il vaudrait mieux comparer la sophrologie à des séances non-spécifiques de relaxation. Pour évaluer un médicament, cette attitude peut être améliorée par l'**administration d'un placebo** dans le bras contrôle. Cette option permet de **conserver l'aveugle** (nous reviendrons sur cette notion).

Intervention déjà validée ou couramment réalisée :

La meilleure option consiste à comparer le bras d'intervention à une intervention déjà validée ou réalisée couramment. C'est ce qui est fait dans la plupart des essais thérapeutiques médicamenteux : le nouveau médicament est comparé à la molécule de référence dans le traitement de la maladie dont il est question. Cette attitude, très sévère, est la meilleure pour évaluer la supériorité du nouveau traitement. Elle permet aussi d'évaluer si on le souhaite sa non-infériorité, qui peut être appréciée si le traitement est mieux toléré ou moins cher. Lorsqu'il s'agit de médicaments, cette approche permet également le plus souvent de **conserver l'aveugle** au même titre que le placebo évoqué précédemment (nous reviendrons sur cette notion). Enfin, cette approche est **plus éthique** car elle ne laisse aucun participant sans traitement.

Il est fréquent qu'il soit difficile de choisir le bras contrôle et qu'on souhaite néanmoins évaluer les effets de chaque étape de la prise en charge. Il est alors possible de définir **plus de deux bras**, à condition de déterminer ces bras de telle manière que les résultats soient comparables et interprétables.

6.2.2.4 Classification selon la modalité d'affectation à un bras

Pour être crédibles, les études interventionnelles comparatives doivent affecter leurs sujets dans un bras ou l'autre par **randomisation** (tirage au sort). Il existe différentes méthodes, selon qu'on souhaite garantir strictement ou non les effectifs dans chaque bras. On parle d'**essai randomisé contrôlé** (RCT, *randomized controlled trial*).

En l'absence de randomisation, on se retrouverait quasiment dans le cas des études observationnelles, où l'affectation à un bras est essentiellement le fait du biais d'indication :

[6] L'effet placebo peut être très important : dans les douleurs des métastases osseuses, il diminue de 10% ces douleurs, qui sont très importantes et aucunement psychologiques. L'effet placebo n'est pas lié aux convictions du sujet ni à ses connaissances du médicament, il relève du pur conditionnement. Cet effet existe également chez les nourrissons et les animaux.

les patients reçoivent telle ou telle intervention en fonction de leur probabilité d'y survivre… Cela rendrait les résultats ininterprétables.

Il est très important que la randomisation soit réalisée **après l'inclusion** : le patient ne doit pas pouvoir refuser ou accepter de participer à la recherche en fonction du bras qu'on lui propose.

6.2.2.5 Classification selon la connaissance du bras (aveugle, etc.)

Une intervention est réalisée **sans aveugle** si tout le monde (le patient, les soignants, le statisticien) sait dans quel bras se trouve chaque patient.

On parle de **simple aveugle** lorsque le patient lui-même ignore dans quel bras il se trouve. Ceci n'est pas toujours possible. Pour l'évaluation d'un médicament, ce simple aveugle peut être obtenu par l'utilisation d'un placebo ou d'un traitement de référence. Pour l'évaluation d'une chirurgie, cela supposerait la réalisation d'une intervention factice, dont l'acceptabilité dépend de la lourdeur de la chirurgie. Pour l'évaluation d'une intervention autre, cela peut dépendre des connaissances propres du patient (ex : certains savent distinguer une séance de thérapie cognitive et comportementale d'une séance de psychanalyse, d'autres non).

On parle de **double aveugle** lorsque, en plus, les investigateurs ignorent dans quel bras le patient se trouve. Cela permet d'éviter qu'ils influencent le patient, ou qu'ils biaisent, volontairement ou involontairement, le suivi et la réalisation des examens d'évaluation. Le double aveugle est le standard qu'on souhaite atteindre dans la plupart des essais cliniques.

On parle de **triple aveugle** lorsque, en plus, les personnes qui analysent les données, connaissent un numéro de bras auquel les patients sont affectés, sans savoir à quoi correspond ce bras. Ceci est un peu théorique car, en cas de différence importante de résultat entre les bras, le statisticien devinera assez vite lequel est lequel. D'autre part, le protocole d'analyse étant écrit à l'avance, le statisticien ne peut pas le modifier même s'il sait quel est le bras à évaluer.

6.2.3 Etudes comparatives quasi-expérimentales

Les études quasi-expérimentales peuvent permettre d'obtenir des conclusions comparables aux études interventionnelles (en moins bien), tout en restant dans le cadre réglementaire des études observationnelles.

Design **ici-ailleurs** : on peut comparer un service de soins dans lequel un traitement est systématiquement proposé, à un autre service de soins qui propose un autre traitement.

Design **avant-après** : on peut également étudier les prises en charge dans un service, puis ces mêmes prises en charge après mise en place d'un nouveau protocole.

Dans les deux exemples cités, l'étude n'est pas considérée comme interventionnelle du point de vue du patient, au sens où le traitement qui est proposé au patient est le traitement standard du moment, qui est proposé à tous les patients du même service remplissant les mêmes conditions. Cependant, du point de vue des résultats attendus, ces résultats sont réellement comparatifs.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

7 Questionnaires : taux de sondage, taux de réponse

Dans les études sur questionnaire, qui sont un cas particulier d'études quantitatives, la question du taux de réponse est très importante pour apprécier le caractère biaisé ou non des résultats, et permettre leur généralisation de l'échantillon vers la population.

7.1 La théorie, pure, belle et inapplicable

La finalité d'une étude est toujours d'observer un fait dans un échantillon, et de l'extrapoler à une population. En termes mathématiques et stricts, voici comment cela se passe :

- Si l'échantillon est égal à la population, alors la quantité mesurée dans l'échantillon est celle qu'on souhaite connaître en population
- Si l'échantillon est un peu plus petit que la population, il est possible d'extrapoler de manière simple et déterministe (ex : on connait le sexe de 90% des personnes, et la moitié d'entre eux sont des hommes : alors la proportion d'hommes en population est forcément comprise entre 45% et 55%)
- Si l'échantillon est beaucoup plus petit que la population :
    - Si l'échantillon est issu de la population par **alea strict** (ex : tirage au sort), alors l'inférence statistique s'appuie sur un ensemble de méthodes qui permettent d'extrapoler le résultat obtenu dans l'échantillon à la population, avec des outils tels l'estimation, les intervalles de confiance et les tests statistiques.
    - Si l'échantillon n'est pas strictement aléatoirement issu de la population, il n'existe théoriquement pas de méthode d'inférence.

Cet algorithme, s'il était utilisé en pratique, amènerait toujours à conclure qu'il n'est pas possible de réaliser d'inférence de l'échantillon vers la population. Or, il faut bien trouver un moyen d'avancer !

7.2 La pratique en Santé, moins propre mais opérationnelle

Réduisons notre sujet d'étude au cas des questionnaires. Trois ensembles imbriqués peuvent être définis (Figure 14) : la population d'intérêt, l'échantillon des sondés, et l'échantillon des répondants.

Cela nous permet de définir :

- **Le taux de sondage** : proportion de personnes interrogées parmi la population
- **Le taux de réponse** : proportion de répondants parmi les sondés

La taille de l'échantillon final est donnée par l'Équation 4.

$$Taille\ échantillon = Taille\ population \times Taux\ de\ sondage \times Taux\ de\ réponse$$

Équation 4. Taux de sondage et de réponse

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

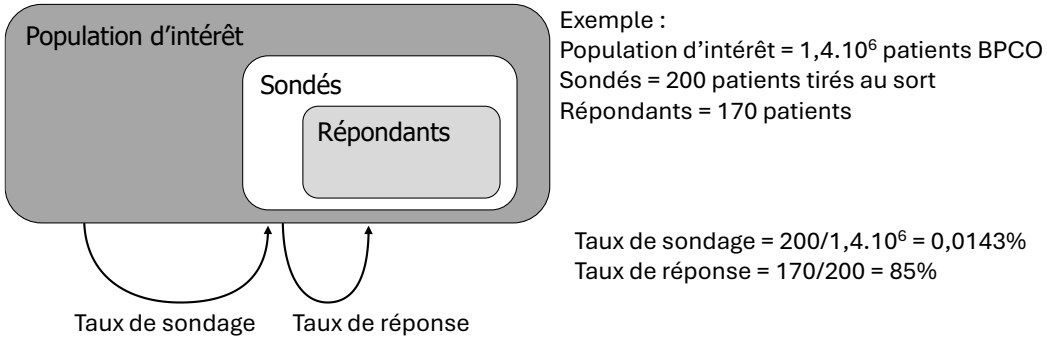


Figure 14. Taux de sondage, taux de réponse

En théorie, pour permettre une inférence statistique, le processus de sélection de la population vers l'échantillon doit être aléatoire. Voici comment cela peut s'appliquer en pratique aux deux étapes énoncées précédemment (sondage, puis réponse).

Le taux de sondage, en pratique, est presque toujours très faible (Figure 14). La seule solution est donc de s'assurer que la sélection des personnes sondées est aléatoire, ou presque. Les meilleures méthodes consistent donc à obtenir des listes exhaustives de personnes puis réaliser un tirage au sort. Ces listes dépendront de la population visée :

- Liste électorale → population des citoyens
- Annuaire téléphonique → population des résidents (groupés par foyer)
- Liste des patients de plusieurs cabinets de groupe → population des patients d'un territoire

En pratique, une première approximation est réalisée au moment de choisir la liste. Il faut ensuite réaliser un tirage au sort sur cette liste (voir le chapitre 2 Sélectionner les sondés par tirage au sort en page 68). D'autres méthodes permettent d'avoir des sélections pseudo-aléatoires, plus ou moins acceptables selon le contexte (ex : les 200 premiers patients consécutifs se présentant aux urgences à compter de telle date, en espérant que la pathologie étudiée ne soit pas saisonnière…).

Le taux de réponse, en pratique, peut être variable. S'il est faible, il n'existe aucun moyen de s'assurer que la sélection des répondants résulte d'un processus aléatoire. Un **taux de réponse faible est inacceptable**, et est jusqu'à preuve du contraire considéré comme annonciateur d'un biais de sélection. Or cette preuve du contraire ne peut jamais être apportée.

Exemple : Vous envoyez par la poste un questionnaire aux habitants sur le respect des consignes de tri des déchets. Vous obtenez 5% de réponses. Raisonnablement, on peut penser que seuls les partisans écologistes ou anti-écologistes ont répondu. En outre, les personnes qui ont le temps de répondre sont sur-représentées. Leurs réponses sont nettement moins neutres que celles de la population générale. Le biais est évident et interdit toute extrapolation des réponses des répondants vers les réponses de la population sondée.

La seule solution est donc d'**obtenir un taux de réponse élevé**, c'est-à-dire en pratique **entre 67% et 100%**. Nous verrons plus tard comment y parvenir (voir section 1 Concevoir un questionnaire en page 59).

7.3 Les échantillons représentatifs et la méthode des quotas, vraiment très sale

Les instituts de sondage ne peuvent pas appliquer les principes énoncés plus haut, car **le taux de réponse est toujours extrêmement faible**, en particulier chez les personnes qui travaillent, tout simplement car elles ne sont pas chez elles aux heures où les sondeurs ont le droit de les appeler !

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Ils ont dû mettre au point des méthodes sans lesquelles ils ne peuvent pas travailler et que nous devons considérer comme **inacceptables en recherche**.

La **méthode de l'échantillon représentatif** consiste à constituer un échantillon de sondés dont on sait d'avance qu'ils accepteront de répondre au sondage. On comprend bien que, par définition, cette sélection ne peut être aléatoire, puisque la grande majorité de la population refuse de se rendre disponible pour répondre aux sondages. Il faudra donc les rémunérer ou, pire d'un point de vue statistique, les sélectionner. Ces instituts chercheront ensuite à ne garder que certains des individus, de manière obtenir un échantillon qu'ils disent « représentatif ». Pour cela, ils feront en sorte de retrouver dans l'échantillon des sondés la même distribution conjointe de variables simples mais connues : âge, sexe, région, etc. Evidemment, cette méthode garantit que l'échantillon des sondés est « représentatif » sur les variables d'inclusion, mais aucunement qu'il le soit sur les questions qu'on s'apprête à lui poser. Le terme d'« échantillon représentatif » est, en soi, une tromperie.

La **méthode des quotas** est similaire mais pire, et s'applique en particulier aux sondages téléphoniques. Cette fois-ci, la sélection n'est pas réalisée dans la constitution de l'échantillon de sondés, mais directement d'après les caractéristiques des répondants. Au fur et à mesure que certaines personnes acceptent de répondre, on n'appelle plus que des personnes dont la catégorie est insuffisamment représentée dans l'échantillon des répondants. Evidemment, cette méthode garantit que l'échantillon des répondants est « représentatif » sur les variables d'inclusion, mais aucunement qu'il le soit sur les questions qu'on s'apprête à lui poser.

Ces méthodes ne sont pas satisfaisantes mais on comprend que les instituts de sondage n'aient pas d'autre possibilité. Cependant, à force de se défendre, ils ont réussi à faire croire à certains que leur méthode était meilleure que celle des chercheurs.

Désormais, contrairement aux journalistes, vous le savez :

1- Les échantillons représentatifs constituent une mauvaise méthode
2- Le tirage au sort en population totale, est LA méthode de référence
3- Le nombre brut de sondés n'a aucune importance, c'est le taux de réponse qui compte

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

8 Calculer le nombre de sujets nécessaires

8.1 Pourquoi calculer le NSN ?

Le **nombre de sujets nécessaires (NSN)** est calculé dans certaines études afin de déterminer combien de sujets devront finalement être inclus.

Les premiers motifs sont d'ordre **éthique** :

Si un certain nombre est suffisant pour atteindre l'objectif fixé, il n'est pas nécessaire d'en inclure plus, et donc il n'est pas éthique de soumettre inutilement un nombre supplémentaire de sujets aux risques inhérents à la recherche.

Inversement, si un certain nombre est nécessaire pour atteindre l'objectif fixé, il n'est pas souhaitable d'en inclure moins, car on n'atteindrait pas l'objectif fixé, et alors tous les patients inclus auraient été soumis à un risque inutilement.

Les motifs sont également d'ordre **économique** :

Si un certain nombre est suffisant pour atteindre l'objectif fixé, il n'est pas nécessaire d'en inclure plus, ce serait un gaspillage de ressources (financières ou en temps de travail).

Inversement, si un certain nombre est nécessaire pour atteindre l'objectif fixé, il n'est pas souhaitable d'en inclure moins, car on n'atteindrait pas l'objectif fixé, et alors toutes les dépenses engagées auraient été inutiles.

Ce calcul n'est cependant pas nécessaire pour toutes les études, comme nous le verrons.

8.2 Dans quelles études calculer le NSN ?

La Figure 15 présente les principaux cadres d'étude. En résumé, le NSN ne doit être calculé que dans les études sur des personnes humaines (RIPH), soumises à l'approbation du CPP. Dans toutes les autres études, ce calcul est inutile ou facultatif. Nous allons détailler cette conclusion.

Dans les **études bibliographiques**, en général les auteurs choisissent une période de durée « ronde », par exemple 10 ans ou 20 ans. Cette durée est choisie sur des arguments scientifiques mais aussi pour atteindre un nombre d'articles à revoir qui soit à la fois suffisamment important, mais pas trop au regard des moyens humains mobilisés. Les méthodes de calcul du NSN ne sont donc **jamais** utilisées.

Dans les études portant sur des **données préexistantes**, il n'y a aucun motif éthique poussant à déterminer le NSN, dans la mesure où les investigateurs ne sont jamais au contact de patients, et où l'étude n'est pas définition pas interventionnelle. Il peut exister cependant parfois des motifs économiques, lorsqu'il faut mobiliser des humains pour relire des dossiers papiers. Le NSN peut alors être calculé pour des motifs économiques. Mais là aussi, le raisonnement est généralement plutôt d'inclure un nombre de dossiers correspondant à une période de durée « ronde » (1 an, 5 ans, 10 ans…), et de l'adapter en fonction du temps disponible. Dans les recherches impliquant des bases de données, il n'y a aucune limitation ni éthique ni économique, et on a tendance à exploiter toutes les données disponibles. Il est très fréquent que des *reviewers* externes exigent un calcul du NSN, mais clairement ils se trompent, et font cette demande uniquement parce qu'ils ne connaissent pas les recherches sur les données. Il s'agit généralement de *reviewers* plus âgés ou issus de la recherche clinique. Il faut alors leur expliquer poliment.

Dans les études portant sur des **personnes** (hormis les professionnels de santé eux-mêmes), si ces études sont **quantitatives**, le calcul du NSN est obligatoire pour des motifs éthiques principalement, et économiques. C'est un des points sur lesquels le CPP se prononcera. Pour

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

les études **qualitatives**, on inclut généralement un nombre faible de personnes (une dizaine), et les règles permettant de déterminer ce nombre sont différentes et spécifiques de ce champ.

Dans les études portant sur les **professionnels de santé**, le calcul du NSN n'est pas obligatoire. Il peut cependant être réalisé pour des motifs économiques (et non éthiques).

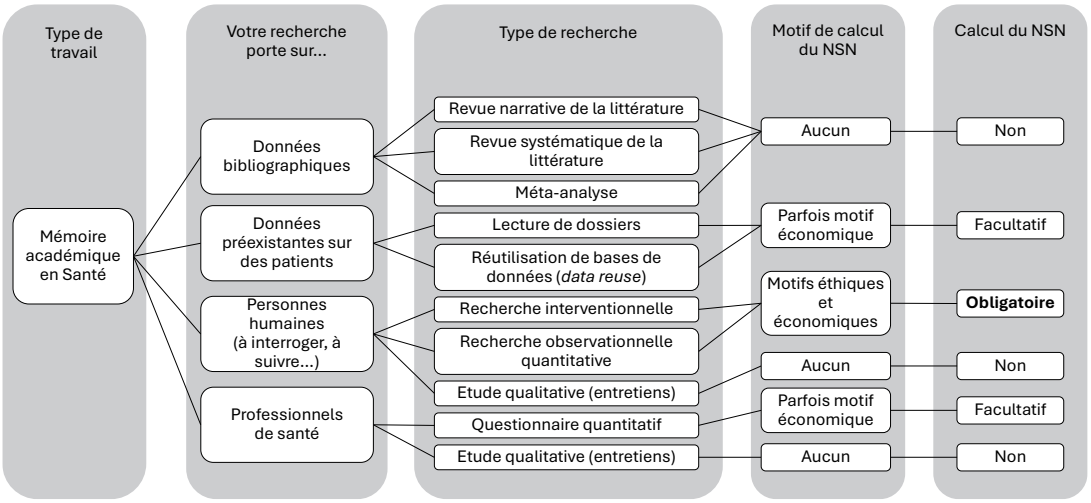


Figure 15. Dans quelles études calculer le nombre de sujets nécessaires

Nous verrons plus bas que les fondements du calcul du NSN sont discutables, et que cela suffit à disqualifier ces méthodes si elles ne sont pas obligatoires.

8.3 Méthode de calcul

Le NSN ne peut être calculé que pour une seule analyse statistique. Cela suppose donc que, dès le début, la recherche poursuivre **un seul objectif principal**, auquel sera adossée **une seule analyse**. Nous verrons ici l'exemple du calcul d'un intervalle de confiance de proportion, puis l'exemple d'un test de comparaison de deux moyennes.

Ces deux exemples sont spécifiques de deux méthodes statistiques, et il existe autant de procédés que de méthodes statistiques. Cependant, après avoir compris ces exemples, vous n'aurez aucun mal à généraliser le procédé, et surtout à comprendre en quoi ces calculs sont discutables.

8.3.1 Exemple : intervalle de confiance d'une proportion

Imaginons qu'on souhaite estimer la proportion d'un caractère en population, et son intervalle de confiance (Figure 16). Cette estimation sera faite depuis un échantillon, dont il faut fixer la taille à l'aide du calcul du NSN.

Une première étape consiste à fixer, arbitrairement, des paramètres statistiques liés à la précision de l'estimation. Il s'agit du risque alpha, généralement 5%, soit un intervalle de confiance à 95%. On décidera arbitrairement également qu'on souhaite une certaine précision de l'intervalle de confiance, par exemple 3% pour une proportion de femmes.

La deuxième étape requiert de se faire une idée assez précise de ce que pourrait être la proportion inconnue π en population. Pour ce faire, on doit s'appuyer sur la littérature scientifique ou l'expérience des investigateurs, et mitiger des résultats incertains et contradictoires.

On choisit la méthode statistique pressentie pour la future étude, et il en découle une méthode de calcul du NSN. Cette méthode tient naturellement compte de l'aléa. Elle fournit directement un NSN théorique.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Des ajustements sont ensuite réalisés, en fonction du nombre attendus de perdus de vue (ou de non-répondants, ou de questionnaires inexploitables, etc.). On tiendra compte, sans le dire, de la capacité des centres investigateurs à recruter des patients, et on ajoutera une marge de manœuvre, pour obtenir le NSN qui sera présenté dans le dossier adressé au CPP.

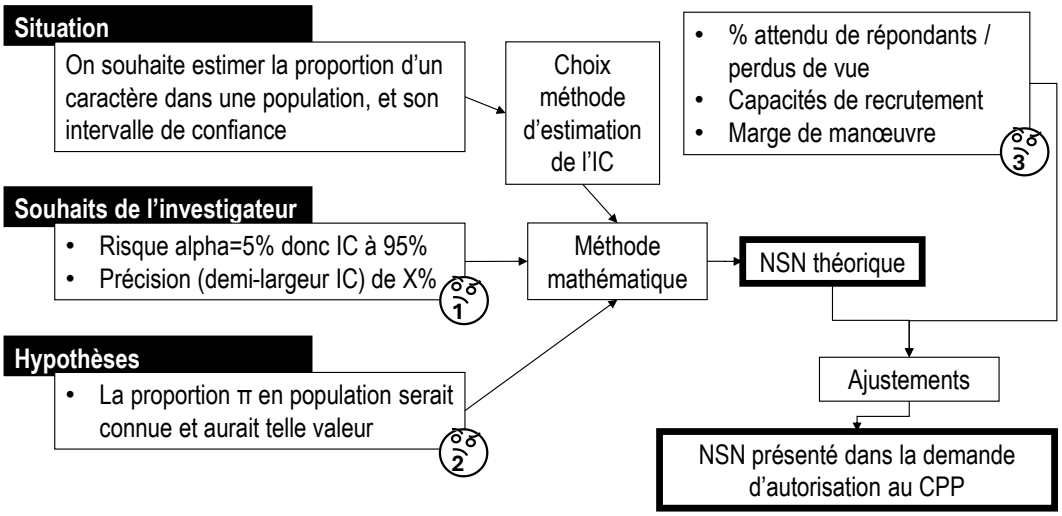


Figure 16. Calcul du NSN pour l'intervalle de confiance d'une proportion

S'il est incontestable que la méthode mathématique qui permet de calculer le NSN théorique est valide, ce procédé pose de nombreux problèmes méthodologiques.

Premièrement (1 sur Figure 16), les souhaits de l'investigateur sont totalement arbitraires. Si le 5% ou 95% est habituel en recherche clinique, aucune règle n'indique quelle doit être la précision d'un intervalle de confiance (1% ? 3% ? 5% ?).

Deuxièmement (2 sur Figure 16), le calcul doit s'appuyer sur une hypothèse de la valeur de la proportion π en population. Or cette proportion est inconnue, puisque justement l'étude vise à l'estimer. On s'appuiera sur des articles, qui tous obtiennent des valeurs différentes. Pour une étude entièrement novatrice, on ne pourra s'appuyer que sur l'intuition des investigateurs.

Troisièmement (3 sur Figure 16), la proportion de perdus de vue est spéculative. En réalité, officieusement, la personne qui calcule le NSN prendra en compte les capacités de recrutement des centres et les ressources financières disponibles. Si le résultat du NSN théorique ne lui convient pas, il suffira de modifier discrètement les paramètres d'entrée (1 et 2 sur Figure 16), pour modifier de manière très importante le NSN obtenu, mais sans que le CPP ait les moyens de contester ces données de départ, parce que certaines sont arbitraires et d'autres incertaines.

8.3.2 Exemple : comparaison de deux moyennes

Imaginons qu'on souhaite à présent réaliser un test statistique comparant deux moyennes, dans le but de prouver à partir d'un échantillon que ces moyennes sont différentes en population (ex : la taille moyenne des hommes serait différente de celle des femmes en population) (Figure 17). Ce test statistique sera réalisé depuis un échantillon, dont il faut fixer la taille à l'aide du NSN.

Une première étape consiste là aussi à fixer arbitrairement les paramètres statistiques liés à la précision de l'estimation : le risque alpha, généralement 5%, et la puissance statistique du test (probabilité, si une différence existe réellement, qu'elle soit mise en évidence), souvent 80%.

La deuxième étape requiert de se faire une idée assez précise de ce que pourraient être en population et dans chaque sous-groupe la moyenne μ et l'écart type σ (dispersion). Pour ce

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

faire, on doit s'appuyer sur la littérature scientifique ou l'expérience des investigateurs, et mitiger des résultats incertains et contradictoires.

On choisit la méthode statistique pressentie pour la future étude, et il en découle une méthode de calcul du NSN. Cette méthode tient naturellement compte de l'aléa. Elle fournit directement un NSN théorique.

Des ajustements sont là aussi réalisés, en fonction du nombre attendus de perdus de vue (ou de non-répondants, ou de questionnaires inexploitables, etc.). On tiendra compte, sans le dire, de la capacité des centres investigateurs à recruter des patients, et on ajoutera une marge de manœuvre, pour obtenir le NSN qui sera présenté dans le dossier adressé au CPP.

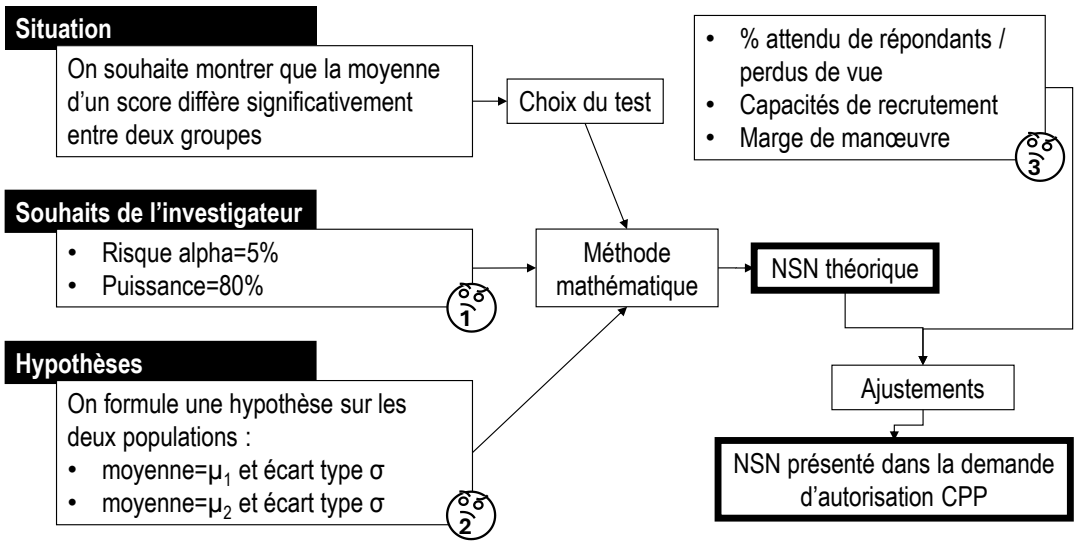


Figure 17. Calcul du NSN pour comparer deux moyennes

S'il est incontestable que la méthode mathématique qui permet de calculer le NSN théorique est valide, ce procédé pose de là encore nombreux problèmes méthodologiques.

Premièrement (1 sur Figure 17), les souhaits de l'investigateur sont arbitraires. Le risque alpha de 5% est habituel en recherche clinique, mais la puissance pourra opportunément être fixée à 80% ou 90%, car il n'y a pas de règle.

Deuxièmement (2 sur Figure 17), le calcul doit s'appuyer sur une hypothèse des valeurs en population de μ_1, μ_2, et σ. Or ces paramètres en population sont inconnus, puisque justement l'étude vise à estimer la différence entre μ_1 et μ_2. On s'appuiera sur des articles, qui tous trouvent des valeurs différentes, et les investigateurs pourront prendre ce qui les arrange. Il faut noter en particulier que des modifications de σ seront assez filoutes car discrètes mais avec un effet majeur sur le NSN obtenu. Pour une étude entièrement novatrice, on ne pourra s'appuyer que sur l'intuition des investigateurs.

Troisièmement (3 sur Figure 17), la proportion de perdus de vue est spéculative. En réalité, officieusement, la personne qui calcule le NSN prendra en compte les capacités de recrutement des centres et les ressources financières disponibles. Si le résultat du NSN théorique ne lui convient pas, il suffira de modifier discrètement les paramètres d'entrée (1 et 2 sur Figure 17), pour modifier le NSN obtenu, mais sans que le CPP ait les moyens de contester ces données de départ, parce que certaines sont arbitraires et d'autres incertaines.

8.3.3 En général

De manière générale, les entrées d'un calcul du NSN sont :

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- Une hypothèse sur la loi de distribution en population (c'est paradoxal car c'est justement ce qu'on souhaite ensuite estimer)
- Un paramétrage arbitraire de précision ou d'erreur : le risque alpha ou le niveau de confiance de l'intervalle, la précision souhaitée ou la puissance statistique du test
- Une méthode mathématique validée
- Une pincée d'opportunisme pour ajuster les premiers items aux résultats qu'on souhaite obtenir

Les spécialistes de la recherche clinique connaissent bien les limites de l'exercice, mais apprécient tout de même de disposer de méthodes donnant un ordre de grandeur du NSN. Inversement, les non-spécialistes de la recherche clinique tendent à « faire une fixette » sur le sujet, et à se laisser impressionner par la complexité mathématique de ces méthodes, en oubliant à quel point les données de départ sont spéculatives.

8.4 Conduite à tenir pour la plupart des mémoires académiques

Voici quelques bons arguments pour ne pas calculer le NSN dans de nombreux mémoires académiques :

- Votre étude n'entre pas dans le champ des RIPH, et vous n'avez donc pas l'obligation légale de le faire
- Vous n'avez pas d'argument scientifique consensuel pour assoir le calcul du NSN
- Vous n'avez pas accès à une cellule de soutien biostatistique

Vous pouvez donc proposer un nombre de sujets basé sur une méthode plus simple. En voici un exemple, que vous pourrez adapter.

L'auteur du travail souhaite proposer un questionnaire à des médecins, et souhaite obtenir *in fine* 80 réponses. Cet auteur est particulièrement impliqué, et expédiera un questionnaire papier, avec enveloppe de réponse prétimbrée, puis réalisera une relance téléphonique systématique. Compte tenu que le sujet est plutôt intéressant, la démarche sérieuse et le questionnaire attractif et court, on table sur un taux de réponses de 70%. Le fichier d'adresses étant plutôt fiable mais pas très récent, on s'attend à un taux de transfert par La Poste de 90%.

Calcul : 80 / 0,7 / 0,9 = 126,98

On enverra donc 127 courriers. Le coût moyen est d'environ 3€ par courrier, soit 381€. On se résigne à dépenser 400€, et on décide donc d'envoyer 133 courriers.

Nous verrons dans le chapitre suivant comment concevoir un questionnaire papier et améliorer son taux de réponse.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Recueillir les données

Nous verrons dans cette section comment acquérir et corriger des données, avant l'analyse statistique. Le début de cette section concernera les personnes qui utilisent un questionnaire, puis la fin de cette section concernera tous les lecteurs, même ceux qui n'utilisent pas un questionnaire.

La première étape pour vous sera peut-être de **concevoir un questionnaire** (voir chapitre 1 Concevoir un questionnaire en page 59). Nous l'aborderons dans la perspective d'un questionnaire diffusé en **format papier**, pour augmenter le taux de réponse, qui constitue un enjeu crucial (voir chapitre 2 Questionnaires auto-administrés : augmenter le taux de réponse des professionnels de santé en page 68).

Nous verrons ensuite comment sélectionner les personnes à interroger (voir chapitre 3 Sélectionner les sondés par tirage au sort en page 71).

Dans tous les cas, que vous utilisiez ou non un questionnaire, il vous faudra ensuite saisir les données obtenues puis les corriger (voir chapitre 4 Saisir des données en page 74 puis chapitre 5 Vérifier, corriger et recoder des données en page 87) Enfin, avant d'analyse les données, se posera la question de la gestion des données manquantes (voir chapitre 6 Gérer les données manquantes en page 93).

1 Concevoir un questionnaire

1.1 Rappel sur les types de variables

Nous reverrons plus précisément les types de variables dans les chapitres dédiés aux analyses quantitatives de données. Pour ce chapitre sur les questionnaires, vous aurez besoin de savoir identifier quelques types de variables. Ces types de variables sont présentés sommairement.

Les **variables quantitatives** correspondent à des nombres sur lesquels il est possible de réaliser des opérations algébriques. Exemple : la taille, le poids, le nombre d'enfants, le nombre d'hospitalisations, etc. Elles sont **discrètes** lorsque le nombre de modalités est fini (ex : nombre d'enfants), et **continues** lorsque le nombre de modalités est infini (ex : poids, qui peut être égal à 81,48615762…).

Les **variables qualitatives (monovaluées)** correspondent à des réponses textuelles (non-numériques) avec une seule réponse possible, parmi un nombre fini de modalités. Exemple : la couleur des cheveux (bruns, blonds, blancs…), le stade d'un cancer (1, 2a, 2b, 3, etc.).

Les **variables binaires** correspondent à des questions auxquelles il n'y a que deux réponses possibles, qu'on peut exprimer par oui/non, 0/1, vrai/faux, etc.

La **variables qualitatives multivaluées** correspondent à des questions auxquelles plusieurs réponses peuvent être simultanément apportées. Ex : « quel(s) moyen(s) de transport utilisez-vous habituellement pour aller travailler (plusieurs réponses possibles) ? à pied / à vélo / en voiture / en transports en commun ». Nous verrons par la suite que, pour le statisticien, ces réponses sont considérées comme autant de variables binaires.

Ces types de variables étant définis, nous verrons ensuite comment les recueillir à l'aide de composants de formulaire adaptés.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

1.2 Composants de formulaires

Nous exposerons plus bas pourquoi il est très fortement recommandé de recourir à des **questionnaires papier**, et non des questionnaires numériques (sur Internet par exemple) (voir chapitre 2 Questionnaires auto-administrés : augmenter le taux de réponse des professionnels de santé en page 68). Pour rester cohérent avec ce conseil appuyé, nous ne présenterons que les **composants de formulaires adaptables sur papier**, et ignorerons des composants uniquement numériques, comme les listes déroulantes.

Si vous utilisez pour votre usage personnel un questionnaire numérique (ex : vous revoyez des dossiers de patients et remplissez vous-même un questionnaire), la fiabilité des saisies et le taux de réponse ne seront pas des préoccupations pour vous, et vous pourrez très bien ignorer les recommandations de ce chapitre.

Bien que nous présentions ici des règles de mise en forme de questionnaires papier, nous utiliserons les noms de composants définis dans le langage HTML (HyperText Markup Language), qui permet de réaliser des pages internet.

1.2.1 Saisie libre

Les saisies libres sont généralement utilisées pour les **variables quantitatives** (ex : Quel âge avez-vous ? Combien de fois avez-vous été hospitalisé ?), notamment les variables quantitatives continues, ou les variables quantitatives discrètes comportant de nombreuses modalités.

Pour les **variables qualitatives**, les saisies libres devraient **rarement être utilisées** dans les questionnaires, pour deux raisons. Premièrement, une saisie libre aura plus de chances d'être de **mauvaise qualité**, et de souffrir du fait que, en l'absence de suggestion, certains répondants pensent à une réponse et d'autres n'y pensent pas, mais le regrettent par la suite. L'évocation n'est pas la même pour tous, ce qui induit une erreur de réponse. Deuxièmement, une saisie libre est souvent utilisée pour des **données directement ou indirectement nominative**, ce qui est la plupart du temps non-souhaité.

La Figure 18 illustre les composants permettant une saisie libre.

La ***textbox*** (exemples 1, 2, 3, 4 sur la Figure 18) permet une saisie mono-ligne. Elle est particulièrement appropriée pour les nombres (2 et 3). On n'hésitera pas à proposer une sorte de grille de saisie (3), ou à indiquer l'unité en commentaire et à droite de la *textbox* (2 et 3), ou à préciser le format en grisé dans la case et en commentaire (4). La *textbox* est idéale pour une variable quantitative discrète ou continue, et permet également la saisie d'une variable qualitative, avec les réserves ci-dessus.

La ***textarea*** (5 sur la Figure 18) permet une saisie multiligne. Elle permet par exemple de saisir une adresse postale ou un commentaire. On comprend qu'il sera difficile de réaliser une analyse quantitative des réponses saisies dans un tel composant de formulaire.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

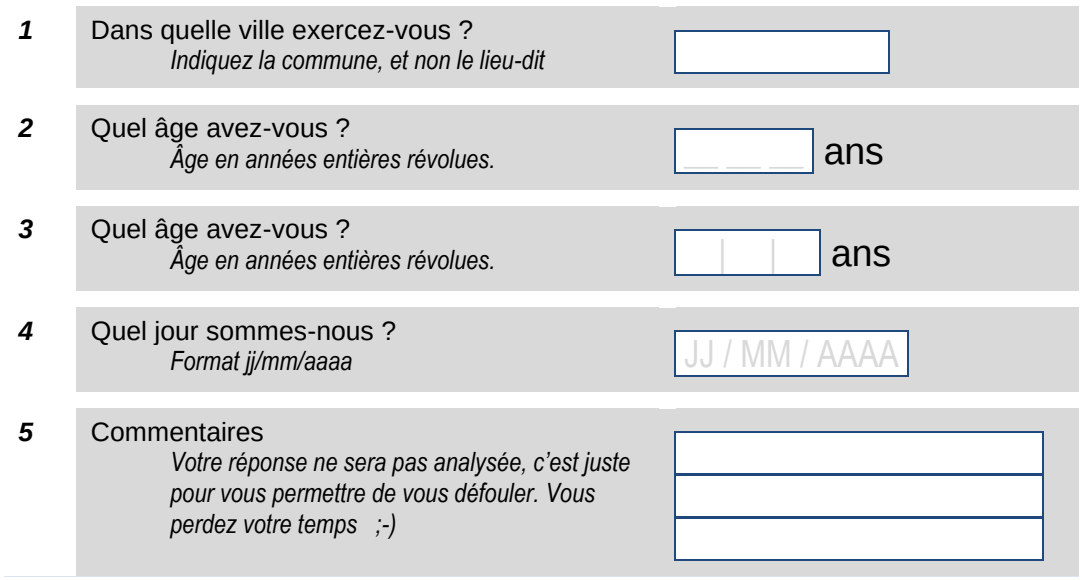


Figure 18. Composants de saisie libre : textbox (1, 2, 3, 4) et textarea (5)

1.2.2 Saisie contrainte à une seule réponse possible

La ***radiobox*** (exemples 1 et 2 sur Figure 19) est le composant idéal pour une seule réponse possible. Elle se présente sous la forme d'une bulle, par opposition aux cases à cocher. Les personnes familiarisées avec les formulaires électroniques savent que l'aspect de bulle suppose une et une seule sélection. Cela n'empêchera pas de le répéter dans la notice de chaque question. Les réponses pourront être présentées sur une même ligne ou sur plusieurs lignes, selon la place disponible.

La *radiobox* est appropriée pour les variables quantitatives discrètes à peu de modalités, les variables qualitatives (ordonnées ou non), et les variables binaires.

La *radiobox* pourrait permettre de recueillir une variable quantitative discrétisée (ex : quel âge avez-vous ? 18-40 ans, 40-60 ans, 60-80 ans) mais, à moins qu'il ne s'agisse de préserver l'anonymat, il sera alors plus approprié de recueillir la valeur exacte, et éventuellement de la discrétiser lors de l'analyse statistique.

On parle de *radiobox* (bouton radio) par évocation des boutons qui permettaient sur les postes radio de choisir le type de fréquence (modulation de fréquence, courtes moyennes ou longues ondes). Lorsqu'on appuie sur un de ces boutons, les autres boutons se relèvent automatiquement. Il n'est donc pas possible d'enfoncer plusieurs boutons en même temps. Il n'est normalement pas possible de n'enfoncer aucun bouton non plus.

L'**échelle visuelle analogique**, **EVA** (exemple 3 sur Figure 19), permet de recueillir une quantité approximative à l'aide d'une coche. Elle est rarement proposée en formulaire numérique, mais toujours efficace en formulaire papier. Elle est appropriée pour une quantification approximative, comme un niveau de douleur par exemple. L'opérateur de saisie pourra simplement mesurer la position de la coche et reporter ce nombre, avec ou sans conversion, dans la grille de saisie. De facto, même si elle est peu précise, l'EVA permet la collecte d'une variable quantitative continue.

La ***checkbox unique*** (exemple 4 sur Figure 19) permet de recueillir une réponse « oui ». C'est une option concurrente avec une *radiobox* à deux réponses, « oui » ou « non ». Elle se présente sous la forme d'un carré unique. Cette présentation consomme peu de place, mais on ne saura pas si la personne a répondu « non » ou a oublié de répondre. En pratique, elle est surtout utilisée pour le consentement, situation dans laquelle il n'est pas nécessaire de distinguer le « non » de l'absence de réponse. Elle est plutôt déconseillée pour les variables binaires qui font l'objet de l'enquête.

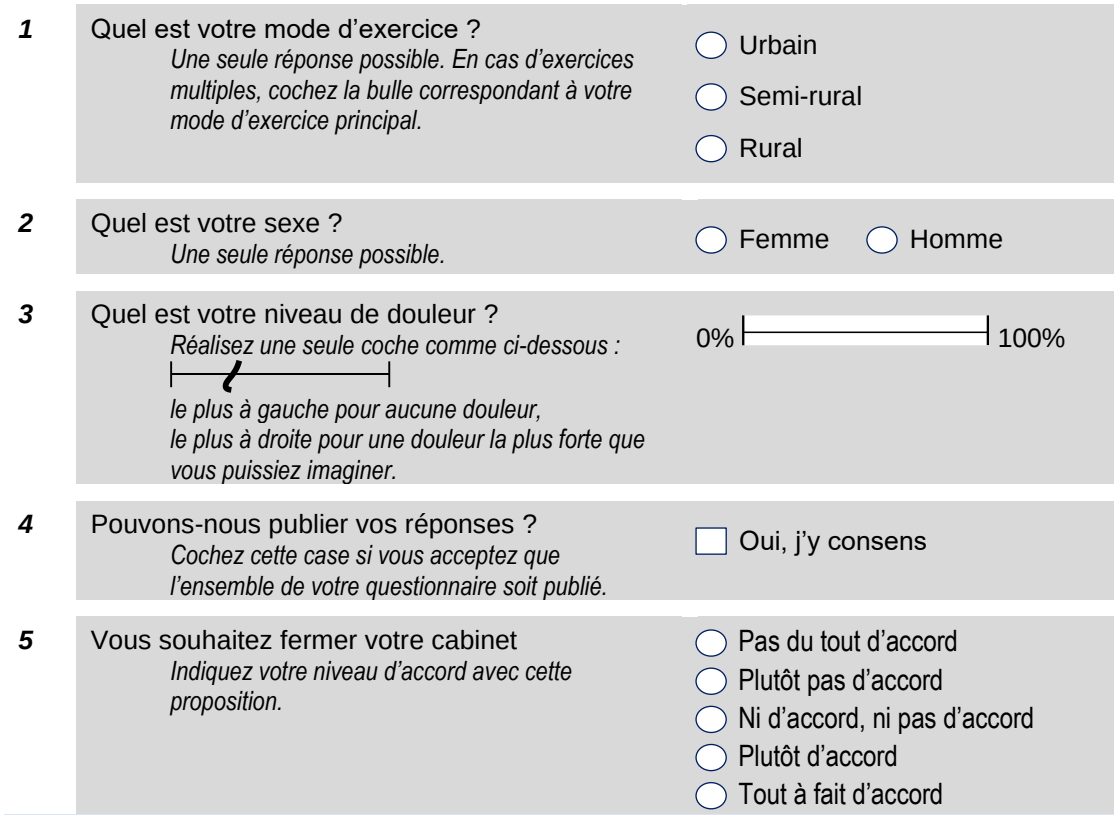


Figure 19. Composants de saisie contrainte à une seule réponse possible : radiobox (1, 2, 5), échelle visuelle analogique (3) et checkbox unique (4)

Un cas particulier d'utilisation des *radiobox* est l'**échelle de Likert** (5 sur Figure 19), du nom de Rensis Likert, psychologue étasunien. Elle consiste en une échelle semi-quantitative. Sa forme la plus fréquent est une affirmation simple, face à laquelle le participant exprime son accord ou son désaccord, allant de « tout à fait d'accord » à « pas du tout d'accord ». Cet outil est particulièrement adapté pour recueillir une opinion, un sentiment, et plus généralement pour la psychométrie. Cela n'a aucun sens de les utiliser pour mesurer un fait objectif.

Les **échelles de Likert impaires** proposent au milieu une modalité entièrement neutre, appropriée pour les répondants qui n'ont aucune opinion sur un sujet. Les échelles paires ne la proposent pas. Nous recommandons d'éviter les échelles paires, tout simplement parce que les réponses doivent correspondre à ce que pensent réellement les répondants, et que cela n'a aucun sens de « tordre le bras » des répondants.

Nous recommandons les libellés les plus fréquemment retrouvés, proposés dans cet ordre :

- Pas du tout d'accord
- Plutôt pas d'accord
- Ni d'accord, ni pas d'accord
- Plutôt d'accord
- Tout à fait d'accord

Pour ce qui est des libellés des propositions, nous recommandons d'utiliser des **libellés simples, univoques et radicaux**, permettant ainsi aisément au sujet de se prononcer. Les libellés « tièdes » ne devraient pas être utilisés : si le libellé est « j'aime plutôt Picsou », les personnes qui adorent picsou répondront « plutôt d'accord », tout comme les personnes qui l'aiment seulement un peu. On ne saura alors pas interpréter leurs réponses. Les propositions « j'adore Picsou » ou « je déteste Picsou » seront nettement préférables.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Aucun effort particulier ne doit être fait pour que certaines personnes répondent systématiquement à droite et d'autre systématiquement à gauche : il n'est pas nécessaire que les libellés soient tous formulés dans le même sens. Il faut simplement que les libellés soient intelligibles, univoques et radicaux.

Les libellés doivent également contenir une seule idée. Soit la proposition « J'adore Picsou car j'adore l'agent » : les personnes qui n'aiment pas Picsou, et celles qui l'adorent mais pour une autre raison (son intelligence et son courage par exemple), répondront la même chose. Il sera alors impossible d'interpréter la réponse.

Nous reviendrons plus bas sur la présentation des échelles de Likert (voir partie 1.4.2 Echelles de Likert en page 65).

1.2.3 Saisie contrainte à plusieurs réponses possibles

Les ***checkbox* multiples** permettent au répondant de sélectionner simultanément possiblement plusieurs réponses à une question, en saisie contrainte (Figure 20). Le répondant pourra, selon la présentation de la question, cochez zéro, une ou plusieurs cases.

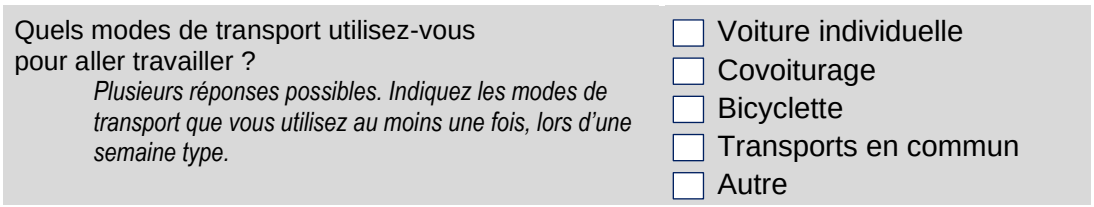


Figure 20. Composants de saisie contrainte à plusieurs réponses possibles : checkbox multiple

Plusieurs questions se posent lors de la conception d'une telle question :

- Que cochera une personne pour laquelle la question est inappropriée ?
- Que cochera une personne pour laquelle aucune des réponses ne convient ?
- Y a-t-il des modalités incompatibles entre elles ?
- Inversement, existe-t-il des modalités qui incluent une autre modalité déjà présente ? (cocher l'une devrait forcément amener à cocher l'autre)

Nous ne pourrons pas apporter de recommandation systématique, mais nous vous invitons à réfléchir aux réponses suivantes qui pourraient apparaître dans l'exemple de la Figure 20 :

- « aucune de ces réponses » : une telle modalité permettra de clairement différencier les personnes qui ont répondu à la question mais négativement, de celles qui n'ont pas répondu. Inversement, elle est par définition incompatible avec les autres, et nécessite donc un contrôle de cohérence lors de l'analyse.
- « autre » : une telle modalité permet là aussi de séparer les répondants insatisfaits des propositions, des non-répondants. Elle peut être un peu frustrante si de nombreuses personnes la choisissent, car on ne saura pas ce qui se cache derrière le libellé.
- « toutes ces réponses » : ce genre de modalité est à éviter, il vaut mieux laisser la personne cocher toutes les cases.
- « un véhicule motorisé » : cette modalité est aussi à éviter car elle inclut trois modalités déjà présentes dans cet exemple. Certains la cocheraient et ne cocheraient pas les autres, rendant l'interprétation impossible.

Il n'existe pas de bonne pratique standardisée, le bon sens doit l'emporter. Il peut également être utile de faire tester le questionnaire par des personnes tatillonnes.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

1.3 Utilisation de terminologies

Les terminologies sont des dictionnaires qui listent des libellés, ou associent des codes à des libellés. Un exemple est constitué par les catégories socio-professionnelles (CSP) de l'Insee (Tableau 5).

Autant que possible vous devez utiliser des terminologies existantes, pour les raisons suivantes :

- Les terminologies publiées font généralement suite à, au minimum, les travaux d'un groupe d'experts. Elles sont habituellement **mieux pensées** que les listes qui nous viennent spontanément (moins ambiguës, sans redondance, sans espace mort).
- Les terminologies publiées sont souvent **connues des répondants**, qui répondent avec plus de précision.
- Utiliser une terminologie connue rendra vos **résultats comparables** avec les résultats déjà publiés.

Tableau 5. Catégories socio-professionnelles de l'Insee

Code	Intitulé
AZ	Agriculture, sylviculture et pêche
BE	Industrie manufacturière, industries extractives et autres
CZ	*dont : industrie manufacturière*
FZ	Construction
GI	Commerce de gros et de détail, transports, hébergement et restauration
JZ	Information et communication
KZ	Activités financières et d'assurance
LZ	Activités immobilières
LI	*dont : loyers imputés des logements occupés par leur propriétaire*
MN	Activités spécialisées, scientifiques et techniques et activités de services administratifs et de soutien
OQ	Administration publique, enseignement, santé humaine et action sociale
RU	Autres activités de services

1.4 Conseils pour la présentation sur papier

1.4.1 Mise en page

Une mise en page soignée permet :

- d'améliorer l'**attractivité** du questionnaire et donc le **taux de réponse**
- de faire comprendre au répondant où une réponse est attendue, et donc d'améliorer la **qualité des réponses**
- de faire comprendre au répondant les éventuelles **dépendances fonctionnelles** entre questions, et donc d'améliorer la qualité des réponses

Un premier conseil est de mettre en place des **zones sémantiques**, permettant de comprendre de quoi on parle dans une question particulière. La partie gauche de la Figure 21 illustre comment les questions peuvent être regroupées en zones sémantiques, permettant de lever toute ambiguïté, et parfois de raccourcir les libellés des questions.

En cas de **dépendances fonctionnelles** entre questions, il est également conseillé de bien le faire comprendre au répondant à l'aide d'une mise en page spécifique, comme l'illustre la partie droite de la Figure 21. Cela permet d'améliorer la qualité des réponses, au regard de la cohérence bivariée entre certaines réponses et de l'exhaustivité des réponses.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

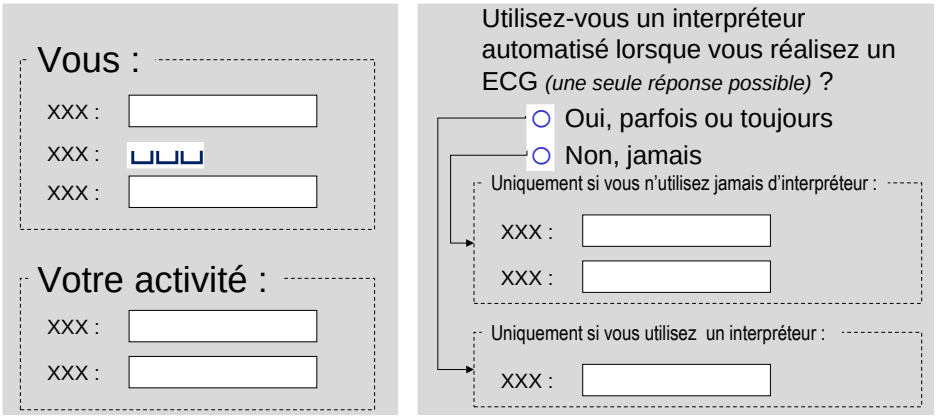


Figure 21. Exemple de formulaire implémentant des zones sémantiques (gauche) et des sauts conditionnels (droite)

Afin d'améliorer l'attractivité du questionnaire et l'exhaustivité des réponses, il est utile de bien faire comprendre au répondant à quel endroit les réponses sont attendues, en utilisant des zones de couleur, et en alignant les composants dans lesquels des réponses sont attendues. La Figure 22 montre un extrait du formulaire de déclaration de l'Impôt sur le Revenu des Personnes Physiques (IRPP, administration fiscale) : les zones de réponse apparaissent sur fond blanc, et sont clairement alignées.

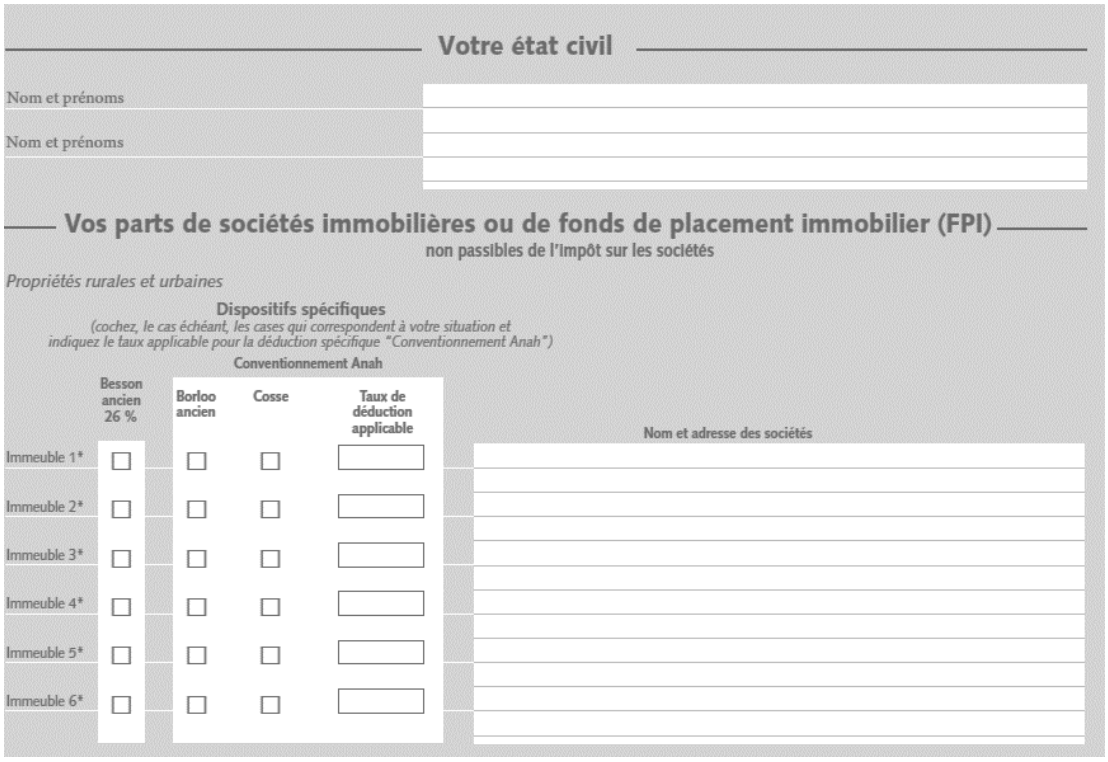


Figure 22. Formulaire IRPP : exemple d'utilisation réussie des couleurs de fond et alignements

1.4.2 Echelles de Likert

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Pour chacune des affirmations suivantes, indiquez votre degré d'accord ou désaccord en cochant une seule case par ligne.	**Pas du tout d'accord**	**Plutôt pas d'accord**	**Ni d'accord, ni pas d'accord**	**Plutôt d'accord**	**Tout à fait d'accord**
Les vaccins protègent des maladies infectieuses	○	○	○	○	○
Les vaccins sont dangereux pour la santé	○	○	○	○	○
La diffusion des vaccins vise à empoisonner délibérément la population	○	○	○	○	○
La diffusion d'opinions antivaccins est motivée par la recherche de profits financiers	○	○	○	○	○

Figure 23. Exemple d'échelles de Likert regroupées et alignées

Il peut être utile de regrouper toutes les échelles de Likert d'un questionnaire au même endroit, même si elles traitent de questions différentes. Ce regroupement permettra deux choses. Tout d'abord, la mise en page sera plus agréable et il sera aisé d'aligner les *radiobox* de réponses. Ensuite, on pourra **mutualiser les libellés** des réponses : en plus de faire gagner de la place, cela fera gagner du temps au répondant, limitant également le découragement que pourrait engendrer une trop grande quantité de texte (Figure 23).

1.4.3 Adaptation à la technique d'impression

La mise en page doit également être adaptée à la technique de reproduction du questionnaire papier.

Les **photocopieuses mécaniques** (anciennes, pas les numériques) ne gèrent que le strict noir et blanc : il faut donc éviter les couleurs intermédiaires, qui deviennent des tâches informes. Si un texte est prévu sur une zone colorée, le texte deviendra illisible.

Les **imprimantes à jet d'encre** diffusent de l'encre dans le papier. Cette encre est hydrosoluble. Il est donc possible d'écrire avec un stylo à bille sur les zones colorées. Il est possible d'écrire au feutre sur ces zones, mais c'est peu recommandé car cela risque de faire baver l'encre. En cas d'exposition à la pluie, le questionnaire sera fortement abîmé.

Les **imprimantes laser** déposent de l'encre et une couche d'enrobage hydrophobe. Le côté négatif est qu'il est généralement difficile d'écrire sur les zones colorées, car elles deviennent hydrophobes. Le côté positif est que l'impression laser résiste très bien à l'eau de pluie.

Si vous utilisez une imprimante ou photocopieuse moderne, il est donc conseillé de **laisser en fond blanc les zones de réponses**, et plutôt de colorer discrètement les zones où aucune réponse n'est attendue, comme dans les exemples montrés dans le présent ouvrage (de la Figure 18 à la Figure 23).

1.5 Mode d'administration du questionnaire

Certains questionnaires peuvent être utilisés comme outil de saisie par un investigateur, sans rencontrer d'individu. C'est le cas par exemple des **e-CRF** (*electronic clinical research form*), ou des questionnaires utilisés par des investigateurs qui réutilisent des données de dossiers patients. La plupart des questionnaires sont cependant utilisés pour sonder un individu (patient ou proche d'un patient dans les RIPH, professionnel de santé dans les évaluations de pratiques professionnelles).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Les **questionnaires hétéro-administrés** sont complétés par l'investigateur qui interroge la personne. L'interrogation peut se faire en face-à-face ou au téléphone.

Les **questionnaires auto-administrés** sont remplis par la personne qui fait l'objet de l'enquête et non par l'investigateur. C'est alors cette personne qui est également la clef du taux de réponse. Une multitude d'organisations sont possibles : questionnaire distribué à un patient lors de son admission, questionnaire en libre-service dans une salle d'attente, questionnaire envoyé par la poste, etc. Dans les questionnaires auto-administrés, la question du taux de réponse est cruciale. Nous y reviendrons dans le chapitre 2 Questionnaires auto-administrés : augmenter le taux de réponse des professionnels de santé en page 68.

1.6 Respect de l'anonymat, le cas échéant

Votre formulaire devra être conforme avec l'autorisation CNIL obtenue (voir chapitre 3 Identifier les autorisations nécessaires en page 33). Si vous déclarez que le questionnaire est anonyme, il devra vraiment l'être, et donc ne comporter aucune information directement nominative ni indirectement nominative (voir le chapitre 3.2.2 Réidentification dans les BDD indirectement nominatives en page 35). Si vous affirmez que votre questionnaire est anonyme, il ne devra comporter :

- Ni information directement nominative[12], notamment : tout ou partie du nom et du prénom, numéro de sécurité sociale, numéro de téléphone, courriel
- Ni information indirectement nominative, notamment si elles sont nombreuses et inutiles à l'étude : date et lieu de naissance, tout ou partie de l'adresse (code postal, commune, etc.), structure de soin fréquentée, date d'admission ou de consultation dans une structure de soin

Le potentiel de réidentification des personnes dépend du contexte. Si votre étude porte sur des professionnels de santé, en désert médical, la notion d'un exercice particulier (ex : médecine du sport) suffit parfois à réidentifier le médecin. Cette même information dans une métropole ne présente aucun risque.

De même, ce potentiel de réidentification est plus ou moins porteur de risque selon les données de santé qui pourraient être découvertes. Si l'étude porte sur la gestion de la douleur dans l'arthrose du sujet âgé, le risque est minime. Mais si l'étude porte sur des mineures qui réalisent une IVG à l'insu de leurs parents dans un quartier violent, l'enjeu de protection de l'anonymat est nettement plus pressant.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2 Questionnaires auto-administrés : augmenter le taux de réponse des professionnels de santé

2.1 Préambule

Nous ne pourrons pas traiter ici tous les cas de questionnaires auto-administrés (ex : papier posé dans une salle d'attente, lien à usage unique envoyé par email, questionnaire nominatif, etc.).

Un cas qui revient très souvent dans les mémoires académiques en santé, est celui des **questionnaires anonymes auto-administrés soumis à des professionnels de santé**. Deux options sont alors fréquemment retrouvées :

- Le questionnaire électronique ouvert à tous, accessible avec un lien générique (par opposition au lien à usage unique)
- Le questionnaire papier envoyé par la poste, avec retour par courrier postal

Ne faisons pas durer le suspense : entre ces deux options, nous recommandons très clairement les **questionnaires papier par voie postale**, qui seuls permettent d'obtenir des résultats valides. Nous expliquerons pourquoi par la suite.

2.2 Questionnaires auto-administrés électroniques (internet)

Nous commenterons les questionnaires électroniques dans leur application typique, qui consiste à envoyer à de nombreux professionnels de santé, par email ou sur un réseau social, un lien générique permettant à tous d'accéder à un questionnaire anonyme.

Ces questionnaires électroniques ont de nombreux avantages :

- La **mise en forme** proposée par de nombreux outils en ligne est satisfaisante
- L'**effort de saisie** est reporté sur le répondant
- Les **contraintes** de saisie sont appliquées, ainsi que les **sauts conditionnels** et, si cela est bien défini, cela permet d'augmenter la qualité des données
- La transformation en **tableau de données** est immédiate et donne des tableaux généralement de bonne qualité
- Ces questionnaires peuvent être diffusés massivement pour un **coût faible ou nul**

Dans le contexte exposé, les inconvénients des questionnaires dématérialisés sont peu nombreux mais **rédhibitoires**. Le problème de fond est qu'en pratique on ne sait pas à qui le questionnaire est envoyé, et on ne sait pas qui répond, à moins que le questionnaire soit nominatif. Cela se traduit par :

- Un taux de réponse **inconnu**
- Un taux de réponse généralement **extrêmement faible** (ex : on obtient 300 réponses mais, d'une manière ou d'une autre, on a peut-être sollicité 10 000 personnes)
- Un **biais de sélection** des répondants qui est majeur : ceux qui répondent sont plus connectés, plus intéressés ou ont plus de temps que les autres. De facto, les personnes très occupées, ou plus âgées, sont exclues de l'enquête.
- L'impossibilité de savoir si une personne répond **une ou plusieurs fois**. A l'extrême, le questionnaire peut être « trollé », c'est-à-dire intentionnellement pollué par des répondants fictifs
- L'impossibilité de **contrôler les critères d'inclusion**, dans la mesure où n'importe qui peut répondre en suivant le lien
- Dans de nombreux cas, l'absence de garantie contre le **vol de données** par l'hébergeur du questionnaire

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

La question du taux de réponse, inconnu ou très bas, suffit à disqualifier totalement ces questionnaires.

2.3 Questionnaires auto-administrés papier

Nous commenterons les questionnaires papier dans leur application typique, qui consiste à envoyer à une liste limitée de professionnels de santé un questionnaire papier par voie postale, à renvoyer à une adresse indiquée.

Ces questionnaires sont envoyés à une liste restreinte de personnes, sélectionnées par tirage au sort sur une liste plus importante (voir le chapitre 3 Sélectionner les sondés par tirage au sort en page 71).

Les questionnaires papier ont des avantages et inconvénients opposés aux questionnaires électroniques :

- Le taux de réponse est précisément **connu**
- Le taux de réponse peut être **élevé** (nous verrons juste après par quels moyens), de l'ordre de 70% à 80%
- Si le taux de réponse est élevé, le **biais de sélection** devient forcément négligeable
- Mécaniquement, chaque répondant répond **au plus une fois**
- Les **critères d'inclusion** sont généralement contrôlés lors de la préparation de la liste d'inclusion
- Les données ne peuvent être **volées** par un hébergeur

Inversement, la mise en forme du questionnaire, la saisie des données, le contrôle qualité des données sont plus fastidieux.

2.4 Améliorer le taux de réponse d'un questionnaire postal

Les facteurs permettant d'augmenter les taux de réponses des questionnaires ont été étudiés scientifiquement [23]. Ces résultats et notre expérience amènent à préconiser ce qui suit.

Tout d'abord, le questionnaire lui-même doit :

- Être imprimé **en couleurs** (oui, c'est prouvé !)
- Avoir une présentation **très soignée**
- Occuper au maximum **une seule feuille recto-verso**, ce qui offre la garantie d'un remplissage rapide ne décourageant pas le répondant

Ensuite, l'enveloppe adressée au sondé doit contenir (Figure 24) :

- Le questionnaire
- Une **lettre d'accompagnement** (là encore, c'est prouvé !)
- Une **enveloppe de retour** comportant déjà l'adresse, et **affranchie** au tarif en vigueur

La **lettre d'accompagnement** doit être polie, courte, et comporter idéalement la photographie de l'investigateur. Cela peut créer une empathie et améliorer le taux de réponse, surtout si l'investigateur est un étudiant de la même profession que le sondé.

L'**enveloppe de retour affranchie** permet d'éviter les erreurs, de diminuer le découragement lié à l'enchaînement des tâches rébarbatives, mais sert également à entraîner une empathie : si l'investigateur est un étudiant et a accepté de dépenser le montant de l'affranchissement, le sondé pourra se sentir obligé de répondre.

A ce propos, il faut noter que la solution d'enveloppes T de La Poste n'est pas adaptée à ce type d'enquête : cette solution est intéressante lorsque le nombre d'envois est très important, mais qu'on espère une très faible proportion de retours (on paiera uniquement les retours effectifs). Dans notre cas, c'est exactement l'inverse : le nombre d'envois est modéré mais on espère avoir tout autant de retours.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Ainsi, le coût total d'un envoi (papèterie, impression, double affranchissement) est actuellement de l'ordre de **3 euros par personne sondée**. Pour respecter ce coût, le poids total de l'ensemble devra atteindre au plus **20 grammes**. On veillera à utiliser par exemple :

- Une enveloppe DL 11x22 en papier 80g/m² pour l'aller (plus le poids de l'étiquette autocollante et du timbre)
- La même enveloppe, pliée en trois, pour le retour (avec là aussi une étiquette et un timbre)
- Une feuille A4 recto-verso en papier 80g/m² pour le questionnaire
- Une feuille A5 recto en papier 80g/m² pour la lettre d'accompagnement. On pourra imprimer 2 lettres par feuille A4, puis couper les feuilles au massicot

Un dernier point est très important : ce sont les **relances téléphoniques** qui améliorent le taux de réponse. Si votre questionnaire est anonyme, alors les relances téléphoniques devront être systématiques.

D'expérience, on peut envoyer les courriers, attendre deux semaines, puis procéder aux relances téléphones, renvoyer le courrier à certains interlocuteurs qui le demanderont, et attendre encore deux semaines. Il est inutile d'attendre plus que cela.

Toujours de notre expérience, l'envoi de courriers permet d'atteindre un taux de réponses de 35%, et les relances téléphoniques permettent de montrer à 70%. L'intérêt du questionnaire, sa mise en page et le courrier d'accompagnement jouent un rôle important dans ces taux de réponses.

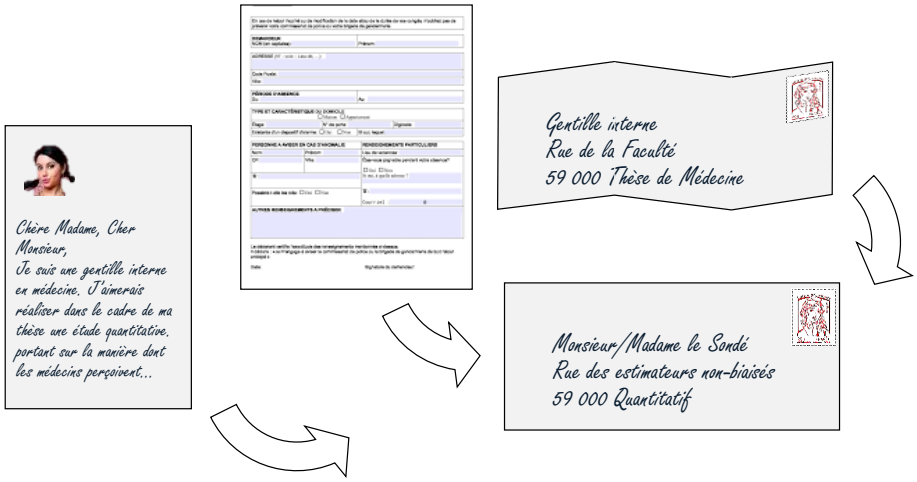


Figure 24. Envoi d'un questionnaire avec lettre d'accompagnement et enveloppe retour

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3 Sélectionner les sondés par tirage au sort

3.1 Préambule

Comme exposé dans le chapitre 7 Questionnaires : taux de sondage, taux de réponse en page 51, dès lors que le taux de sondage n'est pas égal à 100%, l'inférence statistique (tests, intervalles de confiance, etc.) ne pourra théoriquement être réalisée que si les sondés sont sélectionnés aléatoirement parmi la population d'intérêt.

Parfois, on se contentera d'un **pseudo-alea**. Par exemple, si on décide d'inclure tous les **patients consécutifs** qui se présentent aux Urgences durant toute une semaine, ce protocole est recevable car le fait de prendre « tous les patients consécutifs » prévient toute tentative de sélectionner des patients particuliers. Cependant, il faudra s'assurer que le choix de cette semaine en particulier n'ait pas d'impact non-souhaité au regard du problème étudié. Pour étudier la fréquence de phénomènes saisonniers, comme les infections respiratoires, asthmes graves ou poussées d'insuffisance cardiaque, ce serait problématique. Inversement, si on souhaite étudier des facteurs pronostiques en se restreignant à ces pathologies, c'est plutôt souhaitable.

Cependant, la plupart du temps, on aura recours à **un tirage au sort**, par exemple :

- Tirage au sort rétrospectif de dossiers parmi tous les patients pris en charge l'année passée
- Tirage au sort d'individus en temps réel à leur admission
- Tirage au sort de professionnels de santé sur la liste des inscrits dans un ordre professionnel
- Tirage au sort de professionnels de santé dans un annuaire sur internet

Nous illustrerons trois situations pratiques de tirage au sort dans cette section. Si nécessaire, on pourra recourir aux mêmes techniques si on souhaite identifier des sous-groupes au sein de l'échantillon.

3.2 Tirer au sort des éléments d'une liste finie

La situation la plus classique consiste à tirer au sort un nombre fixé d'éléments parmi une liste déjà entièrement constituée. Le procédé le plus simple (Figure 25) consiste à prendre la liste en question dans un tableur, créer une colonne aléatoire à côté des individus, puis trier le tableau par ordre croissant de cette colonne. Il suffit ensuite de prendre les premiers individus de la liste, en nombre souhaité. On notera que, à chaque opération, contrairement à ce que montre la Figure 25, les valeurs aléatoires sont recalculées. Cela n'a aucune importance, il ne faut pas s'en inquiéter car le caractère aléatoire est acquis dès l'exécution du tri du tableau.

(1) Disposer la liste complète

Nom	Prénom
Bio	Man
Hello	Kitty
Tchou	Pi
Petit Ours	Brun
Astro	Le Petit Robot

(2) Ajouter une colonne dont la formule est =alea()

Nom	Prénom	Alea
Bio	Man	0.46534
Hello	Kitty	0.076357
Tchou	Pi	0.360205
Petit Ours	Brun	0.728115
Astro	Le Petit R(	0.531419

(3) Trier par valeurs croissantes de cette colonne

Nom	Prénom	Alea
Hello	Kitty	0.076357
Tchou	Pi	0.360205
Bio	Man	0.46534
Astro	Le Petit R(	0.531419
Petit Ours	Brun	0.728115

(4) Sélectionner les n premiers éléments de cette liste

Nom	Prénom	Alea	Sélection
Hello	Kitty	0.076357	1
Tchou	Pi	0.360205	1
Bio	Man	0.46534	0
Astro	Le Petit R(	0.531419	0
Petit Ours	Brun	0.728115	0

Figure 25. Réalisation d'un tirage au sort sur liste finie, avec Excel ou Calc

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.3 Tirer au sort des éléments d'une liste finie, présentée par pages séparées

Il arrive fréquemment que la liste sur laquelle on souhaite réaliser le tirage au sort soit présentée par un organisme externe sous forme d'un ensemble de pages numérotées, mais que les éléments sur ces pages ne soient pas numérotés, et qu'il ne soit pas possible de récupérer tous ces éléments dans un tableau. C'est le cas par exemple de :

- une liste déjà imprimée sur plusieurs pages, avec un nombre relativement stable d'éléments par page
- un annuaire professionnel en ligne ou site de prise de rendez-vous nécessitant de faire défiler les pages

Le procédé est le même que précédemment, mais nécessite simplement de prendre en compte la présentation par page (Équation 5).

On commence par calculer le nombre approximatif d'individus de l'annuaire, *n* (Équation 5). On réalise ensuite un tableau comme précédemment, si ce n'est que les noms d'individus sont simplement remplacés par des numéros séquentiels allant de *1* à *n* : on identifie ainsi les numéros *i* des individus à conserver. Leur position (numéro de la page et position sur la page) est donnée par l'Équation 5. La troncature est ici l'arrondi à l'entier inférieur ou égal. Dans cet exemple, les numéros de pages ou de positions sur la page s'entendent au sens courant (1=premier, 2=deuxième, etc.).

$$p = nombre\ de\ pages$$

$$k = nombre\ d'éléments\ par\ page$$

$$n = nombre\ total\ d'éléments \approx p \times k$$

$$Numéro\ de\ page\ individu\ i = troncature\left(\frac{i}{k}\right) + 1$$

$$Position\ sur\ page\ individu\ i = i - k \times troncature\left(\frac{i}{k}\right)$$

Équation 5. Localiser un individu sur des pages d'annuaire

Exemple : vous souhaitez identifier le 26ème individu, alors que l'annuaire est présenté avec 10 individus par page. Il se trouve sur la 3ème page, en 6ème position. Si par malchance vous recherchez un individu qui n'existe pas (par exemple, sur la toute dernière page), il suffit de tirer au sort un nouvel individu.

3.4 Tirer au sort des éléments prospectivement

Parfois, la sélection aléatoire doit avoir lieu alors que la liste d'éléments n'est pas encore déterminée : il peut s'agir par exemple de sélectionner des patients se présentant aux urgences, mais d'être capable de dire pour chaque patient s'il est sélectionné ou non au moment où il se présente, alors que la liste totale des patients n'est pas constituée. Deux solutions sont possibles.

Cas n°1 : vous ne disposez pas d'un numéro d'ordre

La sélection (oui/non) peut être réalisée à l'aide de n'importe quel dispositif ayant une réponse binaire, et pouvant être calibrée en fonction de la probabilité souhaitée. Voici tout d'abord 3 exemples simples :

- Pour une probabilité à 50%, vous pouvez lancer une pièce de monnaie. Face=inclusion, Pile=exclusion.
- Pour une probabilité à 25%, vous pouvez tirer une carte dans un jeu complet. Cœur=inclusion, autres couleurs=exclusion.

- Pour une probabilité à 3/13, vous pouvez tirer une carte dans un jeu complet. 2, 3 ou 4 => inclusion, 5 et plus => exclusion (tous les multiples de 1/13 sont alors possibles).

Ce procédé peut aisément être généralisé à plus de 2 modalités si nécessaire.

Pour une probabilité plus exotique, un peut lancer une formule **=alea()** sur un tableur. Si le nombre est inférieur à la probabilité souhaitée, on inclut le patient, sinon on l'exclut.

Exemple : vous souhaitez inclure 1/6 des patients, vous pouvez écrire =alea()<1/6 dans une cellule. Si elle affiche VRAI, le patient est inclus, si elle affiche FAUX le patient est exclu.

Cas n°2 : vous disposez d'un numéro d'ordre

Si vous disposez d'un numéro d'ordre du patient, constituez sur un tableur une liste finie allant de 1 à un nombre très important, supérieur au nombre attendu. Créez une colonne comportant la formule **=alea()<proba** en remplaçant proba par la probabilité souhaitée (1/3 dans l'exemple en Figure 26). Après obtention d'une première sélection, **copiez le résultat en tant que valeur**, ou dans un fichier de traitement de texte par exemple, sinon le tableur recalculera la valeur aléatoire à chaque opération sur l'onglet de calcul. Lorsque le patient numéro *i* se présente à vous, regardez son statut dans le tableau ainsi obtenu.

Num	Sélection
1	=ALEA()<1/3
2	=ALEA()<1/3
3	=ALEA()<1/3
4	=ALEA()<1/3
5	=ALEA()<1/3
6	=ALEA()<1/3

Num	Sélection
1	VRAI
2	FAUX
3	FAUX
4	FAUX
5	FAUX
6	VRAI

Figure 26. Pré-affectation avec une probabilité d'1/3. Gauche : formules. Droite : résultats

Cette solution, à peine un peu plus complexe que la précédente, garantit que l'affectation du patient à l'échantillon ne sera pas arbitrairement influencée par l'opérateur.

Ces deux solutions présentent l'inconvénient, contrairement à celles sur liste finie, de ne pas garantir le nombre de patients inclus (ou inclus dans tel ou tel groupe). La solution suivante, elle, le permet.

3.5 Tirer au sort des éléments prospectivement, en garantissant la proportion finale

Enfin, il se peut que la situation soit comparable à la précédente, mais qu'on souhaite garantir un équilibre prédéterminé entre deux groupes. Il faut alors réaliser un **tirage au sort sans remise**. On peut alors disposer d'une urne ou d'une enveloppe contenant des petits papiers, et réaliser un tirage sans remise. Lorsqu'un papier est tiré il donne le statut du patient. Ce papier est ensuite détruit.

Cette méthode peut cependant être critiquée : les derniers patients auront une affectation déterminée par celle des premiers patients.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

4 Saisir des données

4.1 Cas d'application

Cette section peut se lire de plusieurs manières selon la source de vos données :

- Si vous avez obtenu des **formulaires papier** et devez les saisir, cette section est directement faite pour vous.
- Si vous avez obtenu des données d'un **formulaire en ligne**, les données devraient être propres : survolez cette section sans l'ignorer pour vérifier si les recommandations sont bien appliquées. Vous en tirerez quelques bénéfices à moindre effort.
- Si vous **saisissez directement** des données dans un tableur sans passer par un formulaire papier, imaginez ce qu'aurait dû être ce formulaire papier, et suivez plus scrupuleusement encore ces recommandations.
- Si vous obtenez des données par **réutilisation d'une base de données**, cette section vous indique quel modèle de données cible vos mécanismes d'extraction de caractéristiques doivent permettre d'obtenir.

La présentation proposée ici est optimisée pour l'analyse statistique qui suit : c'est celle dont a besoin un statisticien. Elle n'est pas celle qu'aurait proposé un informaticien, pour gérer et faire vivre un jeu de données normalisé : c'est intentionnel.

4.2 Principes généraux

4.2.1 Présentation générale

En-dehors des exceptions qui suivent, le recueil doit être réalisé dans une **unique table**, comportant **une ligne par individu** statistique, et **une colonne par variable** (deux pour la survie).

Les noms de variables doivent apparaître sur la première ligne uniquement. Il est strictement interdit d'utiliser des cellules **fusionnées ou fractionnées**.

Tableau 6. Exemple de tableau de données bien présenté

id_patient	age	atteinte	grade
1	63	Centrale	0
2	59	Périphérique	0
3	69	Périphérique	1
4	45	Périphérique	1
5	76	Centrale	0
6	46	Périphérique	1
7	86	Centrale	1
8	47	Périphérique	0
9	46	Centrale	2
10	89	Centrale	1
11	97	Généralisée	1
12	64	Centrale	2

4.2.2 Choix de l'individu statistique

L'individu statistique est l'entité sur laquelle porte votre étude. Cet individu statistique peut être une personne physique, un séjour hospitalier (si un même patient vient deux fois, il est alors représenté sur deux lignes), une consultation, mais aussi une dent d'une personne, un œil d'une personne, etc.

Une fois cet individu choisi, il se matérialise comme une ligne, dans un tableau unique.

La Figure 27 montre le contrexemple de deux groupes de patients, saisis dans deux tableaux différents. Ces deux tableaux doivent être réunis verticalement. Une colonne supplémentaire (colonne « deces » en Figure 27) doit être créée pour identifier le sous-groupe auquel les patients appartiennent. Les variables définies seulement dans un sous-groupe doivent bien être créées, mais les valeurs manquantes doivent être indiquées (colonnes « karnofski » et « cause_deces » en Figure 27).

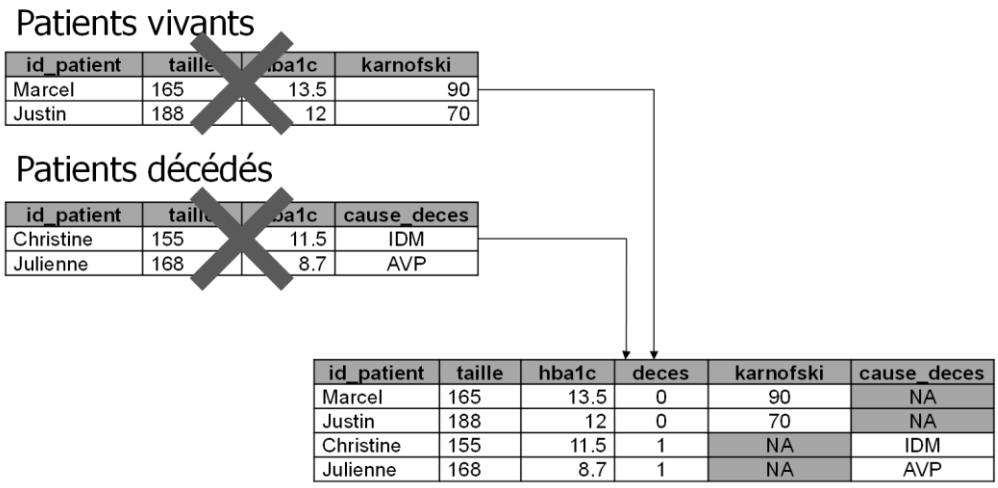


Figure 27. Contrexemple : patients saisis dans deux tableaux différents.

La Figure 28 montre le contrexemple de patients pour lesquels plusieurs consultations sont réalisées. Le tableau initial comporte une ligne par patient, mais les données des consultations (données temporelles, donc) sont répétées de gauche à droite. Le nombre de consultations peut être variable par patient, entraînant une incohérence du tableau de données. Dans ce cas, l'individu statistique est manifestement la consultation et non le patient. Il faut donc une table avec une ligne par consultation, et non une ligne par patient.

La première option, multi-table, consiste à créer une table par patient et une table par consultation. Un identifiant commun du patient permettra, si besoin, de rapatrier les informations du patient dans la consultation lors de l'analyse.

La deuxième option, monotable, suppose de renoncer à toute analyse à l'échelle du patient. Il suffit alors de créer une table comportant une ligne par consultation, et répéter les informations du patient dans cette table.

On notera que le numéro d'ordre de consultation (1, 2, 3…) quitte les noms de colonnes pour devenir une valeur de champ. On notera également que la première représentation faisait apparaître des données manquantes, mais que ce n'est plus le cas par la suite.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

id_patient	sexe	date_consult_1	hba1c_1	date_consult_2	hba1c_2
Patient_1	1	2010-02-03	8.5	2010-06-01	10
Patient_2	0	2009-06-21	10.2	NA	NA

Option 1, multi-table

*Table **data_patients** (2 patients) :*

id_patient	sexe
Patient_1	1
Patient_2	0

*Table **data_consultations** (3 consultations) :*

id_patient	num_consult	date	hba1c
Patient_1	1	2010-02-03	8.5
Patient_1	2	2010-06-01	10
Patient_2	1	2009-06-21	10.2

Option 2, mono-table

*Table **data_consultations** (3 consultations, avec informations répétées sur le patient) :*

id_patient	num_consult	date	hba1c	sexe
Patient_1	1	2010-02-03	8.5	1
Patient_1	2	2010-06-01	10	1
Patient_2	1	2009-06-21	10.2	0

Figure 28. Contrexemple : mesures répétées dans le temps

De la même manière, la Figure 29 montre l'exemple de mesures répétées sur des parties différentes du corps. Il s'agit ici de données classiques d'ophtalmologie, où les yeux peuvent être considérés comme des individus statistiques. Là aussi, il est possible de représenter les données dans deux tables, ou une table d'yeux avec répétition des informations concernant le patient (qui est ici la manière la plus classique). On notera que la latéralité de l'œil (gauche ou droite) quitte les noms de colonnes pour devenir une valeur de champ.

id_patient	sexe	date_diagnostic_og	acuite_og	date_diagnostic_od	acuite_od
Patient_1	1	1990-02-03	8	1995-06-01	10
Patient_2	0	NA	NA	2001-06-01	9

Option 1, multi-table

*Table **data_patients** (2 patients) :*

id_patient	sexe
Patient_1	1
Patient_2	0

*Table **data_yeux** (3 yeux) :*

id_patient	oeil_droit	date_diagnostic	acuite
Patient_1	0	1990-02-03	8
Patient_1	1	1995-06-01	10
Patient_2	1	2001-06-01	9

Option 2, mono-table

*Table **data_yeux** (3 yeux, avec informations répétées sur le patient) :*

id_patient	sexe	oeil_droit	date_diagnostic	acuite
Patient_1	1	0	1990-02-03	8
Patient_1	1	1	1995-06-01	10
Patient_2	0	1	2001-06-01	9

Figure 29. Contrexemple : mesures répétées sur différentes parties du corps

4.2.3 Noms des variables

Dans la table (ou les tables) de données, les noms de variables doivent figurer en première ligne, et il est interdit de fusionner ou fractionner les cellules.

Cependant, on peut imaginer que, en raison d'un grand nombre de colonnes, il soit plus aisé de saisir les valeurs dans un tableau dont les colonnes sont regroupées à l'aide de « chapeaux » (partie haute de la Figure 30). Dans ce cas, il faut bien comprendre que la ligne contenant ces « chapeaux » devra être supprimée : il est donc important que les noms des colonnes, sur la deuxième ligne, se suffisent à eux-mêmes : ils doivent pouvoir être lus sans disposer du chapeau, qui disparaîtra (partie bas de la Figure 30).

visite_anesth			visite_chirurgien		
date	nom	duree	date	nom	duree
…	…	…	…	…	…

↓devient

visite_anesth			*visite_chirurgien*		
vanesth_date	vanesth_nom	vanesth_duree	vchir_date	vchir_nom	vchir_duree
…	…	…	…	…	…

Figure 30. Exemple de tableau avec des "chapeaux", avant (haut) et après (bas) renommage

Les noms de variables doivent respecter les contraintes suivantes :

- Chaque nom de variable doit être unique
- Les noms des variables doivent, de préférence, respecter ces contraintes :
  - Composé lettres non accentuées de a à z en minuscule, et des chiffres de 0 à 9, et éventuellement l'*underscore* « _ » (« tiret du 8 »), à l'exclusion de tout autre caractère
  - Commençant par une lettre
  - Ne comportant pas, en particulier, de : caractères accentués, « e et o collés », « a et e collés », majuscules, espaces, tirets, points, ponctuations, caractères spéciaux
- Ces noms de variables suivent, traditionnellement, ces conventions :
  - Les identifiants figurent parmi les premières colonnes, les plus à gauche, et débutent par « id_ »
  - Les décomptes commencent par « nb_ »
  - Les noms de variables sont plutôt courts, intelligibles, et recourent à l'*underscore* « _ ». Exemples : date_prem_consult, date_der_consult, nb_visites…

4.2.4 Dé-identification des enregistrements

Même si vous êtes autorisés à réaliser une étude nominative, le fichier de données ne doit pas contenir d'information directement nominative (nom, adresse, téléphone, date de naissance…), ni d'information indirectement nominative (dates, structures de soins…) si ce n'est pas nécessaire à l'analyse statistique (voir 3.2.2 Réidentification dans les BDD indirectement nominatives en page 35 et 1.6 Respect de l'anonymat, le cas échéant en page 67).

Pour une étude nominative, il faudra donc stocker toutes ces informations dans un tableau conservé confidentiellement par l'investigateur, comportant un numéro arbitraire d'identification (1, 2, 3, etc.). Ce numéro sera reporté dans le tableau confié pour l'analyse statistique. Ainsi, si une erreur est découverte dans les données (ou un besoin de vérification), il sera possible de retourner au cas réel.

4.3 Détails en fonction du type de variable

4.3.1 Identifiants

Les identifiants sont le plus souvent des nombres entiers, mais pourraient également être des chaînes de caractère. Comme évoqué précédemment, si le fichier de données est amené à être partagé, par exemple pour l'analyse statistique, ces identifiants ne devraient pas être constitués d'une partie du nom ou du prénom.

Particularités :

- Les données manquantes sont interdites
- Si la table représente l'entité à laquelle l'identifiant se rapporte, l'identifiant doit être unique

- Si l'identifiant est numérique et unique, un simple diagramme en bâtons des effectifs ou un histogramme permet de vérifier l'absence de problème

4.3.2 Variables quantitatives, discrètes ou continues

Le format de saisie des variables quantitatives dépend du logiciel utilisé. Dans un tableur, si elles sont correctement saisies, leurs valeurs **s'alignent automatiquement à droite**. Tant que la valeur reste alignée sur la gauche, c'est qu'il y a un problème de saisie. Dans tous les cas, les consignes suivantes permettent de saisir sans difficulté un nombre :

- n'utilisez jamais de **séparateur de milliers**
- n'écrivez jamais l'**unité** dans la case : l'unité doit être la même pour toute une colonne, et documentée dans le cahier des variables, mais jamais saisie dans la case
- utilisez le **séparateur décimal** de votre système d'exploitation (le point ou la virgule). Si vous saisissez directement dans un fichier texte, utilisez plutôt le point
- la **notation scientifique** avec la lettre « E » ou « e » est généralement acceptée (par exemple, 1,23e-05 = 1,23E-05 = $1,23x10^{-5}$)
- pour les **durées**, utilisez toujours des **nombres décimaux**, dans une unité fixée et identique pour toute la colonne (pour mémoire, un mois dure 30,44 jours, et une année 365,25 jours)
- n'utilisez jamais de symbole **pourcentage** « % ». Pour « 32,1% », saisissez « 0,321 »
- pour les nombres à virgule, indiquez toute la **précision** disponible, sans arrondi. Les arrondis seront calculés plus tard pour l'affichage des paramètres statistiques (ex : la moyenne), mais des arrondis prématurés altèreront la qualité de l'information.

Concernant les **données manquantes**, et ce quelle qu'en soit la raison :

- dans un tableur, laisser la case vide
- dans un fichier destiné à un statisticien, indiquez « NA » (deux lettres en majuscule signifiant « *not available* », sans guillemet, sans autre caractère)

4.3.3 Variables binaires

Les variables binaires doivent uniquement être saisies à l'aide d'un des deux nombres entiers suivants :

- **0** (zéro) pour signifier non/faux/absent/jamais (etc.)
- **1** (un) pour signifier oui/vrai/présent/toujours (etc.)

Dans un tableur, si ces valeurs sont correctement saisies, elles s'alignent automatiquement à droite.

Concernant les **données manquantes**, et ce quelle qu'en soit la raison :

- dans un tableur, laisser la case vide
- dans un fichier destiné à un statisticien, indiquez « NA » (deux lettres en majuscule, sans autre caractère)

4.3.4 Dates

Avant de saisir une date, posez-vous ces questions :

- Avez-vous vraiment besoin d'analyser une date ?
- L'anonymat est-il préservé ?

Il faut comprendre qu'une date, si elle est correctement saisie dans un tableur, est automatiquement transformé en nombre entier (par exemple le nombre de jours écoulés depuis la veille du 01/01/1900), et qu'un format de date présentant le même aspect que votre saisie initiale est ajouté par le tableur à la cellule. Ainsi, vous ne voyez pas la différence, mais la saisie est fondamentalement modifiée. Nous ne détaillerons pas le mécanisme sous-jacent

et toutes ses implications en termes de calculs possibles, mais il faut en retenir plusieurs choses :

- Si une date présente un format valide, votre tableur l'alignera automatiquement à droite
- Si une date reste alignée à gauche, c'est que son format n'est pas valide
- Lors de la saisie :
    - utilisez le format standard de votre système d'exploitation
    - faites très attention avec les formats ambigus, comme les formats anglosaxons (m/j/aaaa)
    - si vous avez un doute, testez quelques dates sans ambiguïté (avec un nombre de jours supérieur à 12) et contrôlez l'alignement
    - si vous utilisez un format de date numérique, utilisez toujours 2 chiffres pour les jours (quitte à utiliser un zéro initial), 2 chiffres pour les mois (de même), et 4 chiffres pour les années
    - en cas de doute, utilisez un format explicite comme « 3 fév 2024 »
- Après la saisie, avant l'export des données : sélectionnez toute la colonne et appliquez un format univoque comme « aaaa-mm-jj »

Si vous saisissez les données directement dans un fichier texte, choisissez vous-même un format univoque comme « aaaa-mm-jj » et utilisez le même dans toute la colonne.

Concernant les **données manquantes**, et ce quelle qu'en soit la raison :

- dans un tableur, laisser la case vide
- dans un fichier destiné à un statisticien, indiquez « NA » (deux lettres en majuscule, sans autre caractère)

4.3.5 Variables qualitatives non-ordonnées (nominales)

Pour saisir une variable qualitative, saisissez directement le mot (ou les mots) correspondant à la modalité. Exemples de saisies valides : « bleu », « sage-femme », etc.

Notes :

- Il n'y a pas de contrainte sur la saisie, mais privilégiez un **texte simple et court**, et évitez de préférence les accents, caractères spéciaux, apostrophes, etc. (ex : préférez « SF » à « sage-femme » ou « maïeuticien » ou « maïeuticienne »).
- **Ne remplacez pas ces valeurs par des nombres**, c'est inutile et source d'erreur !
- Dans un tableur, ces valeurs restent alignées à gauche.
- Attention à ne pas introduire de variations typographiques lors de la saisie (ex : Bleu, bleu, [espace]bleu, etc.). L'idéal sera de contrôler toutes les modalités présentes après la saisie, comme nous le verrons par la suite.
- La chaîne de caractères « NA » ne peut pas, en soi, être utilisées pour reproduire une modalité. Elle est réservée aux valeurs manquantes.

Nous verrons que l'analyse statistique n'est possible que s'il y a peu de modalités (maximum douze). Il faudra songer à des regroupements dans le cas contraire.

Concernant les **données manquantes**, et ce quelle qu'en soit la raison :

- dans un tableur, laisser la case vide
- dans un fichier destiné à un statisticien, indiquez « NA » (deux lettres en majuscule, sans autre caractère)

4.3.6 Variables qualitatives ordonnées (ordinales)

Les variables qualitatives ordonnées suivent les mêmes règles de saisie que les variables qualitatives non-ordonnées. Cependant, pour rendre immédiatement cohérentes les différentes sorties (tableaux de contingence, graphiques), il peut être intéressant dès la saisie de choisir des modalités dont l'ordre alphabétique correspond à l'ordre sémantique.

Exemple pour le niveau d'études :

- 0_aucun
- 1_brevet
- 2_bac
- 3_licence
- 4_master
- 5_these

Pour les **échelles de Likert**, on peut faire de même, ou directement saisir un nombre entier défini par convention. Je vous recommande la saisie suivante, allant de -2 à +2, qui sera pratique pour les calculs par la suite :

- Pas du tout d'accord : **-2**
- Plutôt pas d'accord : **-1**
- Ni d'accord, ni pas d'accord : **0**
- Plutôt d'accord : **1**
- Tout à fait d'accord : **2**

Ces saisies permettront un contrôle très rapide des tendances, simplement en calculant la moyenne de la colonne : une moyenne positive indique l'accord et une moyenne négative indique le désaccord.

4.3.7 Variables qualitatives multivaluées

Les variables qualitatives multivaluées sont des réponses à plusieurs réponses possibles. En présence de *k* modalités différentes (initialement k *checkbox*), il faut saisir **k variables binaires dans k colonnes**, à raison d'une colonne par modalité possible. Il sera opportun que les noms des variables binaires rappellent à la fois le nom de l'ensemble de la question (en préfixe) et le nom de la modalité binaire (en suffixe), comme illustré en Figure 31. Pour chaque variable binaire, on saisit « 1 » si la modalité est présente et « 0 » dans les autres cas.

Une réflexion devra être menée sur les valeurs, notamment les valeurs manquantes. Selon les cas, on pourra recommander les attitudes suivantes :

- Si aucune case n'a été cochée :
  - Si une modalité « aucun » était disponible dans le questionnaire, indiquez « NA » pour chaque variable binaire (deuxième ligne du tableau en Figure 31)
  - Si au contraire vous pensez que le répondant a cherché à répondre mais qu'aucune modalité ne lui correspondait, saisissez « 0 » pour chaque variable binaire
- En cas de saisie incohérente (dernière ligne du tableau en Figure 31), indiquez « NA » pour chaque variable binaire
- Si une modalité « tous » a été cochée, il serait cohérent de considérer également que chaque autre élément a été coché
- De manière plus générale, ayez une saisie cohérente si vous avez proposé des modalités « aucun », « tous », « pas concerné », etc.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

id	atcd
1	hta ; tabac
2	NA
3	tabac ; IDM
4	tabac
5	tabac
6	NA
7	aucun
8	hta
9	hta ; IDM
10	Cancer_uterus ; cancer_prostate

id	atcd_hta	atcd_idm	atcd_tab
1	**1**	0	**1**
2	NA	NA	NA
3	0	**1**	**1**
4	0	0	**1**
5	0	0	**1**
6	NA	NA	NA
7	0	0	0
8	**1**	0	0
9	**1**	**1**	0
10	NA	NA	NA

Figure 31. Saisie d'une variable qualitative multivaluée

4.3.8 Variables décrivant un événement temps-dépendant (survie)

Les variables correspondant à un événement binaire temps-dépendant sont très particulières. Elles nécessitent d'abord de savoir si l'événement est survenu ou non. Ensuite, si l'événement est survenu on doit savoir au bout de combien de temps. Enfin, si l'événement n'est pas survenu, on doit savoir pendant combien de temps. Ces deux durées semblent de natures différentes : en réalité elles sont similaires, puisque dans les deux cas il s'agit d'une période sans événement, mais ensuite cette période se clôt tantôt par un événement, tantôt par l'absence d'événement[7]. On créera donc deux colonnes :

- Une colonne binaire indiquant si, oui ou non, l'événement a été observé
- Une colonne quantitative continue (et toujours strictement positive) indiquant la durée pendant laquelle la personne n'a pas présenté l'événement. Cette colonne sera exprimée dans la même unité de temps, qui ne sera pas saisie dans les cellules.

Pour faciliter l'analyse, il est conseillé d'utiliser des noms de colonnes qui commencent de la même manière, et qui se terminent par « _evt » pour la colonne d'événement binaire, et « _delai » pour la colonne quantitative continue de durée.

Ainsi, le questionnaire proposé en Figure 32 se traduit par la saisie en Figure 33. Très souvent, dans un questionnaire, les questions relatives aux délais seront matérialisées dans des emplacements différents du questionnaire. Ainsi, dans la Figure 34, la première question sera renseignée par tous les répondants, mais elle sera également utilisée pour connaître le délai sans événement chez les personnes qui répondront « non » à la deuxième question. Pour ceux qui répondront « oui » à la deuxième question, c'est la troisième question qui permettra de connaître le délai.

Concrètement, l'exemple ici montré peut paraître peu réaliste car il implique une bonne mémorisation des dates et un calcul de différence entre deux dates.

[7] Le fait que la période d'observation se termine sans événement est appelé en statistique une « censure ». Ce terme n'est aucunement péjoratif.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

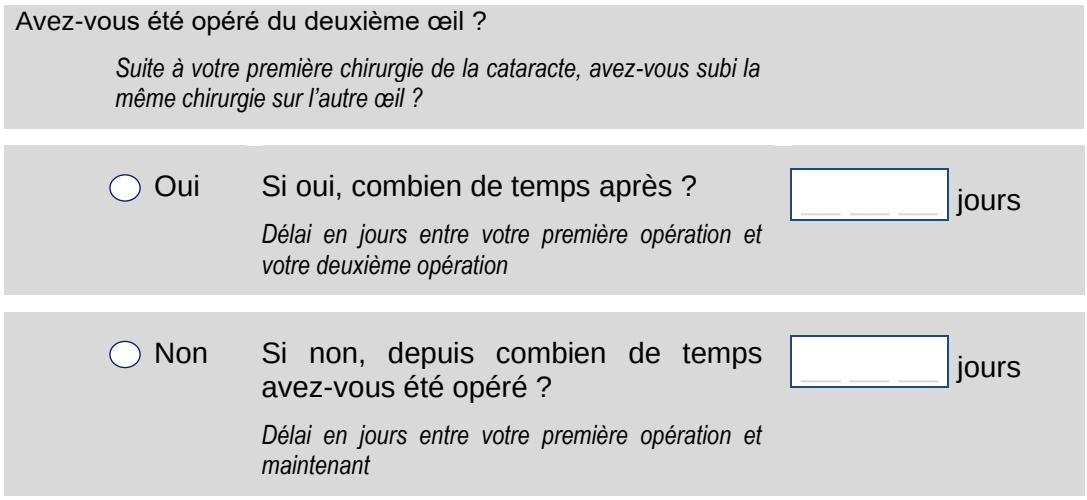


Figure 32. Exemple de présentation en questionnaire d'une question de survie

id	chir_oeil2_evt	chir_oeil2_delai
1	1	4
2	1	5
3	0	8
4	0	10

Figure 33. Exemple de saisie d'une variable de survie
Les individus 1 et 2 ont présenté l'événement au bout de 4 et 5 semaines,
les individus 3 et 4 n'ont pas présenté d'événement malgré un suivi de 8 et 10 semaines

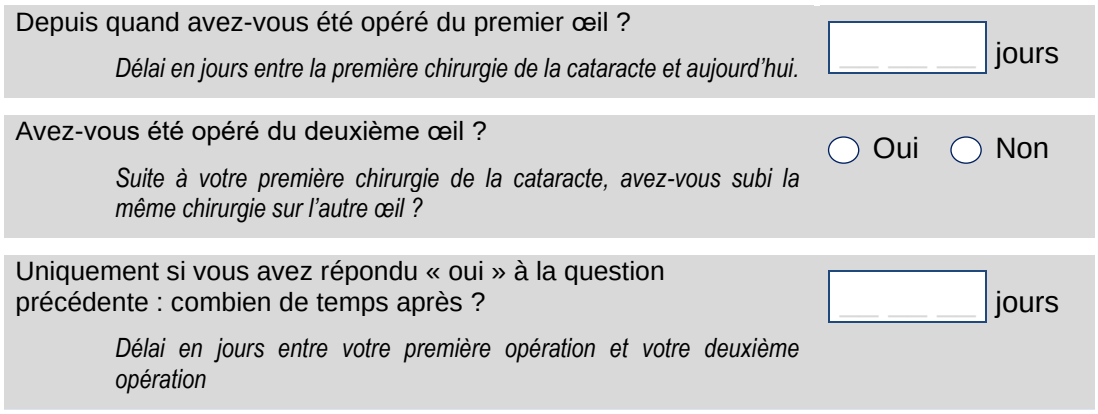


Figure 34. Autre exemple de présentation d'une question de survie :
les informations utiles sont séparées

4.4 Réflexions sur ce qu'est une variable quantitative

Une variable quantitative est un ensemble de nombres qui représentent une quantité. Ces variables quantitatives peuvent être issues de[8] :

- **Décomptes**, ou dénombrements (ex : nombre d'hospitalisations, nombre d'enfants)

[8] Cette distinction didactique n'est pas toujours univoque. Ainsi, lorsqu'une personne paie 3€ à un commerçant, cela peut être perçu comme une mesure, mais aussi comme un décompte (nombre de pièces de 1€) ou un calcul (différence entre le montant de la caisse après et le montant de la caisse avant la transaction). Bref, même si ce n'est pas clair, vous avez compris !

- **Mesures** : nombres produits par un dispositif ou un procédé de mesure (ex : poids, hémoglobinémie, revenu mensuel, etc.)
- **Calculs** : nombres issus d'un calcul réalisé sur plusieurs quantités définies précédemment (ex : l'indice de masse corporelle, le revenu moyen par membre du foyer, etc.)

Toutes les variables représentant des **dates** ou des **heures**, ainsi que les **délais**, quels que soient leur unité, sont également au fond des variables quantitatives issues de mesures ou de calculs (voir section 4.3.4 Dates en page 78). Rappelons au passage que les âges administratifs sont des délais entre la naissance et un événement d'intérêt, exprimés en années et arrondis à l'entier inférieur (en moyenne, ils sous-estiment de 0,5 ans le véritable âge). L'affichage des dates et des délais est rendu plus abordable aux êtres humains, à l'aide de chaînes de caractères structurées. En voici deux exemples :

- La date du 8 novembre 2024 peut être représentée de différentes manières, comme « 2024-11-08 » ou « 8 nov 24 » (etc.) et est généralement considérée par un tableur comme le nombre 45604, qui est le nombre de jours écoulés depuis la veille du 01/01/1900.
- La différence, un même jour, entre 17h00 et 18h15 peut se représenter « 1h et 15 minutes » ou encore « 75 minutes », mais sera considérée par de nombreux logiciels comme le nombre 4500, qui est le nombre de secondes séparant ces deux instants

Toutes les variables quantitatives citées précédemment sont susceptibles de subir des **opérations arithmétiques** : il est clairement licite, selon le contexte, de réaliser certaines opérations comme l'addition (ex : le poids des passagers d'un véhicule, etc.) ou des moyennes pondérées (ex : la concentration du sucre de deux solutions qu'on mélange, etc.).

Dans une perspective d'analyse statistique, **la notion de variable quantitative peut être élargie** à des valeurs qui ne peuvent pas licitement faire l'objet d'opérations arithmétiques : nous les appellerons dans ce chapitre des **variables pseudo-quantitatives**. Ces variables pseudo-quantitatives doivent cependant respecter quelques propriétés fondamentales (qui sont acquises pour les vraies quantités) :

- Elles peuvent être **figurées par des chiffres** (évidemment)
- Elles peuvent être **ordonnées** au vu de leur figuration en chiffre, sans aucune ambiguïté (l'ordre croissant des nombres est cohérent)
- On retrouve, d'une manière ou d'une autre, **une certaine régularité dans leur croissance** : le fait de passer de 0 à 1 a approximativement la même signification que de passer de 1 à 2, ou de 2 à 3, etc.

Nous rappelons néanmoins qu'il n'est pas souhaitable de saisir en numérique des variables purement qualitatives non-ordonnées (ex : une commune, une couleur, etc.).

Nous donnerons ici quelques exemples de saisie numérique de variables pseudo-quantitatives.

Le Tableau 7 illustre comment on peut coder en nombres les réponses à une question fermée ou à une échelle de Likert, correspondant formellement à une variable qualitative ordonnée. Dans ces deux exemples, on retrouve bien la figuration chiffrée, la possibilité de trier les réponses, et la notion de régularité de croissance.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 7. Exemples de saisie en nombre : variables qualitatives ordonnées

Question fermée

Réponse	**Codage**
Non	-1
Je ne sais pas	0
Oui	1

Echelle de Likert

Réponse	**Codage**
Pas du tout OK	-2
Plutôt pas OK	-1
Ni OK, ni pas OK	0
Plutôt OK	1
Tout à fait OK	2

De même, certaines variables quantitatives ont parfois été discrétisées lors du recueil, par souci d'anonymat (ou par erreur), mais pourront néanmoins être saisies en nombre et traitées en variable pseudo-quantitative, par exemple en utilisant les **centres de classes**[9]. Le Tableau 8 montre deux exemples de cela. Dans l'exemple à gauche du Tableau 8, sur un grand nombre d'étudiants, nous pourrions ainsi montrer, par exemple, que les filles obtiennent un classement meilleur que les garçons d'en moyenne 253 places. Dans l'exemple à droite du Tableau 8, nous pourrions ainsi montrer une corrélation linéaire entre les revenus mensuels et la valeur de la résidence principale. On notera que cette approche tolère plus ou moins des classes d'inégales largeurs, en fonction de la distribution précise de la variable.

Tableau 8. Exemples de saisie en nombre : variables quantitatives recueillies par classe, mais saisies par valeur centrale de classe

Rang de classement au concours d'accès au 3ème cycle de médecine

Réponse	Codage
1er – 1000ème	500
1001ème – 2000ème	1500
2001ème – 3000ème	2500
3001ème – 4000ème	3500
4001ème – 5000ème	4500
…	…

Salaire mensuel net

Réponse	Codage
Aucun revenu	0
Moins de 1000€	500
1001 à 2000€	1500
2001 à 3000€	2500
3001 à 4000€	3500
…	…

Nous pourrons procéder de même pour les dates. Si par exemple une valeur à saisir est « novembre 2024 », vous pourrez plutôt saisir le 15/11/2024 qui est le centre de la classe allant du 01/11/2024 au 30/11/2024. Cette date sera comprise par votre tableur comme le nombre 45611.

Enfin, si on réfléchit bien, les **notes** ou les **scores** entrent également dans cette catégorie des variables pseudo-quantitatives. Si on prend l'exemple d'un contrôle de mathématiques mis au point par un enseignant de collège, cet enseignant détermine une liste d'exercices, décide arbitrairement comment noter chaque réponse à un exercice, et décide arbitrairement du barème à appliquer pour obtenir une note sur 20. Il ne s'agit pas d'une vraie quantité, au sens où, par exemple, il n'est pas garanti que faire composer ensemble un élève qui a 15/20 et un élève qui a 5/20 permette d'obtenir la note de 20/20 (somme) ou celle de 10/20 (moyenne). Et pourtant, cela fonctionne bien pour de nombreux usages : on peut par exemple trier les élèves par note, et cela correspond peu ou prou au tri par niveau de compétence dans la discipline. On peut également réaliser diverses analyses statistiques, et évaluer la corrélation linéaire avec le nombre d'heures de travail consacrées à la discipline. Dans toute note ou dans tout score, dès lors que le barème est pensé « intelligemment » par un expert du domaine d'application, la variable qui en découle peut être analysée comme une variable quantitative.

[9] Dans le cas d'une variable représentant un rang, le centre de classe est également la moyenne des rangs, pour toutes les classes hormis la dernière. Ce nombre prend ainsi encore plus de sens.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

De manière générale, lorsque c'est approprié et qu'on hésite, il sera souvent préférable de saisir ces variables pseudo-quantitatives **comme des nombres**. Cela présente plusieurs avantages :

- **Saisie** accélérée, avec moins d'erreurs, et plus facile à contrôler
- **Moindre perte d'information** : la saisie comme un nombre au lieu de grandes classes permet souvent d'augmenter le nombre de modalités, donc la richesse de l'information
- Prise en compte de **l'ordre inter-groupes** : alors qu'il existe des tests statistiques appropriés pour les variables qualitatives ordonnées, ils sont peu utilisés, ce qui amène à tort les analystes à ignorer toute l'information contenue dans l'ordre des modalités. Avec une saisie numérique, la prise en compte de l'ordre est implicite
- Conservation de **l'ordre intra-groupe** : la discrétisation efface l'existence d'un classement à l'intérieur de chaque modalité. La saisie numérique permet souvent, implicitement, d'augmenter le nombre de modalités saisies, et donc de préserver plus d'information sur le classement des individus
- Absence de problème lié au **nombre de modalités** différentes : alors qu'il n'est pas possible de saisir un trop grand nombre de modalités pour une variable qualitative, il n'y a aucune limite pour un nombre

A divers égards, les avantages listés ci-dessus aboutissent généralement à une **augmentation de la puissance des tests statistiques**.

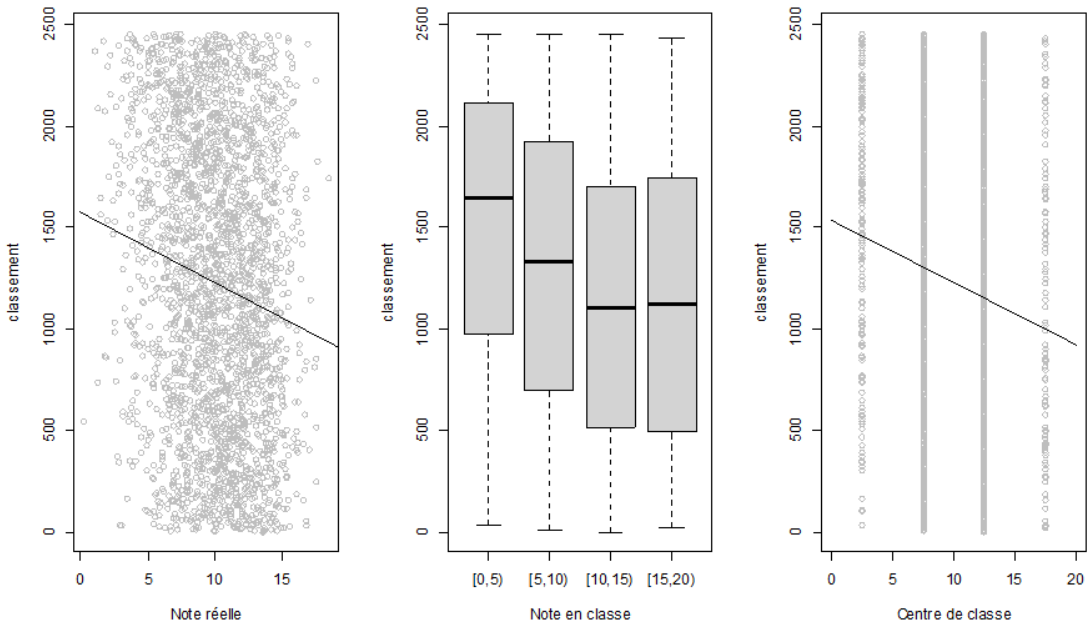
Tout ce qui précède n'empêchera pas d'utiliser durant l'analyse deux procédés inverses, si la linéarité d'une relation entre variables pose un problème :

- Discrétiser de nouveau une variable lors de l'analyse (la remettre en classes)
- Analyser les rangs d'une variable et non les valeurs originales

On retiendra que, si une variable est saisie comme quantitative, on peut la transformer en qualitative si besoin durant l'analyse. Inversement, si une variable est saisie comme qualitative, il ne sera généralement plus possible de la retransformer en variable quantitative.

Pour épiloguer sur ce chapitre, nous montrerons l'exemple du concours de fin de 1ère année commune de santé. Nous nous intéressons au classement à l'ensemble du concours, en fonction de la seule note de SSH (santé société humanité). La partie gauche de la Figure 35 montre le classement final en fonction de cette note, lorsqu'on dispose de toute l'information, à l'aide d'un nuage de points, assorti d'une droite de régression, dont l'équation figure en bas. Si pour une raison ou pour une autre nous discrétisons la note en 4 classes, il reste possible de tracer 4 boxplots juxtaposées, comme au milieu de la Figure 35. On observe alors que l'axe des abscisses n'est plus quantitatif, et nous verrons plus tard qu'il n'existe plus de méthode donnant une relation générique entre X et Y. On devra se contenter de comparer deux-à-deux les classes. Si cependant on remplace chaque classe par son centre de classe (ex : remplacer l'intervalle [0;5[par la valeur 2,5) comme à droite de la Figure 35, on constate qu'il est de nouveau possible de tracer une droite de régression, dont l'équation figure en bas. Dans cet exemple réel, les deux équations de droite sont très semblables, malgré la perte d'information.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.



y= -34,5.x + 1575.5 *(équation incalculable)* *y= -30,5.x+1533.9*

Figure 35. Exemple : classement obtenu au concours, en fonction de la note en SSH
Gauche : note réelle. Milieu : note discrétisée. Droite : avec le centre de classe

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

5 Vérifier, corriger et recoder des données

Nous verrons rapidement dans cette section comment contrôler que les données saisies ont le bon format, et recoder certaines d'entre elles s'il le faut.

5.1 Détecter et corriger les erreurs de saisie

5.1.1 Généralités

Si vous saisissez vos données dans un tableur comme Microsoft Excel®, LibreOffice Calc® ou Google Sheet®, le plus simple est de commencer par ajouter un « filtre automatique » ou un « autofiltre » à votre tableau de données. Pour ce faire, après avoir sélectionné le tableau à modifier :

Avec Microsoft Excel : *menu Données > Filtre*

Avec LibreOffice Calc : *menu Données > Autofiltre*

Pour chaque colonne du tableau, cette fonctionnalité ajoutera une liste déroulante des valeurs rencontrées (voir Figure 36). Initialement, cette liste déroulante est destinée à restreindre l'affichage aux lignes qui remplissent les conditions choisies. Nous l'utiliserons déjà simplement pour visualiser les différentes modalités rencontrées.

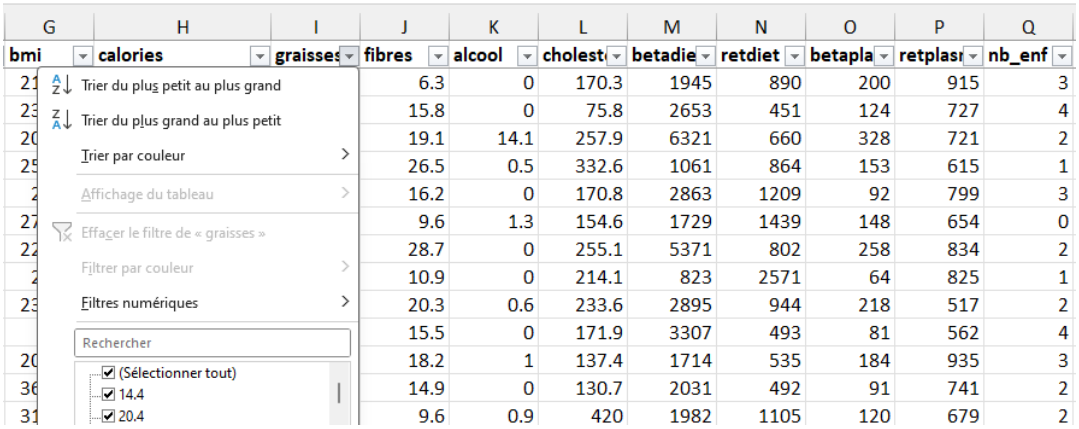


Figure 36. Exemple de filtre automatique sur un tableau (clic sur la flèche de "graisses")

5.1.2 Variables quantitatives

Pour les variables quantitatives, les différentes valeurs apparaissent triées par ordre croissant. Il est ainsi aisé de repérer (Figure 37) :

- Des valeurs extrêmes, au début ou à la fin de la liste
- Des valeurs textuelles saisies par erreur, qui apparaissent après l'ensemble des nombres, donc « hors classement ». Il peut s'agir de saisies avec une lettre, une unité, un mauvais séparateur décimal, un espace, etc.
- Les valeurs manquantes, figurant en fin de liste

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

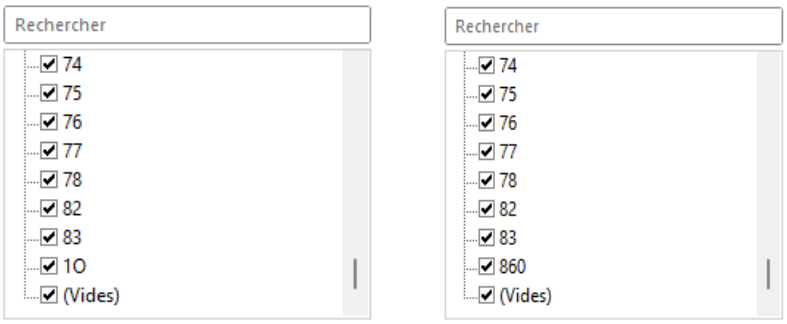
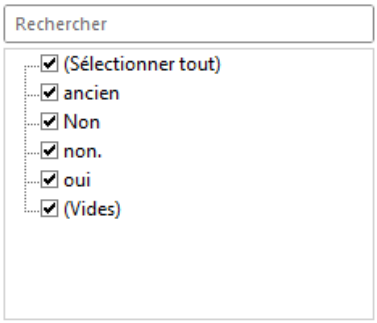


Figure 37. Anomalies sur une variable quantitative. Gauche : un o majuscule a remplacé un zéro, la modalité apparaît en fin de liste. Droite : exemple de valeur extrême (860kg)

5.1.3 Variables qualitatives ou binaires

Pour les **variables qualitatives ou binaires**, il est aisé de vérifier les différentes modalités rencontrées, les valeurs manquantes, et de détecter le cas échéant les variations typographiques (Figure 38, partie gauche).

/!\ Attention : sur les versions récentes d'Excel, certaines modalités différentes sont fusionnées dans cet affichage. Il faudra réaliser un tableau croisé dynamique pour vraiment voir la diversité des modalités, puis les corriger (Figure 38, partie droite).



Étiquettes de lignes	Nombre de tabac2
non	1
ancien	43
Non	154
non.	1
oui	115
(vide)	
Total général	**314**

Figure 38. Anomalies sur une variable qualitative. Les variations typographiques de la réponse « non » sont moins bien détectées par le filtre automatique (gauche) que dans le tableau croisé dynamique (droite).

Pour insérer un tableau croisé dynamique, sélectionnez le tableau de données puis :

Avec Microsoft Excel : *menu Insertion > Tableau croisé dynamique*

Avec LibreOffice Calc : *menu Insertion > Table dynamique*
(ou pilote de données dans les versions plus anciennes)

L'utilisation de ces tableaux n'est pas évidente, mais de nombreux tutoriels se trouvent sur Internet.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

5.1.4 Dates

Selon la version du tableur que vous utilisez, les dates peuvent être particulièrement bien gérées. La Figure 39 montre comment Microsoft Excel représente les dates : bien que représentées numériquement dans les données, elles sont affichées de manière hiérarchique. Les dates qui n'ont pu être décodées par le tableur apparaissent en fin de liste, comme des saisies textuelles.

Figure 39. Représentation des dates en filtre automatique : les dates sont présentées hiérarchiquement, l'anomalie apparait en fin de liste

5.1.5 Dépendances fonctionnelles

Enfin, en utilisant les filtres dans leur but premier, qui est de restreindre l'affichage à certaines lignes, vous pourrez vérifier que certaines variables ne sont saisies que lorsque c'est approprié (ex : « si vous avez répondu oui à la question précédente… »).

5.2 Recoder et agréger les données

Afin de représenter graphiquement et d'analyser statistiquement des variables qualitatives, il est nécessaire que le **nombre de modalités n'excède pas une dizaine**. Or souvent, tel n'est pas le cas. Il faut donc les regrouper.

De même, pour certaines représentations graphiques et analyses statistiques, il est nécessaire de **discrétiser les variables quantitatives** (ou les dates), c'est-à-dire de les représenter en classes (ex : de 0 à 10, de 10 à 20, etc.).

Pour ce faire, nous ne recommandons pas de remplacer les colonnes existantes, mais bien de **créer de nouvelles colonnes**. En effet, il est important de pouvoir revenir aux données initiales, pour corriger le mécanisme d'agrégation, voire pour proposer simultanément différentes stratégies d'agrégation.

Dans un tableau, trois procédés différents peuvent être utilisés, en fonction de votre maîtrise du tableur, du nombre d'enregistrements, et du nombre de modalités.

5.2.1 Option 1 : recodage manuel par lots

Cette option est recommandée si vous n'êtes pas à l'aise avec un tableur, et pour un faible nombre de modalités, quel que soit le type de variable (qualitative ou quantitative).

- Créez une nouvelle colonne en recopiant le contenu de la précédente (gauche sur Figure 40)

- Appliquez un filtre automatique sur tout le tableau
- Itérativement, sur la nouvelle colonne uniquement, sélectionnez tour-à-tour les valeurs (ou plages de valeurs) à modifier (milieu sur Figure 40), et écrasez les anciennes valeurs par de nouvelles valeurs (droite sur Figure 40).

Cette approche fonctionne tant pour les variables qualitatives (par liste de modalités) que quantitatives (par intervalle de valeurs).

NB : si vous souhaitez saisir quelque chose qui pourrait ressembler à une formule, comme « 20-30 », pour éviter un message d'erreur, débutez votre saisie par une **apostrophe** en saisissant « '20-30 ».

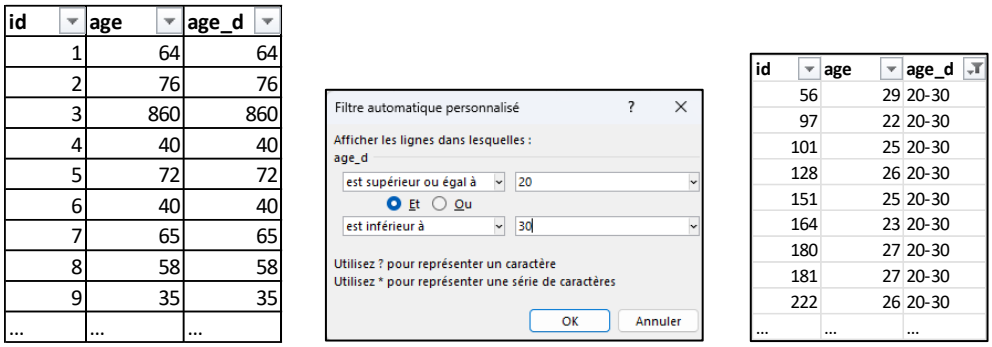


Figure 40. Exemple de recodage manuel d'une variable, par filtre automatique

5.2.2 Option 2 : recodage par formule

Cette approche est recommandée en particulier pour les variables quantitatives, et si vous êtes à l'aise avec les tableurs. Il est possible de recoder les variables en insérant une formule. La Figure 41 illustre cela avec l'âge : en fonction des valeurs de la colonne B, la colonne C affichera « 00-45 », « 45-80 » ou « 80-100 ». Si vous souhaitez vous documenter sur ce type de formules, lisez la documentation de votre tableur (ou un tutoriel) portant sur les **formules** en général, et la fonction **si()** en particulier.

Cette approche permet, en cas de mise à jour des données source, d'appliquer la transformation sans trop d'effort supplémentaire.

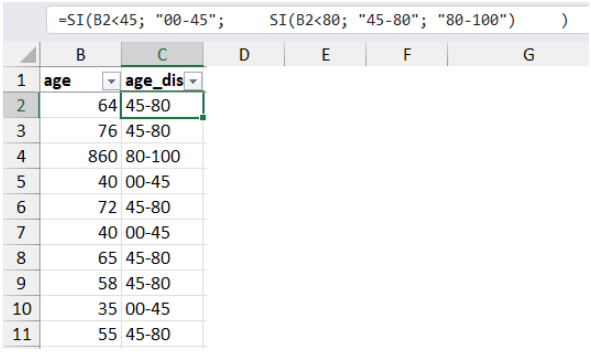


Figure 41. Exemple de recodage d'une variable quantitative par formule

Dans le cas des variables quantitatives, l'opération réalisée en Figure 41 s'appelle une **discrétisation**. Les seuils choisis pour discrétiser sont généralement arbitraires, et réalisent un mélange pertinent entre :

- Des seuils retrouvés dans la **littérature**
- Des seuils intuitifs pour les **experts** du domaine

- Les **tertiles, quartiles ou quintiles** observés dans les données (seuils qui séparent l'échantillon en 3, 4 ou 5 sous-groupes équilibrés)
- Les **arrondis** que tout le monde attend, par habitude

Par exemple, pour discrétiser l'âge de patients hospitalisés en chirurgie orthopédique, on pourra souvent utiliser les catégories suivantes :

- [000 ; 001[nourrissons
- [001 ; 015[pédiatrie
- [015 ; 025[jeunes gens
- [025 ; 065[adultes
- [065 ; 075[jeunes retraités
- [075 ; 110[personnes âgées

Ces catégories seront volontiers fusionnées si leurs effectifs sont faibles. Notez que, dans cet exemple, nous avons utilisé 3 caractères pour chaque nombre, afin que le tri alphabétique des intervalles ainsi générés corresponde au tri numérique.

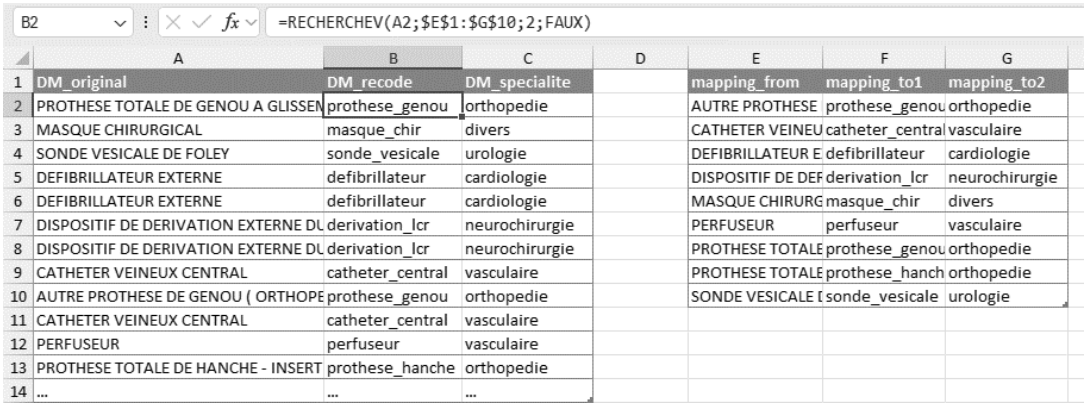
5.2.3 Option 3 : recodage par table de correspondance (*mapping*)

Cette approche est recommandée pour les variables qualitatives, si vous êtes à l'aise avec les tableurs. Il peut être particulièrement adapté de définir une **table de correspondance** (ou **table de *mapping***) contenant tous les libellés d'origine, et proposant à chaque fois un libellé de destination. Ensuite, la nouvelle colonne cherchera à chaque fois le libellé de destination en fonction du libellé d'origine. Si vous souhaitez vous documenter sur ce type de formules, lisez la documentation de votre tableur (ou un tutoriel) portant sur les **formules** en général, et la fonction **recherchev()** en particulier.

Si votre tableau initial contient de très nombreuses lignes, avec de nombreux ex-aequo, cette approche vous fera gagner du temps. Elle présente également l'avantage de proposer une excellente **traçabilité** des choix d'agrégation, et leur **évolutivité** en fonction des erreurs rencontrées ou des changements de stratégie : les données source sont conservées, et si on modifie le mapping, les données calculées sont immédiatement mises à jour. Cette approche permet, en une seule fois, de définir si on le souhaite **plusieurs niveaux** successifs de recodage ou agrégation.

La Figure 42 illustre cela. La colonne A contient les données originales. Les colonnes E, F et G contiennent la table de correspondance (en général, on la placera plutôt dans un onglet différent, ici nous souhaitions simplement l'afficher sur la même figure). Cette table de correspondance contient en colonne E la liste dédoublonnée des modalités rencontrées en A. En colonnes F et G, elle propose deux manières différentes de recoder ces libellés : une par type de matériel, l'autre par spécialité. Dans le tableau initial, les colonnes B et C rapatrient automatiquement ces nouveaux libellés. Comme illustré dans la formule de la cellule B2, on utilise la fonction **recherchev()** de Microsoft Excel ou LibreOffice Calc.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.



B2 : =RECHERCHEV(A2;$E$1:$G$10;2;FAUX)

	A	B	C	D	E	F	G
1	DM_original	DM_recode	DM_specialite		mapping_from	mapping_to1	mapping_to2
2	PROTHESE TOTALE DE GENOU A GLISSEN	prothese_genou	orthopedie		AUTRE PROTHESE	prothese_genou	orthopedie
3	MASQUE CHIRURGICAL	masque_chir	divers		CATHETER VEINEU	catheter_central	vasculaire
4	SONDE VESICALE DE FOLEY	sonde_vesicale	urologie		DEFIBRILLATEUR E	defibrillateur	cardiologie
5	DEFIBRILLATEUR EXTERNE	defibrillateur	cardiologie		DISPOSITIF DE DEF	derivation_lcr	neurochirurgie
6	DEFIBRILLATEUR EXTERNE	defibrillateur	cardiologie		MASQUE CHIRURG	masque_chir	divers
7	DISPOSITIF DE DERIVATION EXTERNE DU	derivation_lcr	neurochirurgie		PERFUSEUR	perfuseur	vasculaire
8	DISPOSITIF DE DERIVATION EXTERNE DU	derivation_lcr	neurochirurgie		PROTHESE TOTALE	prothese_genou	orthopedie
9	CATHETER VEINEUX CENTRAL	catheter_central	vasculaire		PROTHESE TOTALE	prothese_hanch	orthopedie
10	AUTRE PROTHESE DE GENOU (ORTHOPE	prothese_genou	orthopedie		SONDE VESICALE [	sonde_vesicale	urologie
11	CATHETER VEINEUX CENTRAL	catheter_central	vasculaire				
12	PERFUSEUR	perfuseur	vasculaire				
13	PROTHESE TOTALE DE HANCHE - INSERT	prothese_hanche	orthopedie				
14	...	...	...				

Figure 42. Recodage avec une table de correspondance (mapping)

Cette approche permet également, en cas de mise à jour des données source, d'appliquer la transformation sans trop d'effort supplémentaire, car la table de correspondance a déjà été définie. Il faudra simplement détecter les valeurs « #N/A » dans les colonnes B et C, qui témoigneront d'entrées de la colonne A sans équivalence dans la table de correspondance. Il suffira alors de créer ces entrées dans la table de correspondance et, si besoin, d'étendre la plage de définition de cette table dans les formules.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6 Gérer les données manquantes

6.1 Préambule

Les données manquantes constituent un problème souvent abordé par les biostatisticiens, en particulier ceux impliqués dans la recherche clinique. Nous verrons dans cette section comment les gérer. Avant de poursuivre, nous attirons tout de suite votre attention sur un point : bien qu'il existe des méthodes très perfectionnées pour imputer les données manquantes, **le plus probable est que ces méthodes ne soient ni nécessaires, ni même licites pour votre travail**. Nous verrons dans la dernière partie (6.6 Conduite à tenir, arbre décisionnel en page 101) un arbre décisionnel vous indiquant si, oui ou non, vous devez imputer les données manquantes. La réponse sera très probablement « non ». Cela étant posé, vous pouvez maintenant vous détendre et lire ce chapitre.

Nous verrons dans cette section comment sont définies les données manquantes (6.2 Définition page 93) : ce qu'elles sont, en insistant sur le fait que cette notion est toute relative, et comment on peut les typer en théorie. Nous proposerons ensuite trois manières de les gérer : l'analyse de cas complets, qui est l'attitude standard (6.3 Analyse des cas complets en page 96), l'imputation de données manquantes, qui n'a de sens que dans certains cas (6.4 Imputation de données manquantes en page 98), et une approche originale facile à mettre en œuvre et efficace, que je vous propose (6.5 Simple et efficace : identifier le NA comme une modalité en page 101). Nous achèverons ce chapitre avec un arbre décisionnel simple et opérationnel (6.6 Conduite à tenir, arbre décisionnel en page 101).

6.2 Définition

6.2.1 Généralités

Prosaïquement, une donnée manquante est une « case vide » dans votre tableau de données. Cette « case vide » sera matérialisée différemment selon le logiciel utilisé :

- Dans un logiciel de statistique, le codage est habituellement **NA** (les deux lettres en majuscules, comme *not available*). Cela dit, un codage différent pourra être spécifié lors du chargement du fichier.
- Dans un tableur, le mieux est tout simplement de laisser la case vide.
- Dans certains vieux logiciels, certains codent « 99 » ou encore « -1 » : de tels usages sont formellement déconseillés.

6.2.2 Une notion toute relative

Nous insisterons sur le fait que, contrairement à ce qui est souvent enseigné, la notion de « données manquantes » est cependant **toute relative**.

Exemple 1 : *la Figure 43 montre exactement les mêmes données vues par un informaticien (gauche) et un épidémiologiste (droite). L'informaticien utilise une table comportant une ligne par mesure de la kaliémie (taux de potassium) : disposant de 3 mesures, il crée 3 lignes. Aucune donnée manquante n'est visible. L'épidémiologiste analyse des personnes, et souhaitait disposer de 2 mesures par personne : il crée donc 2 lignes, mais attend 2 colonnes de mesures par personne (mesures répétées), ce qui fait apparaître une valeur manquante. La donnée manquant pour l'un, ne manquait pas pour l'autre. Ce problème de représentation des données est fortement présent dans les esprits. Ainsi, certains épidémiologistes affirment que les modèles mixtes « gèrent les données manquantes », alors que ce n'est bien entendu pas le cas.*

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Personne	Date	Valeur
Josette	Lundi	4.3
Josette	Jeudi	3.8
Lucien	Lundi	3.4

Personne	Lundi	Jeudi
Josette	4.3	3.8
Lucien	3.4	**NA**

Figure 43. Kaliémies identiques vues par un informaticien (gauche) ou un épidémiologiste (droite)

Exemple 2 : *un TIM (technicien d'information médicale) recueille, pour chaque patient, les différentes maladies que présente ce patient, sous forme de codes CIM10[24] dans le cadre du PMSI[25]. Pour la patiente Aglaé Müller, il code I10 pour hypertension artérielle, et M16 pour coxarthrose. Cela signifie que ces pathologies sont bien présentes et, implicitement et sans le préciser, que toutes les autres sont absentes. Si ce TIM avait recueilli les diagnostics de manière binaire, il aurait renseigné les valeurs 1 pour ces deux codes, mais pas pour tous les autres codes (par exemple E119 pour un diabète de type 2, E660 pour une obésité). Alors, les variables correspondant à E119 et E660 auraient pu prendre pour valeur 0 ou NA, selon que le TIM s'engage sur l'absence de ces maladies, ou ne sait simplement pas si elles sont présentes. La représentation des données sous la forme d'une liste de codes ne permet donc pas de différencier la réponse 0 de la réponse NA.*

6.2.3 Priorité au bon sens !

Même si certains développements méthodologiques sont complexes, il ne faut pas se départir de bon sens face à des données manquantes. Le bon sens apporte souvent une réponse plus simple.

Tout d'abord, parlons des **dépendances fonctionnelles**, dont voici deux exemples.

Exemple 1 : *dans la Figure 44, la personne interrogée ne doit répondre à la deuxième question (à droite sur Figure 44) que si elle a répondu « oui » à la première question (à gauche sur Figure 44). Naturellement, pour les personnes qui ont répondu « non » à la première question, la valeur à saisir dans le tableur pour la deuxième question est « NA ».*

Exemple 2 : *on note si certains patients ont un médicament et, le cas échéant, à quelle posologie. La posologie n'est évidemment recueillie que lorsque le patient a le médicament.*

Formellement, il s'agit bien dans les deux exemples précédents d'une donnée manquante. Cette donnée manquante est naturelle et obligatoire, il n'est donc pas question de l'imputer. L'ensemble des discussions de cette section **ne concerne pas ce type de donnée manquante**.

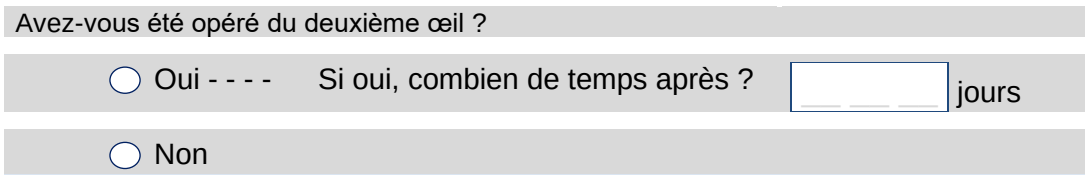


Figure 44. Exemple de question conditionnelle

Il arrive également que la donnée manquante **puisse aisément être imputée** par une valeur qui, **à dire d'expert**, est évidemment la bonne. En voici deux exemples.

Exemple 3 : *une étude porte sur des patients hospitalisés en rapport avec une traumatisme crânien. On souhaite recueillir notamment le score de Glasgow[26], qui va de 15 points pour l'état normal, à 3 points pour l'état de coma profond ou mort clinique. On observe que certains de ces patients sont restés moins de 3 heures aux urgences et sont rentrés directement à domicile, sans que le score de Glasgow n'ait été rapporté dans le dossier. Les circonstances laissent penser que leur score de Glasgow est égal à 15. On peut donc* ***imputer les valeurs manquantes*** *par la valeur 15. Ce n'est pas tout à fait certain, mais c'est certainement beaucoup plus raisonnable que de penser qu'il est inconnu, et pourrait indifféremment prendre n'importe laquelle des valeurs possibles.*

Exemple 4 : *on s'intéresse à différents paramètres biologiques chez des patients hospitalisés, dont la glycémie. Ce paramètre est mesuré de manière très courante, et ce paramètre est notamment mesuré dès qu'un patient présente des symptômes évocateurs d'hyperglycémie ou d'hypoglycémie, ou dès qu'un patient a une probabilité plus élevée de valeur anormale (ex : patients diabétiques). Par conséquent, l'absence de mesure est très probablement liée à une valeur normale. Il sera certainement plus approprié d'imputer toutes les valeurs manquantes par une valeur normale (ex : 5,5 mmol/l) que de procéder autrement.*

6.2.4 Description théorique : MCAR, MAR, MNAR

Au-delà des cas évidents décrits dans le chapitre précédent, il peut être utile de considérer les données manquantes en termes d'alea lié à leur disparition, car cela impacte directement le traitement qu'on pourra leur réserver. Comme nous le verrons dans l'arbre décisionnel, il est probable que cette description ne vous soit pas utile, mais elle est académique et instructive (6.6 Conduite à tenir, arbre décisionnel en page 101).

MCAR :

Le cas le plus favorable est celui des données manquant de façon complètement aléatoire (*missing completely at random*, MCAR). Pour chaque individu, la probabilité de voir sa valeur manquante est la même. Il pourrait s'agir par exemple d'oublis ou erreurs de saisie indépendants du cours réel du soin, de données perdues ou détruites par un vol de documents, un bug informatique, une tache de café, une invasion de souris, etc.

Ce cas est des plus favorables car, si on n'analyse que les cas pour lesquels la variable est définie, cela ne risque pas de biaiser les résultats, car tout se passe comme si on avait tiré au sort des individus pour les évincer au dernier moment. Hormis diminuer le nombre d'individus, cela n'a pas de conséquence sur le résultat de l'analyse. Ce cas est malheureusement très théorique.

MAR :

Un autre cas est celui des données manquant de façon aléatoire (*missing at random*, MAR). Certes, la probabilité qu'une valeur manque ne dépend pas de sa propre valeur, mais elle n'est pas pour autant indépendante des autres variables. En voici un exemple. Certains patients sont hospitalisés en urgence pour une fracture de la jambe. On cherche à savoir si un score fonctionnel quelconque s'améliore par la suite. Ils sont revus dans l'établissement qui accueille les urgences, mais les plus riches d'entre eux sont revus dans un autre établissement. A priori, il n'y a pas de raison que ces patients évoluent plus ou moins bien que les autres (le caractère manquant n'est pas lié à la valeur du score), mais on peut penser que la probabilité que la valeur soit manquante est liée à d'autres variables, comme le revenu.

Ce cas est plutôt favorable. Premièrement, l'analyse de la seule variable d'intérêt, en omettant les valeurs manquantes, n'est pas biaisée. Deuxièmement, pour les analyses multivariées, des méthodes « agnostiques » d'imputation pourront être mises en œuvre.

MNAR :

Le cas le plus gênant est celui des données manquant de façon non-aléatoire (*missing not at random*, MNAR). Dans ce cas, c'est justement parce qu'elle aurait eu une certaine valeur que la variable n'a pas été observée. Les exemples sont très nombreux. Par exemple, on peut s'intéresser à l'évolution de certains patients dans un état grave mais les plus graves d'entre eux décèdent à notre insu et ne sont donc pas réévalués. Autre exemple, on réalise un sondage téléphonique mais les personnes qui souhaitent voter pour une personne en particulier ne se sentent pas assez en confiance pour en faire part[10]. Autre exemple, on

[10] Cet exemple est hautement réaliste : en 2022, les instituts de sondage avaient été incapables de prédire que le 2ème tour des élections présidentielles verrait s'affronter Jacques Chirac et Jean-Marie Le Pen. De nombreux sondés ne se sentaient pas suffisamment en confiance pour annoncer leur intention de voter pour Jean-Marie Le Pen au premier tour.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

souhaite évaluer le revenu de personnes en les appelant au téléphone au domicile, mais aux heures du sondage, ceux qui gagnent le plus sont en train de travailler et ne peuvent pas répondre au téléphone.

Ce cas est très défavorable et trouve difficilement sa réponse dans l'imputation « agnostique » de données manquantes, or ce cas est de loin le plus fréquent dans les études en santé. Nous reviendrons sur ce point dans les parties 6.4.2 Critique de cette approche en page 99 et 6.6 Conduite à tenir, arbre décisionnel en page 101.

6.3 Analyse des cas complets

Face à des données manquantes, la première option, très fréquente, consiste à n'analyser que les données disponibles, sans réaliser d'imputation de données manquantes. Théoriquement, cette approche n'est valide que dans le cas MCAR. En pratique, c'est celle qui est privilégiée lorsque la proportion de cas complets est suffisamment élevée. C'est celle qu'appliquent la grande majorité des procédures statistiques, par défaut. Nous y reviendrons dans la partie 6.6 Conduite à tenir, arbre décisionnel en page 101.

6.3.1 Analyse univariée des cas complets

Les analyses univariées consistent à analyser une seule colonne à la fois. Dans ces analyses, la plupart du temps on réalise une analyse en cas complets, c'est-à-dire qu'on analyse les valeurs disponibles, et on ignore les valeurs manquantes (Figure 45). Cela aboutit simplement à diminuer le nombre de valeurs analysées, ce qui ne pose pas de problème particulier.

Si la proportion de valeurs manquantes est faible (par exemple moins de 5%), on ne le mentionne pas spécifiquement, ou on peut mentionner entre parenthèses la proportion de valeurs manquantes.

Exemple : L'âge moyen est de 84,5 ans (DS=14,8 ; 2,5% de valeurs manquantes)

Si la proportion de valeurs manquantes est importante, il peut être sage de ne pas décrire la variable car, comme dans le cas des questionnaires avec un faible taux de réponses, il est possible que les paramètres mesurés soient biaisés.

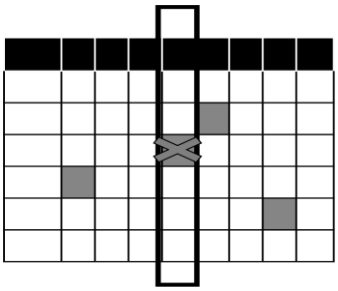


Figure 45. Analyse univariée en cas complets

6.3.2 Analyse bivariée des cas complets

Les analyses bivariées consistent à analyser deux colonnes, généralement pour évaluer leur liaison statistique. Dans ces analyses, la plupart du temps on réalise là aussi une analyse en cas complets, c'est-à-dire qu'on analyse les individus qui ont une valeur connue pour les deux variables (voir Figure 46). On exclut donc, pour cette analyse, les individus qui ont une ou deux valeurs manquantes. Cette exclusion est réalisée automatiquement par la procédure statistique, il n'y a rien à faire. Cela aboutit, de même, à diminuer le nombre de valeurs analysées.

Si la proportion d'individus ainsi exclus est faible, on peut ne pas le mentionner, ou citer rapidement la proportion de cas exclus. Dans le cas contraire, il peut être sage de ne pas réaliser cette analyse.

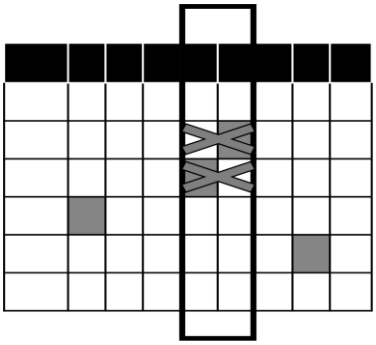


Figure 46. Analyse bivariée en cas complets

6.3.3 Analyse multivariée des cas complets

Une analyse multivariée consiste à analyser, dans la même procédure statistique, plusieurs colonnes à la fois. Très souvent, il s'agit d'expliquer une seule colonne par un ensemble d'autres colonnes (souvent une dizaine de colonnes). L'analyse en cas complets consiste là aussi à exclure tous les individus qui ont au moins une valeur manquante sur l'ensemble des colonnes. Ceci peut vite nous amener à exclure la plupart des individus (Figure 47).

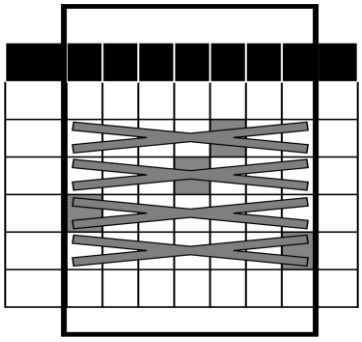


Figure 47. Analyse multivariée en cas complets

Plusieurs attitudes sont valables et permettent d'analyse uniquement des cas complets.

Analyse en cas complets :

Si les cas complets restent majoritaires, on peut réaliser une analyse en cas complets.

Sélection des variables :

Il est généralement souhaitable d'identifier les variables qui font le plus chuter le nombre d'individus, et se poser la question de la pertinence de leur prise en compte. Il sera souvent intéressant d'abandonner certaines variables pour augmenter le nombre de cas complets.

Il est à noter que, même dans certaines procédures qui éliminent automatiquement certaines variables, il se peut que les individus soient évincés au vu des variables initialement incluses, et non des variables qui restent au terme de la procédure (ex : régressions pas-à-pas, arbres de décision). C'est dommage ! Il est donc beaucoup plus efficace de limiter dès le début le nombre de variables incluses dans ces procédures.

Construction de variables composites :

On pourra également construire des variables composites alimentées par différentes variables. Ainsi, on pourra construire la variable « insuffisance rénale », fausse par défaut, et vraie si on trouve un DFG faible, ou un code CIM10 d'insuffisance rénale, ou un acte de dialyse rénale, etc. Ces variables composites sont construites « intelligemment », et ne restent en valeur « NA » que si toutes les variables source sont en NA en même temps.

Dans les cas problématiques que les attitudes simples ne suffisent pas à résoudre, et seulement si elle est valide, il faudra songer à une méthode d'imputation de données manquantes, comme développé par la suite.

6.4 Imputation de données manquantes

L'imputation de données manquantes est possible dans les cas de MCAR et de MAR. Rappelons néanmoins qu'en santé le cas le plus fréquent est le MNAR, nous y reviendrons par la suite dans les parties 6.4.2 Critique de cette approche en page 99 et 6.6 Conduite à tenir, arbre décisionnel en page 101.

6.4.1 Aperçu de méthodes disponibles

Nous rappelons que, si une **imputation à dire d'expert** est possible, comme dans les exemples d'un chapitre précédent (6.2.3 Priorité au bon sens ! en page 94), il sera toujours préférable de procéder ainsi.

Il est également possible d'**imputer par une valeur fixe**. On utilisera alors généralement la moyenne constatée dans l'échantillon, ou la médiane. Dans le cas d'une variable binaire ou qualitative, on tirerait au sort une valeur en respectant la probabilité estimée. Cette approche est valide dans le cas des MCAR, cas très rare.

Il est également possible d'**imputer par une valeur issue d'individus similaires**. On utilisera alors généralement la moyenne, la médiane, ou les probabilités constatées dans le sous-groupe auquel appartient l'individu, ou dans un nuage d'individus considérés comme proches à l'issue d'une étape de classification non-supervisée (ex : *k-means*). Cette approche est valide dans le cas des MCAR, mais peut également être valide dans le cas MAR si les variables liées sont prises en compte.

Il est également possible d'**imputer par un calcul tenant compte des autres variables**. Cela suppose la mise au point d'un modèle statistique expliquant la variable comprenant des valeurs manquantes, par les autres variables disponibles (sans valeur manquante). On pourra ainsi prédire les quelques valeurs manquantes. Différentes approches existent, et certaines introduisent un alea. Cette approche est donc valide dans les cas MCAR et MAR, sans restriction.

Souvent, il arrive également que plusieurs variables présentent simultanément des données manquantes, et pas sur les mêmes individus. Le tableau de données ressemble alors à une tranche d'emmental, présentant de nombreux trous non-alignés. Il existe alors des méthodes d'**imputation multiple**, qui proposent de boucher tous les trous simultanément (ex : Nipals, Mice, etc.). Certaines d'entre elles sont itératives, et proposent d'imputer la variable *j* par des valeurs, y compris des valeurs manquantes, des autres variables (pour le premier tour, toutes les autres valeurs manquantes sont imputées aléatoirement). Puis on utilise notamment ces valeurs imputées de *j* pour la variable *j+1*, etc. Le processus finit par se stabiliser, ou différents résultats peuvent être ensuite moyennés.

Nous ne détaillerons pas ces méthodes pour deux raisons :

- Elles relèvent de biostatisticiens spécialisés : si vous avez besoin de telles méthodes, c'est que vous avez besoin de l'aide du soutien d'**une équipe de biostatistique**
- La plupart du temps, l'hypothèse de MCAR ou MAR requise est **totalement délirante**, nous reviendrons dessus juste après. Donc, dans la plupart des cas, vous avez lu ces paragraphes pour rien 😉

6.4.2 Critique de cette approche

L'approche par imputation de données manquantes est valide dans certains cadres d'études, mais pas dans la majorité des études.

6.4.2.1 Recherches RIPH (ou EPP) protocolisées et financées

Dans les recherches impliquant la personne humaine (ou impliquant des professionnels de santé), qui sont obligatoirement protocolisées avant d'être autorisées et très souvent financées, les mesures et examens que subissent les sujets sont obligatoires, le recueil de données est fait par des professionnels rémunérés et dédiés uniquement à cela (les ARC, attachés de recherche clinique), si bien que l'exhaustivité est élevée. L'hypothèse de MCAR est souvent valide, et celle des MAR également, car il n'y a aucune raison particulière pour laquelle un patient n'aurait pas un examen, alors que cet examen a été planifié, financé, et que le patient s'était d'avance engagé à l'accepter. Ce n'est pas toujours vrai, mais souvent. Les hypothèses sont donc réunies pour une imputation multiple.

On notera également que l'imputation n'est d'ailleurs pas souvent nécessaire car :

- Pour ces mêmes raisons, la proportion de données manquantes est nulle ou faible
- Dans ces recherches, un objectif principal a été établi, et il porte généralement sur une seule analyse univariée ou bivariée, rendant le besoin d'imputation superflu

L'imputation sera le plus souvent réalisée pour les analyses multivariées, qui ne relèvent pas de l'objectif principal mais des analyses secondaires.

6.4.2.2 Recherches observationnelles sur le soin courant

Dans les recherches impliquant la personne humaine, mais observationnelles portant sur le soin courant, certes du personnel sera dédié à la collecte d'informations, mais il n'est pas question de modifier le cours du soin pour la recherche. Autrement dit, **rien n'impose la réalisation d'examens complémentaires** s'ils ne sont pas utiles au patient lui-même. La conséquence est que, non seulement des informations manquent, mais par essence elles ne manquent jamais par hasard : **le MNAR est la règle**. Il existe un **biais d'indication** sur la réalisation des examens complémentaires : on réalise un examen justement parce qu'il a une plus forte probabilité d'obtenir un résultat anormal.

6.4.2.3 Recherches sur des données

Dans les recherches sur les données, on rencontre le même biais majeur que dans le cas précédent. S'y ajoute souvent une erreur supplémentaire liée à la collecte des données, qui n'est pas financée, et ne fait donc pas l'objet d'un effort particulier. Non seulement **le cas MNAR est la règle**, mais en plus l'**exhaustivité du recueil est diminuée**.

6.4.2.4 Le MNAR est souvent la règle, surtout en recherche sur les données

Les mémoires académiques réalisés en santé (thèses d'exercice, mémoires de master, thèses d'université) relèvent très majoritaire de trois types :

- Des recherches observationnelles sur **le soin courant**, non-financées
- Des recherches **sur des données** (par lecture humaine ou extraction de données)
- Des évaluations de pratiques professionnelles basées sur des **questionnaires**

Ces trois types sont **systématiquement** entachés de **données manquantes MNAR**, rendant **impossible** l'utilisation de techniques agnostiques d'imputation de données manquantes.

Pour illustrer cela, nous vous livrons les effectifs rencontrés dans la base nationale du PMSI. On y recherche le code CCAM d'une radiographie de la cuisse, et le code CIM10 d'une fracture de fémur. On obtient les effectifs rapportés sur la Figure 48 :

- Lorsqu'un patient n'a pas de radiographie du fémur, la probabilité d'avoir une fracture est de 0,53% (un patient sur 200)
- Lorsqu'un patient a une radiographie du fémur, la probabilité d'avoir une fracture est de 41,9% (un patient sur 2), soit un odds ratio de 134

Imaginons maintenant la recherche suivante sur les données. On souhaite estimer la proportion de fractures. Si une radiographie est présente et révèle une fracture, on coche « oui ». Si une radiographie est présente et ne révèle pas de fracture, on coche « non ». Dans les autres cas, la valeur est « NA ». On obtient alors le tableau reproduit en bas de la Figure 48 :

- 41,9% de réponses « oui » (parmi les réponses non-manquantes)
- 58,1% de réponses « non » (parmi les réponses non-manquantes)
- Pour information, 99,7% de réponses manquantes (parmi toutes les réponses)

L'hypothèse MCAR, totalement délirante, amènerait à imputer ces valeurs manquantes par 41,9% de réponses « oui ». On conclurait ainsi que 7,49 millions d'individus supplémentaire ont eu une fracture, alors qu'ils ne sont que 95 583 de plus. L'hypothèse MAR est également fausse, car il est évident que la probabilité d'avoir une radio du fémur est beaucoup plus faible si la radiographie est normale. Face à 99,7% de données manquantes, on peut de toute manière conclure qu'aucune hypothèse ne tient la route.

		fracture fémur		
		non	oui	
radio cuisse	non	17 770 581	95 583	*17 866 164*
	oui	28 605	20 666	*49 271*
		17 799 186	*116 249*	*17 915 435*

$$P(fracture_{femur} \mid \overline{radio_{cuisse}}) = 95583/17866164 = 0{,}53\%$$

$$P(fracture_{femur} \mid radio_{cuisse}) = 20666/49271 = 41{,}9\%$$

$$OR = 134{,}3$$

Fracture de hanche attestée par une radiographie	Effectif	Proportion
oui	20 666	41,9% des non-manquants
non	28 605	58,1% des non-manquants
NA	17 866 164	99,7% du total

Figure 48. Exemple dans la base nationale du PMSI (actes CCAM et diagnostics CIM10)

Deux autres options sont possibles.

La première consiste à penser que, en l'absence de radiographie, il n'y a pas de fracture. Cette option n'est pas exacte, puisqu'on voit bien que 0,53% des cas sans radiographie ont une fracture, mais elle est pleine de bon sens. Cependant, pour un examen rare (or en Santé tous les examens et toutes les maladies sont rares), elle n'est pas sans impact : elle ignore tout de même 82,2% des cas de fracture.

La deuxième option est présentée juste après.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6.5 Simple et efficace : identifier le NA comme une modalité

Dans le cas pratique extrême mais très réaliste (et même exact !) de la Figure 48, une deuxième option consiste simplement à décrire la réponse « manquant » **comme une nouvelle modalité**. On obtient alors les réponses suivantes :

- 0,11% de réponses « radiographie positive »
- 0,16% de réponses « radiographie négative »
- 99,7% de réponses « radiographie absente »

Formellement, il n'y a ainsi plus aucune donnée manquante, et toutes les procédures statistiques pourront être réalisées en cas complets. Dans ce cas précis, cette catégorie devient ultra-majoritaire et il s'agit d'une catégorie que tout le monde interprète comme une catégorie à très faible (mais non-nulle) probabilité de fracture. Cette attitude est **toujours valide**. Au pire, si les valeurs manquantes sont rares, ou en cas de MCAR, elle est simplement moins pertinente que l'analyse en cas complets. Vous ne prenez donc aucun risque à l'appliquer.

Ce procédé peut être appliqué à toutes les variables **qualitatives** (qui le restent) et à toutes les variables **binaires** (qui deviennent alors qualitatives à 3 modalités). Il peut également être appliqué aux variables **quantitatives**, comme l'illustre le Tableau 9. Il faudra lors **discrétiser** la variable, de manière à la rendre qualitative. Nous expliquons comment discrétiser une variable avec un tableur dans le chapitre 5.2.2 Option 2 : recodage par formule en page 90.

Tableau 9. Exemple de discrétisation de l'âge, avec données manquantes explicites (lignes 3 et 8)

id_patient	age	age_d
1	63	25-64
2	59	25-64
3	*NA*	inconnu
4	45	25-64
5	76	75-109
6	46	25-64
7	86	75-109
8	*NA*	inconnu

D'un point de vue statistique, de prime abord, la discrétisation d'une variable quantitative peut être perçue comme une perte d'information, qui prive de la possibilité d'utiliser certains outils (ex : régressions, corrélation, etc.). A bien y regarder, au contraire, cette discrétisation est souvent une bonne option, car les hypothèses sous-jacentes des méthodes statistiques applicables aux variables quantitatives sont rarement valides (ex : linéarité, logit-linéarité, log-linéarité, distribution et indépendance des résidus, homoscédasticité, etc.). Nous y reviendrons beaucoup plus tard.

6.6 Conduite à tenir, arbre décisionnel

Pour un mémoire académique en santé, nous proposons l'arbre décisionnel de la Figure 49.

Pour les analyses univariées et bivariées (qui sont de loin les plus fréquentes), si les valeurs manquantes sont peu nombreuses, le plus simple est d'analyser les cas complets uniquement (comme presque tout le monde). En cas de nombreuses valeurs manquantes (ex : ¼ ou plus), il peut être efficace de représenter les valeurs manquantes comme une modalité à part entière.

Pour les analyses multivariées, on analysera les cas complets s'ils sont suffisamment nombreux, en s'aidant des subterfuges énoncés précédemment. On n'aura recours à l'imputation de données manquantes, avec l'appui d'une cellule de biostatistique, que dans les recherches sur les personnes protocolisées et financées, si l'hypothèse MCAR ou MAR

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

semble valide. Dans tous les autres cas, on pourra représenter les données manquantes comme une nouvelle modalité.

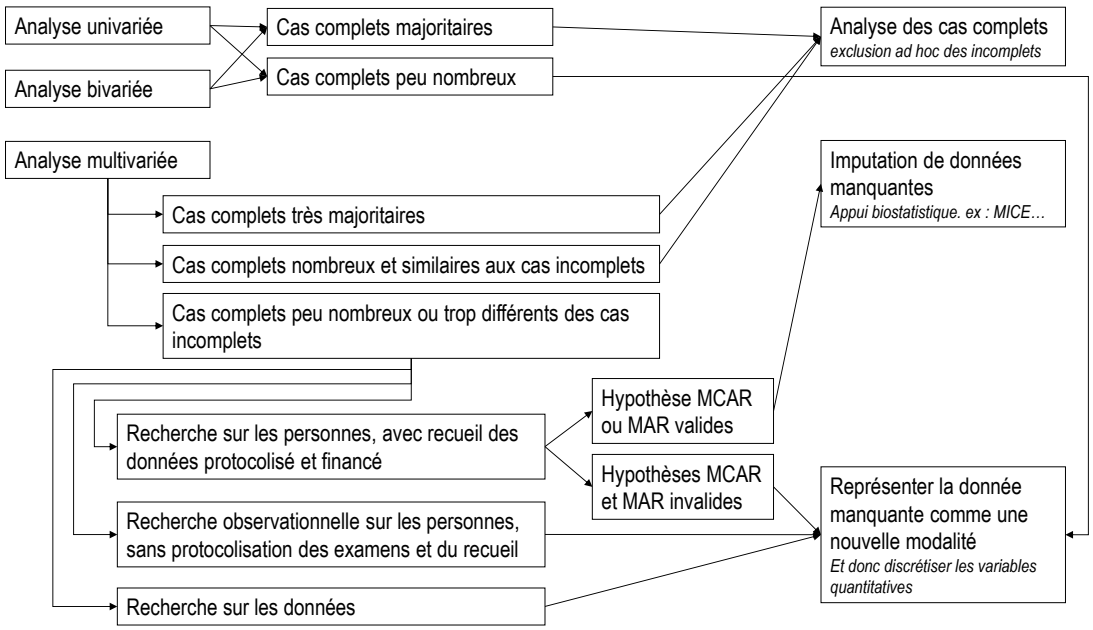


Figure 49. Gestion des données manquantes : conduite à tenir

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Réaliser les analyses statistiques

1 Préambule

Les analyses statistiques débutent toujours par l'analyse descriptive systématique de chaque colonne du fichier (ou d'une paire de colonne pour la survie) : c'est l'**analyse univariée** (voir chapitre 2 Analyses statistiques univariées en page 105). Cette analyse présente plusieurs intérêts :

- Elle termine de détecter les anomalies des données, à corriger.
- Elle permet à l'analyste de mieux comprendre les données, et vérifier tout de suite certaines hypothèses nécessaires pour choisir les méthodes d'analyses bivariées et multivariées.
- Elle apporte, en soi, des réponses très utiles à l'étude. Dans certaines études, les analyses univariées sont suffisantes.

Le cadre de l'analyse univariée inclut :

- la **description** des variables (graphiques, paramètres)
- le calcul d'**intervalles de confiance** des moyennes et des proportions
- les **tests** de comparaison d'une moyenne ou d'une proportion observée à une valeur **attendue**
- les **tests** de comparaison « de deux moyennes **appariées** » ou « de deux proportions **appariées** » (en simple, la description d'une évolution dans le temps, ou d'une différence gauche-droite)

Les analyses peuvent ensuite être poursuivies par des **analyses bivariées** (voir chapitre 3 Analyses statistiques bivariées en page 150), qui explorent le lien entre deux variables (ex : les hommes sont-ils plus lourds que les femmes ? Autrement dit, le poids est-il indépendant du sexe, pour sa tendance centrale ?). Dans cette partie :

- nous verrons comment explorer (plus ou moins) l'indépendance statistique :
  - entre deux variables qualitatives (comparer des proportions), éventuellement en tenant compte d'un appariement
  - entre une variable qualitative et une variable quantitative (comparer des moyennes), éventuellement en tenant compte d'un appariement
  - entre deux variables quantitatives (corrélation, régression)
- nous verrons également certains cas particuliers d'analyse bivariée, où l'on s'intéresse plus à quantifier la force de l'association, qu'à démontrer l'absence d'indépendance, qui est déjà connue :
  - facteurs de risque et facteurs protecteurs en épidémiologie
  - tests diagnostiques à réponse binaire ou quantitative (courbe ROC)
  - concordance entre deux juges (coefficient Kappa)

Les analyses peuvent enfin se conclure par des **analyses multivariées** (voir chapitre 4 Analyses statistiques multivariées, en bref en page 212), qui visent soit à décrire de manière agnostique des individus ou des groupes d'individus (méthodes non-supervisées, parfois relevant des classifications), soit, le plus souvent en santé, à expliquer et prédire une variable en particulier (méthodes supervisées, relevant des classifications ou des régressions).

Dans le chapitre suivant, nous partagerons quelques réflexions sur les tests statistiques, notamment les tests qui n'ont pas vraiment de sens (voir chapitre 5 Réflexions sur certains tests statistiques ou leur paramétrage en page 214).

Enfin, nous verrons comment interpréter une association statistique, et les pièges à éviter, dont les biais (voir chapitre 6 Interpréter une association statistique en général en page 226).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

L'ensemble des chapitres qui suivent s'adressent à une personne qui souhaite :

- **comprendre** les analyses statistique
- **réaliser** elle-même ces analyses **avec un tableur**
- identifier les quelques cas où elle devra **se faire accompagner**
- **interpréter et présenter** les résultats de ces analyses

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2 Analyses statistiques univariées

2.1 Avant de commencer…

Les analyses univariées s'intéressent à chaque colonne du fichier, indépendamment des autres. La première étape est d'identifier le type de variable dont il s'agit. Les indicateurs à calculer sont alors imposés par des normes qui simplifient énormément le travail : nous les présenterons.

2.1.1 Définir le type de variable

Nous avons vu une première fois dans le chapitre 1.1 Rappel sur les types de variables en page 59 comment les variables pouvaient être catégorisées, puis nous avons vu comment les saisir dans un tableau. Nous verrons dans ce chapitre que ce classement peut être simplifié pour les analyses statistiques (voir Figure 50 en page 106).

Les **variables quantitatives** correspondent à des nombres sur lesquels il est possible de réaliser des opérations algébriques. Elles sont **discrètes** lorsque le nombre de modalités est fini (ex : nombre d'enfants), et **continues** lorsque le nombre de modalités est infini (ex : poids, qui peut être égal à 81,48615762…). Nous rappelons à cette occasion que les dates et les délais sont des variables quantitatives (voir chapitre 4.3.4 Dates en page 78). En pratique, nous les traiterons comme des **variables quantitatives** si elles sont « brutes », ou des **variables qualitatives** si elles sont discrétisées (voir Figure 50 en page 106).

La **discrétisation** est une opération qui consiste à transformer des nombres en « étiquettes textuelles » pertinentes du point de vue du spécialiste des données en question. En voici deux exemples :

- Combien avez-vous d'enfants ? La réponse brute appartient à l'ensemble des entiers naturels $\mathbb{N}$ *{0 ; 1 ; 2 ; 3…}*. On pourrait remplacer ces valeurs par les modalités qualitatives « aucun », « un », « plus d'un ».
- Quel âge avez-vous ? La réponse appartient à un sous-ensemble des réels positifs $\mathbb{R}+$: l'intervalle *[0 ; 125]*. On pourra remplacer ces valeurs par les modalités "[15;25[", "[25;65[", "[65;75[", "[75;110[" (voir cet exemple détaillé en partie 5.2.2 Option 2 : recodage par formule en page 90)

La discrétisation peut être réalisée dès la conception du questionnaire : ce n'est pas conseillé d'un point de vue analyse de données, mais cela peut correspondre à une exigence d'anonymat. Idéalement, elle est réalisée durant le nettoyage des données et sans supprimer les données initiales (comme dans la partie 5.2.2 Option 2 : recodage par formule en page 90), ou durant l'analyse statistique. Il est toujours plus approprié de connaître la distribution de la variable dans l'échantillon avant d'entreprendre une discrétisation.

Les **variables qualitatives (monovaluées)** correspondent à des questions pour lesquelles il existe une seule réponse textuelle possible, parmi un nombre fini de modalités. Elles peuvent être non-ordonnées (exemple : la couleur des cheveux « brun », « blond », « blanc », etc.). Elles peuvent être ordonnées (exemple : le stade d'un cancer « 1 », « 2a », « 2b », « 3 »). Dans tous les cas, elles seront simplement traitées comme des **variables qualitatives** (voir Figure 50 en page 106).

Les **variables binaires** correspondent à des questions auxquelles il n'y a que deux réponses possibles, qu'on peut exprimer par oui/non, 0/1, vrai/faux, etc. Elles seront traitées comme des **variables qualitatives** (voir Figure 50 en page 106).

Les **variables qualitatives multivaluées** correspondent à des questions auxquelles plusieurs réponses peuvent être simultanément apportées. Ex : « quel(s) moyen(s) de transport utilisez-vous habituellement pour aller travailler (plusieurs réponses possibles) ? à pied / à vélo / en voiture / en transports en commun ». Elles sont saisies comme autant de variables binaires

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

qu'il y avait de réponses possibles. Pour le statisticien, ces variables n'existent donc plus en tant que telles, il s'agira d'un ensemble de variables binaires qui pourront être analysées séparément les unes des autres, et éventuellement présentées côte-à-côte (pour la présentation des données dans un tableau, voir 4.3.7 Variables qualitatives multivaluées en page 80).

Les **événements binaires dépendants du temps** sont saisis sur deux colonnes couplées. Ils seront décrits comme des **variables de survie** (pour la présentation des données dans un tableau, voir 4.3.8 Variables décrivant un événement temps-dépendant (survie) en page 81).

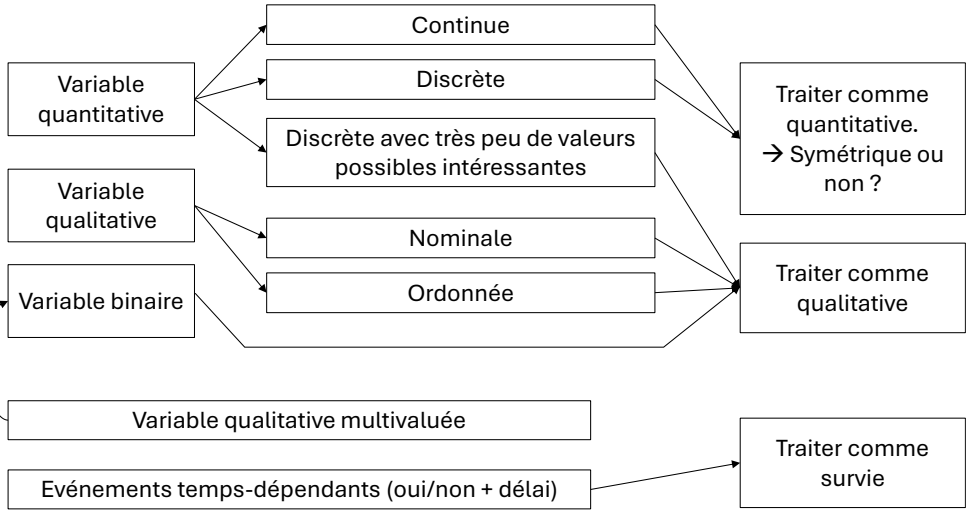


Figure 50. Du type de variable à son traitement pour l'analyse statistique

Cette simplification nous ramène donc à trois types de variables :

- Les variables **quantitatives**
- Les variables **qualitatives**
- Les variables **de survie**

2.1.2 Généralités sur le choix des indicateurs à produire

En théorie, une foule d'indicateurs sont calculables pour chaque type de variable. En pratique, il existe des normes internationales, référencées sur https://www.equator-network.org/, qui indiquent ce qui doit être calculé et présenté dans les résultats des études en santé (Figure 51). Les normes les plus fréquemment utilisées sont :

- **Strobe** pour les études **observationnelles** : cela inclut donc les questionnaires, et c'est de loin la norme la plus utile pour les mémoires académiques en santé
- **Consort** pour les **essais randomisés**
- **Prisma** pour les **revues systématiques** de la littérature

Dans la suite de cet ouvrage, nous vous présenterons uniquement ce qu'il convient de décrire, en fonction du type de variable. Vous constaterez que c'est généralement très succinct !

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

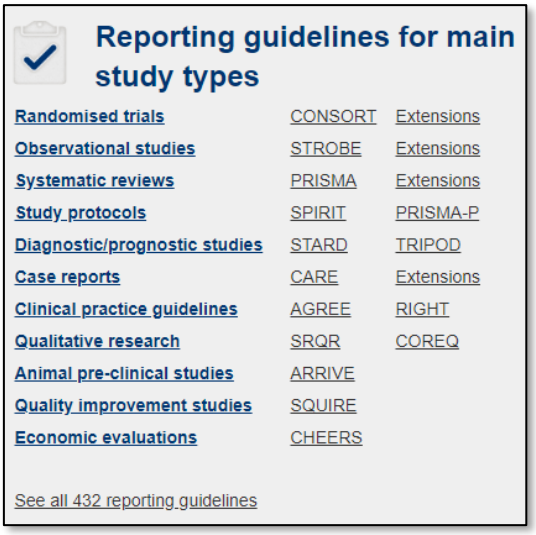


Figure 51. Capture d'écran du site https://www.equator-network.org

2.1.3 Ce que nous verrons ici

Dans cette section « analyses univariées », nous aborderons :

- La **description systématique** de chaque colonne, en fonction de son type. Nous verrons que c'est, de loin, **le plus important**.
- Le calcul d'**intervalles de confiance** des moyennes et des proportions, qui sera abordé en même temps que la description. Nous insisterons sur le fait que, en général, ces intervalles de confiance **ne doivent pas être calculés**.
- Les **tests d'adéquation** entre une moyenne observée et une moyenne attendue, ou entre une proportion observée et une proportion attendue. Nous verrons ensuite que ces tests, bien que hautement valides, ne se trouvent que dans des recherches de piètre qualité, et **ne doivent donc pas être réalisés**.
- Les **tests de comparaison appariées (avant/après, gauche/droite)** dans un seul groupe. Nous verrons ensuite que ces tests, bien que hautement valides, se trouvent généralement dans des recherches de piètre qualité, et **sont généralement déconseillés**.

Bref, seules les parties relatives à la description des variables vous seront réellement utiles. J'espère que cela vous confortera dans l'idée que la statistique en santé est avant tout un ensemble de pratiques simples, reproductibles, et fondées sur le bon sens plutôt que sur des formules mathématiques complexes.

2.2 Variables qualitatives

2.2.1 Description et présentation

Pour décrire une variable qualitative, il est nécessaire et suffisant de présenter un tableau (ou une phrase) listant les **modalités rencontrées** (après recodage), avec à chaque fois leur **effectif brut** et leur **proportion** (pourcentage). La proportion est simplement l'effectif de la modalité divisé par l'effectif de l'échantillon, après éviction des valeurs manquantes. On présentera l'effectif brut comme le résultat principal, et la proportion entre parenthèses.

Cette attitude se justifie ainsi. Premièrement, l'effectif brut est une donnée factuelle, tandis que le pourcentage est déjà une forme d'interprétation, car on peut choisir de rapporter l'effectif à tel ou tel sous-groupe de l'échantillon, notamment lorsqu'une question ne s'adresse qu'à une partie de l'échantillon. Deuxièmement, lorsqu'on dispose de ces deux informations, on peut

recalculer la taille de l'échantillon, et retrouver tout un tas d'informations, comme l'intervalle de confiance d'une proportion. Il n'est donc pas utile de donner plus d'informations que cela.

Si vous utilisez un tableur, vous pourrez aisément retrouver ces effectifs et proportions à l'aide d'un **tableau croisé dynamique**, ou **table dynamique**, ou **pilote de données**, ou ***pivot table*** (selon votre logiciel).

Les règles d'écriture de la langue française imposent de ne jamais débuter une phrase par un nombre écrit en chiffres. Généralement, lorsqu'on souhaite débuter une phrase par un nombre, on l'écrit en toutes lettres lorsqu'il s'écrit en un ou deux mots, sinon on ajoute un complément circonstanciel, du type « Parmi les patients, … ».

Enfin, pour ce qui est de la représentation graphique, tout un tas de graphiques sont à votre disposition. Ces graphiques ont tous un point commun : **une quantité géométrique est proportionnelle à la proportion** de la modalité qu'elle représente. Généralement, cette quantité est la surface d'une figure géométrique. Il peut s'agir de (Figure 52 et Figure 53) :

- un rectangle dont une seule dimension varie proportionnellement à la fréquence
- un secteur de disque dont l'angle au centre est proportionnel à la fréquence
- un disque dont le rayon est proportionnel **à la racine carrée** de la fréquence
- un carré dont le côté est proportionnel **à la racine carrée** de la fréquence
- un rectangle dont les deux dimensions varient, de telle manière que leur produit reste proportionnel à la fréquence

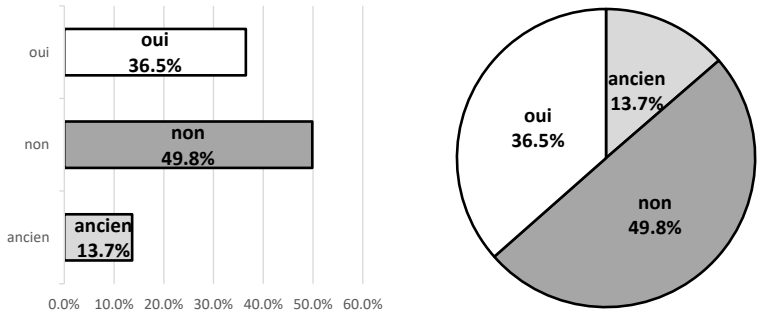


Figure 52. Représentation des modalités d'une variable qualitative par des graphiques représentant la fréquence sur une dimension : diagramme barres (gauche), diagramme en secteurs (camembert, droite)

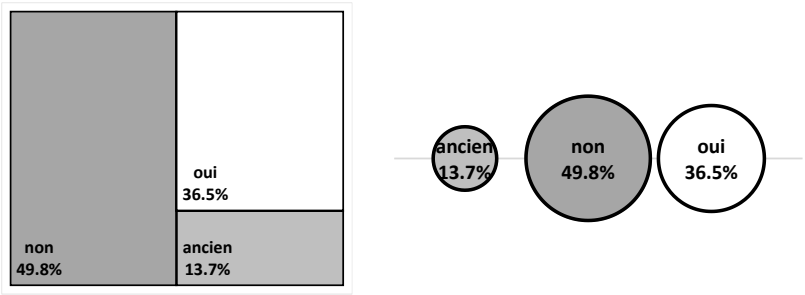


Figure 53. Représentation des modalités d'une variable qualitative par des graphiques représentant la fréquence sur deux dimensions : treemap (gauche), diagramme en bulles (droite)

De manière générale, il faudra se méfier des graphiques dont deux ou trois dimensions varient en même temps, et s'assurer que la quantité (surface ou volume projeté) ne soit pas proportionnelle au carré ou au cube de la proportion, mais bien à la proportion représentée (Figure 53).

Tous ces graphiques se construisent sur un tableau de contingence, qui présente la fréquence par modalité. On veillera également à exclure du graphique un éventuel total du tableau.

2.2.2 Exemples de présentation selon le sous-type de variable

Pour une **variable qualitative quelconque**, on choisira volontiers un ordre par fréquence décroissante. On veillera à présenter toutes les proportions avec la même précision (souvent un pourcentage avec un chiffre après la virgule), et en indiquant explicitement « ,0 » si nécessaire. Voici un exemple de rédaction :

Parmi les patients, 22 (44,0%) sont bruns, 16 (32,0%) sont blonds et 12 (24,0%) ont les cheveux blancs.

Pour une **variable qualitative ordonnée**, le principe est le même, mais il faut s'astreindre à respecter l'**ordre** croissant ou décroissant des modalités. Voici un exemple :

Parmi les patients, 13 (26,0%) ont une atteinte de grade I, 15 (30,0%) une atteinte de grade II et 22 (44,0%) une atteinte de grade III.

Pour une **variable binaire**, il n'est pas nécessaire de mentionner les deux modalités. Avec bon sens, on mentionnera seulement la modalité la plus frappante, ou la plus fréquente si elles sont au même niveau sémantique. On notera que, si la variable est codée « 1 » pour la modalité qui nous intéresse et « 0 » pour l'autre, alors il suffit de calculer **la moyenne** de la colonne pour obtenir la proportion de la modalité d'intérêt. Voici deux exemples :

Seulement 2 patients (4,0%) sont décédés.

L'échantillon comporte 36 (72,0%) femmes.

Enfin, les différentes variables binaires issues d'une **variable qualitative multivaluée** pourront être décrites dans la même phrase, mais en faisant comprendre au lecteur que plusieurs modalités peuvent cohabiter chez les mêmes sujets :

Parmi les antécédents, on retrouve 5 cas de diabète (10,0%), 3 cas d'infarctus (6,0%) (…), étant entendu qu'un patient peut présenter plusieurs antécédents. Douze patients (24,0%) n'avaient aucun antécédent.

2.2.3 Calcul de l'intervalle de confiance d'une proportion

2.2.3.1 Préambule

Les pourcentages que nous avons calculés précédemment sont formellement des **proportions mesurées dans l'échantillon**, et ce sont aussi, de facto, une **estimation de la proportion (inconnue) en population**. Cet ouvrage n'étant pas un ouvrage de statistique, nous ne définirons pas lourdement ce qu'est l'estimation et quels sont ses fondements mathématiques. Sachez simplement qu'il est possible de calculer l'**intervalle de confiance à 95% (IC95)** de chaque proportion, c'est-à-dire en pratique l'intervalle dans lequel la vraie proportion inconnue en population a 95% de chances de se trouver, compte tenu de l'observation que vous avez réalisée dans l'échantillon. Ce pourcentage, 95%, s'est imposé comme une convention en santé, et correspond donc à un risque de première espèce de 5% (risque alpha, 5% de chances de se tromper).

D'après les normes internationales citées précédemment (Consort, Strobe, Prisma), cet **IC95 ne doit pas être calculé, sauf** si votre objectif principal était justement d'estimer une proportion. Il y a deux raisons à cela :

- Le calcul d'un intervalle de confiance est, en soi, une inférence statistique, et on souhaite éviter l'inflation du risque de première espèce, lié à la répétition des inférences statistiques (nous en reparlerons)
- De toute manière, l'effectif brut et la proportion sont suffisants au lecteur pour recalculer cet intervalle de confiance, s'il le souhaite vraiment

Si vraiment vous souhaitez calculer un intervalle à 95% de la proportion en population, alors il est donné par une méthode qui utilise la Loi Normale. Nous donnerons directement le cas des intervalles de confiance à 95%, qui correspondent aux bonnes pratiques en santé. Le

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

coefficient donné par la Loi Normale est alors 1,96. Pour appliquer cette méthode, vous devrez avoir observé au moins 5 individus dans chaque modalité de votre variable. Cette même formule fonctionne pour chaque modalité d'une variable qualitative ou binaire, quel que soit le nombre de ces modalités.

2.2.3.2 Mise en œuvre avec un tableur

La mise en œuvre dans un tableur comme Microsoft Excel ou LibreOffice Calc est très simple (Figure 54). Il nous faut disposer de la fréquence du caractère étudié (nombre compris entre 0 et 1, ou entre 0% et 100% si vous préférez), et de l'effectif total de l'échantillon. La première étape consiste à calculer la demi-largeur de l'intervalle de confiance (en cellule P5, avec la formule reproduite en cellule Q5). Les deux bornes de l'intervalle sont alors données en retranchant ou en ajoutant ce demi-intervalle à la proportion observée (cellules P6 et P7 en Figure 54).

	O	P	Q
1	Effectif total :	315	
2	Fréquence :	13.70%	
3			
4	IC à 95% :		
5	demi-intervalle :	3.80%	=1.96*RACINE(P2*(1-P2)/P1)
6	borne basse :	9.90%	=P2-P5
7	borne haute :	17.50%	=P2+P5

Figure 54. Calcul de l'intervalle de confiance d'une proportion avec un tableur

On vérifiera par la même occasion que l'effectif de la modalité et l'effectif des autres modalités est bien supérieur ou égal à 5.

2.2.3.3 Quelques précisions sur le calcul de l'IC95 d'une proportion

L'intervalle à 95% de la proportion en population est donné par la méthode de Wald[27,28], qui utilise la Loi Normale (Équation 6).

$$IC_{0,95} = p \pm 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}$$

Équation 6. Intervalle de confiance d'une proportion, méthode de Wald
(p = proportion mesurée, n = effectif)
Conditions : n.p≥5 et n.(1-p)≥5

Cette méthode d'intervalle de confiance pose problème lorsqu'une des modalités comporte moins de 5 individus dans votre échantillon. Si tel est le cas, on lui préfèrera la méthode de Clopper et Pearson[29], que vous ne pourrez pas mettre en œuvre avec un simple tableur.

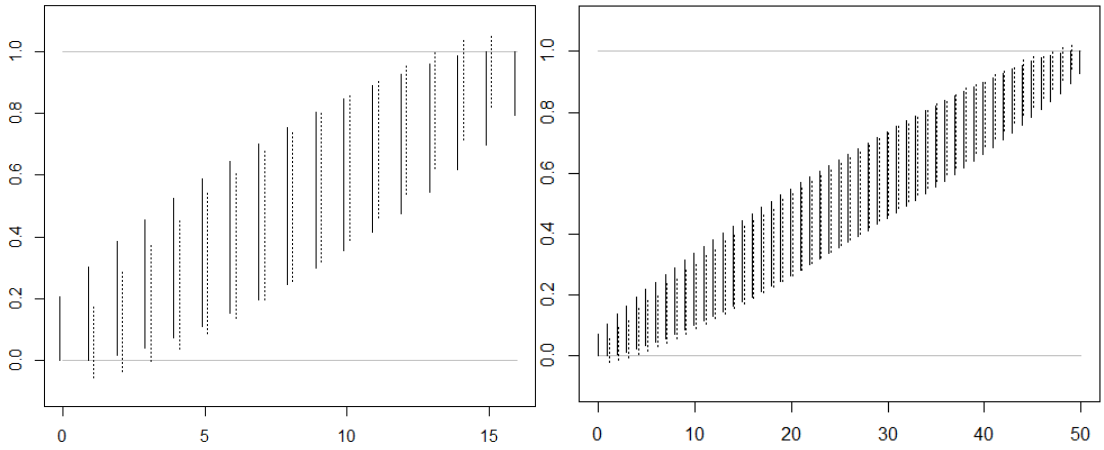


Figure 55. Calcul d'un IC95 (en Y) en fonction du nombre d'individus positifs (en X), parmi 16 (à gauche) ou 50 (à droite) individus. Segments pleins : Clopper et Pearson. Pointillés : loi normale

Sur la Figure 55 à gauche, nous supposons avoir un échantillon de 16 individus et, selon le nombre X d'individus (entre 0 et 16) observés ayant le caractère étudié, on obtient un intervalle de confiance à 95% de la proportion (projeté en Y) : en trait plein nous représentons la méthode de Clopper et Pearson (qui est réputée exacte) et en pointillés la méthode de Wald utilisant la loi normale. Sur la partie droite de la Figure 55, nous faisons de même avec un échantillon de 50 individus. On observe que sur les extrémités, notamment lorsqu'on a de 0 à 4 individus positifs, ou de n-4 à n individus positifs, la méthode de Wald donne des résultats :

- très différents, généralement avec un IC95 plus étroit
- parfois erronés, avec une borne inférieure négative, ou une borne supérieure excédant 1

2.2.4 Tests de comparaison d'une proportion observée à une proportion attendue

2.2.4.1 Introduction

Maintenant que l'on sait calculer l'IC95 d'une proportion inconnue en population, des méthodes similaires permettent de savoir si l'observation qu'on a réalisée dans l'échantillon (exemple : « sur *n* individus, *n.p* sont porteurs du caractère ») est **compatible avec une hypothèse externe** (exemple : « en population, la proportion π inconnue de ce caractère vaut p_0 »). Cette hypothèse peut provenir d'articles scientifiques, de données publiées par un institut national de statistique, d'une affirmation infondée qu'on souhaite réfuter, etc. On l'appelle **H_0, hypothèse nulle**, et on la formule comme suit :

$$H_0 : \pi = p_0$$
$$\pi = proportion\ inconnue\ en\ population$$
$$p_0 = proportion\ alléguée\ par\ une\ source\ quelconque$$
$$p = x/n = proportion\ mesurée\ dans\ notre\ échantillon\ de\ taille\ n$$

Équation 7. Hypothèse nulle d'un test de comparaison d'une proportion observée à une proportion attendue

Si notre observation $p=x/n$ est **trop éloignée** de la valeur attendue p_0, alors **on rejette H_0** : cela signifie que notre échantillon n'est pas issu aléatoirement d'une population dans laquelle la probabilité individuelle du caractère est de p_0. En pratique, on peut l'interpréter de deux manières : on peut dire que $\pi \neq p_0$ au sens « ils racontaient n'importe quoi », ou on peut dire que l'échantillon n'est pas issu par tirage au sort de la population dont il est question. Autrement dit, si on s'intéresse à une sous-population particulière (ex : les femmes enceintes, versus les autres humains), on pourra conclure que cette sous-population est différente du reste de la population. D'un point de vue mathématique, ces deux conclusions sont similaires,

l'interprétation dépendra donc du contexte. On conclura quelque chose comme « la proportion observée *p* est significativement différente de p_0 ».

Si notre observation *p* est **assez proche** de la valeur attendue p_0, alors **on ne rejette pas H_0**. Observer un phénomène qui est compatible avec une hypothèse ne suffit pas à la confirmer (ex : vous trouvez 55% de femmes dans un échantillon, et untel affirme que la proportion est de 55% en population… peut-être qu'en réalité cette proportion était de 53% en population ? on ne peut pas savoir !). On conclura quelque chose comme « on ne met pas en évidence de différence significative entre la proportion observée *p* et la proportion attendue p_0 ».

De manière générale, on comprend intuitivement que, pour un effectif *n* fixé, plus la différence observée $p - p_0$ est importante, plus on sera confiant en rejetant H_0.

De même, pour une différence observée $p - p_0$ donnée, plus l'effectif de l'échantillon est élevé, plus on sera confiant en rejetant H_0.

Par convention en santé, on rejettera H_0 si une méthode mathématique, appelée **test statistique**, nous permet de le faire avec une confiance suffisamment élevée. Pour ce faire, nous nous attacherons à calculer une **p valeur (ou « petit p », ou *p value*)**, qui est, en supposant que H_0 est vraie, la probabilité qu'on avait d'observer quelque chose comme notre échantillon, ou quelque chose d'encore plus éloigné de H_0. Ce « encore plus éloigné » s'entend « d'un côté comme de l'autre », « au-dessous comme au-dessus », on parlera de test bilatéral. Ces notions deviendront plus claires quand nous verrons ensemble un exemple.

Si on souhaite comparer une proportion observée à une proportion attendue, plusieurs options s'offrent à nous :

- Dans tous les cas, on pose $H_0 : \pi = p_0$
- Option 1 : on pourra simplement **calculer l'IC95** de la proportion en population : si la valeur attendue p_0 n'appartient pas à cet IC95, on pourra rejeter H_0, avec moins de 5% de chances de se tromper, tout simplement ! Nous avons vu deux méthodes pour calculer cet IC95 :
  - Option 1A : la **méthode de Wald**, basée sur la loi normale
  - Option 1B : La **méthode de Clopper et Pearson**, toujours valide, mais que vous ne pourrez pas mettre en œuvre avec un tableur
- Option 2 : on pourra réaliser un **test statistique**, et rejeter H_0 si sa **p valeur** est inférieure à 5%. Plusieurs tests sont accessibles :
  - Option 2A : le **test binomial**, qui est un test exact : nous le présenterons en détail car il permet de très bien comprendre comment les tests statistiques fonctionnent
  - Option 2B : le **Khi-deux (ou Khi² ou X²) d'adéquation**, qui est de loin le test le plus populaire
  - Option 2C : un test basé sur la **Loi Normale**, qu'en pratique on n'utilise pas

Dans les deux chapitres qui suivent, nous détaillerons le **test binomial**, que nous vous conseillons de bien comprendre, et le **test du Khi² d'adéquation**, que vous utiliserez peut-être en pratique. Nous verrons cependant plus tard (voir chapitre 5.1 Tests de comparaison à une norme en page 214) qu'il n'est pas vraiment conseillé d'utiliser ces tests. Nous vous proposons ci-dessous un algorithme décisionnel simple et efficace (Figure 56) : il tient compte à la fois de vos souhaits et aptitudes, et des conditions de validité des outils proposés.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

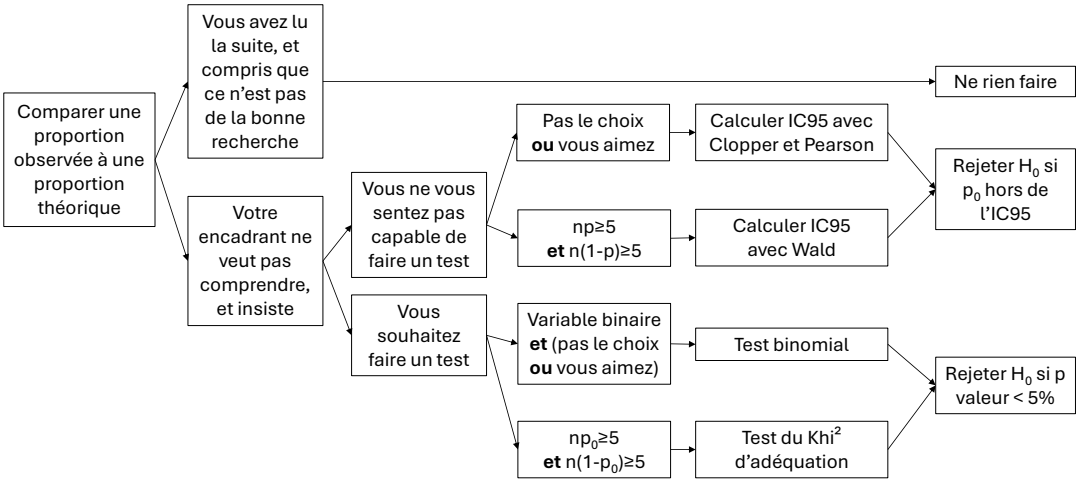


Figure 56. Arbre décisionnel : comparer une proportion observée à une proportion attendue

2.2.4.2 Test binomial

Le test binomial est un **test exact**, et **non-paramétrique** (car, formellement, il ne s'intéresse pas à la proportion mais aux effectifs). Nous reviendrons sur ces classifications (purement théoriques) dans la section 5.4 Test paramétrique ou non-paramétrique ? en page 217.

2.2.4.2.1 Introduction

On notera :

π = proportion théorique en population

p = proportion observée dans un échantillon

p_0 = une proportion attendue

n = taille de votre échantillon

On s'intéresse à une **variable binaire**, et à la survenue de sa modalité « 1 ». Vous avez observé, dans UN échantillon de taille n, une proportion p. Une source externe vous affirme que, en population, cette proportion devrait valoir une certaine valeur p_0.

On pose l'hypothèse nulle $H_0 : \pi = p_0$

Si p est suffisamment éloignée de p_0, ce test vous permettra de rejeter H_0, et d'affirmer que, dans la population dont votre échantillon est issu, π ne vaut pas p_0.

Pour ce faire, en supposant que H_0 est vraie, on calculera la « p valeur ». La p valeur est la probabilité d'observer cette valeur observée de p ou une valeur de p plus éloignée encore de p_0. La p valeur est donc un indicateur de plausibilité de notre observation, en supposant que H_0 est vraie. Nous utiliserons deux conventions qui se sont imposées dans le monde de la recherche en santé (hormis pour les tests de non-infériorité) :

- Le test sera **bilatéral** :
  - On ne cherche pas à montrer que p est spécifiquement supérieur, ou spécifiquement inférieur, à p_0 : on cherche à montrer que p est différent de p_0
  - Concrètement, la p valeur sera calculée de manière **bilatérale**
- Le test sera réalisé au **risque alpha de 5%** :
  - C'est, de manière consensuelle, le seuil retenu en Santé
  - Une observation qui avait moins d'une chance sur vingt de se produire est considérée comme peu vraisemblable
  - On **rejettera donc H_0 si p valeur <5%**

2.2.4.2.2 Rappels mathématiques niveau lycée

Dans ce test, nous calculerons des probabilités de manières exactes. Pour ce faire nous aurons besoin de méthodes mathématiques définies au lycée. Nous les rappelons ici.

Vous disposez en main de *x* éléments. Combien de suites ordonnées de ces *x* éléments pouvez-vous réaliser ? Vous devez d'abord choisir un premier élément (*x* possibilités) puis un deuxième élément (*x-1* possibilités restantes), puis un troisième, etc. La réponse est une **factorielle**, « factorielle de x », qui se note avec la lettre x suivie d'un point d'exclamation (Équation 8).

$$x! = x \times (x-1) \times (x-2) \times \ldots \times 2 \times 1$$

Équation 8. Factorielle de x

A présent, vous disposez en main de *n* éléments. Vous souhaitez n'en utiliser que *x*, avec $0 \leq x \leq n$. Combien de suites ordonnées de *x* éléments pouvez-vous réaliser ? Vous devez d'abord choisir un premier élément (*n* possibilités) puis un deuxième élément (*n-1* possibilités restantes), puis un troisième, etc. Simplement, à la différence du cas précédent, vous vous arrêterez dès que vous aurez *x* éléments. Pour le x^{ième} élément, vous aviez encore *n-x+1* possibilités. La réponse est **un arrangement**, « arrangement de x éléments parmi n » (Équation 9).

$$A_n^x = n \times (n-1) \times \ldots \times (n-x+1)$$

$$A_n^x = \frac{n \times (n-1) \times \ldots \times (n-x+1) \times (n-x) \times \ldots \times 2 \times 1}{(n-x) \times \ldots \times 2 \times 1}$$

$$A_n^x = \frac{n!}{(n-x)!}$$

Équation 9. Arrangement de x éléments parmi n

Enfin, vous disposez en main de n éléments, vous souhaitez en utiliser x pour réaliser des *mains* et non des *suites*, c'est-à-dire que l'ordre des éléments une fois sélectionnés n'a pas d'importance. Si nous utilisions l'arrangement, chaque main de *x* éléments serait représentée par *x!* suites différentes. La réponse à cette question est donc une **combinaison**, la « combinaison de x éléments parmi n ». Nous obtenons sa valeur en divisant l'arrangement par *x!*. La combinaison peut être notée avec une écriture française utilisant la lettre *C*, ou avec l'écriture internationale utilisant les grandes parenthèses (Équation 10).

$$C_n^x = \binom{n}{x} = \frac{A_n^x}{x!} = \frac{n!}{(n-x)!.x!}$$

Équation 10. Combinaison de x éléments parmi n

La combinaison étant définie, nous pouvons définir ce qu'est la **loi binomiale**.

Imaginons que, en population, la probabilité d'un caractère soit π. Dans un échantillon de taille *n*, quelle est la probabilité d'observer exactement *x* cas portant ce caractère ? On peut raisonner ainsi (on parlera de cas positif et de cas négatif pour simplifier la rédaction). Souhaitons tout d'abord que le premier cas soit positif (probabilité π), puis que le deuxième cas aussi (probabilité π) et ainsi de suite jusqu'au x^{ième} : cela est associé à une probabilité de π^x. Il faut ensuite que tous les suivants soient négatifs : *n-x* éléments avec une probabilité de $1-\pi$, cela est associé à une probabilité de $(1-\pi)^{n-x}$. Ayant ainsi imposé que les *x* premiers soient positifs et les *n-x* suivants soient négatifs, nous avons fortement restreint les possibilités par rapport à la question posée. La question est donc de savoir, dans mon échantillon, sur quelles positions disposer les *x* malades, parmi les *n*. La réponse est C_n^x : en remultipliant la probabilité calculée précédemment, on retire l'effet de cet ordre arbitraire. La probabilité, sur *n* individus,

qu'exactement *x* soient porteurs du caractère étudié, est donc donnée par la Loi Binomiale, comme l'indique l'Équation 11.

$$P(X = x) = \pi^x . (1 - \pi)^{(n-x)} . C_n^x$$
$$x \in \{0,\ 1,\ \ldots,\ n\}$$

Équation 11. Loi binomiale : probabilité d'observer x cas parmi n portant un caractère, si la probabilité de ce caractère en population est π.

2.2.4.2.3 Exemple de mise en œuvre du test

A présent, ces rappels étant faits, mettons en œuvre le test dans une situation concrète. Nous montrerons dans le même temps comment réaliser ce test avec un tableur.

Nous nous intéressons au cancer du larynx. Dans un échantillon de 16 individus, 12 sont des hommes. Or, d'après des données populationnelles, on s'attendait à trouver 49% d'hommes. Cette observation (12/16), est-elle compatible avec l'hypothèse (49%) ?

L'hypothèse nulle se formule ainsi :
H_0 : π = 0,49

Pour la forme, on peut également formuler une hypothèse alternative[11] :
H_1 : π ≠ 0,49

Dans la Figure 57, nous disposons tout d'abord l'effectif total (16 en cellule B1) et la probabilité (sous H_0) qu'un individu soit un homme, ou proportion attendue en population (0,49 en B2). Dans un échantillon de 16 individus, nous aurions pu en théorie observer 0, 1, 2,… 15 ou 16 hommes. Nous disposons ces hypothèses en cellules A5 à A21. Nous ajoutons en colonne B la probabilité que nous aurions eue d'observer chacun de ces nombres. Cette probabilité est donnée par l'Équation 11, en remplaçant *n* par 16 et *π* par 0,49 (ici on la note p_0 et non *π*, car ce n'est pas la vraie probabilité en population, mais une probabilité d'hypothèse). Le tableur utilisé nous permet d'aller un peu plus vite, en utilisant la formule Excel **loi.binomiale()** ou Calc **loibinomiale()** comme visible en haut de la Figure 57 (la formule est celle en cellule B5). La somme de ces probabilités vaut bien 1 (voir cellule B22). La situation observée avait exactement 2,36% de chances de se produire. C'est déjà assez peu, mais ceci n'est pas suffisant pour conclure : si nous avions eu un échantillon plus grand, chaque ligne du tableau aurait été associée à une probabilité très faible, y compris les lignes les plus probables (celles présentant une proportion observée proche de 0,49), pour que la somme reste égale à 100%. Pour conclure, nous regardons plutôt toutes les lignes associées à une probabilité inférieure ou égale à 2,36% : nous incluons les observations plus extrêmes (12, 13, 14, 15 et 16) mais aussi les observations moins probables, mais « de l'autre côté », pour des valeurs très faibles (0, 1, 2, 3, 4). Le fait d'aller chercher des observations possibles des deux côtés est ce qui caractérise un **test bilatéral**. Nous calculons la somme des probabilités associées à tous ces événements (colonne E) : cette somme est 4,53%, **c'est la p valeur**. Elle est inférieure à 5%, nous pouvons rejeter l'hypothèse nulle et conclure que l'échantillon est issu d'une population dans laquelle la probabilité d'être un homme n'est pas égale à 49%. On peut donc écrire que **la proportion observée d'hommes est significativement différente de 49%**. En termes d'interprétation médicale, on peut aller plus loin et penser qu'il s'agit probablement d'une maladie à prédominance masculine.

[11] La formulation de l'hypothèse alternative est habituelle dans les cours de statistique. En pratique, cela n'apporte rien (hormis le cas très particulier des essais de non-infériorité). Nous vous épargnerons la formulation des hypothèses H_1 dans cet ouvrage.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

B5 : =LOI.BINOMIALE(A5;$B$1;$B$2;FAUX)

	A	B	C	D	E
1	n	16			
2	p0	0.49			
3					
4	x	P(x)	Situation observée	Situation observée, ou moins probable ?	P(x) pour sélection
5	0	0.00%		1	2.0947E-05
6	1	0.03%		1	0.00032201
7	2	0.23%		1	0.002320363
8	3	1.04%		1	0.010403718
9	4	3.25%		0	0
10	5	7.49%		0	0
11	6	13.19%		0	0
12	7	18.11%		0	0
13	8	19.58%		0	0
14	9	16.72%		0	0
15	10	11.24%		0	0
16	11	5.89%		0	0
17	12	2.36%	ici !	1	0.023588757
18	13	0.70%		1	0.006973448
19	14	0.14%		1	0.00143571
20	15	0.02%		1	0.000183921
21	16	0.00%		1	1.10443E-05
22	Somme	100.00%		9	4.53%

Figure 57. Réalisation d'un test binomial avec un tableur

2.2.4.2.4 Le test binomial, en général

On dispose d'une probabilité attendue p_0 d'un caractère (autrement dit, proportion attendue en population).

Dans un échantillon de taille *n*, on observe *x* fois ce caractère, soit une proportion observée de *x/n* qui peut donc différer de p_0.

Trois questions, au fond identiques, peuvent se poser :

- La différence observée entre p_0 et *x/n* dépasse-t-elle le simple effet du hasard (fluctuations aléatoires liées à l'échantillonnage) ?
- La proportion observée *x/n* diffère-t-elle *significativement* de la proportion attendue p_0 ?
- Notre échantillon est-il plausiblement issu par tirage au sort d'une population caractérisée par $\pi=p_0$?

Nous répondons à ces questions par un **raisonnement par l'absurde**.

- Posons l'hypothèse nulle H_0 : « la probabilité du caractère est bien de p_0 »
- Calculons, sous l'hypothèse H_0, la **p valeur bilatérale** de l'observation, autrement dit la probabilité sous H_0 d'observer ce qu'on a observé OU d'observer une situation encore moins probable. Pour ce faire :
  - Pour chaque valeur possible de *x*, calculer sa probabilité sous H_0 avec une loi binomiale
  - Identifier la valeur de *x* observée, puis toutes les valeurs plus extrêmes en termes de probabilités (donc plus éloignées de l'hypothèse nulle), qu'il s'agisse de valeurs proches de *0* ou proches de *n*
  - Calculer la somme des probabilités associées à ces valeurs : c'est la p valeur
- Concluons :
  - Fixons le seuil d'interprétation de la p valeur à 5% (risque α, défini plus bas)
  - Si la p valeur est inférieure à 5%, on rejette l'hypothèse H_0
  - Sinon, on ne peut rien conclure

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Par définition, si H_0 est vraie, le **risque α** (alpha) ou **risque de première espèce** est la probabilité de rejeter H_0 à tort. Par convention en santé, on le fixe à 5%.

2.2.4.2.5 Quelques précisions sur le test binomial

Le test binomial, également appelé test binomial exact, a été publié par Ronald Fisher en 1925[30]. Il peut être classé ainsi :

- C'est un **test exact** : la p valeur calculée est toujours exacte, il ne suppose pas un effectif supérieur à un certain seuil par exemple (contrairement aux tests asymptotiques)
- C'est un **test non-paramétrique** car, formellement, il ne cherche pas à estimer un paramètre (la proportion) mais observe directement la probabilité associée à chaque effectif potentiellement rencontré. Cette distinction est un peu artificielle car cet effectif est proportionnel à la proportion, mais, formellement, x/n ne prétend pas toujours être une estimation valable de la proportion.

Le test binomial est un test qui est **toujours valide** si la variable étudiée est **binaire** (et non qualitative à plus de deux modalités).

La Figure 58 montre les effectifs limites permettant de rejeter H_0 (avec p valeur<5%) dans un test binomial. En haut de la Figure 58, on a fixé $p_0=50\%$. On voit que pour $n<6$ il n'est pas possible de rejeter H_0, quelle que soit la valeur de x. Pour des échantillons de taille *6, 7* ou *8*, on peut rejeter H_0 si $x=0$ ou $x=n$. Pour des échantillons de taille *9, 10* ou *11*, on peut rejeter H_0 pour x valant *0, 1, n-1* ou n. Pour des échantillons de taille *12, 13* ou *14*, on peut rejeter H_0 pour x valant *0, 1, 2, n-2, n-1* ou n. Et ainsi de suite.

En bas de la Figure 58, on montre ces limites de significativité pour $p_0=10\%$. On voit alors qu'il devient très facile de rejeter H_0, même avec des effectifs faibles, par exemple pour 2 individus portant le caractère sur un total de 2, ou 2 sur 3, ou 3 sur 4, mais qu'il faut de plus gros échantillons pour pouvoir rejeter H_0 pour des proportions observées plus faibles que la proportion attendue (lorsqu'on obtient 0 cas sur 34 et plus). La figure du milieu en Figure 58 montre ces limites de significativité pour $p_0=25\%$.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

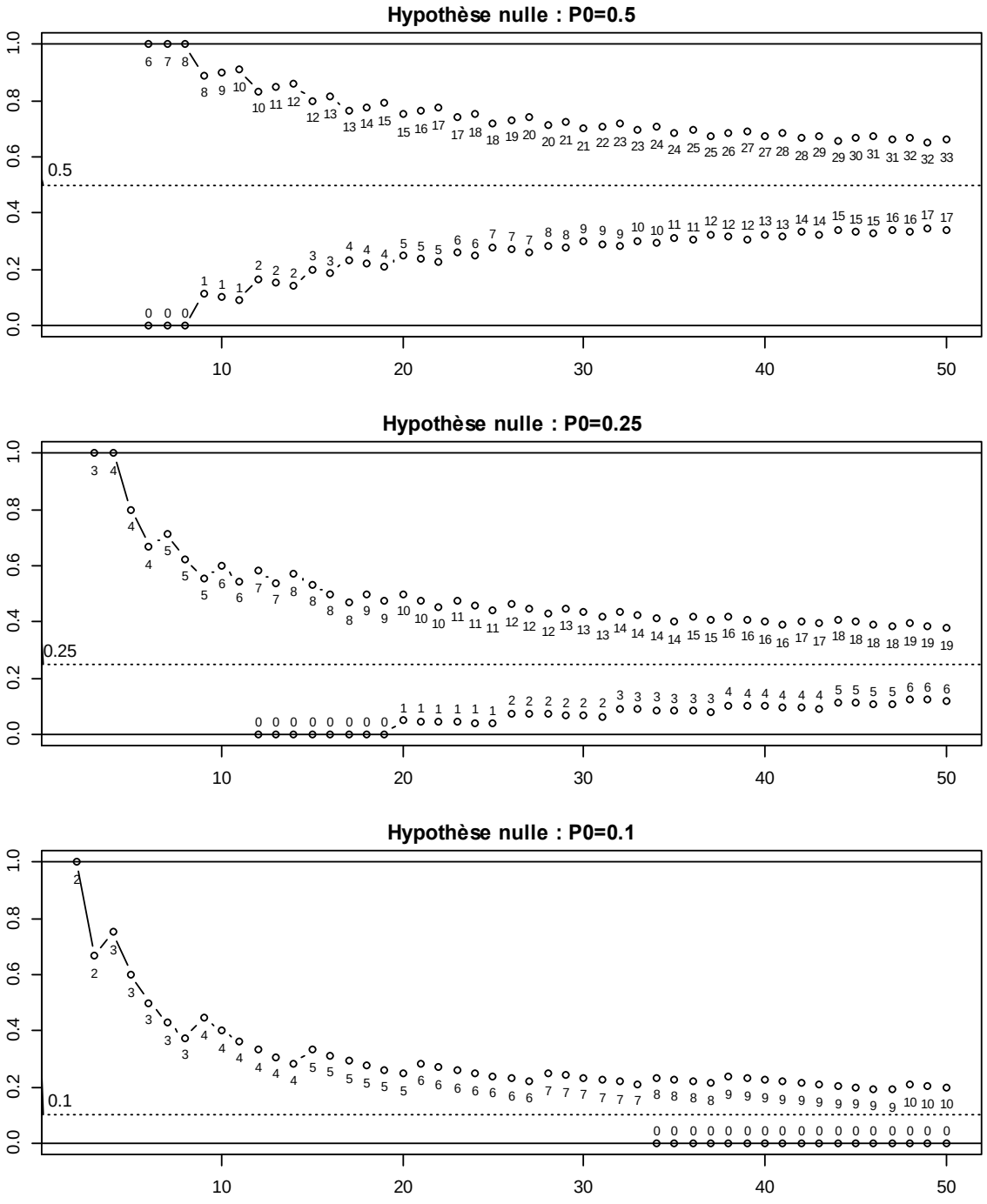


Figure 58. Limites de significativité d'un test binomial
Abscise : n, taille de l'échantillon
Ordonnée : x/n, proportion observé
Points : valeurs de x à partir desquelles on rejette H_0 avec p<5%
En haut : pour p_0=50%. Au milieu : pour p_0=25%. En bas : pour p_0=10%

2.2.4.3 Test du Khi² d'adéquation

2.2.4.3.1 Introduction

Le test binomial était un test exact : il présentait l'avantage didactique que la p valeur fût calculée « à la main », mais c'était aussi son inconvénient en pratique, car le calcul était un peu long. Dans une situation similaire, on peut avoir recours du **test du Khi² d'adéquation** (ou Chi-deux, ou X^2, etc.), qui présente :

- deux avantages :
    - un **calcul rapide et facile**
    - la possibilité de s'intéresser aux variables **qualitatives à plus de 2 modalités** si nécessaire
- deux inconvénients :
    - ce test n'est plus un test exact mais **asymptotique**, car il nécessite que l'échantillon soit suffisamment grand (nous verrons plus précisément de quelle manière) pour être valide
    - nous verrons qu'en pratique cet effectif minimal peut restreindre son application aux **échantillons moyens ou grands**

Le **test du Khi² d'adéquation** n'est pas un test exact, contrairement au test précédent : c'est un **test asymptotique**. Concrètement, nous ne pourrons pas calculer directement sa p valeur. Au lieu de cela, nous calculerons **une statistique de test**, qui est ici le X^2 (khi²). Ensuite, une table de distribution du X^2 sous l'hypothèse nulle nous donnera la **p valeur** correspondante. Cette étape a clairement été décrite par Pearson et, en tant qu'utilisateur du test, il nous suffira d'admettre cette étape supplémentaire du test. Nous verrons cependant que ce raisonnement n'est valide qu'au-delà de certaines conditions d'effectifs. Nous reviendrons sur la classification de ce test dans la section 5.4 Test paramétrique ou non-paramétrique ? en page 217.

2.2.4.3.2 Exemple de mise en œuvre du test

Soit l'exemple suivant.

A un âge donné, parmi les bébés normaux : 50% marchent, 12% ont une ébauche de marche, 38% ne marchent pas. Sur un échantillon de 80 bébés prématurés, nous observons que 35 marchent, 4 ont une ébauche de marche, 41 ne marchent pas. Concernant l'acquisition de la marche, on cherche à savoir si les bébés prématurés ont une distribution différente des bébés de la population générale.

Le test se construit comme suit :

- Population étudiée : les bébés prématurés
- Variable étudiée : la marche, décrite comme une variable qualitative à 3 classes. On s'intéresse à la proportion de chaque classe
- Echantillon : n=80, avec les effectifs 35, 4, 41
- Hypothèse nulle H_0 : la variable suit les probabilités attendues (50%, 12%, 38%)
- Choix du test : Test du Khi² d'adéquation
- Paramètres du test : bilatéral, risque alpha 5% (comme d'habitude)

Nous traçons tout d'abord (Figure 59) un tableau des effectifs observés, sur la gauche. Nous traçons ensuite un deuxième tableau, de même structure, comportant les effectifs théoriques sous l'hypothèse H_0. En supposant que H_0 soit vraie, nous aurions dû trouver des effectifs théoriques de 80x50%=40,00 dans la première case, 80x12%=9,60 dans la deuxième case, et 80x38%=30,40 dans la troisième case. Ces effectifs ont mécaniquement la même somme, 80. On notera, et c'est important de le souligner, que ces effectifs ne sont pas nécessairement des nombres entiers. Il faut donc garder une précision suffisante.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

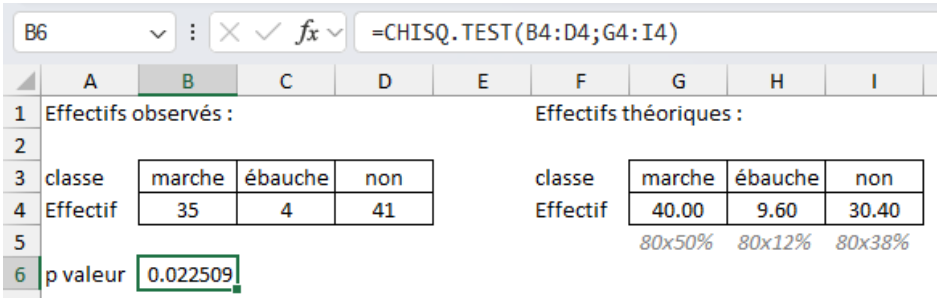


Figure 59. Réalisation d'un Khi d'adéquation avec Excel

La suite est rapidement mise en œuvre : en cellule B6, la formule Excel **chisq.test()** (ou **test.khideux()** anciennement, ou **test.loi.khideux()** avec Calc) prend en paramètre les deux plages de données : la réelle à gauche, et la théorique à droite. Nous devons vérifier que, dans le tableau de droite, **chacun des effectifs théoriques est bien supérieur ou égal à 5** (condition de validité du test) : c'est ici le cas.

En une seule opération, cette formule calcule une métrique d'écart entre les deux plages (la statistique de test X^2), puis la p valeur associée à cet écart. Cette **p valeur** s'affiche directement. Dans notre exemple, elle est inférieure à 5%, on rejette donc H_0. Formellement, on peut conclure que la distribution de la marche dans notre échantillon est significativement différente de la distribution de la marche attendue. D'un point de vue métier, l'interprétation sera que la distribution de la marche pour les prématurés est différente de la distribution de la marche pour les enfants normaux (en l'occurrence, on pourra dire qu'ils présentent un retard d'apprentissage).

Vous savez désormais exécuter un test du Khi² d'adéquation. Dans la partie suivante, nous reviendrons rapidement sur comment, de deux plages d'effectifs, on obtient une p valeur.

2.2.4.3.3 Le test du Khi² d'adéquation, en général

Revenons sur le déroulement d'un test du Khi² d'adéquation, avec quelques formules, mais sans trop de détails.

La première étape consiste à tracer le tableau des effectifs observés, puis le tableau des effectifs théoriques. Ce deuxième tableau comporte, dans chaque case, le produit entre l'effectif total et la proportion attendue, avec le plus de précision possible (**pas d'arrondi** à l'entier notamment).

A cette étape, il faut vérifier que chaque **effectif théorique** est supérieur ou égal à 5. Sinon, il sera toujours possible d'appliquer la **correction de continuité de Yates** (hélas indisponible sur les tableurs), et chaque effectif théorique devra être supérieur ou égal à 3.

Si H_0 est vraie, l'ensemble des effectifs théoriques (T_i) devraient être relativement proches de l'ensemble des effectifs observés (O_i). Autrement dit, chaque différence $(O_i - T_i)$ devrait être relativement faible. La somme de ces différences serait toujours nulle, par construction. Pour évaluer l'ampleur de cette différence, l'idée est d'élever ces différences au carré et de les standardiser, après quoi on calcule leur somme. La statistique de test est donnée par l'Équation 12 : on voit qu'il s'agit d'une somme qui sera toujours positive. En supposant que H_0 est vraie, cette statistique sera proche de zéro, et aura peu de chances de s'en écarter fortement. La **p valeur** nous donnera, en supposant H_0 vraie, la probabilité d'observer ce que nous avons observé, ou d'observer une situation plus éloignée encore de H_0. Nous y reviendrons.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

Équation 12. Calcul de la statistique de test du X^2

Cependant, comme les effectifs T_i sont potentiellement non-entiers et les effectifs O_i sont entiers, on observera en moyenne un écart de 0,5 pour chaque couple de cases, même si H_0 est vraie. Cet écart sera négligeable pour de grands effectifs mais deviendra gênant pour de petits effectifs. La **correction de continuité de Yates** (Équation 13) propose une approche conservatrice : en diminuant les écarts calculés, elle empêche de rejeter à tort H_0. Nous verrons plus bas qu'il est possible de l'appliquer systématiquement, ce que propose le logiciel R. Elle est hélas absente des tableurs.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - T_i - 0{,}5)^2}{T_i}$$

Équation 13. Calcul de la statistique de test du X², avec correction de continuité de Yates

Enfin, la Loi du Khi² ayant été décrite par Pearson, cette quantité du X^2 est directement transformée en **p valeur**. Pour ce faire, il est nécessaire de prendre en compte le nombre de cases du tableau (on appelle « nombre de degrés de liberté » ce nombre de cases moins un). Tout ceci est fait automatiquement.

En pratique :

- Si **p valeur < 5% : on rejette H_0**, on peut conclure que les proportions observées sont **significativement différentes** des proportions théoriques
- Si **p valeur > 5%** : on ne rejette pas H_0 : on se trouve face à une **indétermination**, il est interdit de conclure.

2.2.4.3.4 Quelques précisions sur le test du Khi² d'adéquation

Le test du Khi² d'adéquation peut être classé ainsi :

- C'est un **test asymptotique** : la p valeur est approximative, et n'est fiable que lorsque les effectifs théoriques atteignent ou dépassent 5. Avec la correction de Yates, il faut que ces effectifs atteignent ou dépassent 3
- C'est un **test non-paramétrique** car, formellement, il ne cherche pas à estimer un paramètre (la proportion) mais s'intéresse directement aux effectifs observés. Comme pour le test binomial, cette distinction est un peu artificielle.

La Figure 60 montre les effectifs limites permettant de rejeter H_0 (avec p valeur<5%) dans un test du **Khi² d'adéquation** avec **seulement 2 modalités** et **sans correction de Yates**. Ce graphique est directement comparable à celui du test binomial (Figure 58 page 118).

En haut de la Figure 60, on a fixé p_0=*50%*. On voit que pour *n<10* il n'est pas possible de rejeter H_0, quelle que soit la valeur de *x*. Pour des échantillons de taille *10*, on peut rejeter H_0 si *x* vaut 0, 1, 9 ou 10. Pour des échantillons de taille *11, 12* ou *13*, on peut rejeter H_0 pour *x* valant *0, 1, 2, n-2, n-1* ou *n*. Et ainsi de suite.

En bas de la Figure 60, on montre ces limites de significativité pour p_0=*10%*. On constate que le test est fortement handicapé par ses conditions de validité, qui ne sont atteintes que pour des échantillons de 50 individus ou plus.

La figure du milieu en Figure 60 montre ces limites de significativité pour p_0=*25%*.

La Figure 61 montre le même graphique, pour un test du Khi² d'adéquation d'une variable binaire, mais cette fois-ci avec la **correction de continuité de Yates**. Cette correction permet de prendre en charge des effectifs théoriques supérieurs ou égaux à 3. Bien qu'elle ne soit plus nécessaire à partir de 5, nous l'avons maintenue dans toute la Figure 61. Il est intéressant d'observer son effet :

- Cette correction **étend nettement l'utilisabilité du Khi²** pour des effectifs faibles, ce qui est particulièrement sensible pour des valeurs faibles de p_0

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- Néanmoins, lorsque le Khi² sans correction de Yates devient valide, **les seuils limites sont inchangés** : cette correction n'entraîne pas en pratique de perte de puissance statistique

Pour ces deux raisons, dans certains logiciels de statistiques comme R[31], la correction de Yates est activée par défaut. Hélas, dans les tableurs, la correction de Yates n'est pas directement disponible.

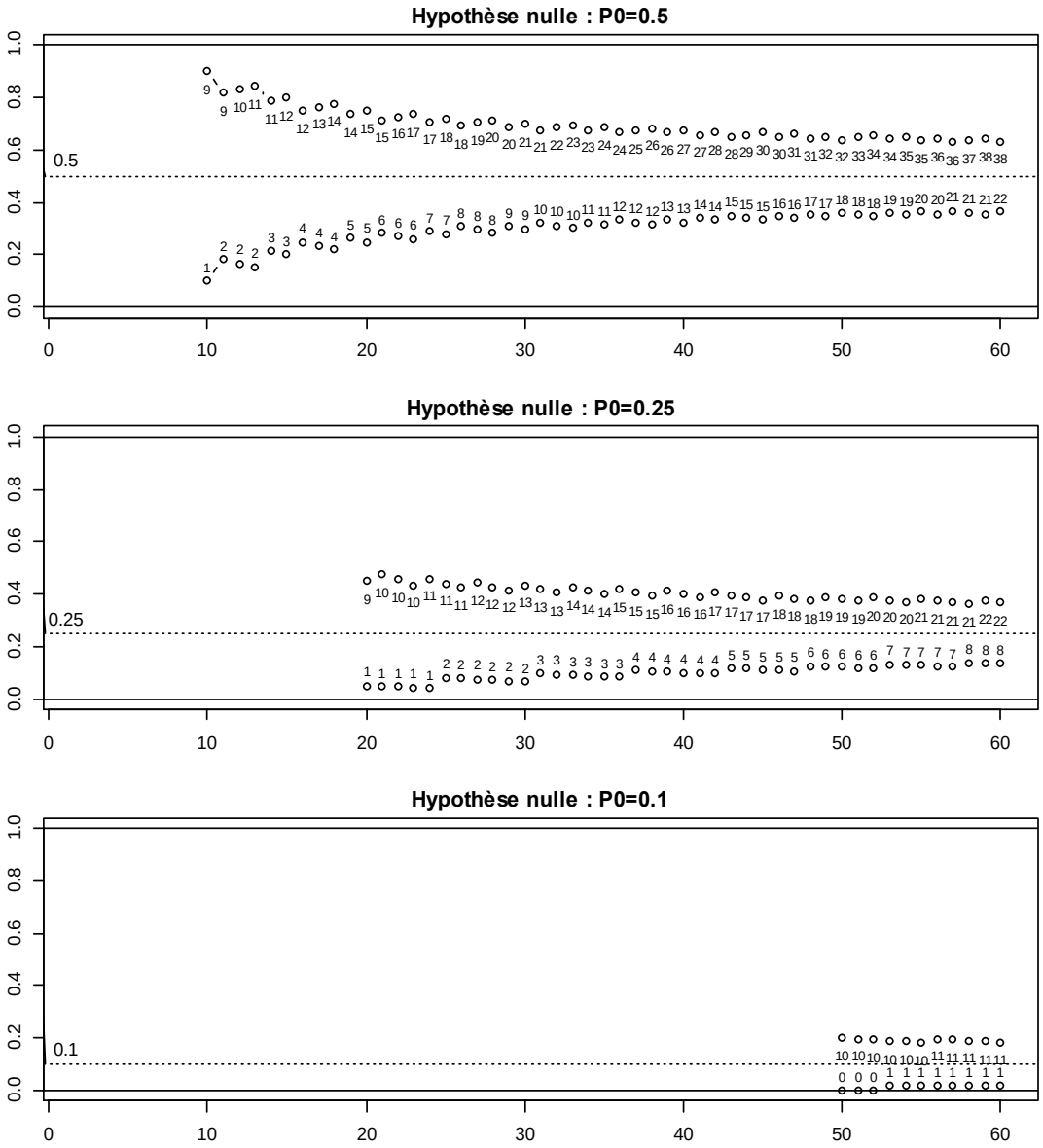


Figure 60. Limites de significativité d'un test du Khi² d'adéquation (sans correction de Yates)
Abscisse : n, taille de l'échantillon
Ordonnée : x/n, proportion observé
Points : valeurs de x à partir desquelles on rejette H_0 avec p<5%
En haut : pour p_0=50%. Au milieu : pour p_0=25%. En bas : pour p_0=10%

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

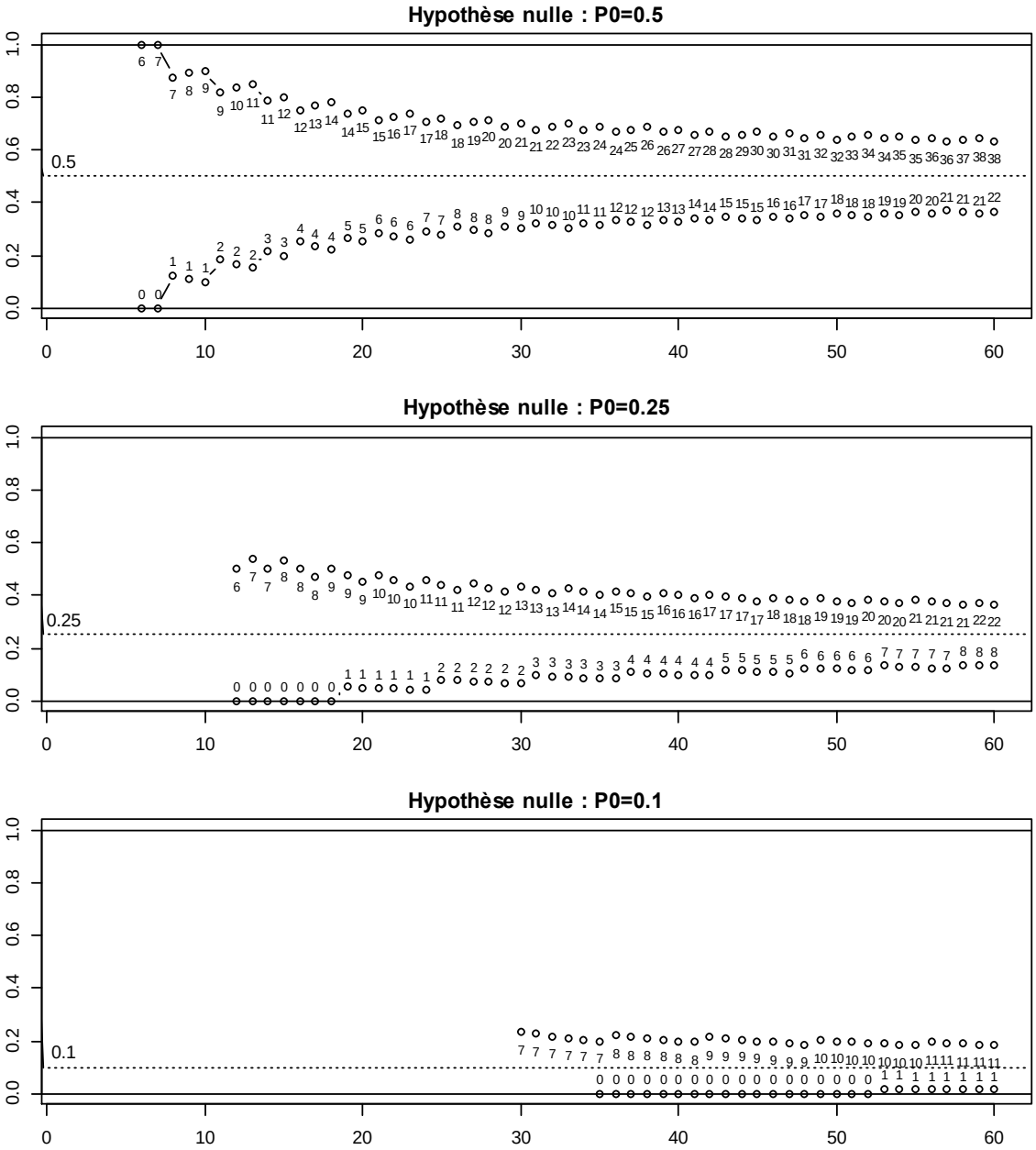


Figure 61. Limites de significativité d'un test du Khi² d'adéquation avec correction de Yates (voir légende du graphique précédent)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.2.5 Tests de comparaison « de deux proportions appariées »

2.2.5.1 Préambule

Dans certaines situations, un même phénomène binaire est mesuré deux fois chez les mêmes individus : il peut s'agir de mesures avant-après, ou gauche-droite.

Le plus souvent, ce phénomène est mesuré une fois à une certaine date, puis une deuxième fois à une autre date, on parle de **mesures appariées avant-après**.

Exemple : on dispose d'un échantillon de participants, qui sont enrhumés ou non (statut avant : 0 ou 1). On les soumet tous au même traitement. Après un certain délai, ils ont un nouveau statut, enrhumé ou non (statut après : 0 ou 1). On prétend vouloir comparer la proportion de personnes enrhumées avant et après. En réalité, on s'intéresse à l'évolution de chaque personne. La question est plutôt de savoir si les individus qui guérissent sont plus ou moins nombreux que les individus qui s'enrhument.

Plus rarement, il peut s'agir d'une mesure à gauche, et d'une mesure à droite, on parle de **mesures appariées gauche-droite**.

Exemple : on dispose d'un échantillon de participants, qui sont capables ou non de porter une certaine charge avec le bras droit (statut à droite : 0 ou 1). On évalue leur capacité à porter cette même charge du bras gauche (statut à gauche : 0 ou 1). On prétend vouloir comparer la proportion de personnes capables de porter la charge à droite, avec celle de personnes capables de porter la charge à gauche. En réalité, on s'intéresse à l'asymétrie de chaque personne. La question est plutôt de savoir si les individus qui peuvent porter à droite mais pas à gauche, sont plus ou moins nombreux que les individus qui peuvent porter à gauche mais pas à droite.

Les problèmes « avant-après » ou « gauche-droite » paraissent relever des analyses bivariées, car il s'agirait d'analyse simultanément deux colonnes. En réalité, ce sont des **analyses univariées** : ils s'agît plutôt d'étudier la **variation individuelle**, à l'aide d'une nouvelle variable. On calcule une nouvelle variable d'intérêt, la variation de x (Équation 14). Passé ce point, pour simplifier la rédaction, nous nous limiterons au cas avant-après.

$$\Delta x_i = x_{i,après} - x_{i,avant}$$

ou

$$\Delta x_i = x_{i,droite} - x_{i,gauche}$$

Équation 14. Calcul d'une nouvelle colonne, indiquant la variation individuelle de x

Cette variable prend les valeurs exposées en Tableau 10.

Tableau 10. Analyse de "proportions appariées"

x_{avant}	$x_{après}$	Δ_x	Interprétation
0	0	0	Stabilité
1	1	0	Stabilité
0	1	+1	Augmentation
1	0	-1	Diminution

Derrière la question « les proportions sont-elles identiques » se cache en réalité la question « comment nos sujets évoluent-ils ? ». Une manière très simple de répondre à cette question est tout simplement de **décrire la nouvelle variable ainsi créée**. On peut également réaliser un **test de McNemar**, présenté immédiatement après ceci.

Nous verrons cependant dans le chapitre 5.2 Tests appariés dans un seul groupe, avant-après en page 215, que cette approche pose souvent des problèmes méthodologiques et devrait être évitée.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

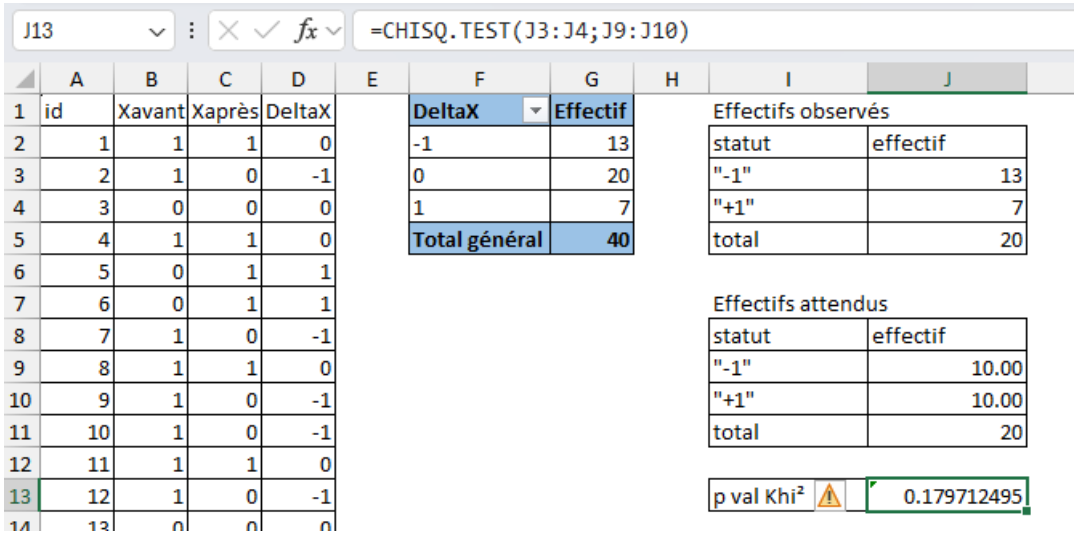
2.2.5.2 Le test de McNemar

Face à la situation précédente, le test de McNemar propose de mettre en évidence une différence systématique (asymétrique), si elle existe. Ce test propose d'omettre les effectifs des individus stables (Δ_x=0), et de tester si les effectifs des individus qui diminuent (Δ_x=-1) et qui augmentent (Δ_x=+1) sont significativement différents. Ceci est réalisé à l'aide d'un **test du Khi²** à 1 degré de liberté. Il s'agit donc, comme le Khi² d'adéquation, d'un test **non-paramétrique**, **asymptotique**.

L'hypothèse nulle H_0 est que la probabilité de diminuer est identique à la probabilité d'augmenter (50% dans les deux cas, uniquement chez les individus dont le statut change).

Pour ce faire, on peut procéder ainsi (Figure 62) :

- On dispose les individus dans un tableau (colonnes A à C, ici tronquées) et on calcule pour chaque individu sa variation (colonne D)
- On réalise un tableau de contingence des effectifs rencontrés (colonnes F et G)
- On recopie **uniquement** les lignes relatives aux changements de statuts (-1 et +1) dans un tableau d'effectifs observés
- On crée un tableau d'effectifs attendus, qui répartit le total en deux moitiés. On vérifie à cette occasion que chaque effectif théorique est bien supérieur ou égal à 5 (autrement dit, il faut observer **au moins 10 individus qui changent** de statut).
- On compare les deux matrices ainsi obtenues à l'aide d'un **test du Khi²** (voir cellule J13 et la barre de formule ; on utilise **chisq.test()** avec Excel, et **test.loi.khideux()** avec Calc)



	A	B	C	D
1	id	Xavant	Xaprès	DeltaX
2	1	1	1	0
3	2	1	0	-1
4	3	0	0	0
5	4	1	1	0
6	5	0	1	1
7	6	0	1	1
8	7	1	0	-1
9	8	1	1	0
10	9	1	0	-1
11	10	1	0	-1
12	11	1	1	0
13	12	1	0	-1

DeltaX	Effectif
-1	13
0	20
1	7
Total général	40

statut	effectif
"-1"	13
"+1"	7
total	20

statut	effectif
"-1"	10.00
"+1"	10.00
total	20

Figure 62. Réalisation d'un test de McNemar avec Excel

Ce test de McNemar est souvent présenté avec une formule différente (qui résulte de la simplification du test du Khi² dans ce cas très précis), et est enseigné à l'aide d'un tableau de contingence qui croise les deux variables initiales. Pour des raisons pédagogiques, il nous a semblé plus approprié de formaliser ce test avec une seule colonne, décrivant la différence de statut pour chaque individu.

Le test de McNemar, comme le test du Khi², est **non-paramétrique** et **asymptotique**.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.3 Variables quantitatives

2.3.1 Description et présentation

2.3.1.1 Arbre décisionnel

La description des variables quantitatives dans les résultats d'une étude suppose généralement de communiquer deux indicateurs :

- un **indicateur de tendance centrale** :
  Les valeurs sont-elles plutôt hautes ou plutôt basses ?
- un **indicateur de dispersion** :
  Les valeurs sont-elles plutôt resserrées ou dispersées autour de leur tendance centrale ?

Le choix de ces indicateurs dépendra de l'aspect symétrique ou non de la distribution, comme indiqué dans l'arbre décisionnel de la Figure 63.

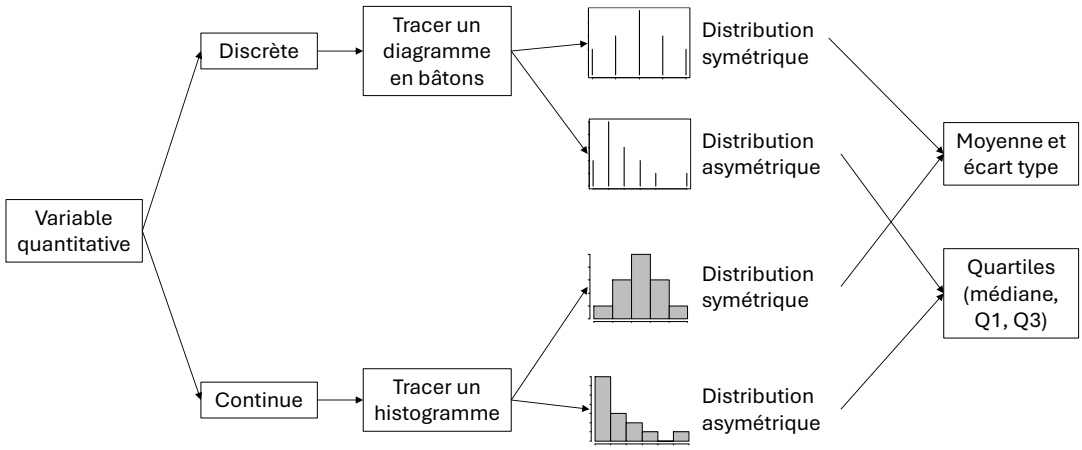


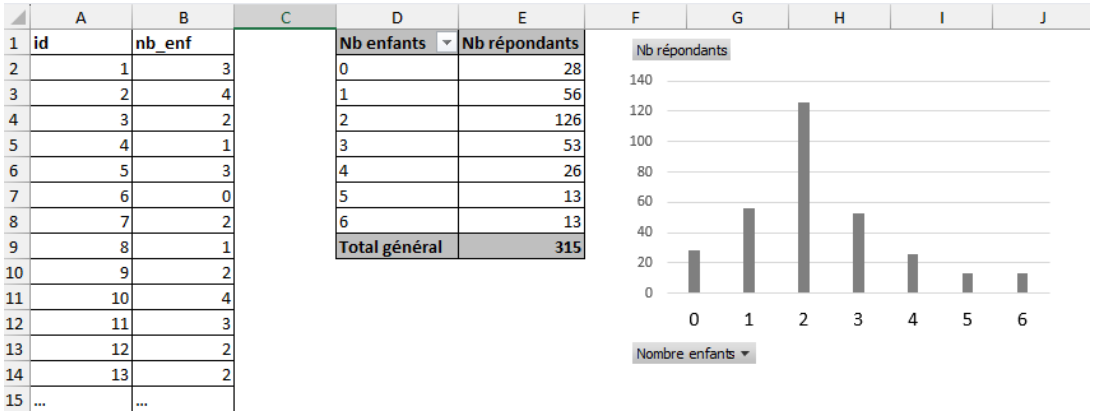
Figure 63. Arbre décisionnel : description d'une variable quantitative

On commencera donc par tracer un graphique adapté en fonction du type de variable quantitative.

2.3.1.2 Variables quantitatives discrètes

Pour les variables quantitatives discrètes, on tracera un **diagramme en bâtons**. Pour faire cela avec un tableur, en partant du tableau de données (Figure 64, colonnes A et B), on réalise tout d'abord un tableau croisé dynamique (Figure 64, colonnes D et E), puis un graphique assis sur ce tableau de contingence (Figure 64, à droite). Ce graphique est parfois appelé à tort histogramme. Il s'agit en réalité d'un diagramme en barres. Il peut être utile de le paramétrer pour diminuer l'épaisseur des barres, qui est en théorie nulle, puisque seules certaines valeurs très précises sont associées à une probabilité non-nulle. Les barres de ce graphique devraient avoir la même largeur que les points pour lesquels les probabilités sont non-nulles, autrement dit la largeur d'un point, autrement dit zéro. En pratique, on utilise une largeur faible mais visible, comme en Figure 64. La hauteur des barres devrait idéalement correspondre à la fréquence (proportion) de chaque modalité. En réalité, les graduations de l'axe des ordonnées n'ont pas vraiment d'importance et seul l'aspect du graphique compte : on peut donc le réaliser en effectifs comme illustré sur la Figure 64.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.



id	nb_enf
1	3
2	4
3	2
4	1
5	3
6	0
7	2
8	1
9	2
10	4
11	3
12	2
13	2
...	...

Nb enfants	Nb répondants
0	28
1	56
2	126
3	53
4	26
5	13
6	13
Total général	315

Figure 64. Tracer un diagramme en bâtons avec un tableur

Attention : dans ce type de graphique, le tableur considère les valeurs en abscisses comme des étiquettes de texte. S'il venait à manquer des valeurs (ex : aucun individu n'a 7 enfants mais un dernier individu a 8 enfants), le tableur juxtaposerait à tort des valeurs non-consécutives (6 et 8 dans l'exemple). Pour éviter cela, il conviendra de réaliser soi-même le tableau de contingence, en incluant les valeurs sans occurrence, et en leur associant un effectif nul.

2.3.1.3 Variables quantitatives continues

Pour les variables quantitatives continues, on tracera idéalement un **histogramme en densité de fréquence**. L'histogramme en densité de fréquence agrège les valeurs de la variable étudiée en classes (éventuellement fixées arbitrairement par l'opérateur) et trace des rectangles dont la surface est proportionnelle à la proportion observée. La largeur de chaque rectangle est la largeur de la classe, sa surface est sa proportion, et la hauteur est ainsi calculée comme étant la proportion divisée par la largeur de classe. La valeur numérique de la hauteur des rectangles (l'axe Y) n'est pas destinée à être lue et interprétée. Cela permet d'obtenir un graphique dont la surface totale est égale à 1, et dont tout le monde sait intuitivement interpréter la forme. La Figure 65 illustre deux exemples d'histogramme en densité de fréquence : celui de droite comporte des classes de largeurs inégales. On observe que cela n'affecte pas l'aspect général de l'histogramme. Néanmoins, les ordonnées ne sont pas parlantes, on pourra même les cacher au lecteur.

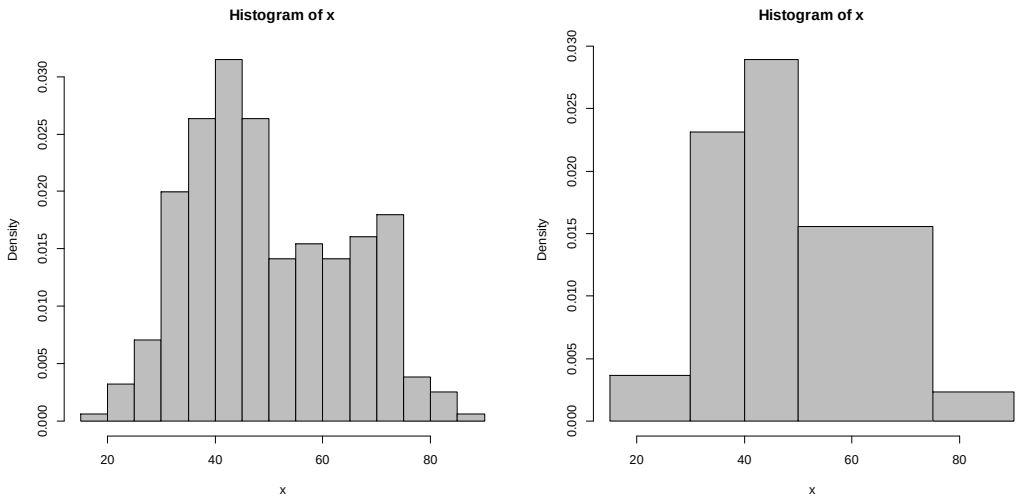


Figure 65. Histogramme en densité de fréquence, représenté par un logiciel de statistique (gauche : classes égales, droite : classes inégales imposées par l'analyste)

Hélas, les tableurs ne sont pas actuellement capables de dessiner des histogrammes en densité de fréquence : il faudra se contenter d'**histogrammes en effectifs** (l'ordonnée décrit directement les effectifs, la surface n'a plus de signification), ce qui suppose obligatoirement **que les classes soient d'égales largeurs** (dans le cas contraire, l'aspect du graphique serait distordu car les surfaces ne seraient plus proportionnelles à la proportion).

La première solution, sur des versions récentes de Microsoft Excel uniquement, est de demander directement le tracé d'un véritable histogramme en effectifs (Figure 66). Diverses options permettent alors de fixer les classes, mais toutes imposent des classes d'égales largeurs. Ce graphique présente cependant l'avantage d'être réalisé immédiatement.

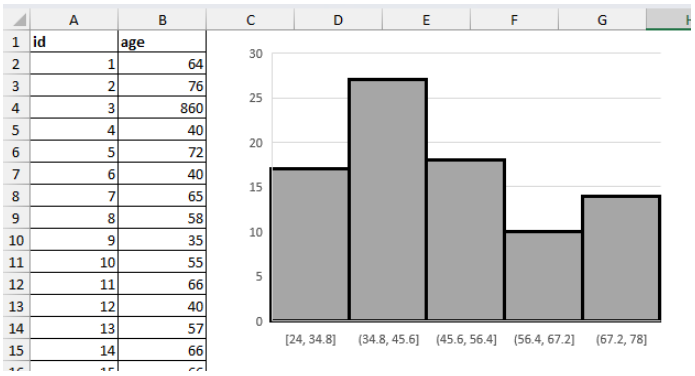


Figure 66. Histogramme en effectifs, tracé par un tableur

On notera ici l'exemple de Microsoft Excel : le mot histogramme est utilisé à de nombreuses reprises pour des graphiques différents, mais un seul d'entre eux est un véritable histogramme (Figure 67).

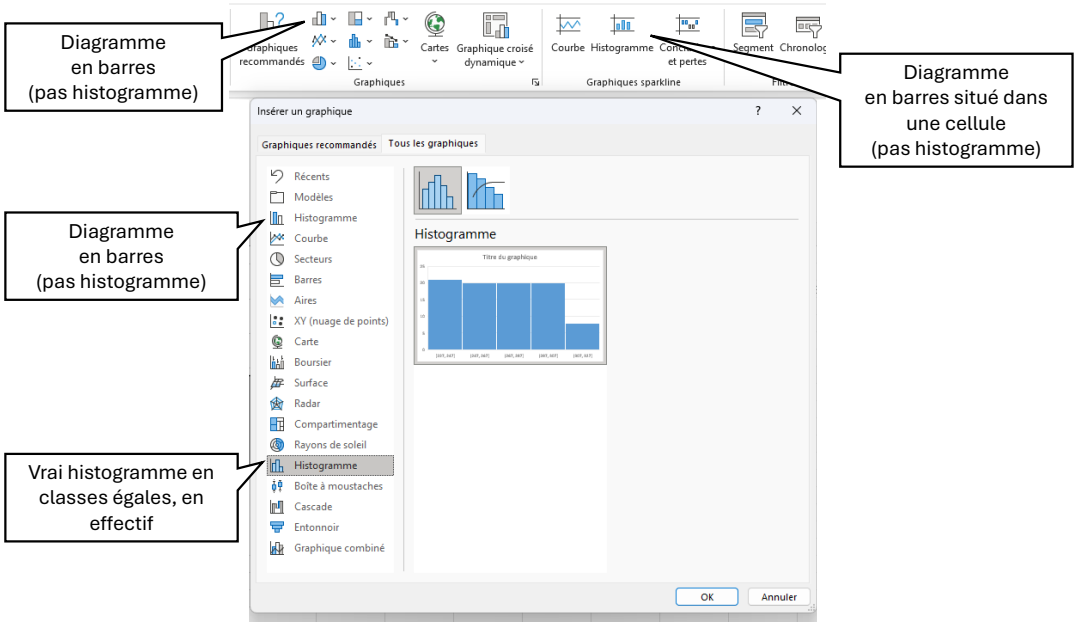


Figure 67. Types de graphiques proposés par Microsoft Excel :
ne pas confondre le diagramme en barres et le véritable histogramme

Avec LibreOffice Calc, ou des versions plus anciennes de Microsoft Excel, ou pour ceux qui souhaitent maîtriser précisément la définition des classes, il reste toujours possible de réaliser un graphique ressemblant à un histogramme, de manière semi-automatisée.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

La première étape consiste à **discrétiser** la variable d'intérêt (Figure 68 à gauche ; en haut on peut lire la formule relative à la cellule C2). À cette étape, on veillera obligatoirement à définir des **classes de largeurs égales**, ceci est vraiment indispensable. On n'hésitera pas à **étendre** les classes pour respecter ce principe d'égales largeurs (ex : si la valeur la plus faible est 28, ne pas proposer l'intervalle [28;40[, mais bien [20;40[). Afin d'obtenir un tri alphabétique qui corresponde également au tri numérique, il peut être utile de forcer des **zéros initiaux** (ex : ne pas écrire [5;15[, mais plutôt [05;15[, ou [005;015[).

La deuxième étape consiste à utiliser un **tableau croisé dynamique** pour obtenir un tableau de contingence (Figure 68 au milieu).

On peut ensuite tracer un diagramme en barres (Figure 68 à droite). L'espace entre les barres devra être nul, de manière à mieux figurer la continuité de la distribution. Les classes étant égales, ce graphique aura la même forme qu'un histogramme en effectifs.

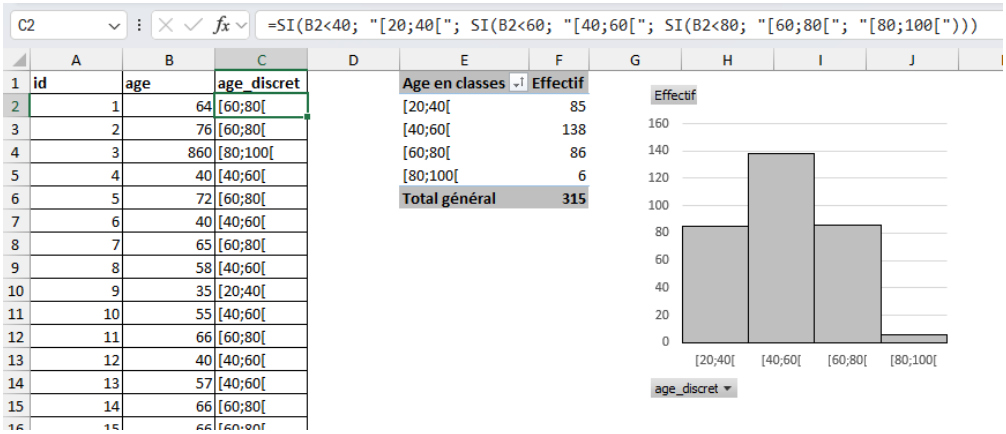


Figure 68. Histogramme en effectifs, tracé par discrétisation manuelle

2.3.1.4 Définition des indicateurs

Parmi les très nombreux indicateurs descriptifs possibles, nous utiliserons en pratique seulement 5 indicateurs. Nous verrons un peu plus bas pourquoi ces indicateurs doivent être choisis selon l'arbre décisionnel présenté en Figure 63 page 126.

La **moyenne** dans l'échantillon se calcule comme la somme des valeurs observées, divisée par l'effectif (Équation 15). Dans un tableur, elle se calcule simplement avec la fonction **moyenne()**.

$$\bar{x} = \frac{\sum x}{n}$$

Équation 15. Calcul de la moyenne dans l'échantillon

Pour calculer l'estimation biaisée de la variance dans l'échantillon, on s'intéresse au carré de l'écart entre chaque valeur et la moyenne. On calcule la moyenne de ces écarts au carré. L'estimation biaisée de l'écart type est la racine carrée de cette moyenne (Équation 16). Il prend donc la même unité que la variable d'intérêt.

$$s_{ech} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

Équation 16. Calcul de l'écart type dans l'échantillon : estimation biaisée

L'estimation présentée en Équation 16 présente cependant un biais : son résultat est, en moyenne, trop faible par rapport à l'écart type en population. On peut corriger très simplement ce biais avec un facteur multiplicateur, qui est proche de 1 pour de grands échantillons, et plus

grand pour de petits échantillons. L'écart type ainsi corrigé s'appelle la **déviation standard (DS)**, ou *standard deviation* (SD), ou **estimation non-biaisée de l'écart type**. Cette quantité se calcule immédiatement dans un tableur à l'aide de la fonction **stdeva()** ou **ecartype()** ou **ecartype.standard()**.

$$SD = DS = s = \sqrt{\frac{\sum(x-\bar{x})^2}{n} \times \frac{n}{n-1}} = \sqrt{(\frac{\sum x^2}{n} - \bar{x}^2) \times \frac{n}{n-1}}$$

Équation 17. Calcul de l'estimation non-biaisée de l'écart type, ou déviation standard

Les quartiles sont les 3 valeurs seuils qui séparent l'échantillon en 4 sous-groupes d'effectifs équilibrés. Si l'échantillon est petit, des règles sont définies pour proposer 3 valeurs réalistes, sachant qu'on obtiendra rarement des équilibres stricts entre les 4 sous-groupes. Ainsi :

- **Q1, le premier quartile**, est la valeur en-dessous de laquelle se trouve 25% de l'effectif
- **Q2, le deuxième quartile**, est la valeur en-dessous de laquelle se trouve 50% de l'effectif. C'est également **la médiane**, parfois notée $\tilde{x}$
- **Q3, le troisième quartile**, est la valeur en-dessous de laquelle se trouve 75% de l'effectif
- **L'intervalle interquartile [Q1 ; Q3]** est un indicateur de la dispersion de la distribution. Parfois, c'est sa largeur, Q3 – Q1, qui est rapporté.e

Les quartiles peuvent se comprendre sur un histogramme (partie gauche de la Figure 69) : ce sont les droites verticales qui séparent la surface de l'histogramme en 4 surfaces égales à 0,25 (si l'histogramme est suffisamment lisse). Autrement dit, si on impose le minimum, les 3 quartiles et le maximum comme bornes des rectangles de l'histogramme, les 4 rectangles ainsi définis ont la même surface (milieu de la Figure 69). Enfin, si on observe le diagramme des fréquences cumulées (partie droite de la Figure 69), les quartiles s'obtiennent en partant des ordonnées ¼, ½ et ¾, en les projetant sur la courbe, puis en redescendant sur l'axe des abscisses.

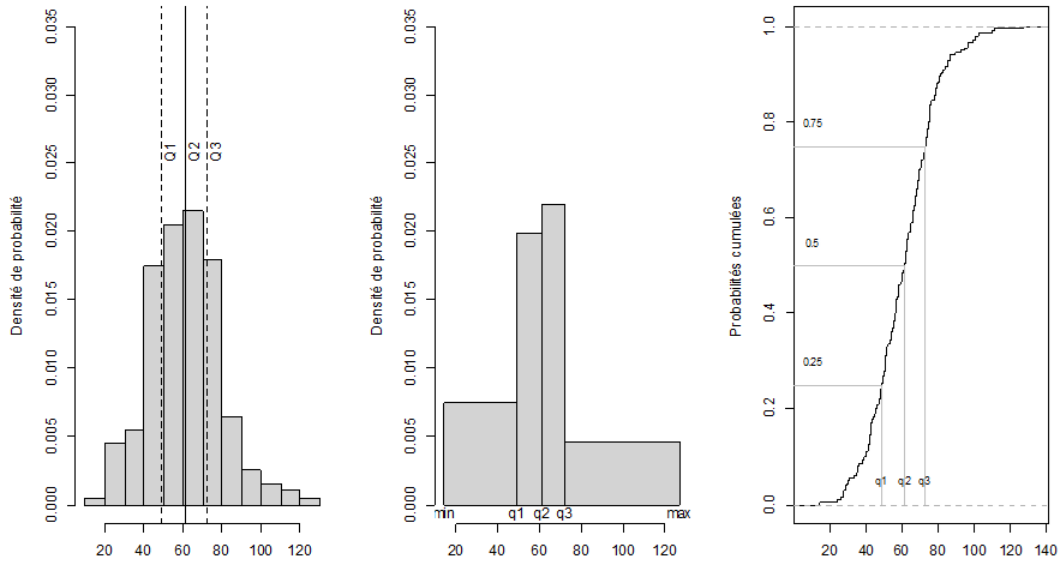


Figure 69. Représentation des quartiles (Q1, médiane Q2, Q3)

2.3.1.5 Choix des indicateurs, exemple de présentation

Revenons à présent à l'arbre décisionnel présenté en Figure 63. S'il fallait résumer une distribution de variable quantitative à l'aide de trois nombres, les quartiles (notés Q1, Q2 ou médiane, et Q3) seraient assurément la meilleure solution, du point de vue descriptif. Leur

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

inconvénient est la complexité. Inversement, la moyenne et l'écart type sont des indicateurs très connus, légèrement plus simple (car l'écart type, noté SD, s'interprète comme une dispersion symétrique) et, surtout, permettant de nombreux calculs et tests statistiques.

Passons en revue 4 situations typiques.

Pour les variables quantitatives discrètes asymétriques (Figure 70), la médiane et les quartiles (à droite) donnent une bonne idée de la distribution. La moyenne et l'écart type (à gauche) sont trompeurs, donnant un ensemble trop étalé vers la gauche, alors que, clairement Q2–Q1 < Q3–Q2 dans cette distribution étalée vers la droite.

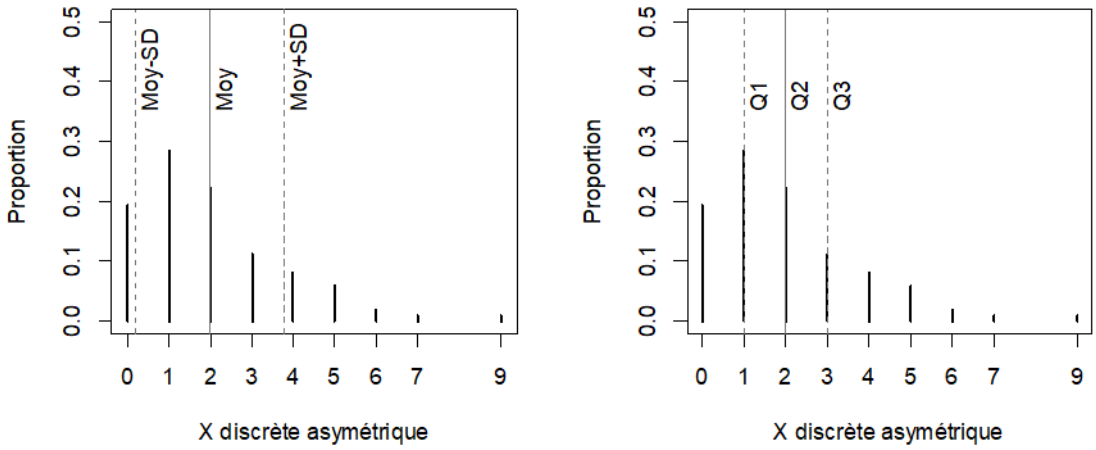


Figure 70. Variable discrète asymétrique : médiane et quartiles (à droite) fournissent la meilleure description de la distribution

Pour les variables quantitatives discrètes symétriques (Figure 71), la médiane et les quartiles (à droite) donnent comme toujours une bonne idée de la distribution. On observe que la moyenne et l'écart type (à gauche) donnent une information très similaire, et sans doute un peu plus précise (il n'y a pas d'arrondi) et plus simple. On privilégiera donc la moyenne et d'écart type dans ce cas.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

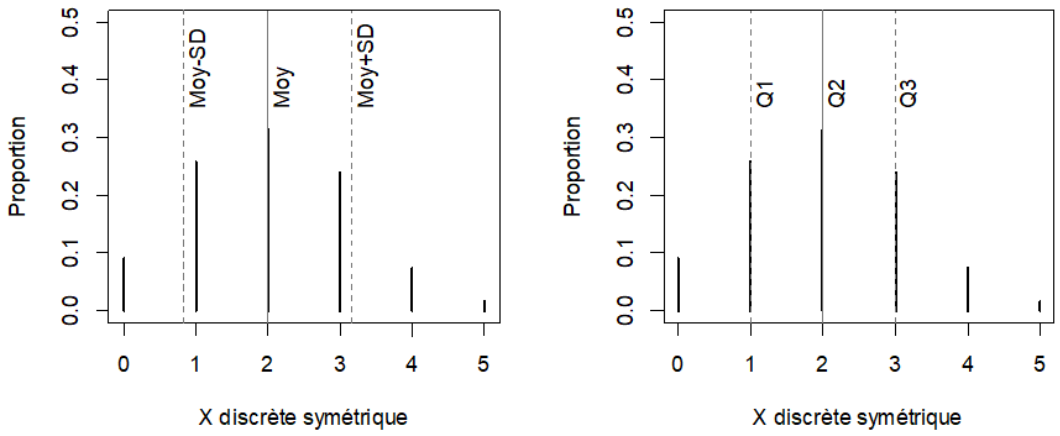


Figure 71. Variable discrète symétrique : les deux options donnent une bonne description

Pour les variables quantitatives continues asymétriques (Figure 72), la médiane et les quartiles (à droite) donnent une bonne idée de la distribution. La moyenne et l'écart type (à gauche) sont trompeurs, donnant un ensemble trop étalé vers la gauche alors que, clairement, Q2–Q1 < Q3–Q2 dans cette distribution étalée vers la droite.

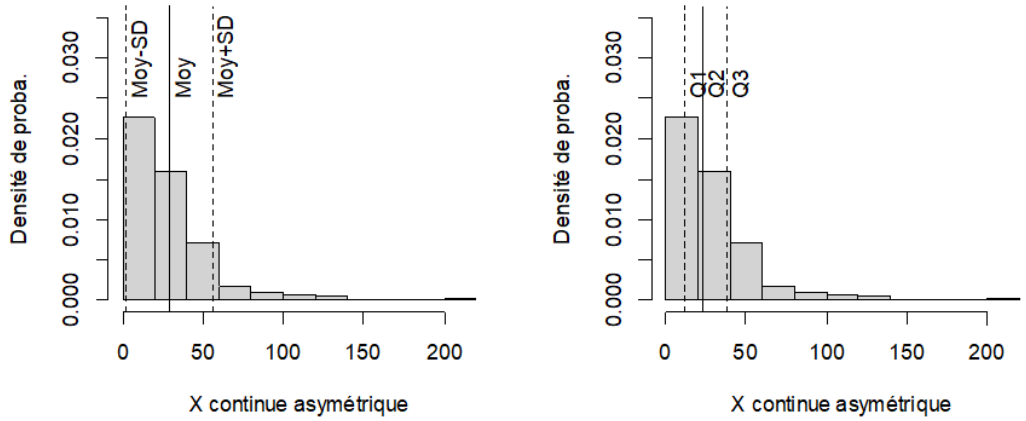


Figure 72. Variable continue asymétrique : médiane et quartiles (à droite) fournissent la meilleure description de la distribution

Pour les variables quantitatives continues symétriques (Figure 73), la médiane et les quartiles (à droite) donnent comme toujours une bonne idée de la distribution. On observe que la moyenne et l'écart type (à gauche) donnent une information assez similaire et plus simple. On privilégiera donc la moyenne et d'écart type dans ce cas.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

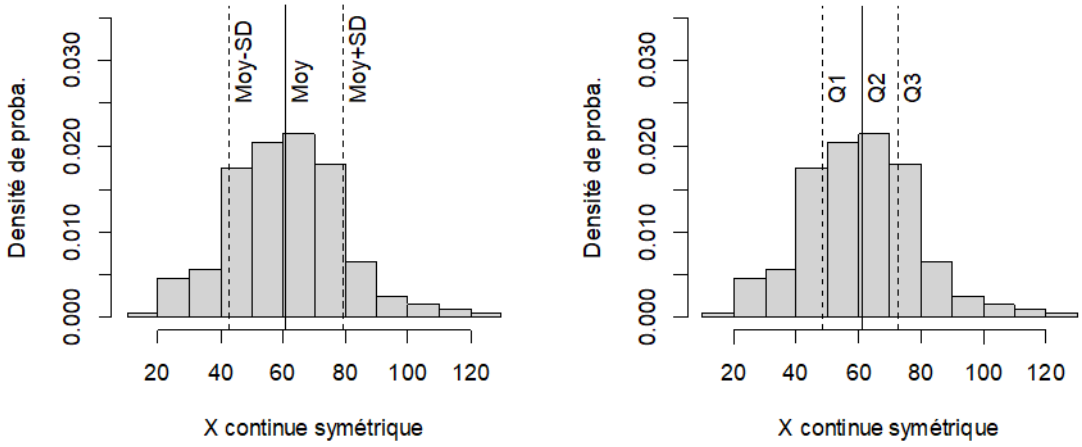


Figure 73. Variable continue symétrique : les deux options donnent une bonne description

Il existe des mesures (*skewness*) ou des tests statistiques permettant de savoir si une distribution est symétrique ou non. Nous ne les utiliserons pas (le caractère inapproprié de ces tests est discuté en chapitre 5.3 Tests qu'on réalise en espérant ne pas rejeter H0 en page 216).

Nous recommandons simplement **d'observer le graphique** le plus adapté (diagramme en bâtons ou histogramme, selon le cas) et de déduire que la distribution est symétrique ou asymétrique.

Si la distribution a une **allure symétrique**, on présentera la **moyenne** et l'**écart type**.

Exemple : « Les participants sont en moyenne âgés de 32,3 ans (SD=5,8) ».

Si la distribution a une allure **clairement asymétrique**, on présentera la **médiane, le premier et le troisième quartile**.

Exemple : « Les répondants ont en médiane 1 enfant (Q1-Q3 : [0 ; 3]) ».

En cas de distribution intermédiaire, on aura tout de même tendance à privilégier la conclusion en faveur de l'aspect symétrique, car le lecteur est plus habitué à lire une moyenne et un écart type, que des quartiles.

On notera que tous ces indicateurs ($\bar{x}$, SD, Q1, Q2, Q3 et même Q3–Q2) s'expriment dans la même unité que la variable d'intérêt.

2.3.1.6 Un dernier graphique utile : la boîte à moustache ou *boxplot*

Un autre graphique est très utilisé, surtout lorsqu'on souhaite juxtaposer plusieurs de ses instances : il s'agit de la ***boxplot***, ou **boîte à moustache**. La *boxplot* présente principalement les quartiles. Sa construction repose sur l'identification préalable de « valeurs extrêmes » (voir plus bas). On peut y voir traditionnellement, de bas en haut :

- Les valeurs extrêmes basses (bulles)
- Le minimum, après exclusion de valeurs extrêmes
- Le premier quartile (bord inférieur du rectangle)
- La médiane (trait épais)
- Le troisième quartile (bord supérieur du rectangle)
- Le maximum, après exclusion de valeurs extrêmes
- Les valeurs extrêmes hautes (bulles)

Cette notion de « **valeurs extrêmes** » est purement algorithmique, et ne doit aucunement faire penser que ces valeurs sont anormales, ni que les individus correspondants devraient être exclus des analyses. Ce n'est qu'un outil de visualisation. Le rendu diffère selon le logiciel employé (Figure 74).

Dans l'exemple de la Figure 74, contrairement à R, Microsoft Excel n'affiche pas les valeurs extrêmes, affiche la médiane avec une croix, et la moyenne avec une ligne horizontale. Ces différences ne doivent pas vous perturber : l'objectif de ces graphiques est de donner une vision générale de la distribution, en termes de tendance centrale, de dispersion et de symétrie.

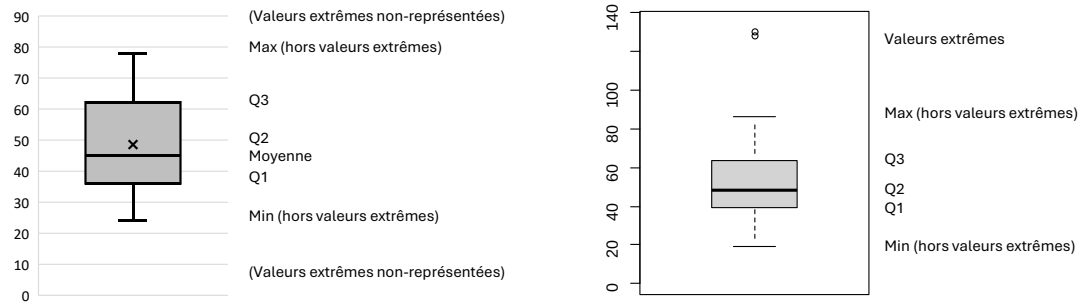


Figure 74. Exemples de boîtes à moustache. Gauche : Microsoft Excel. Droite : R.

2.3.2 Calcul de l'intervalle de confiance d'une moyenne

2.3.2.1 Préambule

Parmi les paramètres mesurés dans l'échantillon, classiquement, un seul d'entre eux fait l'objet d'extrapolation dans la population : il s'agit de la moyenne.

La moyenne $\bar{x}$ définie en Équation 15 est une **moyenne mesurée dans l'échantillon**, et c'est aussi, de facto, une **estimation de la moyenne μ (inconnue) en population**. Cet ouvrage n'étant pas un ouvrage de statistique, nous ne définirons pas lourdement ce qu'est l'estimation et quels sont ses fondements mathématiques. Sachez simplement qu'il est possible de calculer l'**intervalle de confiance à 95% (IC95)** d'une moyenne, c'est-à-dire en pratique l'intervalle dans lequel la vraie moyenne μ inconnue en population a 95% de chances de se trouver, compte tenu de l'observation que vous avez réalisée dans l'échantillon. Le choix du pourcentage, 95%, est une convention en santé, et correspond à un risque de première espèce de 5% (risque alpha, 5% de chances de se tromper).

D'après les normes internationales citées précédemment (Consort, Strobe, Prisma), cet **IC95 ne doit pas être calculé, sauf** si votre objectif principal était justement d'estimer une moyenne. Il y a deux raisons à cela :

- Le calcul d'un intervalle de confiance est, en soi, une inférence statistique, et on souhaite éviter l'inflation du risque de première espèce, lié à la répétition des inférences statistiques
- De toute manière, la moyenne, l'écart type et l'effectif sont suffisants au lecteur pour recalculer cet intervalle de confiance, s'il le souhaite vraiment

Si vraiment vous souhaitez calculer un intervalle à 95% de la moyenne en population, alors il est donné par une formule utilisant la Loi de Student. Cette méthode requiert des conditions de validité que nous détaillerons par la suite.

2.3.2.2 Calcul avec un tableur

La mise en œuvre dans un tableur comme Microsoft Excel ou LibreOffice Calc est assez simple (Figure 75). Il nous faut disposer de l'effectif total, la moyenne et l'écart type (estimateur

non-biaisé). Pour mémoire, ces nombres sont obtenus avec les formules nb(), moyenne() et ecartype.standard() par exemple. La première étape consiste à calculer la demi-largeur de l'intervalle de confiance (en cellule N6, avec la formule reproduite en cellule O6). Ce calcul nécessite un coefficient multiplicateur (supérieur ou égal à 1,96) qui est donné par une fonction liée à la loi de Student, qui admet pour paramètres 0,05 et l'effectif moins un ; ainsi que l'écart type et l'effectif. Les deux bornes de l'intervalle sont alors données en retranchant ou en ajoutant ce demi-intervalle à la moyenne observée (cellules N7 et N8 en Figure 75).

	M	N	O
1	Effectif total :	311	=NB(B2:B316)
2	Moyenne :	50.49	=MOYENNE(B2:B316)
3	Ecart type :	15.13911	=ECARTYPE.STANDARD(B2:B316)
4			
5	IC à 95% :		
6	demi-intervalle :	1.69	=LOI.STUDENT.INVERSE.BILATERALE(0.05;N1-1) * N3 / RACINE(N1)
7	borne basse :	48.80	=N2-N6
8	borne haute :	52.18	=N2+N6

Figure 75. Calcul de l'intervalle de confiance d'une moyenne avec un tableur

Cette méthode requiert des conditions de validité que nous détaillerons dans le paragraphe suivant.

2.3.2.3 Quelques précisions sur le calcul de l'IC95 d'une moyenne

L'intervalle à 95% de la moyenne en population est donné par une formule utilisant la Loi de Student (Équation 18). Le coefficient *t* est donné par la table de la loi de Student, qui prend ici deux entrées : $\propto = 5\%$ par convention, et $\nu = n - 1$ (« nu », le nombre de degrés de liberté, est l'effectif moins 1). Ce coefficient vaut 1,96 pour les échantillons de très grande taille, mais augmente pour les échantillons plus petits (par exemple, 2,2 pour 10 individus). Il permet de corriger le fait que l'écart type est lui-même issu d'une estimation, nécessairement entachée d'erreur.

$$IC_{0,95} = \bar{x} \pm t_{0,05\,;\,\nu} \times \frac{DS}{\sqrt{n}}$$

$$\nu = n - 1$$

Équation 18. Intervalle de confiance d'une moyenne
($\bar{x}$ = moyenne observée, DS=déviation standard, n = effectif, t : coefficient donné par la table de Student)

Cette méthode n'est pas toujours applicable. Dans la Figure 76, nous présentons un arbre décisionnel.

Dans les cours de statistique, en théorie, il faudrait appliquer l'arbre représenté en partie gauche de la Figure 76. Cette conduite à tenir est en pratique peu opérationnelle. La première raison est que, en pratique, la variance n'est jamais connue en population, il est donc inutile de le mentionner. La deuxième raison est que la condition de normalité de la distribution de la variable étudiée est inutilement sévère.

Nous vous proposons donc l'arbre décisionnel figurant en partie droite de la Figure 76. Vous pourrez donc utiliser la méthode proposée en Équation 18 si votre échantillon comporte **au moins 30** individus, **ou** si la variable étudiée semble suivre une **distribution symétrique**. Il n'est pas nécessaire que la variable étudiée suive une loi normale : on peut aisément montrer par simulation que des distributions unimodales symétriques permettent un calcul approprié de l'IC95 de la moyenne, avec la Loi de Student, et ce même pour des effectifs très faibles (Figure 77). Les distributions multimodales symétriques sont également acceptables, pour des

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

effectifs un peu moins faibles. Les distributions asymétriques posent problème, car la moyenne est très sensible aux valeurs extrêmes (Figure 77). Nous recommandons d'étudier la distribution de la variable étudiée de manière visuelle. Il existe des tests statistiques dédiés à cette question, mais nous recommandons de ne pas les utiliser (voir le chapitre 5.3 Tests qu'on réalise en espérant ne pas rejeter H0 en page 216).

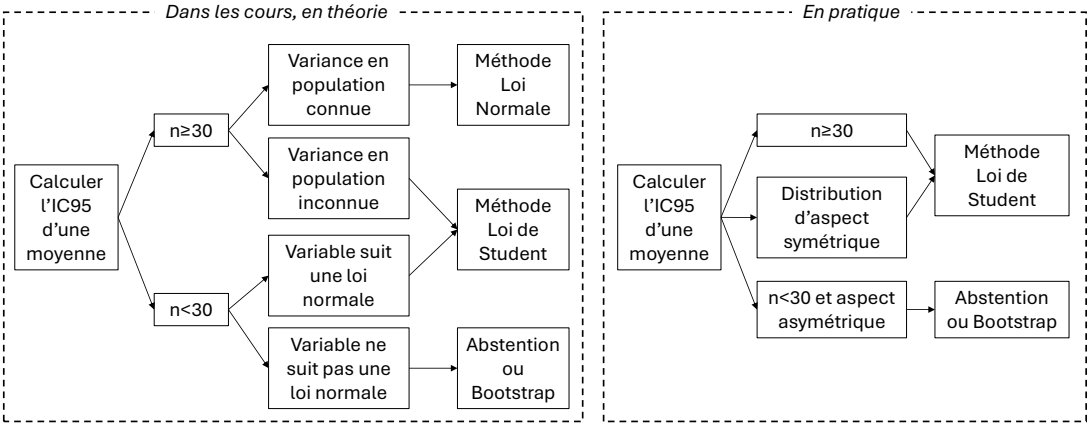


Figure 76. Arbre décisionnel pour le calcul de l'IC95 d'une moyenne

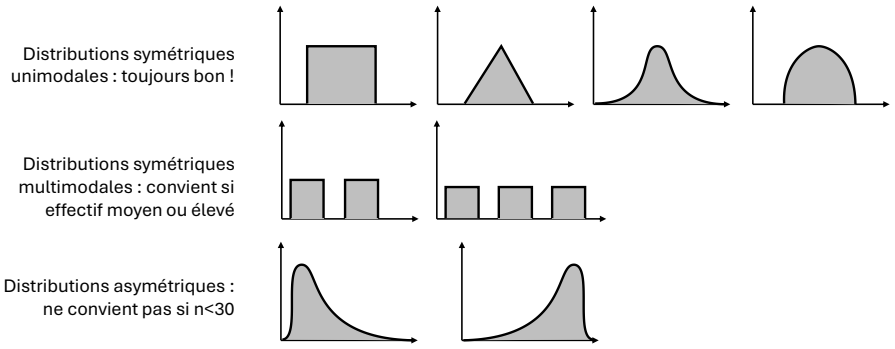


Figure 77. Effectifs inférieurs à 30 : distributions compatibles ou non avec méthode de Student

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.3.3 Tests de comparaison d'une moyenne observée à une moyenne attendue

2.3.3.1 Préambule

En théorie, il serait possible de comparer la distribution observée d'une variable quantitative à une distribution attendue, de diverses manières :

- Comparaison de la **tendance centrale** observée à une valeur attendue :
  - la **moyenne** : c'est ce que nous ferons ci-après
  - la **médiane** : c'est théoriquement possible mais en pratique cela n'est pas fait
- Comparaison de la dispersion observée à une valeur attendue :
  - la **variance** ou **l'écart type** : c'est théoriquement possible mais en pratique cela n'est pas fait
- Comparaison de la forme observée de la distribution, à une forme attendue :
  - Cette comparaison est théoriquement possible, notamment pour comparer la distribution observée à une loi normale. En pratique, ces tests sont réalisés en espérant ne pas rejeter l'hypothèse nulle. Nous verrons les problèmes que cela pose dans le chapitre 5.3 Tests qu'on réalise en espérant ne pas rejeter H0 en page 216

Conclusion : en pratique, on se contente de tester si la moyenne observée dans l'échantillon est compatible avec une moyenne attendue en population. Les autres situations ne seront pas abordées.

2.3.3.2 Introduction

Maintenant que l'on sait calculer l'IC95 d'une moyenne inconnue en population, des méthodes similaires permettent de savoir si l'observation qu'on a réalisée dans l'échantillon (exemple : « sur *n* individus, on observe une moyenne $\bar{x}$ et un écart type *DS* ») est **compatible avec une hypothèse externe** (exemple : « en population, la moyenne μ inconnue vaut m_0 »). Cette hypothèse peut provenir d'articles scientifiques, de données publiées par un institut national de statistique, d'une affirmation infondée qu'on souhaite réfuter, etc. On l'appelle **H_0, hypothèse nulle**, et on la formule ainsi (Équation 19) :

$$H_0 : \mu = m_0$$
$$\mu = moyenne\ inconnue\ en\ population$$
$$m_0 = moyenne\ alléguée\ par\ une\ source\ quelconque$$
$$\bar{x}\ et\ DS = moyenne\ et\ écart\ type\ dans\ notre\ échantillon\ de\ taille\ n$$

Équation 19. Hypothèse nulle d'un test de comparaison d'une moyenne observée à une moyenne attendue

On notera que cette hypothèse nulle ne contient aucun a priori sur l'écart type en population : on utilisera l'écart type estimé depuis l'échantillon pour poursuivre le raisonnement.

Si notre observation $\bar{x}$ est **trop éloignée** de la valeur attendue m_0, alors **on rejette H_0** : cela signifie que notre échantillon n'est pas issu aléatoirement d'une population dans laquelle la moyenne est de m_0. En pratique, on peut l'interpréter de deux manières : on peut dire que $\mu \neq m_0$ au sens « ils racontaient n'importe quoi », ou on peut dire que l'échantillon n'est pas issu par tirage au sort de la population dont il est question. Autrement dit, si on s'intéresse à une sous-population particulière (ex : les diabétiques de type 2, versus les autres humains), on pourra conclure que cette sous-population est différente du reste de la population. D'un point de vue mathématique, ces deux conclusions sont similaires, l'interprétation dépendra donc du contexte. On conclura quelque chose comme « la moyenne observée $\bar{x}$ est significativement différente de m_0 ».

Si notre observation $\bar{x}$ est **assez proche** de la valeur attendue m_0, alors **on ne rejette pas H_0**. Observer un phénomène qui est compatible avec une hypothèse ne suffit pas à la confirmer

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

(ex : vous calculez une hémoglobinémie moyenne de 13,0 dans un échantillon, et untel affirme qu'elle vaut 13,2 en population, ce qui est compatible… mais peut-être qu'en réalité cette moyenne vaut 13,1 ? on ne peut pas savoir !). On conclura quelque chose comme « on ne met pas en évidence de différence significative entre la moyenne observée $\bar{x}$ et la moyenne attendue m_0 ».

De manière générale, on comprend intuitivement que, pour un effectif *n* fixé, plus la différence observée $\bar{x} - m_0$ est importante, plus on sera confiant en rejetant H_0. De même, pour une différence observée $\bar{x} - m_0$ donnée, plus l'effectif de l'échantillon est élevé, plus on sera confiant en rejetant H_0. Un nouveau paramètre entre en jeu ici : il s'agit de la dispersion des valeurs observées autour de leur moyenne. En effet, pour une même différence $\bar{x} - m_0$, à effectif *n* égal, moins la variable est dispersée (étalée) autour de sa moyenne, plus on sera confiant en rejetant H_0.

Comme précédemment, par convention en santé, on rejettera H_0 si une méthode mathématique (appelée test statistique) nous permet de le faire avec une confiance très élevée. Pour ce faire, nous nous attacherons à calculer une **p valeur (ou « petit p », ou *p value*)**, qui est, en supposant que H_0 est vraie, la probabilité qu'on avait d'observer quelque chose comme notre échantillon, ou quelque chose d'encore plus éloigné de H_0. Ce « encore plus éloigné » s'entend « d'un côté comme de l'autre », « au-dessous comme au-dessus », on parlera de test bilatéral.

Si on souhaite comparer une moyenne observée à une moyenne attendue, plusieurs options s'offrent à nous :

- Dans tous les cas, on pose $H_0 : \mu = m_0$
- Option 1 : on pourra simplement **calculer l'IC95** de la moyenne en population : si la valeur attendue m_0 n'appartient pas à cet IC95, on pourra rejeter H_0, avec moins de 5% de chances de se tromper, tout simplement !
- Option 2 : on pourra réaliser un **test statistique**, et rejeter H_0 si sa **p valeur** est inférieure à 5%. Plusieurs tests sont accessibles :
    - Option 2A : le **test de Student**, qui est de loin le test le plus populaire
    - Option 2B : un test basé sur la **Loi Normale**, qu'en pratique on n'utilise pas

Dans le chapitre suivant, nous détaillerons le **test de Student**, que vous utiliserez peut-être en pratique. Nous verrons cependant plus tard, dans le chapitre 5.1 Tests de comparaison à une norme en page 214, qu'il n'est pas vraiment conseillé d'utiliser ce test. Nous vous proposons ci-dessous un algorithme décisionnel simple et efficace (Figure 78) : il tient compte à la fois de vos souhaits et aptitudes, et des conditions de validité des outils proposés.

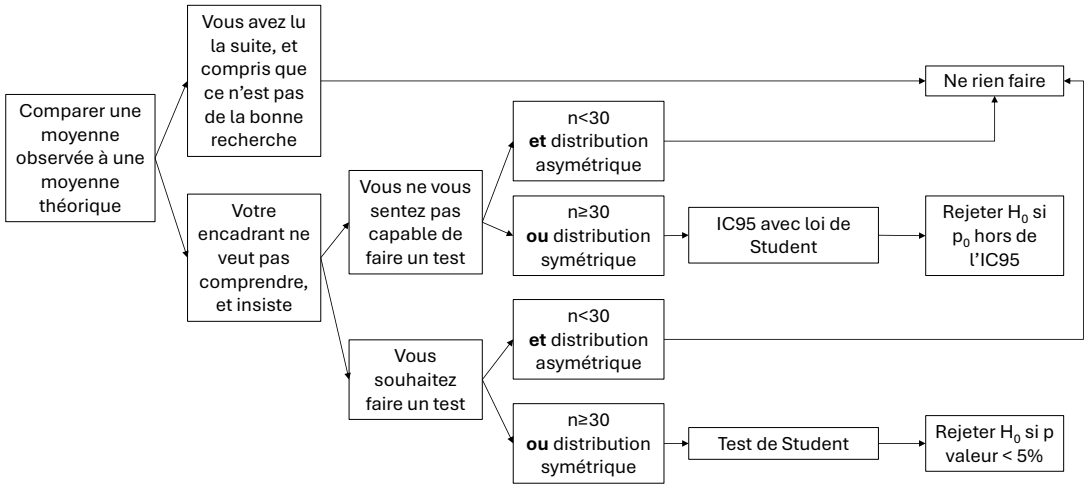


Figure 78. Arbre décisionnel : comparer une moyenne observée à une moyenne attendue

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.3.3.3 Test de Student observé-attendu

Pour simplifier, nous appellerons le test de Student de comparaison d'une moyenne observée à une moyenne attendue, « Student observé-attendu ». Ce test a été publié par William Sealy Gosset sous le pseudonyme de Student[32].

2.3.3.3.1 Exemple de mise en œuvre

Soit l'exemple suivant.

Dans un hôpital psychiatrique, 27 adolescents ayant commis un crime sont hospitalisés. Leur QI est de 82.9 (DS=14.8). En population, le QI suit une loi normale, de moyenne 100. Est-ce significativement différent ?

Le test se construit comme suit :

- Population étudiée : adolescents criminels
- Variable étudiée : soit X le quotient intellectuel (QI). X suit une loi normale d'après l'énoncé.
- Paramètre étudié : µ, moyenne de X en population
- Echantillon : $n = 27$; $\bar{x} = 82{,}9$; $DS = 14{,}8$
- Hypothèse nulle H_0 : $\mu = 100$
- Test : Test de Student observé-attendu
- Paramétrage du test : test bilatéral, α=5% (comme toujours)
- Conditions de validité du test : ici X suit une loi normale, donc le test est valide bien que l'effectif soit inférieur à 30

Nous saisissons les données de l'énoncé dans un tableur (Figure 79). Ici, nous recopions l'effectif, la moyenne et l'écart type, mais en pratique, si vous disposez des données source, il vous sera plus aisé de les calculer directement dans la cellule à l'aide des fonctions **nb()**, **moyenne()** et **stdeva()** (ou **ecartype()** avec Calc, ou **ecartype.standard()** avec Excel). Il est à noter que Microsoft Excel et LibreOffice Calc ne proposent pas directement de fonction pour ce test, c'est pourquoi deux étapes sont nécessaires.

	A	B	C	D
1	Echantillon :			
2	effectif	n	27	
3	moyenne	x_bar	82.9	
4	écart type	DS	14.8	
5				
6	Hypothèse :			
7	moyenne	m0	100	
8				
9	Test :			
10	statistique de test	|t|	6.003662597	=ABS((C3-C7)/C4*RACINE(C2))
11	p valeur	p	2.43884E-06	=LOI.STUDENT.BILATERALE(C10;C2-1)
12	significatif ?	p<5%	VRAI	=C11<0.05

Figure 79. Réalisation d'un test de Student observé-attendu avec un tableur

La première étape consiste à calculer la statistique de test, appelée « t » : c'est la différence entre les deux moyennes, divisée par l'écart type, multipliée par la racine carrée de l'effectif (voir Équation 20 plus bas). Pour la suite, Excel et Calc ne prennent que des valeurs positives de *t* : il sera donc utile de préciser que c'est la valeur absolue qui nous importe (voir formule recopiée en D10 en Figure 79).

Il faut ensuite retrouver la p valeur correspondant avec la formule **loi.student.bilaterale()** (voir formule recopiée en D11 en Figure 79).

Dans notre exemple, cette p valeur est inférieure à 5%, on peut donc rejeter l'hypothèse nulle, et affirmer que le QI moyen dans l'échantillon est significativement différent de 100.

Concrètement, cette observation est en faveur de l'hypothèse selon laquelle la population d'adolescents criminels hospitalisés en psychiatrie a un QI moyen différent de 100 (et donc ici inférieur à 100).

2.3.3.3.2 Le test de Student observé-attendu, en général

Revenons sur le déroulement d'un test de Student observé-attendu, avec quelques formules, mais sans trop de détails.

On dispose d'une moyenne attendue m_0 d'une variable en population. Dans un échantillon de taille *n*, on observe une moyenne $\bar{x}$ et un écart type DS. Naturellement, $\bar{x}$ peut différer de m_0.

Trois questions, au fond identiques du point de vue du test, peuvent se poser :

- La différence observée entre m_0 et $\bar{x}$ dépasse-t-elle le simple effet du hasard (fluctuations aléatoires liées à l'échantillonnage) ?
- La moyenne observée $\bar{x}$ diffère-t-elle *significativement* de la moyenne attendue m_0 ?
- Notre échantillon est-il réellement issu d'une population caractérisée par $\mu=m_0$?

Nous répondons à ces questions par un **raisonnement par l'absurde**.

- Vérifions les conditions de validité *a priori* : un effectif suffisant (n≥30) ou une distribution symétrique (voir Figure 77 page 136)
- Posons l'hypothèse nulle H_0 : « la moyenne μ vaut bien m_0 »
- Calculons, sous l'hypothèse H_0, la **p valeur bilatérale** de l'observation, autrement dit la probabilité sous H_0 d'observer ce qu'on a observé OU d'observer une situation encore moins probable. Pour ce faire :
  - Calculons la statistique de test *t* (Équation 20)
  - Trouvons la p valeur correspondante dans la table de Student, avec $\nu = n - 1$ degrés de liberté
- Concluons :
  - Fixons le seuil d'interprétation de la p valeur à 5% (risque α)
  - Si **p valeur < 5% : on rejette H_0**, on peut conclure que la moyenne observée est **significativement différente** de la moyenne théorique
  - Si **p valeur > 5%** : on ne rejette pas H_0 : on se trouve face à une **indétermination**, il est interdit de conclure.

$$t = \frac{\overline{x} - m_0}{\frac{DS}{\sqrt{n}}}$$

Équation 20. Calcul de la statistique du test de Student observé-attendu

2.3.3.3.3 Quelques précisions sur le test de Student

Le test de Student observé-attendu permet de comparer une moyenne observée à une moyenne attendue. Il peut être classé ainsi :

- C'est un **test paramétrique** car il s'appuie sur l'estimation d'un paramètre, qui est la moyenne
- C'est un **test asymptotique** : la p valeur est calculée en se référant à une table de loi, qui n'est vraie que si la variable suit une loi normale, ou si l'effectif est suffisant

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Ce test a été publié par William Sealy Gosset sous le pseudonyme de Student[32] car son employeur, la brasserie Guiness, ne souhaitait pas que ses concurrents sachent quelles méthodes étaient employées chez eux.

On comprend de la formule précédente (Équation 20) que le test se fonde essentiellement sur la quantité $|(\bar{x} - m_0)/DS|$. Cette quantité répond à la question « de combien d'écarts-types la moyenne observée s'écarte-t-elle de la moyenne attendue ? ». Nous l'appellerons « écart standardisé » ci-après.

Par exemple, si vous attendez une moyenne $m_0 = 10$ et que vous observez une moyenne $\bar{x} = 8$ avec un écart type $DS = 1$, alors on peut affirmer que votre observation s'écarte de deux écarts-types de l'hypothèse nulle (en valeur absolue). L'écart standardisé vaut 2.

Par la suite, la significativité du test dépend également de l'effectif de l'échantillon, d'une part parce que le terme $\sqrt{n}$ multiplie la statistique de test, d'autre part parce que l'effectif détermine le nombre de degrés de liberté, qui modifie la valeur limite dans la table de Student. Ainsi, la Figure 80 présente l'écart standardisé correspondant à une p valeur de 5%, en fonction de l'effectif. Si, pour un effectif donné, vous obtenez un écart standardisé supérieur à celui de cette courbe, votre test sera significatif avec p<5%.

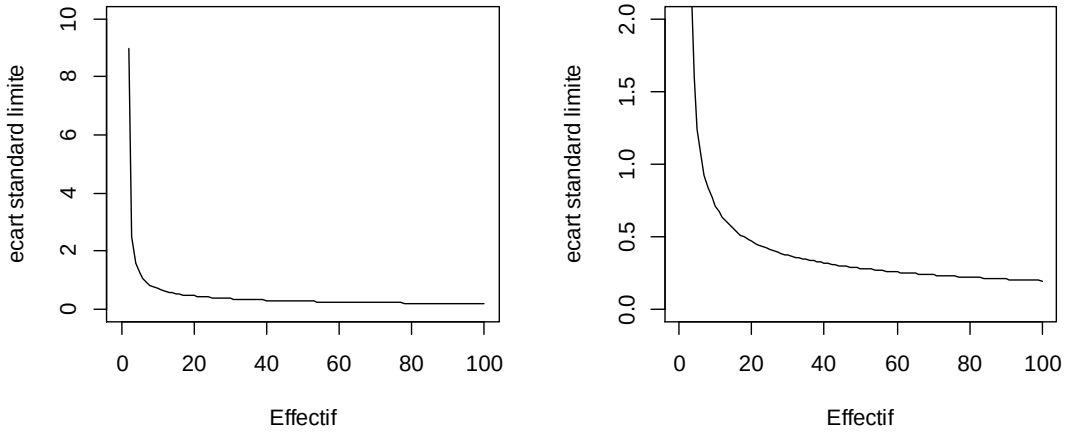


Figure 80. Limites de significativité du test de Student (à droite : zoom) : valeur de $\left|\frac{\bar{x}-m_0}{DS}\right|$ permettant d'obtenir p=5%, en fonction de la taille d'échantillon

On peut donc utiliser la Figure 80 pour réaliser graphiquement le test : il suffit de positionner un seul point, correspondant au couple $\left\{x = n \ ; \ y = \left|\frac{\bar{x}-m_0}{DS}\right|\right\}$. Si ce point est au-dessus de la courbe, on rejette H_0, sinon on ne rejette pas H_0. Le Tableau 11 présente quelques-uns de ces points. On observera que pour un effectif de 7 ou plus, un écart standardisé de 1 est suffisant pour rejeter H_0. Pour un effectif de 100 ou plus, un écart standardisé de 0,2 est suffisant pour rejeter H_0.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 11. Quelques limites de significativité du test de Student (voir figure précédente)

n	Valeur Limite de $\left\|\frac{\bar{x}-m_0}{DS}\right\|$
2	8,985
5	1,242
10	0,715
15	0,554
20	0,468
25	0,413
30	0,373
40	0,320
50	0,284
60	0,258
70	0,238
80	0,223
100	0,198

2.3.4 Tests de comparaisons « de deux moyennes appariées »

2.3.4.1 Préambule

Dans certaines situations, un même phénomène quantitatif est mesuré deux fois chez les mêmes individus : il peut s'agir de mesures avant-après, ou gauche-droite (tous les autres cas sont possibles, vous l'aurez compris).

Le plus souvent, ce phénomène est mesuré une fois à une certaine date, puis une deuxième fois à une autre date, on parle de **mesures appariées avant-après**.

Exemple : on dispose d'un échantillon de participants, pour lesquels on mesure la taille le matin (x_{avant}, en centimètres) et le soir ($x_{après}$ en centimètres). On prétend vouloir comparer la taille moyenne le matin, à la taille moyenne le soir. En réalité, on s'intéresse aux individus, et on veut connaître leur évolution matin-soir. Si rien ne change, c'est la moyenne de cette évolution qui vaudra zéro.

Plus rarement, il peut s'agir d'une mesure à gauche, et d'une mesure à droite, on parle de **mesures appariées gauche-droite**.

Exemple : on dispose d'un échantillon de participants, pour lesquels on mesure la longueur du membre inférieur droit (x_{droit} en centimètres) et la longueur du membre inférieur gauche (x_{gauche} en centimètres). On prétend vouloir comparer la longueur moyenne à droite et à gauche. En réalité, on s'intéresse aux individus, et on veut connaître leur différence droite-gauche. S'il n'y a pas de tendance particulière, la moyenne de cette différence vaudra zéro.

Les problèmes « avant-après » ou « gauche-droite » paraissent relever des analyses bivariées, car il s'agirait d'analyser simultanément deux colonnes. En réalité, il s'agit plutôt d'étudier la variation individuelle, à l'aide d'une nouvelle variable. On calcule une nouvelle variable d'intérêt, la variation de x (Équation 21). Passé ce point, pour simplifier la rédaction, nous nous limiterons au cas avant-après.

$$\Delta x_i = x_{i,après} - x_{i,avant}$$

$$ou$$

$$\Delta x_i = x_{i,droite} - x_{i,gauche}$$

Équation 21. Calcul d'une nouvelle colonne, indiquant la différence individuelle de x

Derrière la question « les moyennes appariées sont-elles identiques » se cache en réalité la question « comment nos sujets évoluent-ils ? ». Une manière très simple de répondre à cette question est tout simplement de **décrire la nouvelle variable ainsi créée**, qui est une variable

quantitative. Elle sera intéressante tant d'un point de vue de la tendance centrale (est-ce que, en moyenne, X augmente ou diminue ?) que de la dispersion (X est-elle plutôt stable ou labile ?).

Si on souhaite réaliser un test statistique pour savoir si X reste stable, ou change (augmente ou diminue), on peut se référer à l'arbre décisionnel de la Figure 81.

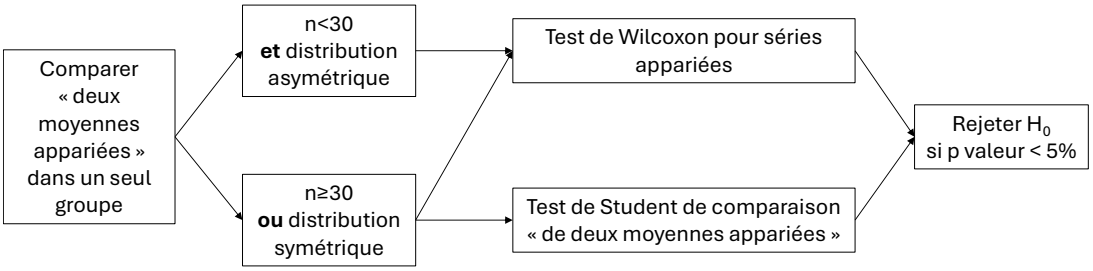


Figure 81. Arbre décisionnel : « comparer deux moyennes appariées »

Nous verrons cependant dans le chapitre 5.2 Tests appariés dans un seul groupe, avant-après en page 215, que cette approche pose souvent des problèmes méthodologiques et devrait être évitée.

2.3.4.2 Le test de Student de comparaison « de deux moyennes appariées »

Lorsqu'on souhaite comparer deux moyennes appariées, l'hypothèse nulle peut être formulée ainsi : « La moyenne de la variation est nulle ». On comprend immédiatement qu'il s'agit simplement de réaliser un test de **Student observé-attendu, avec m_0=0** (sous l'hypothèse nulle, intuitivement, la moyenne de la variable individuelle devrait être nulle). Ce test n'implique donc pas de comparer deux moyennes. La dénomination commune de ce test (« comparaison de deux moyennes appariées ») est donc impropre et peut induire en erreur. Il s'agit bien, au final, d'une **analyse univariée**.

Pour mettre en œuvre un tel test, rendez-vous en section 2.3.3.3 Test de Student observé-attendu en page 139, en gardant simplement en tête l'hypothèse nulle H_0 : m_0=0.

Ce test est, comme nous l'avons vu précédemment, un test **paramétrique** et **asymptotique**.

2.3.4.3 Le test des rangs signés de Wilcoxon pour séries appariées

Le test des rangs signés de Wilcoxon pour séries appariées suit approximativement le même objectif que le précédent. Il s'agit cependant d'un test **non-paramétrique**, car il ne s'intéresse pas à la moyenne, qui est un paramètre, mais à des sommes de rangs, qui ne sont pas des paramètres. D'autre part, il ne s'agit pas d'un test asymptotique, mais d'un **test exact** : l'auteur du test a calculé des probabilités de permutations, et fournit une table permettant directement d'obtenir une p valeur. Les valeurs critiques (qui correspondent à p<5%) sont rendues dans une table prête à l'emploi, pour épargner un effort de calcul inutile à l'utilisateur du test.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 12. Limites de significativité du test de Wilcoxon pour séries appariées, bilatéral à 5%
Rejet de H_0 si $W^+ \leq seuil\_bas$ ou $W^+ \geq seuil\_haut$

Effectif restant	Seuil bas	Seuil haut
6	1	20
7	2	26
8	4	32
9	6	39
10	8	47
11	11	55
12	14	64
13	17	74
14	21	84
15	25	95
16	30	106
17	35	118
18	40	131
19	46	144
20	52	158
21	59	172
22	66	187
23	73	203
24	81	219
25	90	235
26	98	253
27	107	271
28	117	289
29	127	308
30	137	328

Principe du test (nous verrons un exemple par la suite) :

- Hypothèse nulle : H_0 : W+=W-
  (somme des rangs positifs = somme des rangs négatifs)
- Pour chaque individu *i*, calculer sa différence,
  par exemple $\Delta x_i = x_{i,après} - x_{i,avant}$
- Exclure les cas où $\Delta x_i = 0$
- Classer les valeurs $|\Delta x_i|$ par ordre croissant, en déduire le rang de chaque individu vis-à-vis de cette variable
- Calculer W^+, la somme des rangs pour les valeurs positives de Δx_i (ou W^-, la somme des rangs pour les valeurs négatives de Δx_i : cela n'a pas d'importance)
- Utiliser la table des valeurs limites (Tableau 12), chercher la ligne correspondant à l'effectif restant après exclusion des cas stables, et
  rejeter H_0 si $W^+ \leq seuil\_bas$ ou $W^+ \geq seuil\_haut$

Mettons ces principes en œuvre.

Exemple : Sept patients souffrant d'une pathologie articulaire sont évalués avant et après mise en place d'un traitement. Un score fonctionnel est calculé. Son résultat est compris entre 0 et 24, 0 pour l'absence d'altération détectable, et 24 pour une altération fonctionnelle maximale.

Dans un tableau (Figure 82) :

- On commence par reporter les individus, avec leurs valeurs avant et après (colonnes A, B, C)
- On calcule, pour chaque individu, la variation (colonne D) puis sa valeur absolue (colonne E)
- On calcule (colonne F) le rang statistique de cette valeur absolue. Attention :
  - Il faut exclure les individus dont la variation est nulle

    - En cas d'ex-aequo, c'est le rang moyen qui est attribué. Ici par exemple, les individus 1 et 3 sont 2ème et 3ème ex-aequo : on leur attribue donc le rang moyen de 2,5ème et 2,5ème [12]
- On reporte uniquement (colonne G) les rangs associés à des valeurs positives, on calcule la somme de ces rangs (cellule G8) : c'est la statistique de test W+. Ici, W+=8,5 (on pourrait, indifféremment, ne s'intéresser qu'aux rangs associés à des valeurs négatives, et calculer leur somme W-)
- On compte également les effectifs restants (colonne H ; cellule H9)

On se réfère au Tableau 12, à la ligne correspondant à un effectif résiduel de 6 : aucune des conditions n'est remplie car $8{,}5 \in]1; 20[$ donc on ne rejette pas H_0.

	A	B	C	D	E	F	G	H
1	**individu**	**x_{avant}**	**$x_{après}$**	**DeltaX**	**|DeltaX|**	**Rang de |DeltaX|**	**Rangs positifs**	**Effectif restant**
2	1	17	13	-4	4	2.5		1
3	2	14	11	-3	3	1		1
4	3	11	15	4	4	2.5	2.5	1
5	4	14	14	0	0	XXXXXXXXXX		0
6	5	9	0	-9	9	5		1
7	6	8	18	10	10	6	6	1
8	7	18	12	-6	6	4		1
9						Somme :	8.5	6

Figure 82. Réalisation d'un test de Wilcoxon pour séries appariées

2.4 Variables de survie

2.4.1 Description et présentation

2.4.1.1 Définition des données de survie

Les données de survie sont relatives à un événement dépendant du temps. Cet événement peut être :

- Un décès, c'est cet événement qui a donné son nom à ces analyses
- Un autre événement péjoratif : accident, rechute, survenue d'un cancer, infection d'une prothèse, etc.
- Un autre événement mélioratif : retour à domicile, rémission, reprise de la marche, fin d'un arrêt de travail, etc.

Cet événement se produit une fois, au bout d'un certain temps. Heureusement ou malheureusement, pour de nombreux individus, le suivi n'est pas suffisamment long pour observer l'événement. Trois cas de figure se présentent (Figure 83) :

- Des individus pour lesquels l'événement est observé pendant le suivi (premier individu sur la Figure 83)
- Des individus qui, à la fin de l'étude, n'ont toujours pas présenté l'événement (deuxième individu sur la Figure 83) : on parle d'**exclus vivants** (mais, bien sûr, ces individus ne seront pas exclus de l'analyse !)
- Des individus qui quittent l'étude, pour une raison ou une autre, mais n'ont pas présenté l'événement durant leur suivi (troisième individu sur la Figure 83) : on parle de **perdus de vue** (eux aussi feront partie de l'analyse)

[12] en vie courante on dirait qu'ils sont 2ème ex-aequo, mais cette approche altèrerait les propriétés calculatoires des rangs. Exemples de propriétés :
- la somme des rangs doit rester égale à n(n+1)/2
- les rangs suivent une loi uniforme
- le dernier, s'il n'est pas ex-aequo, est nième

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

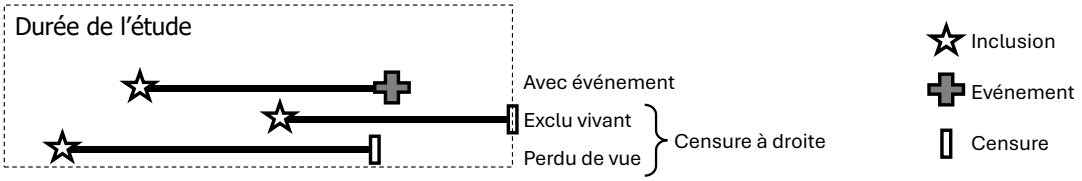


Figure 83. Description de données de survie

Les « perdus de vue » et les « exclus vivant » ne pourront pas être traités différemment pendant l'analyse : ce sont des individus pour lesquels on sait que la durée avant événement est supérieure au délai d'observation : on parle de **censure à droite**. Le mot « censure » désigne ici seulement une information incomplète, et aucunement l'éviction de ces individus. La contribution de ces individus n'est pas anodine : même si on n'observe pas d'événement, on sait tout de même que, pendant un certain temps, ils n'ont pas présenté d'événement.

Les analyses de survie proposent un cadre statistique pour analyser de telles données. Elles permettent également de prendre en compte des événements répétés (nous n'en parlerons pas) et des censures à gauche (nous n'en parlerons pas non plus). Dans la majorité des applications en santé, la gestion d'événements uniques avec censure à droite est amplement suffisante.

Formellement, la variable d'intérêt est la durée entre le début du suivi et l'événement. Cette durée est parfois parfaitement connue, ou parfois partiellement connue, au sens où elle est strictement supérieure à un nombre connu.

2.4.1.2 Rappel sur la présentation de données

Rappelons comment ces informations de survie sont représentées dans le tableau de données (voir 4.3.8 Variables décrivant un événement temps-dépendant (survie) en page 81) :

- Une première colonne binaire indique si, oui ou non, l'événement a été observé
- Une deuxième colonne indique le délai sans événements :
  - Si l'événement a été observé, c'est le délai au bout duquel l'événement a eu lieu
  - Si l'événement n'a pas été observé, c'est le délai pendant lequel l'individu a été suivi (il s'agit alors d'une censure à droite, qu'elle soit liée à la fin de l'étude [exclu vivant] ou au départ, expliqué ou non, de l'individu [perdu de vue])

Le Tableau 13 montre comment ces données sont représentées pour un humain (gauche), sont saisies dans un tableau (milieu) puis seront traitées durant l'analyse (droite).

Tableau 13. Données de survie : pour un humain (gauche), pour la saisie (milieu), puis pendant l'analyse (droite)

Informations		*Saisie de données*			*Durant l'analyse*	
id	**Description littérale**	**Id**	**DC_evt**	**DC_delai**	**Id**	**DC_delai**
Marcel	Décès après 6 semaines	1	1	42	1	42
Justine	Suivie 2 mois, vivante à la fin de l'étude	2	0	61	2	>61
Marceline	Perdue de vue après 1 mois	3	0	30	3	>30

2.4.1.3 Description inappropriée : moyennes, médianes et proportions

Il ne faut surtout pas décrire les données de survie comme des variables traditionnelles. On rappelle que la variable d'intérêt est le temps qui s'écoule jusqu'à un événement. Pour l'illustrer, nous utiliserons deux séries de données de survie. T_a comporte une censure en troisième position, et T_b en deuxième position :

$$T_a = \{1\ ;\ 2\ ;\ >5\} \qquad T_b = \{1\ ;\ >2\ ;\ 5\}$$

Il n'est pas possible de calculer la moyenne, car chaque série comporte une donnée incomplète. On sait seulement que leur moyenne est supérieure, dans les deux cas, à 8/3.

Il n'est pas toujours possible de calculer leur médiane. La médiane vaut 2 pour la série T_a, mais est supérieure à 2 pour la série T_b. Il en est de même pour les autres quantiles.

Il n'est pas toujours possible de calculer la proportion de survie à une certaine date. Dans la série T_a, 1/3 des individus survivent plus de 3 unités, mais dans la série T_b on ne sait pas si c'est 1/3 ou 2/3.

Il faudra mettre en œuvre une méthode de description spécifique aux données de survie.

2.4.1.4 Description appropriée : courbe de Kaplan-Meier

Il fait aujourd'hui consensus que, dans les recherches usuelles en santé, la méthode descriptive de référence est l'**estimateur de Kaplan-Meier**[33], dont la représentation graphique est souvent appelée « courbe de survie ». Cette courbe représente, au fil du temps écoulé depuis l'inclusion, la proportion de survivants, en tenant compte des censures de manière appropriée. Autrement dit, pour un temps donné, la courbe donne la **probabilité estimée de survivre au moins** ce temps-là.

Nous vous montrons ici comment réaliser une courbe de Kaplan-Meier avec un tableur (Figure 84). On atteint ici les limites de l'utilisation simple d'un tableur pour réaliser une analyse statistique.

Nous disposons tout d'abord un tableau de données correctement présentées (en haut à gauche). Ce tableau peut, si besoin, être très long.

De préférence dans un onglet différent, nous traçons un tableau croisé dynamique (en haut au milieu) permettant, pour chaque délai et pour chaque statut d'événement, d'avoir l'effectif concerné. Ce tableau devra être modifié comme suit :

- Propriétés du tableau : « disposition classique », pas de sous-total ni de total
- Propriétés du champ « surv_delai » : « répéter les étiquettes d'éléments », pas de sous-total
- Tri par surv_delai croissant, puis par sur_evt décroissant (pour avoir les décès, puis les censures, et non l'inverse)

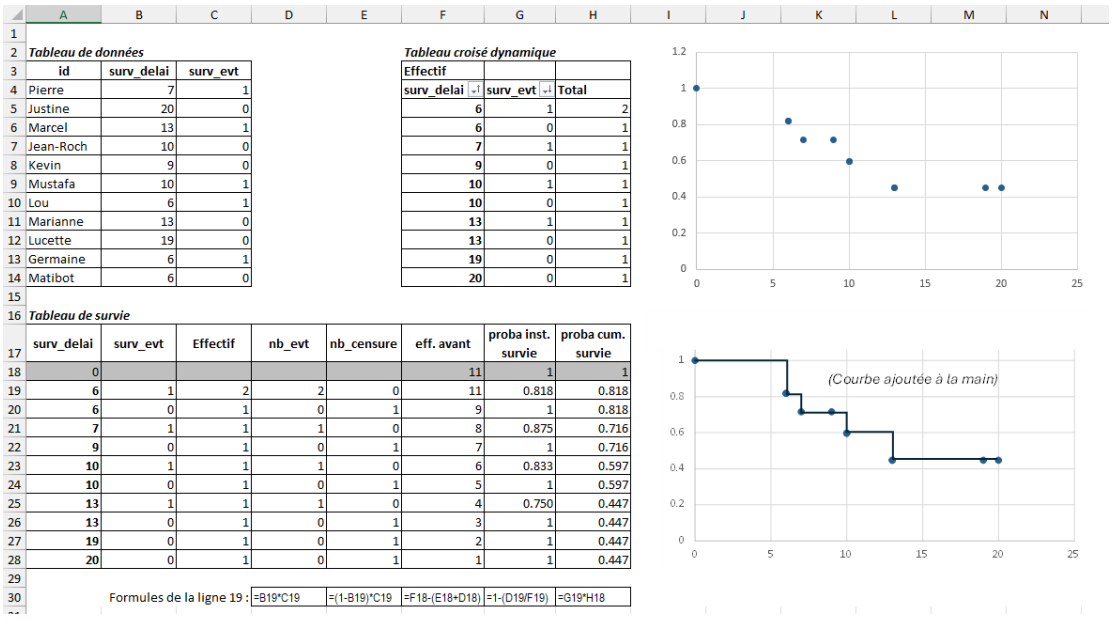


Tableau de données

id	surv_delai	surv_evt
Pierre	7	1
Justine	20	0
Marcel	13	1
Jean-Roch	10	0
Kevin	9	0
Mustafa	10	1
Lou	6	1
Marianne	13	0
Lucette	19	0
Germaine	6	1
Matibot	6	0

Tableau croisé dynamique

Effectif		
surv_delai	surv_evt	Total
6	1	2
6	0	1
7	1	1
9	0	1
10	1	1
10	0	1
13	1	1
13	0	1
19	0	1
20	0	1

Tableau de survie

surv_delai	surv_evt	Effectif	nb_evt	nb_censure	eff. avant	proba inst. survie	proba cum. survie
0					11	1	1
6	1	2	2	0	11	0.818	0.818
6	0	1	0	1	9	1	0.818
7	1	1	1	0	8	0.875	0.716
9	0	1	0	1	7	1	0.716
10	1	1	1	0	6	0.833	0.597
10	0	1	0	1	5	1	0.597
13	1	1	1	0	4	0.750	0.447
13	0	1	0	1	3	1	0.447
19	0	1	0	1	2	1	0.447
20	0	1	0	1	1	1	0.447

Formules de la ligne 19 : =B19*C19 | =(1-B19)*C19 | =F18-(E18+D18) | =1-(D19/F19) | =G19*H18

Figure 84. Tracer une courbe de Kaplan-Meier avec un tableur

Ensuite, on recopie ces données de manière statique dans un tableau (plage A17:H28 dans l'exemple Figure 84). Il faut insérer manuellement une ligne correspondant au délai zéro : elle contient un effectif qui est la taille de l'échantillon, et des probabilités instantanées et cumulées de survie égales à 1. On complète la deuxième ligne avec les formules ici reproduites en ligne 30, puis on étend ces formules aux lignes suivantes du tableau. Voici le raisonnement contenu dans ces formules :

- Le nombre d'individus correspondant au délai qu'on regarde est séparé en nombre d'événements, et en nombre de censures (colonnes D et E)
- Effectif_avant correspond à l'effectif résiduel observé sur la ligne immédiatement au-dessus : c'est l'effectif_avant de la ligne au-dessus, diminué du nombre de décès et de censures de la ligne au-dessus
- La probabilité instantanée de survie est calculée sur la ligne en cours, par 1-(nombre_deces/effectif_avant)
- La probabilité cumulée de survie est la probabilité cumulée de la ligne précédente, multipliée par la probabilité instantanée de la ligne en cours

Pour tracer un graphique, on utilise un « nuage de points », avec pour abscisse le délai, et pour ordonnée la probabilité cumulée de survie (en haut à droite Figure 84). Pour terminer ce graphique, on pourra relier les points à la main par une courbe en marches d'escalier. La courbe ne descend que lorsqu'elle se trouve au-dessus d'un point situé plus bas. Sinon, elle se poursuit horizontalement vers la droite.

2.4.1.5 Comprendre intuitivement la courbe de Kaplan-Meier

La Figure 85 illustre intuitivement la construction de la courbe de Kaplan-Meier. Nous partons d'un tableau de données triées par délai croissant, puis par événement (dans l'ordre décès puis censure) (1 en Figure 85). Dans cet exemple, nous n'avons pas représenté d'événements simultanés.

Nous traçons un diagramme dont la hauteur vaut 1 (ce sera la probabilité de survie, qui est toujours de 100% au début), et dont l'abscisse décrit le temps qui passe, échelonné de 0 (toujours) à 5 unités de temps (dans cet exemple). Au début du suivi, l'échantillon est composé de 5 patients, que nous représentons donc avec 5 bandes horizontales, de hauteur 1/5 chacune (2 en Figure 85). Ces bandes sont représentées dans le même ordre que dans le tableau. Lorsque surviennent les deux premiers décès, nous interrompons chaque bande à la date du décès (3 en Figure 85). On observe que le bord supérieur du graphique descend proportionnellement à la probabilité cumulée de survie. Le troisième individu ne présente pas de décès, mais une censure au temps 3. En l'absence de décès, il faut que la courbe reste horizontale. Or, passé ce point, nous n'aurons plus que 2 individus. La **censure** sur ce troisième individu se matérialise par **une réallocation de l'espace sous la courbe** : la même hauteur est désormais partagée par deux bandes élargies. Leur largeur n'est plus de 0,2 mais de 0,3 dans ce cas (4 en Figure 85). L'individu suivant décède au temps 4, sa bande est donc interrompue et la courbe de survie baisse d'autant. Mais, au lieu de baisser de 0,2, elle baisse désormais de 0,3 car la bande des individus restants était devenue plus large à l'étape précédente (5 en Figure 85). La censure du dernier individu interrompt la courbe de survie, là encore sans la faire baisser.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

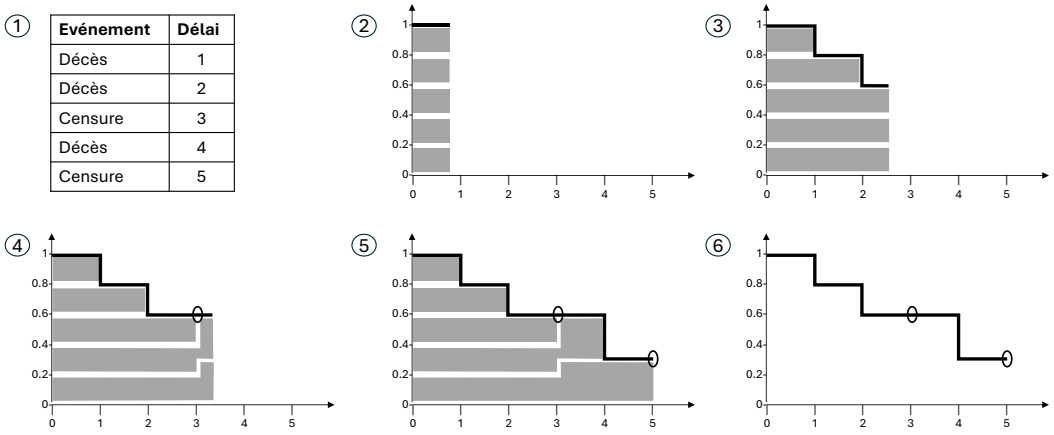


Figure 85. Construction intuitive de la courbe de Kaplan-Meier

Lorsqu'on efface les bandes, il reste la courbe de survie, qui correspond exactement à la méthode de Kaplan-Meier (6 en Figure 85). Elle décrit, pour chaque valeur du temps, la probabilité de survivre au moins ce temps-là.

Cette explication intuitive permet de mieux comprendre la censure :

- Elle n'est pas un décès, donc elle ne fait pas baisser la courbe de survie
- Avant la censure, on sait que l'individu est encore vivant : il participe donc au calcul des probabilités de mourir, en apportant sa contribution au dénominateur (il stabilise la courbe)
- Après la censure, on ne sait pas ce qu'il advient de l'individu : il ne participe donc plus au calcul des probabilités de mourir. La réallocation de l'espace a un impact visible : les décès suivants feront plus baisser la courbe de survie

2.4.2 Calcul de l'intervalle de confiance d'une survie

L'intervalle de confiance d'un estimateur de Kaplan-Meier peut être tracé sur la courbe de survie, sous forme de courbe en pointillés. Son calcul s'appuie sur la loi normale. Il permet de rédiger des phrases du type « Au bout de m mois, la survie sans événement est estimée à X%[Y ; Z] ».

Il faut souligner que :

- Ce calcul est complexe, et n'est pas celui d'une simple proportion de survivants, car l'effectif utilisé pour estimer la survie à un instant donné n'est ni l'effectif de départ, ni le nombre de survivants à ce temps-là
- Ce calcul d'intervalle de confiance est inexact au début du suivi, lorsque la survie est proche de 100%
- Graphiquement, la représentation de l'IC95 n'est pas toujours très lisible

Si vous en êtes à un point où vous souhaitez calculer l'IC95 d'une courbe de survie, il est peut-être temps de vous rapprocher d'un biostatisticien et de vous faire aider. Sinon, pour de nombreux mémoires académiques, ce calcul n'est pas indispensable.

2.4.3 Tests univariés

En pratique, aucun test statistique univarié n'est réalisé sur une variable de survie seule :

- pas de test d'adéquation à une norme
- pas de test apparié dans un seul groupe

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3 Analyses statistiques bivariées

3.1 Préambule

La plupart des analyses inférentielles en santé cherchent à montrer la présence d'une liaison statistique entre deux variables, ou à quantifier son importance. L'analyse conjointe de deux colonnes relatives à un phénomène différent s'appelle l'**analyse bivariée**.

Cela exclut les analyses descriptives de survie : bien qu'on représente ces variables en utilisant deux colonnes, il s'agit en réalité d'une seule et même variable, donc d'une analyse univariée.

La plupart du temps, et cela vaut quelle que soit la signification des variables étudiées, on cherche à montrer qu'il existe une **association statistique entre deux variables**, en rejetant l'hypothèse d'indépendance statistique. L'ensemble des méthodes mises en œuvre à cette fin seront présentées en section 3.2 Cas général : liaison statistique entre deux colonnes 151.

On peut également s'intéresser à **deux colonnes relatives à la même information**, afin de savoir si cette information varie (ex : comparaisons avant-après, gauche-droite), mais dans un seul groupe d'individus. **Il ne s'agit alors pas d'analyse bivariée**. Les méthodes à mettre en œuvre ont déjà été abordées dans la section des analyses **univariées** : voir le chapitre 2.2.5 Tests de comparaison « de deux proportions appariées » en page 124, et le chapitre 2.3.4 Tests de comparaisons « de deux moyennes appariées » en page 142.

Cette question des mesures appariées peut néanmoins se poser dans plusieurs groupes. Il s'agit alors bien d'analyses bivariées, et nous les présenterons dans le chapitre 3.3 Deux variables appariées, dans plusieurs groupes en page 186.

Enfin, dans une dernière partie, nous aborderons les cas particuliers où des méthodes visant plutôt à **quantifier la force de l'association** sont traditionnellement employées (épidémiologie analytique avec risque relatif et odds ratio, tests diagnostiques, courbe ROC, accord inter-juges et coefficient Kappa, outils de détection, intelligence artificielle, etc.) : 3.4 Cas particuliers d'analyses bivariées en page 188.

L'ensemble de ces possibilités est présenté en Figure 86 en page 150.

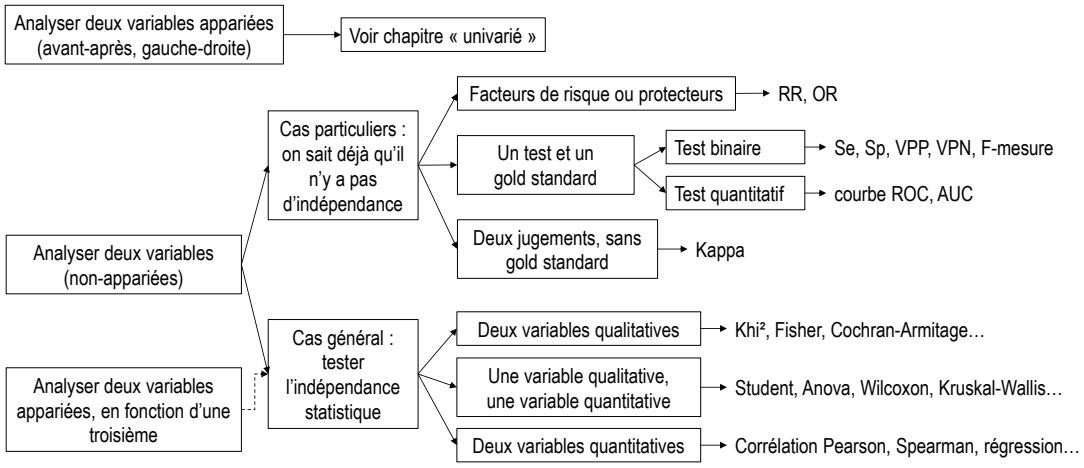


Figure 86. Présentation générale des analyses bivariées

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.2 Cas général : liaison statistique entre deux colonnes

3.2.1 Préambule

Nous avons vu, en introduction de la section dédiée aux analyses descriptives, que nous pouvions finalement décrire **trois types de variables** statistiques :

- Les variables qualitatives (dont les variables binaires)
- Les variables quantitatives
- Les variables de survie

Les combinaisons de ces variables deux-à-deux nous donne **cinq cas de figure** possibles (voir Tableau 14 par la suite) :

- Deux variables qualitatives
- Une variable quantitative et une variable qualitative
- Deux variables quantitatives
- Une variable de survie et une variable qualitative
- Une variable de survie et une variable quantitative

Nous ne proposerons aucune méthode pour deux variables de survie. Nous détaillerons ces cinq cas de figure dans les sections suivantes.

L'ensemble de ces analyses s'intéresse à l'association statistique entre ces deux colonnes et cherchera, d'une manière ou d'une autre, à rejeter cette indépendance, au moins sur certains aspects.

Certaines de ces analyses sont **traditionnellement présentées comme des « comparaisons »**, ce qui peut perturber les étudiants. Certains croient même qu'une analyse bivariée compare deux paramètres. Non, une analyse bivariée analyse conjointement deux variables. Nous en verrons deux exemples.

Exemple 1 : *on dispose de 3 groupes d'habitants, et on s'intéresse à la proportion de malades dans ces trois groupes. On cherche à montrer que ces proportions sont différentes.*

Formulation traditionnelle du problème : on cherche à *comparer trois proportions*. Mais, de même, on aurait pu chercher à comparer les proportions des trois groupes chez les malades, versus chez les non-malades. On pourrait croire qu'il y a autant de tableaux de données que de sous-groupes.

Formulation préférée dans cet ouvrage : dans un tableau de données contenant tous ces habitants, on s'intéresse à *deux variables* : une définissant le statut malade (malade / non-malade), et une autre définissant le groupe auquel les habitants appartiennent (variable qualitative à 3 modalités).

Ainsi, comparer des proportions peut plutôt être décrit comme **analyser la relation entre deux variables qualitatives**, dans un tableau de données unique.

Exemple 2 : *on dispose de 3 groupes d'habitants, et on s'intéresse à leur taille moyenne. On cherche à montrer que ces tailles moyennes sont différentes.*

Formulation traditionnelle du problème : on cherche à *comparer trois moyennes*. On pourrait croire qu'il y a trois tableaux de données.

Formulation préférée dans cet ouvrage : dans un tableau de données contenant tous ces habitants, on s'intéresse à *deux variables* : une variable quantitative décrivant la taille, et une autre définissant le groupe auquel les habitants appartiennent (variable qualitative à 3 modalités).

Ainsi, comparer des moyennes peut plutôt être décrit comme **analyser la relation entre une variable qualitative et une variable quantitative**, dans un tableau de données unique.

Globalement, ces analyses bivariées s'intéresseront à la liaison statistique entre deux variables, et les tests mis en œuvre chercheront à rejeter l'indépendance entre les variables. En réalité, ce n'est pas vraiment le cas lorsqu'au moins une variable quantitative est impliquée.

Tableau 14. Nature de l'hypothèse nulle des analyses bivariées courantes

	Qualitative	**Quantitative**	**De survie**
Qualitative	Indépendance	Egalité des moyennes	Indépendance
Quantitative	Egalité des moyennes	Absence de relation linéaire	Absence de relation log-linéaire
De survie	Indépendance	Absence de relation log-linéaire	[sans objet]

Le Tableau 14 résume la nature des hypothèses nulles que nous pourrons tester.

Entre deux variables qualitatives, on peut bien parler de tests d'indépendance entre les deux variables.

Entre une variable qualitative X et une variable quantitative Y, dans les tests utilisés en pratique, nous nous intéresserons principalement à la tendance centrale de Y (moyenne, médiane…), et non à sa dispersion. Nous répondrons principalement à la question « est-ce que la tendance centrale de la Y est influencée par la valeur de X ? ». Il ne s'agit pas stricto sensu d'indépendance statistique. Ainsi, nous serons capables de détecter les cas où la tendance centrale de Y dépend bien de X (Figure 87 partie gauche), mais nous ne nous intéresserons pas aux cas où seule la dispersion de Y dépend de X (Figure 87 partie droite)[13].

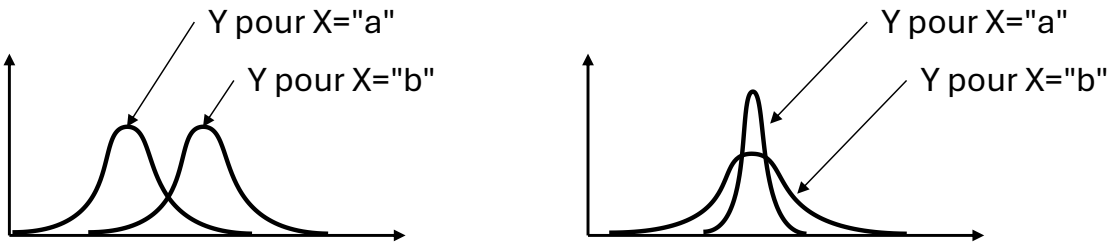


Figure 87. Exemples de situations de non-indépendance qualitatif-quantitatif :
Gauche : la tendance centrale de Y dépend de X
Droite : la dispersion de Y dépend de X

Entre une variable quantitative X et une variable quantitative Y, dans les tests utilisés en pratique, nous nous intéresserons principalement à la détection d'une relation linéaire, qu'elle soit sur les valeurs réelles ou sur leurs rangs. Cette approche nous permettra de détecter des situations fréquentes de non-indépendance (exemple au milieu de la Figure 88), mais il faudra garder à l'esprit qu'il existe des situations de non-indépendance, qui se traduisent néanmoins par l'absence de relation linéaire (exemple à droite de la Figure 88).

[13] Le test de Fisher-Snedecor, par exemple, s'intéresse à cette question. Cette question ne correspondant pas à un cas classique d'analyse de données, nous ne l'aborderons pas dans cet ouvrage.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

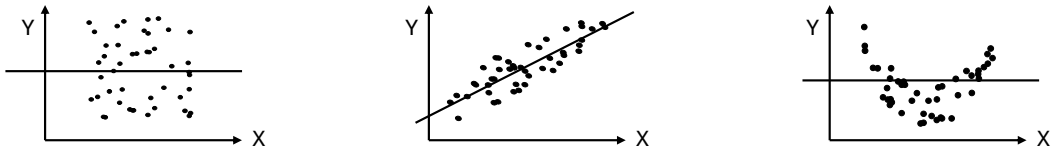


Figure 88. Exemples de situations quantitatif-quantitatif :
Gauche : indépendance, donc absence de relation linéaire
Milieu : liaison statistique monotone, donc présence (notamment) d'une relation linéaire
Droite : liaison statistique, mais impossible de détecter une relation linéaire

Enfin, lorsqu'une variable de survie est impliquée, les tests que nous décrirons pourront bien s'intéresser à l'indépendance entre une variable qualitative et une variable de survie. En revanche, on ne pourra pas évaluer l'indépendance entre une variable quantitative et une variable de survie, mais plutôt l'absence de log-linéarité de l'effet de la variable quantitative sur le risque instantané d'événement représenté dans la variable de survie.

3.2.2 Deux variables qualitatives

3.2.2.1 Préambule

Tester l'indépendance entre deux variables qualitatives A et B revient à se poser plusieurs questions, similaires :

- Est-ce que la distribution de A est identique, dans tous les sous-groupes d'individus définis selon leur valeur de B ?
- Est-ce que la distribution de B est identique, dans tous les sous-groupes d'individus définis selon leur valeur de A ?
- Est-ce que le fait de connaître la valeur de A modifie les hypothèses qu'on peut faire sur la valeur de B, lorsqu'elle est inconnue ?
- Est-ce que le fait de connaître la valeur de B modifie les hypothèses qu'on peut faire sur la valeur de A, lorsqu'elle est inconnue ?

Le sens de la relation A→B ou B→A n'a aucune importance d'un point de vue statistique. Cependant, pour une présentation cohérente des graphiques et des résultats, il est plus approprié d'identifier quelle variable pourrait être la cause potentielle de l'autre, ou au moins quelle variable est connue avant l'autre.

Voici un exemple. Soit la variable « sexe » (H/F), et la variable « prisonnier » (oui/non). Elles ne sont pas indépendantes car :

- *La proportion de prisonniers est plus forte chez les hommes que chez les femmes (1)*
- *La proportion d'hommes est plus forte chez les prisonniers que chez les non-prisonniers (2)*
- *Le fait d'être un homme augmente la probabilité d'être en prison (3)*
- *Le fait d'être en prison augmente la probabilité d'être un homme (4)*

Toutes les affirmations ci-dessus sont vraies. Cependant, comme on sait que le sexe est déterminé dès la naissance alors que le statut prisonnier oui/non l'est plutôt à l'âge adulte, on présentera plus volontiers les affirmations (1) et (3), et on tracera des graphiques dans ce sens.

3.2.2.2 Représentation graphique

L'objectif de la représentation graphique est de voir quelles sont les proportions des modalités de la variable B, au sein de chaque modalité de la variable A.

La première étape est de tracer un tableau de contingence présentant les effectifs en fonction des deux variables simultanément. La Figure 89 montre le statut tabagique en fonction du

sexe dans une étude. Il aurait été moins intuitif, mais aucunement inexact, de présenter le sexe en fonction du statut tabagique.

Le graphique de gauche en Figure 89, réalisé avec un tableur, n'illustre pas les proportions d'hommes ou de femmes, mais permet de voir que dans ce fichier il y a plus de fumeurs chez les femmes que chez les hommes.

Le graphique de droite en Figure 89 a été réalisé avec un logiciel de statistique. Il réalise un premier partitionnement proportionnel sur le sexe, puis un deuxième sur le statut tabagique. Il permet aussi de voir que le fichier comprend nettement moins de femmes que d'hommes.

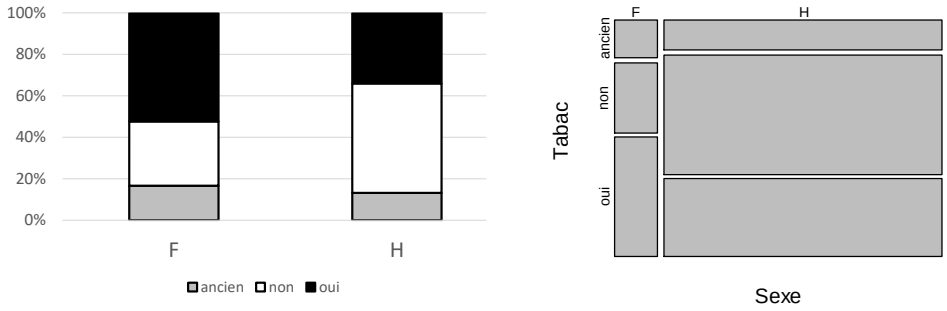


Figure 89. Statut tabagique en fonction du sexe
gauche : barres empilées avec un tableur
droite : diagramme en mosaïque avec un logiciel de statistique

3.2.2.3 Description par des paramètres

Pour décrire la relation entre les deux variables qualitatives, il suffit de décrire des proportions d'une variable, conditionnellement à l'autre variable. En voici un exemple :

« Pami les femmes de l'échantillon, 22 (52,4%) sont fumeuses, 7 (16,7%) sont anciennes fumeuses et 13 (31,0%) sont non-fumeuses. Parmi les hommes, 93 (34,1%) sont fumeurs, 36 (13,2%) sont anciens fumeurs, et 144 (52,7%) sont non-fumeurs. »

On notera que seuls les effectifs sont objectifs, tandis que les pourcentages affichés sont ici calculés au sein de chaque modalité du sexe. C'est un choix, qui est ici cohérent pour le lecteur, mais qui n'est en rien déterminé par une règle mathématique. C'est la tournure de phrase « parmi les femmes (…) », et un peu de calcul mental sur les pourcentages (52,4%+16,7%+31,0%≈100,0%), qui permettent de comprendre comment les pourcentages sont calculés.

Il nous faut à présent utiliser un test statistique pour savoir si cette apparente liaison statistique pourrait être le fait du seul hasard, ou si elle reflète nécessairement une non-indépendance en population.

3.2.2.4 Tests statistiques : arbre décisionnel

La Figure 90 montre l'arbre décisionnel pour tester l'indépendance entre deux variables qualitatives (ou binaires). Nous verrons plus bas que le principal test utilisé est le test du **Khi² d'indépendance**. Ce test peut aisément être réalisé avec un tableur, mais il requiert une condition d'effectif théorique minimal. Lorsque cette condition n'est pas remplie mais qu'une condition plus souple est remplie, on peut réaliser ce même test avec la **correction de continuité de Yates**. Hélas, sa réalisation avec un tableur nécessitera plus d'opérations, ce que nous ne montrerons pas ici. Enfin, lorsqu'aucun des tests n'est réalisable, un **test exact de Fisher** pourra être réalisé avec un logiciel de statistique.

Dans la plupart des études, si la condition d'effectif n'est pas remplie mais qu'au moins une variable a plus de deux modalités, une solution simple et efficace consistera à regrouper les

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

modalités les plus rares, de manière à remplir la condition d'effectifs. Ce regroupement devra être guidé par l'expertise métier, et devra être cohérent.

Par exemple, si vous analysez une variable « diagnostic » qui a pour modalités « oui » « non » et « douteux », vous pourrez regrouper les modalités en « oui » et « non ou douteux », ou encore « non » et « oui ou douteux », de manière à équilibrer les effectifs, et selon les implications en pratique clinique. Il n'aurait pas de sens de regrouper les modalités en « douteux » d'une part et « oui ou non » d'autre part.

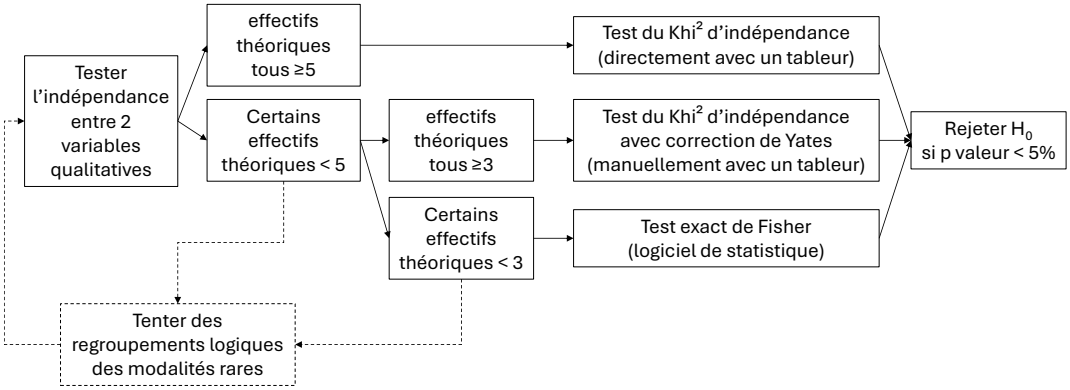


Figure 90. Arbre décisionnel : tester l'indépendance entre deux variables qualitatives

3.2.2.5 Test du Khi² d'indépendance

3.2.2.5.1 Exemple de mise en œuvre

Dans cet exemple, en partant d'un échantillon de 315 individus, nous souhaitons savoir si le statut tabagique est indépendant du sexe.

Nous posons **l'hypothèse nulle H_0** : le statut tabagique et le sexe sont indépendants.

Après avoir évincé les individus qui comportent des valeurs manquantes, nous traçons un premier tableau de contingence croisé, permettant de visualiser les **effectifs observés** pour les modalités des deux variables en même temps (partie gauche de Figure 91). Nous traçons ensuite un deuxième tableau, décrivant les **effectifs théoriques**, ceux qu'on aurait pu obtenir si les deux variables avaient été strictement indépendantes, et **sans arrondir** à l'entier le plus proche (partie droite de Figure 91). Voici comment le tracer :

- Reproduire exactement la même structure que celle du tableau initial
- Reporter le total général et les totaux des lignes et des colonnes, à l'identique
- Calculer chaque effectif théorique, comme étant le produit du total de la ligne et du total de la colonne, divisé par l'effectif total (en haut de la Figure 91, on peut voir la formule utilisée dans la cellule L37). Conserver la partie non-entière de ces nombres (2 chiffres après la virgule)
- Vérifier au passage que chaque valeur est bien supérieure ou égale à 5

La question est : la plage d'effectifs observés (ici 6 cellules, à gauche) est-elle très écartée de la plage d'effectifs théoriques (ici 6 cellules, à droite) ? Pour y répondre, dans une dernière cellule (ici D41), utiliser la formule **chisq.test()** avec Excel ou Calc pour comparer les deux plages de données (cette formule est recopiée en E41 dans l'exemple).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

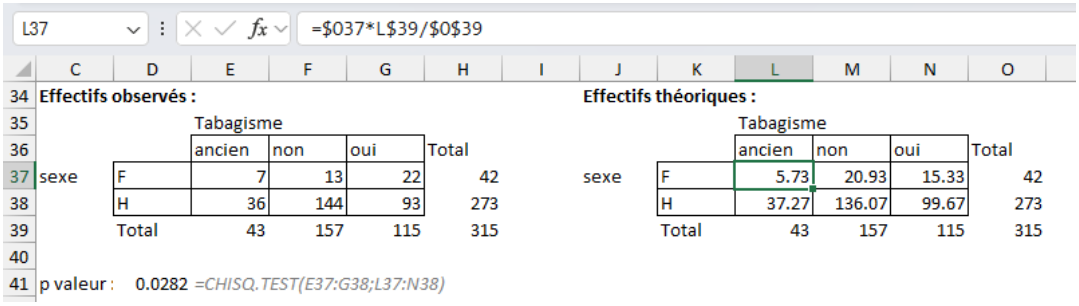


Figure 91. Réalisation d'un Khi² d'indépendance avec un tableur

Nous obtenons dans ce cas une p valeur inférieure à 5%, et pouvons conclure que le statut tabagique n'est pas indépendant du sexe. On peut tourner cette conclusion comme on le souhaite :

- Ces deux variables sont associées statistiquement
- La répartition du sexe n'est pas la même dans les différentes modalités de statut tabagique
- La répartition du statut tabagique n'est pas la même entre les hommes et les femmes
- Etc.

3.2.2.5.2 Le Khi² d'indépendance, en général

Revenons sur le déroulement d'un test du Khi² d'indépendance, avec quelques formules, mais sans trop de détails.

Nous nous intéressons à deux variables qualitatives :

- A, variable qualitative à *l* modalités avec *l*≥2
- B, variable qualitative à *c* modalités avec *c*≥2

Nous posons l'**hypothèse nulle H_0** : les variables A et B sont indépendantes

Nous avons obtenu des effectifs observés, pour chaque modalité jointe de A et B (Équation 22) :

$$\forall i,j \quad O_{i,j} = \#[(A = A_i) \cap (B = B_j)]$$

Équation 22. Effectifs observés

En supposant que H_0 soit vraie, nous calculons les effectifs théoriques. Chaque effectif est le produit du total de la ligne et du total de la colonne, divisé par le total général. Nous devons vérifier que chaque effectif théorique est supérieur ou égal à 5 (Équation 23).

$$\forall i,j \quad T_{i,j} = \frac{L_i \times C_j}{n}$$

$$vérifier : \forall i,j \quad T_{i,j} \geq 5$$

Équation 23. Effectifs théoriques sous H_0

Nous obtenons ainsi deux matrices de dimensions identiques : une pour les effectifs observés, l'autre pour les effectifs théoriques.

Nous calculons ensuite la quantité du Khi² (Équation 24). Pour ce faire, nous prenons chaque couple de cases (observé et théorique) et calculons leur écart au carré, divisé par l'effectif théorique. Nous calculons simplement la somme de toutes les valeurs obtenues.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - T_{i,j})^2}{T_{i,j}}$$

Équation 24. Statistique de test du Khi² d'indépendance

Lorsque le plus faible des effectifs théoriques est égal à 3 ou 4, la condition de validité précédente n'est pas atteinte. Il est cependant possible de poursuivre en appliquant la **correction de continuité de Yates**, qui consiste à retrancher 0,5 de chaque écart observé-théorique constaté. Le reste du test est inchangé.

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - T_{i,j} - 0{,}5)^2}{T_{i,j}}$$

$$vérifier : \forall i,j \quad T_{i,j} \geq 3$$

Équation 25. Correction de continuité de Yates

La quantité X^2 est toujours positive ou nulle. Si H_0 est vraie, alors cette quantité (toujours positive) est plutôt proche de zéro, et sa loi de probabilité suit une loi du Khi², assortie d'un paramètre supplémentaire, appelé nombre de degrés de liberté. Le nombre de degrés de liberté (ddl), noté ν (nu), est le nombre de cases de chaque matrice, en omettant la dernière ligne et la dernière colonne. Ainsi, dans le cas typique de deux variables binaires, il vaut 1. La référence à la table de loi du Khi² nous permet d'obtenir la *p valeur*.

$$\nu = (l-1) \times (c-1)$$

Équation 26. Nombre de degrés de liberté du test du Khi² d'indépendance

Nous pouvons conclure :

- Fixons le seuil d'interprétation de la p valeur à 5% (risque α)
- Si **p valeur < 5% : on rejette H_0**, on peut conclure que les deux variables sont statistiquement liées (elles ne sont pas indépendantes)
- Si **p valeur > 5%** : on ne rejette pas H_0 : on se trouve face à une **indétermination**, il est interdit de conclure.

3.2.2.5.3 Quelques précisions sur le test du Khi² d'indépendance

Le test du Khi² d'indépendance (avec ou sans correction de Yates) peut être classé ainsi :

- C'est un **test non-paramétrique** car, formellement, il ne cherche pas à estimer un paramètre, mais observe des effectifs. Ceci peut paraître artificiel, car les effectifs sont proportionnels aux proportions qui, elles, sont bien des paramètres
- C'est un **test asymptotique** : la p valeur est calculée en se référant à une distribution théorique, qui n'est vraie que si certains effectifs sont suffisants

Dans la Figure 92, nous avons représenté les limites de significativité du test du Khi² d'indépendance. Imaginons deux variables A et B, dont les effectifs sont équilibrés à 50% dans un échantillon de taille n (Tableau 15). On peut fixer un effectif x (compris entre 0 et n/2) dans la case en haut à gauche, et en déduire les autres effectifs. On imagine que si x=n/4, un test ne pourra jamais rejeter l'hypothèse nulle. Inversement, si x se rapproche de 0 (ou de n/2), au-delà d'un certain effectif, il deviendra possible de rejeter l'hypothèse nulle. Ce rejet surviendra avec une différence de proportions de B observées dans chaque sous-groupe de A (ou l'inverse). La Figure 92 représente, en fonction de la taille d'échantillon, la plus petite différence de proportion qu'on peut mettre en évidence.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 15. Exemple de tableau de contingence, prévalences de A et B fixées à 50%

	A=0	A=1	*Totaux*
B=0	x	n/2-x	*n/2*
B=1	n/2-x	X	*n/2*
Totaux	*n/2*	*n/2*	*n*

La Figure 93 fait de même avec une correction de Yates toujours appliquée. On constate que cette correction permet de traiter des échantillons plus petits. Néanmoins, dans certains cas, elle aboutit à une perte de puissance. Par exemple, pour n=22, le test avec correction de Yates rejette H_0 à partir de 2+9+2+9 (soit une différence de proportions de 63%) alors que le test sans correction de Yates rejetait H_0 dès 3+8+3+8 (soit une différence de proportions de 45%).

Enfin, la Figure 94 présente les limites de significativité du Test Exact de Fisher, que nous présenterons plus bas. Ce test permet de rejeter H_0 avec des effectifs très faibles (n=8, dans la configuration 0+4+0+4) et présente ensuite les mêmes limites de significativité que le Khi² avec correction de Yates.

Lorsqu'on sait que le test exact de Fisher s'appuie sur un calcul exact de la p valeur, on peut en déduire que c'est le test du Khi² sans correction de Yates qui, dans certains cas, rejette à tort l'hypothèse nulle. Pour cette raison, en plus de l'utilisabilité, **la correction de Yates est activée par défaut** dans la fonction chisq.test() du logiciel R.

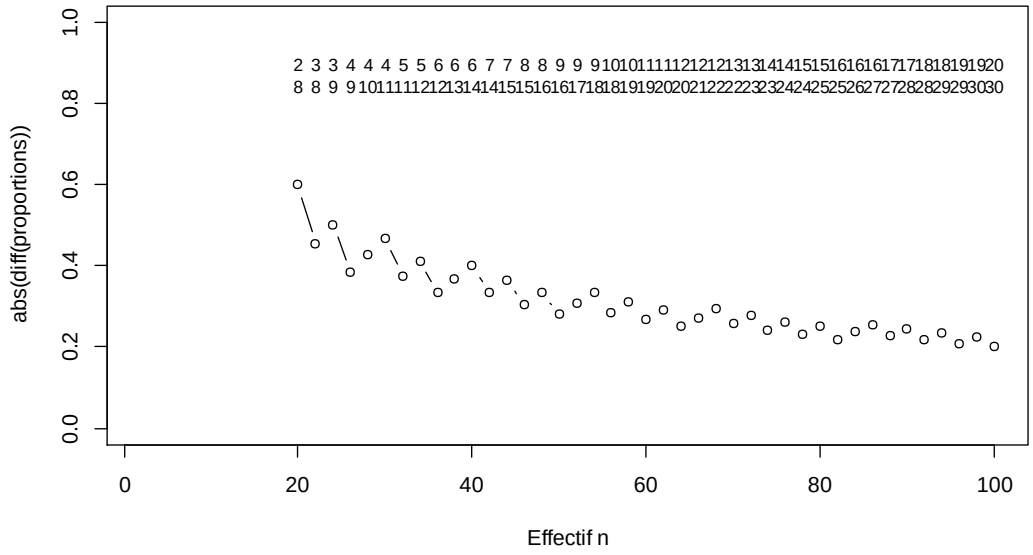


Figure 92. Limite de significativité d'un Khi² d'indépendance, exprimée en différence de proportions, selon le scénario défini dans le texte
Ex : pour n=20, on rejette H0 pour des effectifs de 2+8+2+8 (et donc aussi 1+9+1+9 et 0+10+0+10)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

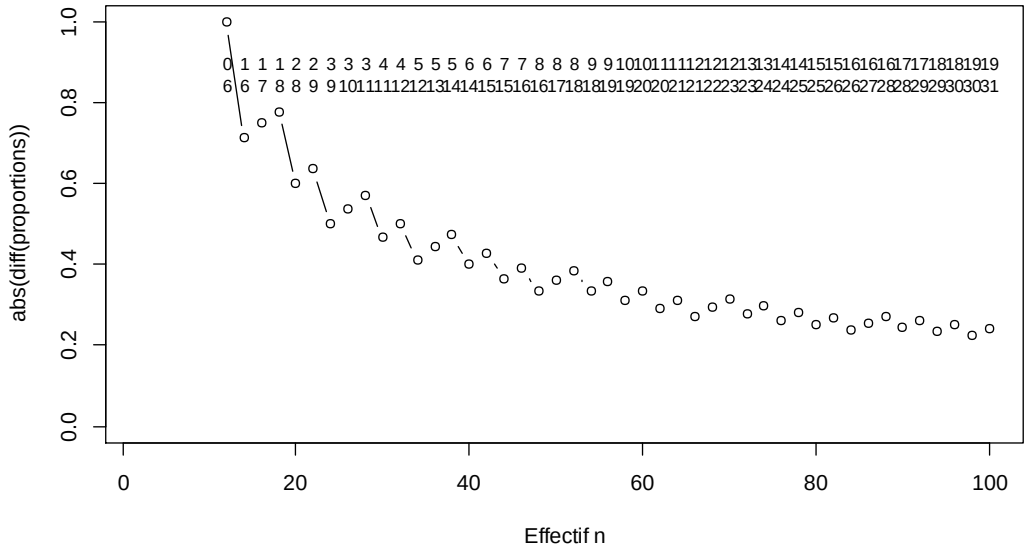


Figure 93. Limite de significativité d'un Khi² d'indépendance avec correction de Yates

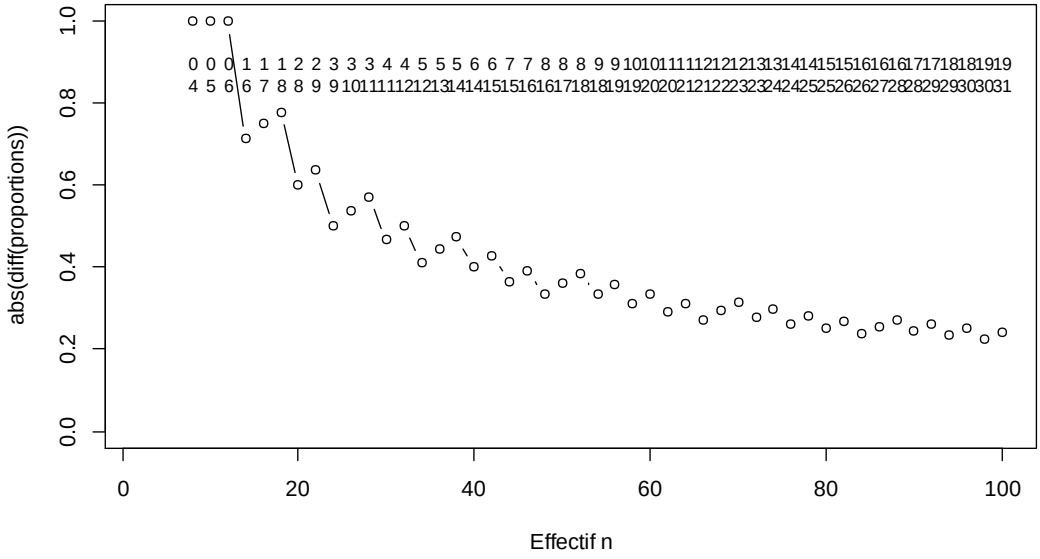


Figure 94. Limite de significativité d'un test exact de Fisher

Une autre situation également décrite est le **test du Khi² d'homogénéité**. Dans ce test, on dispose de deux échantillons différents et, pour chacun, des effectifs des modalités d'une variable qualitative unique. On cherche à savoir si les proportions observées sont significativement différentes entre les deux échantillons. Ce test est, en apparence, conceptuellement très différent du Khi² d'indépendance. Cependant, pour le mettre en œuvre, on pourra très bien regrouper les deux échantillons dans un tableau unique de données, dans lequel on ajoutera une colonne binaire mentionnant de quel échantillon la ligne est issue. Cette manière de présenter les données correspond aux bonnes pratiques, et a été présentée en Figure 27 en page 75. Cela étant fait, le test revient à tester l'indépendance entre la variable binaire qui définit l'échantillon, et la variable qualitative étudiée. Ce test se met en œuvre exactement comme précédemment. La différence n'est que conceptuelle, et ne persiste pas dans la mise en œuvre.

3.2.2.6 Autres tests

D'autres tests sont classiquement enseignés, mais soit ils ne sont pas utilisés en pratique, soit ils nécessitent d'utiliser un logiciel de statistique.

Le test de **comparaison de deux proportions avec la loi normale** :

Ce test est paramétrique (il s'intéresse aux proportions) et asymptotique (il nécessite certains effectifs minimaux). Ce test pourrait concurrencer le Khi² dans le cas d'un tableau de contingence à 2x2 cases. Il n'est pas utilisé en pratique, car il est mathématiquement proche du test du Khi², mais couvre une plage d'utilisation plus restreinte.

Le **test exact de Fisher**, mis au point par Ronald Fisher en 1935[34] :

Ce test est non-paramétrique (il s'intéresse aux effectifs) et exact (sa p valeur a été calculée de manière exacte par son créateur). Comme nous l'avons dit plus haut, ce test est toujours valide et, lorsque sa p valeur diffère de celle du Khi², c'est le test exact de Fisher qui a raison. Ce test a également été étendu aux tableaux de contingence comportant plus de 2x2 cases. C'est donc le test parfait ! Malheureusement, il ne peut pas simplement être réalisé avec un tableur. De plus, pour de grands effectifs, même s'il est éminemment valide, le temps de calcul est plus important.

Le **test de tendance de Cochran–Armitage**, mis au point par Cochran et Armitage en 1954[35,36] :

Ce test est non-paramétrique et asymptotique. Il teste l'indépendance entre une variable binaire, et une variable qualitative ordonnée (3 modalités ou plus). Il devrait être préféré lorsque la variable non-binaire présente une gradation (ex : « non », « un peu », « beaucoup »). En pratique, il est très fréquent que même dans de telles circonstances, les chercheurs continuent d'utiliser le test du Khi² d'indépendance.

3.2.3 Une variable quantitative et une variable qualitative

3.2.3.1 Préambule

Comme nous l'avons énoncé précédemment (voir 3.2.2.1 Préambule en page 153), lorsqu'on s'intéresse à l'indépendance entre une variable qualitative et une variable quantitative, c'est en réalité en termes de **tendance centrale** de la variable quantitative, et non en termes de dispersion.

Tester l'indépendance entre la tendance centrale d'une variable quantitative Y et une variable qualitative X revient à se poser plusieurs questions, similaires :

- Est-ce que la tendance centrale (la moyenne, le plus souvent) de Y est identique, dans tous les sous-groupes d'individus ayant la même valeur de X ?
- Est-ce que la proportion de X=1 est identique selon que Y prend des valeurs faibles ou élevées ?
- Est-ce que le fait de connaître la valeur de X (0 ou 1) modifie les hypothèses qu'on peut faire sur la valeur de Y, lorsqu'elle est inconnue ?
- Est-ce que le fait de connaître la valeur de Y (nombre réel) modifie les hypothèses qu'on peut faire sur la valeur de X, lorsqu'elle est inconnue ?

Le sens de la relation X→Y ou Y→X n'a aucune importance d'un point de vue statistique. Cependant, de manière générale, nous emploierons le sens X→Y, et présenterons la distribution de Y pour chaque modalité de X.

Voici un exemple. Soit la variable « sexe » (H/F), et la variable « poids » (en kg). Elles ne sont pas indépendantes car :

- *Le poids moyen est plus important chez les hommes (1)*
- *La proportion d'hommes augmente lorsque le poids augmente (2)*
- *Le fait d'être un homme augmente la probabilité d'avoir un poids élevé (3)*

- *Le fait s'avoir un poids élevé augmente la probabilité d'être un homme (4)*

Toutes les affirmations ci-dessus sont vraies. Cependant, pour des raisons de causalité mais surtout de simplicité de la formulation, on présentera plus volontiers les affirmations (1) et (3), et on tracera des graphiques dans ce sens.

De manière générale et pour simplifier, nous décrirons toute cette section comme étant un problème de **comparaisons de plusieurs moyennes**.

3.2.3.2 Représentation graphique

Le graphique par excellence dans cette situation est **un ensemble de boxplots**, représentant la distribution de la variable quantitative Y, pour chaque modalité de la variable qualitative X (voir Figure 95). Cette représentation est largement suffisante car nous avons besoin de deux informations :

- La tendance centrale (médiane ou moyenne), car les tests que nous utiliserons s'intéresseront uniquement à cet indicateur
- La dispersion (Q1 et Q3), simplement parce que certains tests que nous utiliserons nécessiteront des dispersions du même ordre de grandeur

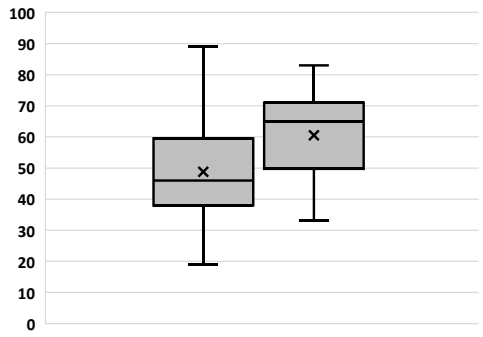


Figure 95. Exemple de boxplot avec Microsoft Excel : âge en fonction du sexe dans un échantillon (hommes à gauche et femmes à droite)

Pour réaliser ce graphique avec un tableur, il faudra séparer la série Y en plusieurs séries présentées indépendamment (Figure 96) : Y pour $X=x_1$, Y pour $X=x_2$, etc. On créera ensuite un graphique en boxplot pour une seule des séries puis, une fois le graphique créé, on ajoutera des séries de données dans le même graphique.

	A	B	C	D	E	F	G
1	age	sexe		age hommes		age femmes	
2	90	H		64		57	
3	89	H		76		66	
4	83	H		89		64	
5	83	F		40		62	
6	83	F		72		75	
7	82	H		40		57	
8	78	H		65		56	
9	78	H		58		65	
10	77	H		35		71	

Figure 96. Exemple de séparation manuelle de l'âge en deux séries : âge en fonction du sexe

Il serait possible de présenter côte-à-côte deux histogrammes mais premièrement, les histogrammes apportent trop d'information. Deuxièmement, pour qu'ils soient comparables, il faudrait que le découpage en classes soit le même, et que les bornes des axes soient identiques. Ceci n'est pas possible avec l'histogramme automatique d'un tableur par exemple. Il nous faudrait réaliser de nombreuses manipulations pour juxtaposer deux histogrammes réalisés manuellement (voir Figure 68 en page 129).

Dans le cas particulier du sexe et de l'âge, la pyramide des âges est appréciée car tout le monde sait la lire. Si vous souhaitez réaliser un tel graphique avec un tableur et que vous êtes prêt à y consacrer une demi-heure, de nombreux tutoriels peuvent être trouvés sur le web (Figure 97).

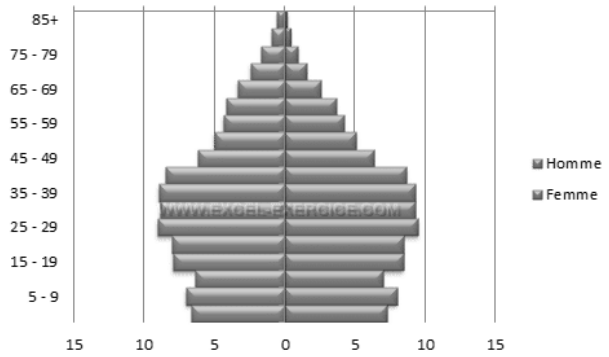


Figure 97. Exemple de pyramide des âges sur https://excel-exercice.com/pyramide-ages/

3.2.3.3 Description par des paramètres

On se contentera de décrire la variable comme nous l'avons vu dans la section dédiée aux analyses univariées, mais en décomposant la description de Y pour chaque modalité de X. La stratégie de description (moyenne et écart type, versus quartiles) est décidée une fois pour toute pour une variable donnée, et non pour chaque modalité.

Exemple : « Dans notre échantillon, les hommes sont en moyenne âgés de 32,3 ans (SD=5,8), tandis que les femmes sont en moyenne âgées de 30,8 ans (SD=5,6) ».

Exemple : « Dans notre échantillon, les hommes ont en médiane 2 enfants (Q1-Q3 : [0 ; 3]), et les femmes ont en médiane 1 enfant (Q1-Q3 : [0 ; 2]) ».

/!\ On notera que, si vous souhaitez réaliser un test statistique, il ne sera pas nécessaire de décrire les distributions de manière aussi complète : les tendances centrales et la p valeur seront suffisantes. Après réalisation d'un test statistique, le premier exemple pourrait être formulé comme suit :

Exemple : « L'âge moyen diffère significativement en fonction du sexe : 32,3 ans pour les hommes, 30,8 ans pour les femmes (p=0,028). ».

3.2.3.4 Tests statistiques : arbre décisionnel

Pour analyser la relation statistique entre une variable qualitative et une variable quantitative, nous vous proposons l'arbre décisionnel de la Figure 98. Cet arbre est sensiblement différent de ce qui est classiquement enseigné, mais nous allons justifier les choix réalisés. Ces choix sont scientifiquement valides et vous faciliteront grandement la suite des travaux.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

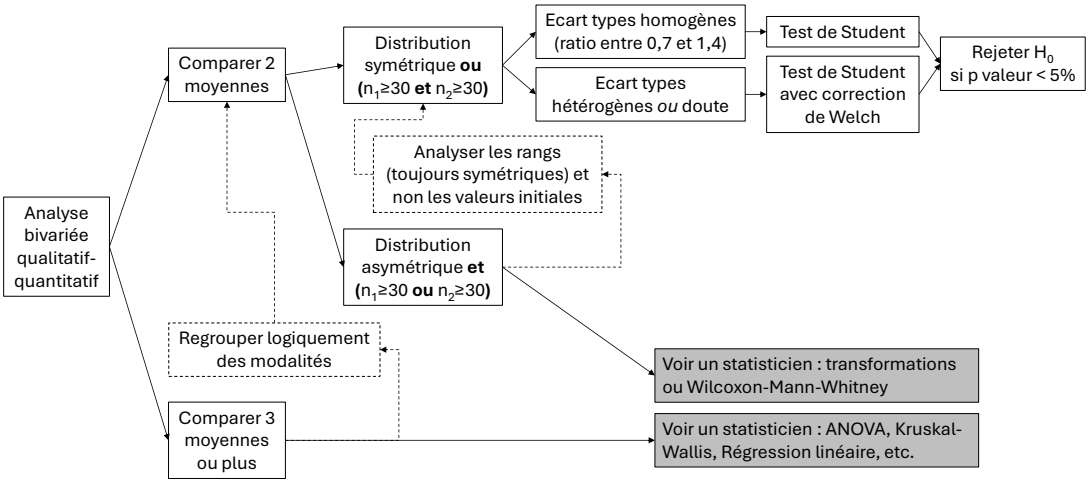


Figure 98. Analyse bivariée qualitatif-quantitatif : arbre décisionnel

Commentons l'arbre décisionnel de la Figure 98.

Si vous souhaitez comparer plus de deux moyennes (3, 4, etc.), sachez que les méthodes disponibles avec un tableur ne le permettent pas. Une possibilité très simple pour vous sera de faire des **regroupements pertinents** d'un point de vue métier. Il est également possible de **comparer les modalités deux-à-deux**, mais en appliquant la **correction de Bonferroni** : au lieu d'interpréter les p valeurs au seuil de 5%, vous les interprèterez au seuil de 5/k%, k étant le nombre de tests réalisés. Si vous tenez absolument à analyser simultanément 3 sous-groupes ou plus, vous devrez faire appel à un statisticien qui réalisera une Anova (analyse de la variance), un test de Kruskal-Wallis, une régression linéaire, etc. Nous dirons un mot sur ces méthodes dans le chapitre 3.2.3.6 Autres tests statistiques en page 169.

Si vous souhaitez comparer deux moyennes (analyser la relation entre une variable binaire et une variable quantitative), vous utiliserez un **test de Student**. Académiquement, ce test requiert soit que la variable quantitative suive une loi normale dans les deux sous-groupes, soit que les deux sous-groupes de l'échantillon contiennent **chacun au moins 30 individus**. La condition de normalité est en réalité sévère : si la variable initiale a, dans chaque sous-groupe, **une distribution symétrique**, c'est en pratique suffisant. Si tel n'est pas le cas, il sera toujours possible, au lieu d'étudier les valeurs de la variable initiale, d'étudier les **rangs des individus**[37]. Nous l'illustrerons plus bas. Les rangs suivent une distribution uniforme par définition, et ont donc toujours une distribution symétrique.

Une dernière question se posera. Le test de Student initial suppose que les dispersions de la variable quantitative sont similaires entre les deux sous-groupes analysés. On parle d'**homogénéité des variances**, ou encore d'**homoscédasticité** (par opposition à l'hétéroscédasticité). Il est classiquement enseigné que ce point doit être testé à l'aide du test de Fisher-Snedecor. Nous déconseillons cette attitude, pour les raisons expliquées dans le chapitre 5.3 Tests qu'on réalise en espérant ne pas rejeter H0 en page 216. Nous suggérons plutôt de calculer les variances dans chaque sous-groupe : si elles sont du même ordre de grandeur (ratio compris entre 1/2 et 2), on réalisera le test de Student classique. Dans le cas contraire ou en cas de doute, on ajoutera une correction de Welch. Si vous calculez ce ratio entre les écarts types et non les variances, il devra être approximativement compris entre 0,7 et 1,4).

On notera que dans le logiciel de statistique R, la correction de Welch est activée par défaut dans la fonction t.test(). Cette correction n'est jamais fausse, mais entraîne une diminution de la puissance statistique du test (aptitude à rejeter H_0 dans certains cas limites).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.2.3.5 Test de Student pour échantillons indépendants, avec ou sans correction de Welch

3.2.3.5.1 Exemple de mise en œuvre classique

La mise en œuvre du test de Student pour échantillons indépendants avec un tableur est simple et rapide (Figure 99).

Dans cet exemple, en partant d'un échantillon de 315 individus, nous souhaitons savoir si l'âge est indépendant du sexe, en termes de tendance centrale.

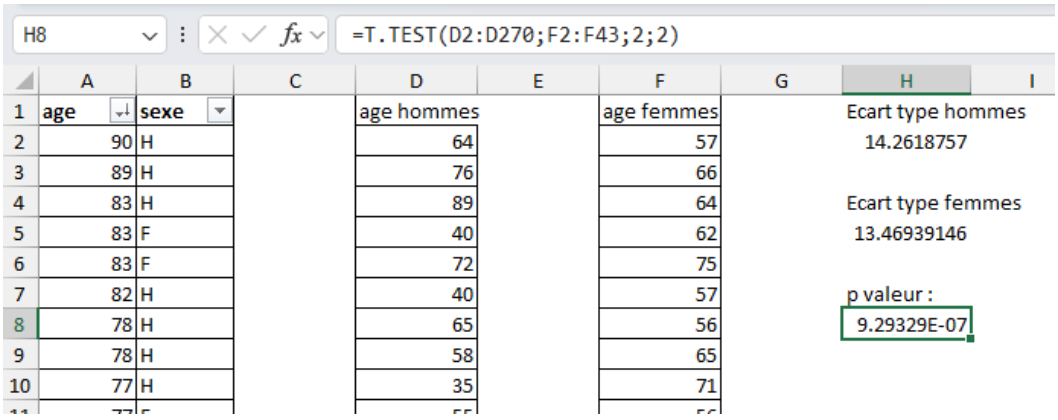
Nous posons **l'hypothèse nulle H_0** : la moyenne de l'âge ne dépend pas du sexe.

La première étape consiste à séparer les deux sous-groupes, et présenter chacun dans une colonne séparée (leurs longueurs sont donc souvent inégales ; voir colonnes D et F en Figure 99).

Ensuite, on calcule l'écart type dans chaque sous-groupe avec la fonction **stdeva()** de Microsoft Excel, ou **ecartype()** de LibreOffice Calc, ou **ecartype.standard()** (voir cellules H2 et H5 en Figure 99). Dans notre exemple on observe que ces deux valeurs sont très proches.

Enfin, on calcule directement la p valeur du test à l'aide de la formule **t.test()** de Microsoft Excel, ou **test.student()** de LibreOffice Calc. Cette formule, en H8 sur la Figure 99, est lisible dans la barre de formule. Elle comprend 4 paramètres :

- La plage des données du premier sous-groupe
- La plage des données du deuxième sous-groupe
- Le nombre « 2 » pour demander un test bilatéral
- Le nombre « 2 » pour le test de Student standard (homoscédasticité : écarts types proches) ou le nombre « 3 » pour appliquer la correction de Welch (hétéroscédasticité : écarts types très différents)



	A	B	C	D	E	F	G	H	I
1	age	sexe		age hommes		age femmes		Ecart type hommes	
2	90	H		64		57		14.2618757	
3	89	H		76		66			
4	83	H		89		64		Ecart type femmes	
5	83	F		40		62		13.46939146	
6	83	F		72		75			
7	82	H		40		57		p valeur :	
8	78	H		65		56		9.29329E-07	
9	78	H		58		65			
10	77	H		35		71			

Figure 99. Réalisation d'un test de Student (ici sans correction de Welch) avec Excel

Nous obtenons dans ce cas une p valeur inférieure à 5%, et pouvons conclure que l'âge moyen n'est pas indépendant du sexe. On peut tourner cette conclusion comme on le souhaite :

- L'âge et le sexe ne sont pas indépendants
- L'âge et le sexe sont associés statistiquement
- L'âge moyen est différent entre les hommes et les femmes
- La proportion d'hommes (ou de femmes) varie en fonction de l'âge
- Etc.

3.2.3.5.2 Exemple de mise en œuvre sur les rangs

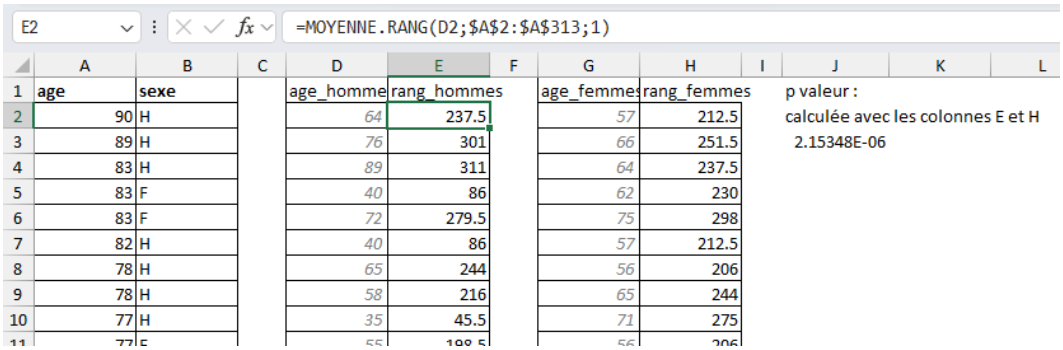
Si on souhaite réaliser un test de Student sur les rangs, le procédé est similaire, mais il faut créer une nouvelle variable décrivant le rang. On peut pour ce faire utiliser la fonction

moyenne.rang() d'Excel ou Calc, qui calcule le rang comme on le fait dans les procédures statistiques : si deux individus sont ex-aequo, on leur attribue à tous deux la moyenne des rangs qui leur auraient été attribués si on les avait classés de force. Exemple : si Marcel et Julien arrivent ex-aequo après le 2ème, ils sont 3ème et 4ème ex-aequo, dont on leur attribue le rang 3,5. Si vous utilisez une ancienne version de ces tableurs, la seule fonction disponible est la fonction **rang()** qui gère moins bien les ex-aequo, mais cela ne devrait pas avoir d'impact majeur sur votre analyse.

La fonction moyenne.rang() nécessite 3 paramètres :

- La cellule dont on veut calculer le rang
- La plage de toutes les valeurs. Avant d'étendre la formule, il faudra figer ses références à l'aide de dollars « $ »
- Un dernier paramètre indique si l'ordre des croissant ou décroissant. Ceci n'aura aucun impact sur la p valeur, mais devra être pris en compte dans l'interprétation des résultats

Dans l'exemple de la Figure 100, nous avons conservé toutes les valeurs de départ dans la colonne A, et avons recopié en colonne D les valeurs des hommes et en colonne G celles de femmes. La cellule E2, dont la formule est visible, illustre comment on retrouve le rang correspondant à la cellule D2, en partant de sa valeur et en la comparant **à l'ensemble de la colonne A originale**. En étendant cette formule, on obtient les rangs d'âges de tous les hommes. On fait de même pour les femmes en colonne H. On observe que certains rangs ne sont pas entiers. La p valeur calculée en J3 l'est comme précédemment, mais en utilisant les colonnes E et H (et non D et G). On notera que cette p valeur est ici plus importante (le test est moins puissant car on perd de l'information), mais reste inférieure à 5%. Lorsqu'on travaille sur des rangs, il n'est généralement pas nécessaire d'utiliser la correction de Welch car les écart types ont des valeurs proches.


E2 =MOYENNE.RANG(D2;$A$2:$A$313;1)

	A	B	C	D	E	F	G	H	I	J	K	L
1	age	sexe		age_homme	rang_hommes		age_femmes	rang_femmes		p valeur :		
2	90	H		64	237.5		57	212.5		calculée avec les colonnes E et H		
3	89	H		76	301		66	251.5		2.15348E-06		
4	83	H		89	311		64	237.5				
5	83	F		40	86		62	230				
6	83	F		72	279.5		75	298				
7	82	H		40	86		57	212.5				
8	78	H		65	244		56	206				
9	78	H		58	216		65	244				
10	77	H		35	45.5		71	275				

Figure 100. Réalisation d'un test de Student sur les rangs avec un tableur

3.2.3.5.3 Le test de Student pour échantillons indépendants, en général

Revenons sur le déroulement d'un test de Student dit « pour deux échantillons indépendants », avec quelques formules, mais sans trop de détails.

Nous nous intéressons à deux variables :

- « X » une variable quantitative
- « A » une variable binaire, qui détermine deux sous-groupes, notés 1 et 2

Nous posons l'**hypothèse nulle H_0**, qui peut être formulée de différentes manières :

- la moyenne de X ne dépend pas de la valeur de A
- $\mu_{x1} = \mu_{x2}$ où x_1 représente la distribution X pour une modalité de A, et x_2 représente la distribution de X pour l'autre modalité de A

Nous observons l'effectif, la moyenne et la déviation standard dans chaque sous-groupe : respectivement : $n_1, \bar{x}_1, SD_1$ dans le premier groupe, et $n_2, \bar{x}_2, SD_2$ dans le deuxième groupe.

Nous vérifions la condition de validité du test de Student :

$$\begin{pmatrix} (n_1 \geq 30 \; et \; n_2 \geq 30) \\ ou \; (X_1 \; d'allure \; symétrique \; et \; X_2 \; d'allure \; symétrique) \end{pmatrix} et \; (SD_1 \approx SD_2)$$

Équation 27. Condition de validité du test de Student sans correction de Welch

La troisième condition inclut un ratio des écarts types compris entre 0,7 et 1,4 (ou ratio des variances compris entre 0,5 et 2). Si elle n'est pas remplie, nous appliquerons la correction de Welch, détaillée par la suite.

Nous calculons la variance sur l'ensemble de l'échantillon :

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Équation 28. Calcul de la variance poolée

Nous calculons ensuite la statistique de test *t* et le nombre de degrés de liberté *v* (« nu » ddl) :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{s.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$v = n_1 + n_2 - 2$$

Équation 29. Calcul de la statistique de test t sans correction de Welch, et du nombre de ddl

Sous l'hypothèse H_0, la quantité *t* devrait suivre une loi de Student à *v* degrés de liberté. Concrètement, la quantité *t* devrait tourner autour de zéro, sans trop s'en écarter. La Loi de Student ayant été décrite, cela permet de calculer la p valeur associée.

Nous pouvons conclure :

- Fixons le seuil d'interprétation de la p valeur à 5% (risque α)
- Si **p valeur < 5% : on rejette H_0**, on peut conclure que les deux variables sont statistiquement liées (elles ne sont pas indépendantes)
- Si **p valeur > 5%** : on ne rejette pas H_0 : on se trouve face à une **indétermination**, il est interdit de conclure.

Si cependant la condition d'homoscédasticité n'est pas remplie, il est possible de réaliser le test de Student avec la correction de Welch, ou correction de Welch-Satterthwaite [14].

Nous vérifions la condition de validité du test de Student avec correction de Welch :

$$(n_1 \geq 30 \; et \; n_2 \geq 30)$$
$$ou \; (X_1 \; d'allure \; symétrique \; et \; X_2 \; d'allure \; symétrique)$$

Équation 30. Condition de validité du test de Student avec correction de Welch

Nous conserverons les deux variances estimées dans les deux sous-groupes.

Nous calculons ensuite la statistique de test *t* et le nombre de degrés de liberté *v* (« nu » ddl). On notera que la quantité *t* tient désormais compte des deux variances, mais que le nombre de degrés de liberté du test est calculé de manière plus complexe. Ce nombre de degrés de liberté est désormais potentiellement un nombre non-entier :

[14] Parfois appelé test de Student-Welch, ou test de Welch

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Équation 31. Calcul de la statistique de test t avec correction de Welch, et du nombre de ddl

Le reste du test se déroule de la même manière, et l'interprétation est la même.

Cette correction est toujours valide, même si les variances sont égales. Elle engendre dans ce cas une perte de puissance. On notera que la correction de Welch est activée par défaut dans le logiciel R lorsqu'on exécute un test de Student avec la fonction t.test().

3.2.3.5.4 Quelques précisions sur le test de Student pour échantillons indépendants

Le test de Student pour deux échantillons indépendants (avec ou sans correction de Welch) peut être classé ainsi :

- C'est un **test paramétrique** car il utilise l'estimation de la moyenne et de l'écart type, qui sont des paramètres
- C'est un **test asymptotique** : la p valeur est calculée en se référant à une distribution théorique, qui n'est vraie que si certains effectifs sont suffisants, ou si la variable initiale suit une loi particulière

En cas d'égalité des variances, on comprend de la formule précédente (Équation 29) que le test se fonde essentiellement sur la quantité $|(\bar{x}_1 - \bar{x}_2)/DS|$ (où DS est l'écart type commun aux deux sous-groupes). Cette quantité répond à la question « de combien d'écarts-types les deux moyennes sont-elles écartées ? ». Nous l'appellerons « écart standardisé » ci-après.

Par exemple, si vous observez dans un groupe une moyenne de 8 et dans l'autre une moyenne de 10, avec un écart type de 1 dans les deux sous-groupes, alors on peut affirmer que les deux moyennes s'écartent de deux écarts-types (en valeur absolue). L'écart standardisé vaut 2.

Par la suite, la significativité du test dépend également de l'effectif de l'échantillon, d'une part parce que le terme $\sqrt{n}$ multiplie la statistique de test, d'autre part parce que l'effectif détermine le nombre de degrés de liberté, qui modifie la valeur limite dans la table de Student. Ainsi, la Figure 101 présente l'écart standardisé correspondant à une p valeur de 5%, en fonction de l'effectif total de l'échantillon. Nous l'appellerons ici « écart standardisé limite » (ce terme n'est pas officiel). Si, pour un effectif donné, vous obtenez un écart standardisé supérieur à celui de cette courbe, votre test sera significatif avec p<5%. La courbe pleine représente le cas où les deux sous-groupes comportent chacun la moitié des individus, les autres courbes représentent des cas de déséquilibres. Ces courbes doivent être comparées en termes d'écart vertical (pour un effectif donné, comment l'écart standardisé limite est-il affecté ?). On observe que le déséquilibre entre les deux sous-groupes affecte la puissance du test, de manière importante pour les petits effectifs, et modérée par la suite.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

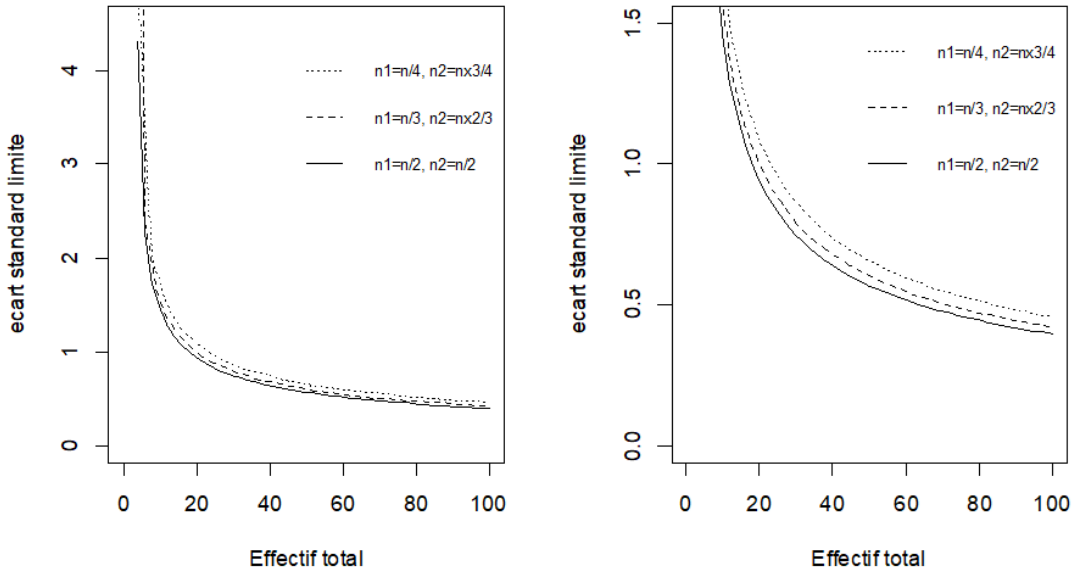


Figure 101. Limites de significativité du test de Student bivarié (à droite : zoom) : valeur de $\left|\frac{\bar{x}_1-\bar{x}_2}{DS}\right|$ permettant d'obtenir p=5%, en fonction de la taille totale de l'échantillon lorsque les variances sont égales. De bas en haut : effectifs 1/2+1/2, puis 1/3+2/3, puis 1/4+3/4

On peut donc utiliser la Figure 80 pour réaliser graphiquement le test : il suffit de positionner un seul point, correspondant au couple $\left\{x = n_{total}\ ;\ y = \left|\frac{\bar{x}_1-\bar{x}_2}{DS}\right|\right\}$. Si ce point est au-dessus des courbes, on rejette H_0. Nous ne présenterons pas de tableau de valeur, car elles dépendent de l'équilibre des effectifs.

En cas d'inégalité des variances, la correction de Welch doit être appliquée. De la même manière que précédemment, la Figure 102 montre la limite de significativité du test de Student avec correction de Welch, en représentant l'écart standardisé limite en fonction de l'effectif total *n*, dans le cas uniquement où les deux sous-groupes ont le même effectif (*n/2* dans chaque sous-groupe). Cet écart standardisé limite est calculé avec l'écart type de tout l'échantillon. Les différentes courbes montrent différents ratios d'écart types entre les deux sous-groupes. On observe que le déséquilibre entre les deux écarts types a un impact modéré pour les effectifs faibles, et quasi-inexistant pour les effectifs élevés.

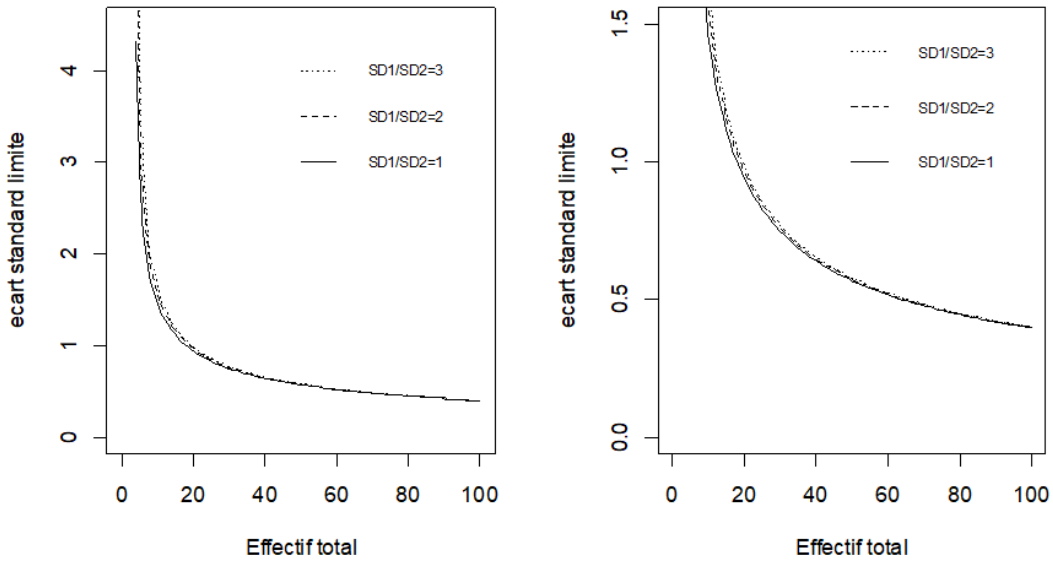


Figure 102. Limites de significativité du test de Student avec correction de Welch (à droite : zoom) : valeur de $\left|\frac{\bar{x}_1-\bar{x}_2}{DS}\right|$ permettant d'obtenir p=5%, en fonction de la taille totale de l'échantillon lorsque les effectifs sont équilibrés. De bas en haut : ratio des écarts types à 1, 2 ou 3

En superposant la courbe pleine de la Figure 101 et la courbe pleine de la Figure 102, on s'intéresse au cas où les deux écarts types sont égaux, et les deux effectifs équilibrés. Ces courbes sont quasiment superposées, illustrant le fait que la correction de Welch, si elle est appliquée alors qu'elle est inutile, diminue faiblement la puissance statistique du test. Cela explique que, dans le logiciel R, elle soit activée par défaut dans la fonction t.test().

3.2.3.6 Autres tests statistiques

D'autres tests sont classiquement décrits, mais vous ne pourrez pas les réaliser avec un tableur, ou n'aurez aucune raison d'y recourir. Les voici, succinctement.

3.2.3.6.1 Comparaison de deux moyennes avec la loi normale

Il s'agit d'un test paramétrique et asymptotique. On enseigne classiquement qu'il est possible de comparer deux moyennes en utilisant un test basé sur la loi normale. Ce test suppose de connaître la variance de chaque sous-groupe en population : évidemment ce n'est jamais le cas dans une analyse statistique, car la variance est toujours estimée dans l'échantillon, et est donc elle-même sujette à erreur. Le test de Student a été imaginé justement pour pallier ce problème. Pour cette raison, ce test n'est pas utilisé en pratique.

3.2.3.6.2 Analyse de la variance, ANOVA

L'analyse de la variance ou ANOVA est un test paramétrique et asymptotique. Elle peut être perçue comme une généralisation du test de Student à plus de deux sous-groupes. Elle s'interprète de la même manière. Elle ne peut hélas pas être réalisée avec un tableur. Vous pourrez solliciter un statisticien, ou contourner ce problème soit en regroupant des modalités, soit en réalisant des comparaisons deux-à-deux. Il faudra alors appliquer une correction de Bonferroni.

3.2.3.6.3 Test de Wilcoxon-Mann-Whitney

Le Test de Wilcoxon-Mann-Whitney est un test non-paramétrique et exact. Il analyse non pas les valeurs natives de la variable quantitative, mais les rangs correspondants. Pour cette

raison, il n'est pas altéré par les valeurs extrêmes, et ne requiert donc aucune condition de validité. C'est un excellent test, qui peut être réalisé avec un tableur, pas directement mais au prix de quelques manipulations. Cependant, il a été montré que le test de Student, s'il était réalisé sur les rangs et non sur les valeurs, donnait des résultats très proches[37]. Vous n'en aurez donc pas besoin.

3.2.3.6.4 Test de Kruskal-Wallis

Le test de Kruskal-Wallis est un test non-paramétrique et exact. Lui aussi analyse les rangs et non les valeurs natives de la variable quantitative. Pour cette raison, il n'est pas altéré par les valeurs extrêmes, et ne requiert donc aucune condition de validité. C'est un excellent test, que vous ne pourrez hélas pas réaliser avec un tableur. Pour contourner cela, vous pourrez réaliser un test de Student sur les rangs, après avoir regroupé des modalités, ou plusieurs tests de Student sur les rangs, en appliquant la correction de Bonferroni.

3.2.3.6.5 Régression logistique simple

Il est théoriquement possible d'étudier la relation entre une variable quantitative et une variable binaire avec une régression logistique simple. Il s'agit d'un test paramétrique et asymptotique. L'esprit d'un tel test serait, non pas de comparer les moyennes de X en fonction de la valeur de A, mais plutôt de visualiser la proportion de A=1 en fonction des valeurs de X. En pratique, personne ne fait cela.

3.2.3.6.6 Tests post-hoc

Lorsqu'on réalise une ANOVA ou un test de Kruskal-Wallis pour comparer 3 tendances centrales ou plus, si le test revient significatif, cela signifie qu'au moins une tendance centrale est significativement différente des autres. Cela n'indique pas laquelle ou lesquelles. Certains se contenteront de le constater visuellement, d'autres voudront réaliser des tests statistiques deux-à-deux pour mieux décrire cette différence. On appelle « tests post-hoc » ces tests statistiques. Il peut s'agir de tests traditionnels comparant deux moyennes (Student, Wilcoxon-Mann-Whitney), ou d'autres tests moins connus. Ils seront interprétés avec une correction de Bonferroni : non pas au seuil 5%, mais au seuil de 5/k%, k étant le nombre de tests réalisés. Si vous réalisez d'emblée des comparaisons par couples des modalités, vous n'aurez bien sûr pas besoin de tels tests.

3.2.4 Deux variables quantitatives

3.2.4.1 Préambule

Comme nous l'avons énoncé précédemment (voir 3.2.2.1 Préambule en page 153), nous nous intéresserons, au fond, à l'indépendance entre deux variables quantitatives, que nous pourrons explorer graphiquement. En revanche, lorsqu'il s'agira de calculer des paramètres ou de réaliser des tests statistiques, nous nous contenterons de **mettre en évidence une relation linéaire** entre ces deux variables : est-ce que, globalement, les deux variables croissent en même temps, ou en sens inverse ? Au sein de l'ensemble de la relation qui existe entre les deux variables, ces méthodes seront capables d'extraire, quantifier et tester **la part de relation linéaire** qui existe.

Rechercher la part de relation linéaire qui existe entre deux variables quantitatives X et Y revient à se poser plusieurs questions, similaires :

- Est-ce que la tendance centrale (la moyenne, le plus souvent) de Y augmente ou diminue lorsque X augmente ?
- Est-ce que la tendance centrale (la moyenne, le plus souvent) de X augmente ou diminue lorsque Y augmente ?
- Est-ce que le fait de connaître la valeur de X permet de mieux prédire la valeur inconnue de Y (valeur différente, et marge d'erreur diminuée) ?

- Est-ce que le fait de connaître la valeur de Y permet de mieux prédire la valeur inconnue de X (valeur différente, et marge d'erreur diminuée) ?
- Peut-on résumer la relation entre X et Y à l'aide d'une équation $Y = aX + b$?

Certaines méthodes considèreront les vraies valeurs de X et Y (corrélation de Pearson, régression), tandis que d'autres s'intéresseront aux rangs de X et Y (corrélation de Spearman), rendant plutôt compte de la tendance croissante ou décroissante, sans attester d'une linéarité à proprement parler.

Le sens de la relation X→Y ou Y→X n'a aucune importance d'un point de vue statistique, mais doit rester cohérente en termes d'interprétation. Certaines méthodes sont symétriques (ex : le coefficient de corrélation), mais d'autres nécessitent d'exprimer une variable en fonction de l'autre (ex : la régression linéaire simple). Par convention, nous exprimerons **Y en fonction de X**. Voici, par ordre de préférence décroissant, des arguments pour choisir quelle variable sera Y et laquelle sera X :

- Y est la conséquence présumée de X
- X préexiste à Y, dans la physiopathologie des individus étudiés
- On cherchera ensuite à prédire Y en connaissant X (par exemple parce que Y est plus difficile à mesurer, ou connue plus tard)
- En l'absence d'argument clair, tout argument de bon sens est valable

Voici un exemple. Soit la variable « taille » (en cm), et la variable « poids » (en kg). Nous nommerons la taille X, et le poids Y, car de toute évidence, c'est parce qu'un individu est grand qu'il est lourd, et non l'inverse. Si nous hésitions, cela n'affecterait aucunement la validité des méthodes employées mais seulement l'interprétation de leurs résultats.

Ces variables ne sont pas indépendantes, et ont une part de relation linéaire croissante, car dans une population ou un échantillon :

- *Le poids devient en moyenne plus important lorsque la taille augmente (1)*
- *La taille devient en moyenne plus importante lorsque le poids augmente (2)*
- *Le fait d'avoir une petite taille augmente la probabilité d'être léger (3)*
- *Le fait d'être léger augmente la probabilité d'avoir une petite taille (4)*

Toutes les affirmations ci-dessus sont vraies. Cependant, pour des raisons de causalité mais surtout de simplicité de la formulation, on présentera plus volontiers les affirmations (1) et (3), et on tracera des graphiques dans ce sens.

Nous garderons à l'esprit que **l'absence de relation linéaire n'est pas équivalente à l'indépendance** statistique :

- X et Y sont en relation linéaire ⇒ X et Y ne sont pas indépendants
- X et Y ne sont pas en relation linéaire ⇒ on ne sait pas si X et Y sont indépendants
- X et Y sont indépendants ⇒ X et Y ne sont pas en relation linéaire
- X et Y ne sont pas indépendants ⇒ on ne sait pas si X et Y sont en relation linéaire

Nous vous renvoyons à la Figure 88 en page 153 pour illustrer cette non-équivalence.

3.2.4.2 Représentation graphique

Lorsque X et Y sont des variables continues (donc théoriquement sans ex aequo), la représentation graphique par excellence est le nuage de points, dans lequel chaque individu est représenté par une marque individuelle (à l'inverse des autres graphiques que nous avons vus précédemment, où les individus étaient regroupés). Ce graphique peut être tracé sans aucun a priori, et permet de visualiser immédiatement le type de relation géométrique entre les deux variables. La Figure 103 illustre cela. La proposition de droite est plus lisible, car les pictogrammes utilisés sont des cercles vides. Ils peuvent paraître disgracieux, mais sont très appréciés car ils permettent de mieux percevoir la densité lorsque de nombreux individus sont rassemblés sur une petite surface. Les tableurs permettent également d'ajouter en un clic une

droite de tendance linéaire (obtenue par régression linéaire simple, méthode que nous expliquerons par la suite).

Si une des deux variables est une variable discrète, le nuage de points peut encore être utile (Figure 104 à gauche), mais le nombre de superpositions de pictogrammes peut induire en erreur. Dans ce cas, on préfèrera une série de boxplots (Figure 104 à droite), en s'assurant cependant que l'axe représentant la variable discrète soit réellement quantitatif. Dans cet exemple, nous aurions préféré représenter le nombre d'enfants en Y et l'âge en X, car le premier est la conséquence du second, mais le tableur utilisé ne permet pas de dessiner des boxplots horizontales.

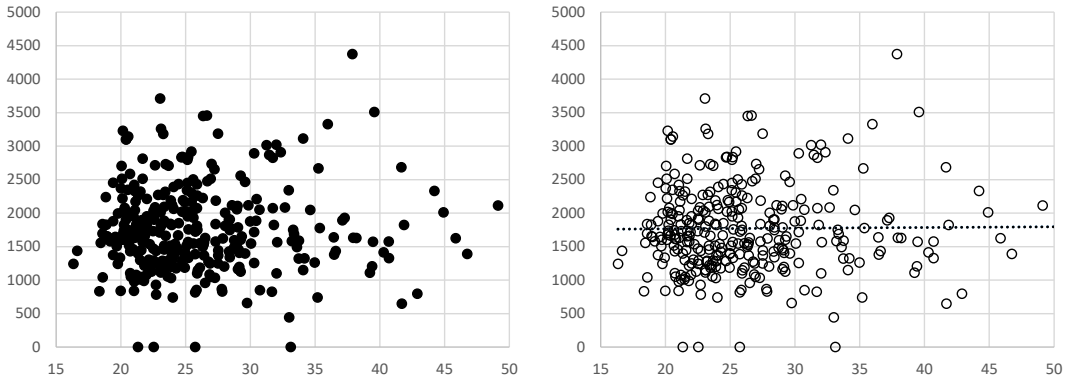


Figure 103. Exemple de nuage de points.
Gauche : disques pleins. Droite : cercles, avec droite de tendance linéaire.

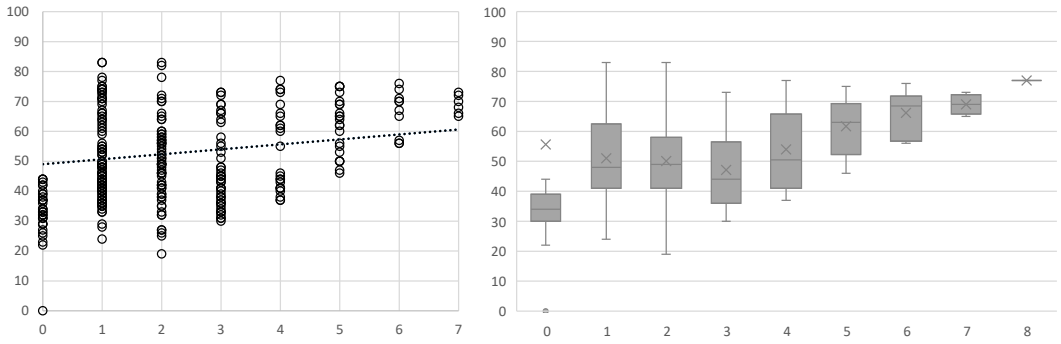
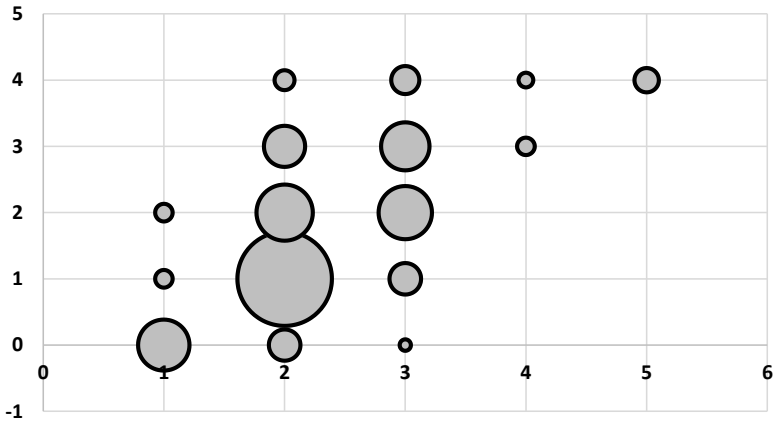


Figure 104. Age (Y) en fonction du nombre d'enfants (X).
Gauche : nuage de points. Droite : boxplots

Si les deux variables sont discrètes, les superpositions de pictogrammes rendront le graphique ininterprétable. Il faudra recourir à un graphique représentant l'effectif en surface, pour chaque couple de valeurs discrètes X et Y. Le **graphique en bulles** (l'effectif étant proportionnel à la surface mais surtout pas au diamètre) est particulièrement adapté (voir Figure 105). Pour réaliser un tel graphique, on trace tout d'abord un tableau croisé de contingence, à l'aide d'un tableau croisé dynamique. Ce tableau devra être modifié comme suit :

- Propriétés du tableau : « disposition classique », pas de sous-total ni de total
- Propriétés du champ qui apparait le plus à gauche : « répéter les étiquettes d'éléments », pas de sous-total

Une fois le tableau bien présenté (à droite sur la Figure 105), la réalisation du graphique en bulles est immédiate. Dans ce graphique, les axes X et Y sont réellement quantitatifs, ce qui nous convient parfaitement.



Nb enfants	Nb chambres	Effectif
0	1	27
0	2	11
0	3	2
1	1	4
1	2	86
1	3	11
2	1	4
2	2	32
2	3	29
3	2	18
3	3	24
3	4	4
4	2	5
4	3	9
4	4	3
4	5	7
5	4	14
5	5	8
6	5	4
6	6	4
6	7	2
7	6	3
7	7	3
8	6	1

Figure 105. Exemple de graphique en bulles (données à droite).
X=nombre d'enfants du foyer. Y=nombre de chambres du domicile

Dans tous les cas, la représentation graphique est **essentielle pour comprendre** la forme de relation qui unit les deux variables, comme nous l'avions vu en Figure 88 en page 153.

3.2.4.3 Description par des paramètres

Dans certains cas de variables discrètes, lorsque cela est parlant pour le lecteur, rien n'interdit de procéder à une description par catégorie, comme nous l'avions vu précédemment. Le bon sens doit primer.

L'indicateur de choix reste le **coefficient de corrélation linéaire de Pearson**, ou le **coefficient de corrélation des rangs de Spearman**. L'équation de droite de la régression linéaire, que nous introduirons plus bas, peut parfois être produite à ce stade, mais ne renseigne pas sur la force de l'association statistique entre X et Y. Nous verrons plus bas lequel de ces deux coefficients choisir, et comment le calculer.

Ces coefficients de corrélation (Pearson ou Spearman) sont des nombres sans unité, toujours compris entre -1 et +1, qui s'interprètent ainsi (nous utiliserons des termes plus précis par la suite, car il existe des nuances entre les deux coefficients) :

- En cas d'association parfaite croissante ou décroissante, il vaut exactement -1 ou +1
- En cas d'absence totale d'association monotone, il vaut 0
- Les valeurs négatives indiquent une relation décroissante (quand X augmente, Y diminue)
- Les valeurs positives indiquent une relation croissante (quand X augmente, Y augmente)
- La corrélation est forte au-delà de 0,5 en valeur absolue, et faible en-deçà

On peut synthétiser ces différentes interprétations, par valeurs croissantes du coefficient de corrélation *r* de Pearson (de même pour r_s de Spearman) :

- -1 association parfaite et décroissante
- $]-1;-0{,}5]$ association décroissante, forte
- $]-0{,}5;0[$ association décroissante, faible
- 0 absence d'association monotone (pas forcément indépendance statistique)
- $]0;0{,}5[$ association croissante, faible
- $[0{,}5;1[$ association croissante, forte
- 1 association parfaite et croissante

Bien évidemment, comme ce coefficient est estimé sur un échantillon (on parle de **coefficient de corrélation empirique**), la confiance qu'on aura dans cette estimation ne sera jamais totale, mais sera d'autant plus grande que l'échantillon sera de grande taille. Les bornes

proposées ici ne doivent donc pas être interprétées strictement (en particulier aux alentours de zéro…). Seuls les tests statistiques décrits plus bas permettront de préciser cette interprétation. Nous reverrons en détail l'interprétation du coefficient de corrélation dans le chapitre 3.2.4.7.3 Quelques précisions, synthèse de l'interprétation en page 180.

La description sera ainsi très simple (nous verrons par la suite comment calculer la p valeur, ou petit p).

Exemple : « Dans notre échantillon, le poids et la taille sont corrélés positivement et fortement (r=0,77, p=0,012). »

3.2.4.4 Paramètres et tests statistiques : arbre décisionnel

Une fois encore, l'arbre décisionnel que nous proposerons ici sera plus simple et pragmatique que celui traditionnellement décrit dans les cours de statistiques. Il est présenté en Figure 106.

La première opération sera toujours de réaliser une **représentation graphique** de la relation X-Y, par exemple avec un nuage de points augmenté d'une droite de régression (comme vu précédemment). Dans la plupart des cas, on pourra calculer le coefficient de corrélation de **Pearson**. Cependant, si on voit apparaître quelques individus avec des **valeurs extrêmes** et qui influent la droite de régression, il faudra calculer le coefficient de **Spearman** à la place. Ce coefficient ne pourra être calculé que si les variables X et Y sont clairement **continues**. En cas de variable discrètes, il vaudra mieux changer de stratégie et, quitte à regrouper encore des modalités, traiter au moins une des deux variables comme une variable qualitative. Dans tous les cas, si on peut calculer le coefficient de Spearman ou celui de Pearson, on pourra ensuite réaliser un **test de nullité** de ce coefficient. Ce test est le même dans les deux cas. Enfin, lorsqu'il est possible de calculer le coefficient de Pearson, on peut compléter la description à l'aide d'une régression linéaire simple, qui nous fournira l'**équation de la droite** de régression, qui peut être un élément descriptif intéressant, et qui permet également de prédire des valeurs inconnues de *y*. Celle-ci n'a pas de sens si c'est le coefficient de Spearman qui a été calculé.

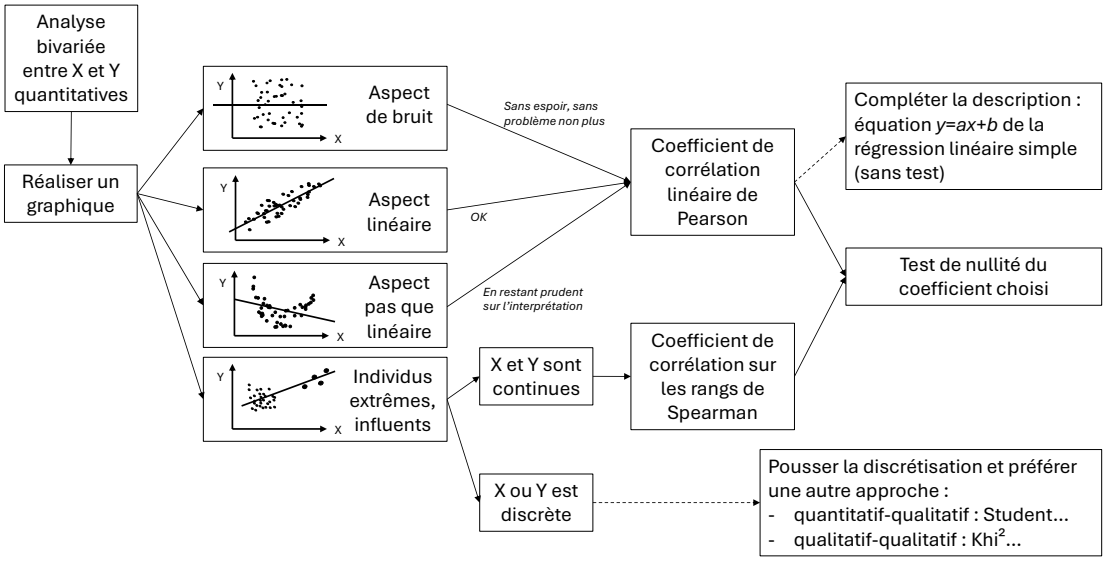


Figure 106. Arbre décisionnel : analyse bivariée de deux variables quantitatives

Nous discuterons ici les motivations de cet arbre décisionnel, au regard de ce qu'on lit habituellement. La question centrale est celle de la validité du coefficient de Pearson, et de son test statistique. Les cours traditionnels partent du principe qu'on souhaite calculer ce coefficient pour décrire exhaustivement la relation entre X et Y, et posent donc comme prérequis que la relation entre X et Y soit réellement linéaire, ou totalement bruité, à l'exclusion de toutes autres relations. On peut lire par exemple que :

- X et Y devraient suivre une loi normale : c'est faux, car une relation linéaire peut exister entre deux distributions pourtant quelconques
- L'effectif devrait être supérieur à un certain seuil : c'est faux, tout simplement
- Les résidus de la régression linéaire correspondante devraient être normalement distribués, et indépendamment de X : cette condition devient superflue si on change de cadre

Si on part du principe que le **coefficient de corrélation de Pearson permet de quantifier la part linéaire de l'association entre X et Y**, non plus de décrire exhaustivement leur relation, alors aucune condition de validité n'est requise, hormis une condition technique : X et Y ne doivent pas être des constantes. En revanche, il est indéniable que ce coefficient est susceptible d'être influencé par des **valeurs extrêmes**, ce que le nuage de points permettra de visualiser. Dans ce cas on lui préfèrera le coefficient de Spearman, calculé sur les rangs. Ce coefficient lui-même sera mis en difficultés pour les variables discrètes, en raison de la manière dont les rangs sont calculés : autour des ex-aequo, ce mode de calcul fait apparaître des sauts trop importants, d'une manière arbitraire qui ne rend pas compte de la réalité des données. Il n'y a pas de critère simple pour juger ce problème. Si le caractère discret semble problématique, il vaudra mieux poursuivre la discrétisation pour outrepasser ce problème. On peut par exemple discrétiser plus encore la variable discrète, pour aboutir à 2 ou 3 modalités, et réaliser un test de Student ou une ANOVA. Si les deux variables, X et Y, sont réduites à 2 ou 3 modalités, on pourra réaliser un test du Khi².

3.2.4.5 Coefficient de corrélation linéaire de Pearson

3.2.4.5.1 Exemple de mise en œuvre

Pour calculer le coefficient de corrélation linéaire de Pearson dans Microsoft Excel ou LibreOffice Calc, après avoir réalisé un graphique, il suffit d'utiliser la fonction **coefficient.correlation()**, ou la fonction **pearson()** comme illustré en Figure 107. Cette formule prend en premier argument la plage occupée par une des variables, et en deuxième argument la plage occupée par la deuxième variable. L'ordre des variables n'a aucune importance. Les deux plages doivent être issues d'un même tableau où chaque individu figure sur une ligne. C'est tout !

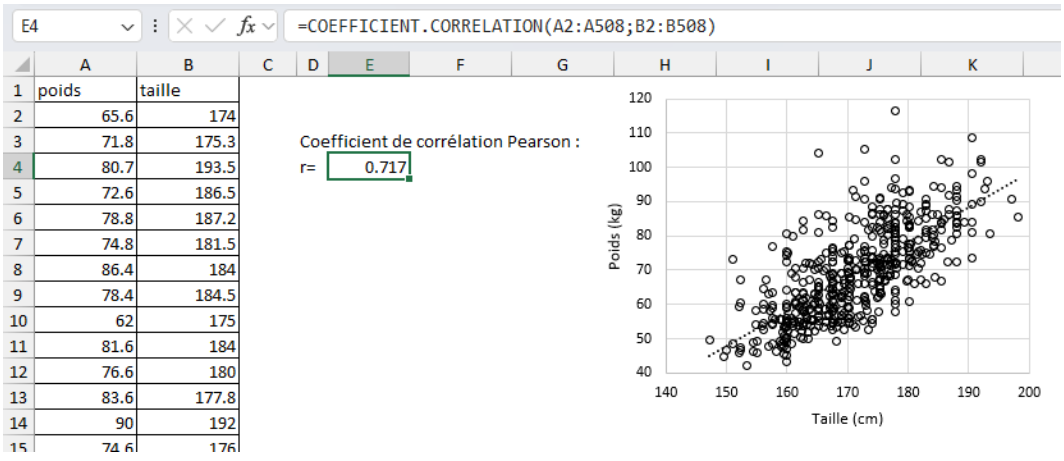


Figure 107. Calcul du coefficient de corrélation linéaire de Pearson avec un tableur

3.2.4.5.2 Le calcul du coefficient de corrélation linéaire de Pearson, en détail

Le calcul du coefficient de corrélation de Pearson se fait en plusieurs étapes. Tout d'abord, pour chaque individu, on peut calculer la quantité $(x_i - \bar{x}).(y_i - \bar{y})$. Par exemple, si Y est le poids et X la taille, cette valeur est le produit de l'écart entre le poids d'un individu et le poids moyen, et de l'écart entre la taille de ce même individu et la taille moyenne. Cette quantité, comme l'illustre la Figure 108, est :

- nulle pour les individus qui sont au niveau de la moyenne pour X ou Y
- fortement positive pour :
  - ceux qui ont des valeurs nettement supérieures à la moyenne en même temps pour X et Y
  - ceux qui ont des valeurs nettement inférieures à la moyenne en même temps pour X et Y
- fortement négative pour :
  - ceux qui ont des valeurs nettement supérieures à la moyenne pour X et des valeurs nettement inférieures à la moyenne pour Y
  - ceux qui ont des valeurs nettement supérieures à la moyenne pour Y et des valeurs nettement inférieures à la moyenne pour X

Plus simplement, cette quantité est positive pour les individus qui rendent compte d'une relation croissante, et négative pour les individus qui rendent compte d'une relation décroissante.

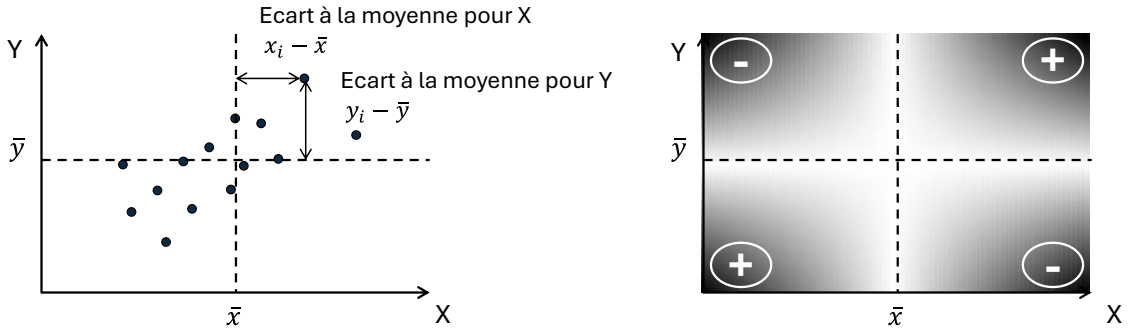


Figure 108. Comportement du produit $(x_i - \bar{x}).(y_i - \bar{y})$

On définit la covariance dans l'échantillon, comme étant la moyenne arithmétique de la quantité précédemment établie :

$$cov_{ech}(X,Y) = (\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y}))/n$$

Équation 32. Covariance calculée dans un échantillon

Cette covariance est un nombre dont l'unité est le produit de l'unité de X par celle de Y. Elle est proche de zéro dans les situations d'indépendance, fortement positive en cas de relation clairement croissante, et fortement négative en cas de relation clairement décroissante. Cependant, son ordre de grandeur dépend de l'ordre de grandeur des variables X et Y. Ainsi, un simple changement d'unité de X ou Y affecte d'autant la valeur de la covariance, alors que, en vrai, la force de l'association entre X et Y reste identique. Dans l'absolu par exemple, personne ne peut dire si une covariance de 5000 est forte ou faible : on sait seulement qu'elle est positive.

Afin de standardiser, la covariance, nous la divisons par le produit de l'écart type de X et de l'écart type de Y (nous utilisons leurs estimateurs biaisés). Ainsi, nous obtenons le coefficient de corrélation empirique de Pearson.

$$r = \frac{\mathrm{cov}_{ech}(X,Y)}{s_{ech_X}.s_{ech_Y}}$$

$$r \in [-1\,;1]$$

Équation 33. Coefficient de corrélation linéaire de Pearson
(« coefficient empirique de corrélation », car calculé dans l'échantillon)

Ce nombre est sans unité, toujours compris entre 1 et -1, et s'interprète comme indiqué dans le chapitre 3.2.4.3 Description par des paramètres en page 173 :

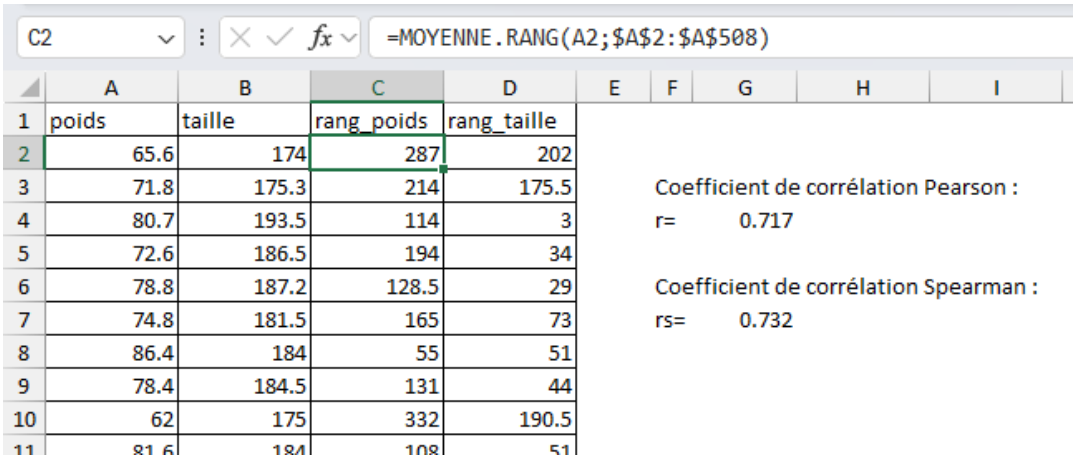
- $r = -1$ association strictement linéaire et décroissante
- $r \in]-1; -0{,}5]$ association linéaire décroissante, forte
- $r \in]-0{,}5; 0[$ association linéaire décroissante, faible
- $r = 0$ absence d'association linéaire (pas forcément indépendance statistique)
- $r \in]0; 0{,}5[$ association linéaire croissante, faible
- $r \in [0{,}5; 1[$ association linéaire croissante, forte
- $r = 1$ association strictement linéaire et croissante

Le calcul de ce coefficient est techniquement possible avec seulement deux individus, pourvu que leurs valeurs de X et Y diffèrent, et vaut alors 1 ou -1, ce qui n'a donc aucun intérêt.

3.2.4.6 Coefficient de corrélation des rangs de Spearman

3.2.4.6.1 Exemple de mise en œuvre

C'est très simple : le coefficient de **Spearman** est simplement un coefficient de **Pearson** qui, au lieu d'être calculé sur les valeurs initiales de X et Y, est calculé sur les **rangs de X** et les **rangs de Y**. Pour le calculer (Figure 109), on ajoute dans le même tableau une colonne des rangs de X et une colonne des rangs de Y, à l'aide de la fonction **moyenne.rang()**. Si votre tableur ne gère pas cette fonction mais une fonction de rangs plus basique, si les ex-aequo sont rares, l'impact devrait être faible et cela ne doit pas vous empêcher de procéder. Cette fonction de rangs a été détaillée dans le chapitre 3.2.3.5.2 Exemple de mise en œuvre sur les rangs en page 164. Il suffit ensuite d'utiliser la fonction **coefficient.correlation()**, ou la fonction **pearson()** comme vu précédemment.



	A	B	C	D
1	poids	taille	rang_poids	rang_taille
2	65.6	174	287	202
3	71.8	175.3	214	175.5
4	80.7	193.5	114	3
5	72.6	186.5	194	34
6	78.8	187.2	128.5	29
7	74.8	181.5	165	73
8	86.4	184	55	51
9	78.4	184.5	131	44
10	62	175	332	190.5
11	81.6	184	108	51

Figure 109. Calcul du coefficient de corrélation des rangs de Spearman

3.2.4.6.2 Le calcul du coefficient de corrélation des rangs de Spearman, en détail

Comme nous l'avons indiqué plus haut, le coefficient de corrélation de Spearman est simplement un coefficient de corrélation de Pearson calculé sur les rangs de X et les rangs de Y.

Spearman a démontré que, en l'absence d'ex-aequo, une formule nettement plus simple pouvait être utilisée :

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

$$r_s = 1 - \frac{6.\sum {d_i}^2}{n(n^2-1)}$$

$$d_i = rang_{yi} - rang_{xi}$$

Équation 34. Coefficient de corrélation des rangs de Spearman, en l'absence d'ex-aequo

Cette formule ne vous sera probablement pas utile, car nous avons vu plus haut comment calculer ce coefficient, même en présence d'ex-aequo.

Ce nombre est sans unité, toujours compris entre 1 et -1, et s'interprète comme indiqué dans le chapitre 3.2.4.3 Description par des paramètres en page 173 :

- $r_s = -1$ association strictement décroissante
- $r_s \in]-1; -0{,}5]$ association décroissante, forte
- $r_s \in]-0{,}5; 0[$ association décroissante, faible
- $r_s = 0$ absence d'association monotone (pas forcément indépendance statistique)
- $r_s \in]0; 0{,}5[$ association croissante, faible
- $r_s \in [0{,}5; 1[$ association croissante, forte
- $r_s = 1$ association strictement croissante

Le calcul de ce coefficient est techniquement possible avec seulement deux individus, pourvu que leurs valeurs de X et Y diffèrent, et vaut alors 1 ou -1, ce qui n'a donc aucun intérêt.

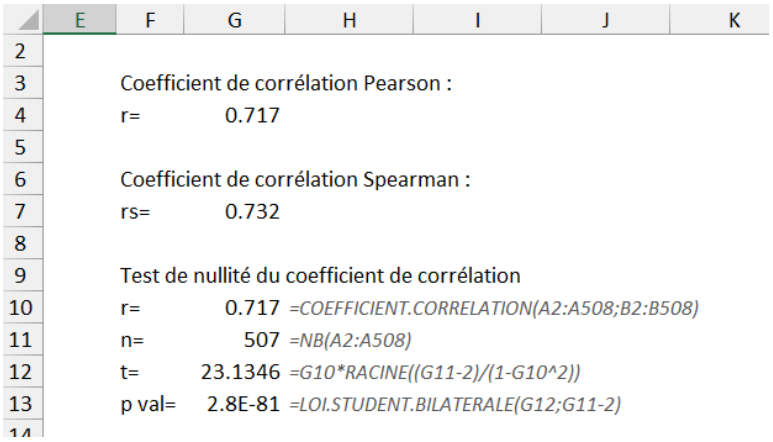
3.2.4.7 Test de nullité du coefficient de corrélation (Pearson ou Spearman), interprétation complète du coefficient

3.2.4.7.1 Exemple de mise en œuvre

Prenons la suite de l'exemple précédent. Nous avions calculé le coefficient de corrélation de Pearson (cellule G4 en Figure 110) et le coefficient de Spearman (cellule G7 en Figure 110). Les tableurs ne proposent pas de fonction toute faite, mais deux formules nous permettront d'y remédier. Voici comment procéder (nous commentons la Figure 110) :

- En cellule G10, nous calculons ou recopions le coefficient de corrélation (ici, celui de Pearson ; formule reproduite sur la droite)
- En cellule G11, nous calculons ou saisissons la taille de l'échantillon (formule reproduite sur la droite)
- En cellule G12, nous calculons une quantité *t* qui est la statistique de test. La formule, reproduite sur la droite, utilise uniquement deux nombres : *r* et *n*. La formule mathématique est exposée plus bas, en Équation 35
- En cellule G13, nous calculons la p valeur à l'aide d'une loi de Student à *n-2* degrés de liberté. La formule est reproduite sur la droite

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.



	E	F	G	H	I	J	K
2							
3		Coefficient de corrélation Pearson :					
4		r=	0.717				
5							
6		Coefficient de corrélation Spearman :					
7		rs=	0.732				
8							
9		Test de nullité du coefficient de corrélation					
10		r=	0.717	=COEFFICIENT.CORRELATION(A2:A508;B2:B508)			
11		n=	507	=NB(A2:A508)			
12		t=	23.1346	=G10*RACINE((G11-2)/(1-G10^2))			
13		p val=	2.8E-81	=LOI.STUDENT.BILATERALE(G12;G11-2)			
14							

Figure 110. Test de nullité du coefficient de corrélation (Pearson ou Spearman) avec un tableur

Si vous n'êtes pas à l'aise avec ces formules, vous pouvez également vous référer directement à un graphique unique de significativité et d'interprétation du coefficient de corrélation, proposé plus bas sur la Figure 111 en page 181.

3.2.4.7.2 Le test de nullité du coefficient de corrélation (Pearson ou Spearman), en détail

Dans un échantillon de taille *n*, nous avons calculé le coefficient de corrélation (*r* pour Pearson, r_s, pour Spearman).

Les conditions de validité du test sont débattues. Vous pouvez considérer qu'il n'y en a pas, mais qu'il faudra être prudent sur l'interprétation, notamment en cas de relation non-linéaire, ou d'individus influents. Un nuage de points vous permettra de déceler immédiatement les difficultés.

Hypothèse nulle : en population, la corrélation entre X et Y est égale à zéro

Nous choisissons d'exécuter un test de nullité du coefficient de corrélation de Pearson ou de Spearman, selon le cas. Nous réaliserons un test bilatéral. Nous calculons la statistique de test *t* (Équation 35).

$$t = r.\sqrt{\frac{n-2}{1-r^2}}$$

$$\nu = n - 2$$

Équation 35. Test de nullité du coefficient de corrélation (Pearson ou Spearman)
Statistique de test et nombre de degrés de liberté de la loi de Student utilisée

Sous l'hypothèse H_0, la quantité *t* devrait suivre une loi de Student à $\nu=n-2$ degrés de liberté. Concrètement, la quantité *t* devrait tourner autour de zéro, sans trop s'en écarter. La Loi de Student ayant été décrite, cela permet de calculer la p valeur associée.

Nous pouvons conclure :

- Fixons le seuil d'interprétation de la p valeur à 5% (risque α)
- Si **p valeur < 5% : on rejette H_0**, on peut conclure que la corrélation entre les deux variables est significativement différente de zéro (elles ne sont pas indépendantes)
- Si **p valeur > 5%** : on ne rejette pas H_0 : on se trouve face à une **indétermination**, il est interdit de conclure.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.2.4.7.3 Quelques précisions, synthèse de l'interprétation

Nous avons vu que deux coefficients de corrélation pouvaient être calculés : Pearson, qui recherche une corrélation linéaire, et Spearman, qui s'intéresse aux rangs uniquement, et identifie plutôt une relation monotone (croissante ou décroissante, sans aller plus loin). Ces deux coefficients ont des avantages et des limites. Nous les comparons sur leurs différences dans le Tableau 16.

Tableau 16. Comparaison des coefficients de corrélation linéaires de Pearson et Spearman

Caractéristique	Pearson	Spearman
Met en évidence et quantifie une relation…	linéaire	monotone
Sensible aux valeurs extrêmes (individus influents)	Oui ☹️	Non 😊
Adapté pour des variables discrètes	Oui 😊	Non ☹️
Permet de calculer une équation de droite (régression linéaire simple)	Oui 😊	Non ☹️

Quoi qu'il en soit, ces coefficients partagent également de nombreux avantages, comme le fait d'être aisément calculables avec un tableur, d'avoir un test de nullité simple et efficace, et d'être facilement interprétables par un grand nombre de lecteurs.

Nous vous prodiguons un conseil simple : commencez toujours par **observer une représentation graphique** de X et Y pour bien comprendre leur relation. Puis **calculez les deux** coefficients. Ils sont souvent très proches. Si les deux coefficients diffèrent, la représentation graphique vous donnera l'explication. Très souvent, le coefficient de Pearson est altéré par quelques individus qui ont des valeurs extrêmes. Le coefficient de Spearman devient alors plus pertinent. En cas de variable discrète, le coefficient de Spearman peut être pris en défaut, et il faut souvent lui préférer le coefficient de Pearson.

Enfin, nous avons vu que ces deux coefficients pouvaient être interprétés selon trois axes : positif/négatif, fort/faible, significatif/non-significatif. Pour chaque réalisation du test statistique, nous avons seulement deux paramètres d'entrée : le coefficient de corrélation (r pour Pearson ou r_s pour Spearman) et la taille de l'échantillon (n). Il vous suffira de vous reporter à la Figure 111 : positionnez un point unique correspondant à votre expérience, et vous obtiendrez en une seule opération le résultat du test de nullité et l'interprétation de votre coefficient de corrélation.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

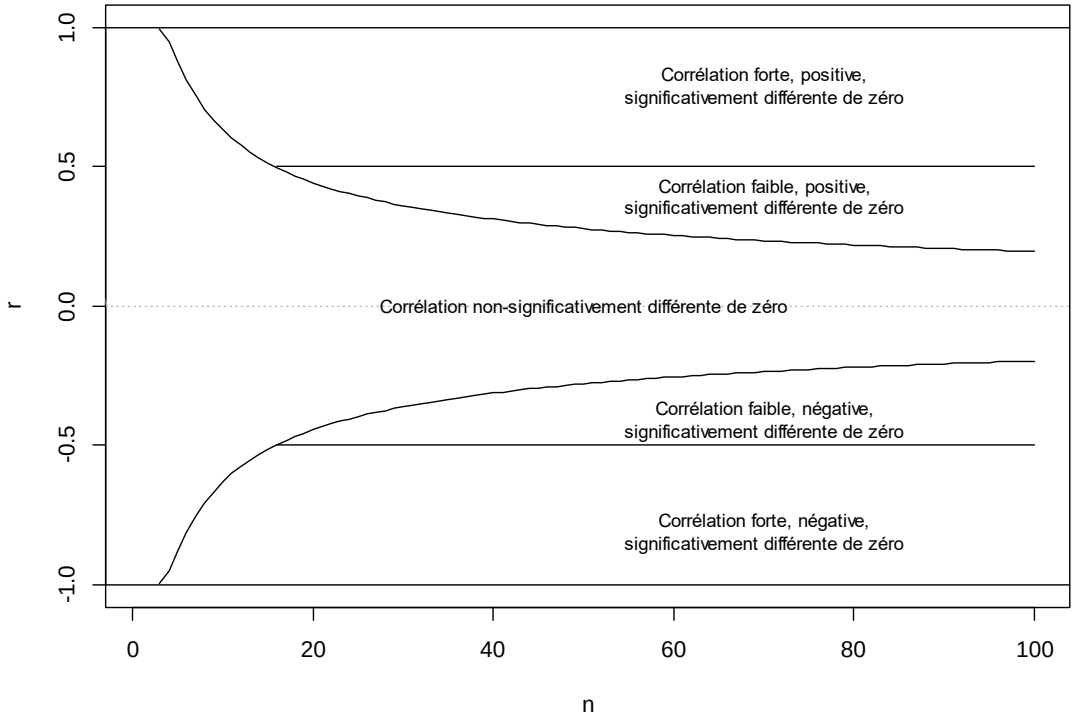


Figure 111. Interprétation d'un coefficient de corrélation de Pearson ou Spearman
Positionnez un point pour votre expérience : effectif en X, coefficient de corrélation en Y

Au cas où vous seriez trop proche de la courbe, le Tableau 17 vous donne les valeurs limites positives de r (ou r_s) en fonction de n : il s'agit des points par lesquels passe la courbe supérieure de la Figure 111.

Tableau 17. Valeurs limites positives du coefficient de corrélation correspondant à p=5%, en fonction de l'effectif

Effectif	Corrélation limite
3	0,997
4	0,950
5	0,878
6	0,811
7	0,754
8	0,707
9	0,666
10	0,632
11	0,602
12	0,576
13	0,553
14	0,532
15	0,514
16	0,497
17	0,482
18	0,468
19	0,456
20	0,444
21	0,433
22	0,423
23	0,413
24	0,404

Effectif	Corrélation limite
25	0,396
26	0,388
27	0,381
28	0,374
29	0,367
30	0,361
32	0,349
34	0,339
36	0,329
38	0,320
40	0,312
45	0,294
50	0,279
55	0,266
60	0,254
65	0,244
70	0,235
75	0,227
80	0,220
90	0,207
100	0,197

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.2.4.8 Régression linéaire simple, et test de la pente

La régression linéaire simple permet de calculer l'équation de la droite qui résume le mieux le nuage de points. La pente de la droite est alors proportionnelle au coefficient de corrélation de Pearson. S'il n'est pas valide de calculer et interpréter le coefficient de corrélation de Pearson, il ne faut pas réaliser de régression linéaire. Cette régression permet de répondre à trois questions :

- Peut-on décrire simplement la relation entre X et Y ?
- Peut-on prédire une valeur Y lorsqu'on connait la valeur de X ?
- Peut-on exprimer la force de la liaison entre X et Y ? Cette question reçoit déjà une réponse suffisante des coefficients de corrélation vus précédemment.

3.2.4.8.1 Exemple de mise en œuvre

En partant des plages de données X et Y, il est très aisé d'obtenir l'équation de la droite de régression avec un tableur. Avec Microsoft Excel et LibreOffice Calc, on peut utiliser les fonctions **pente()** et **ordonnee.origine()** ou plus récemment la fonction **droite.reg()**. Il suffit d'indiquer en paramètres la plage des valeurs de Y puis la plage des valeurs de X. Nous l'illustrons en Figure 112 en poursuivant notre exemple précédent. Les formules sont reproduites à droite de leur case d'exécution.

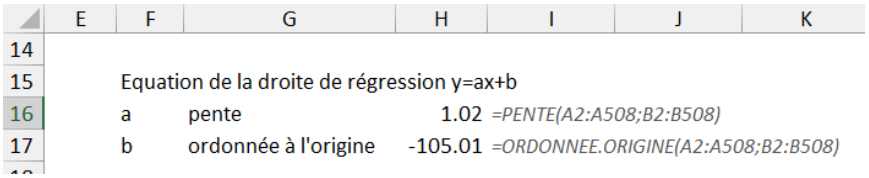


Figure 112. Obtenir l'équation de la droite de régression avec un tableur
Ici, $y = 1,02.x - 105,01$

Ces nombres, contrairement au coefficient de corrélation, tiennent bien compte des unités des deux variables. Ainsi, dans l'exemple précédent, on obtient l'équation de droite $y = 1{,}02.x - 105{,}01$. Cela signifie que, pour un individu dont la taille est connue et le poids inconnu, on peut prédire son poids en injectant sa taille dans l'équation, en respectant les mêmes unités que l'échantillon d'apprentissage du modèle.

3.2.4.8.2 La régression linéaire simple, en détail

La régression linéaire simple fait directement suite au calcul d'un coefficient de corrélation de Pearson (et non de Spearman !).

Il est **toujours valide** en soi de calculer une équation de droite de régression linéaire simple. Cependant, tout comme le coefficient de corrélation de Pearson, si on souhaite utiliser cette équation pour décrire un nuage de points et faire une prédiction, il faudra s'assurer de deux points (nous verrons comment les vérifier) :

- Il faut, au fond, que la relation entre les deux variables soit principalement **linéaire**
- Il faut que le calcul de r ne soit pas **influencé** par quelques individus extrêmes

On souhaite **modéliser** la relation entre X et Y par une équation $Y = aX + b + \varepsilon$ où ε représente l'erreur du modèle. On peut également voir $Y = aX + b$ comme une équation de la **droite de régression linéaire simple**. Alors, l'erreur ε est une erreur « verticale » (voir Figure 113). On l'appelle également **résidu** : c'est la part de Y qui ne peut pas être prédite par le modèle. La méthode d'ajustement vise à minimiser ces résidus. On utilise pour ce faire une méthode mathématique appelée « méthode des moindres carrés », qui vise à minimiser la somme des carrés de tous ces résidus.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

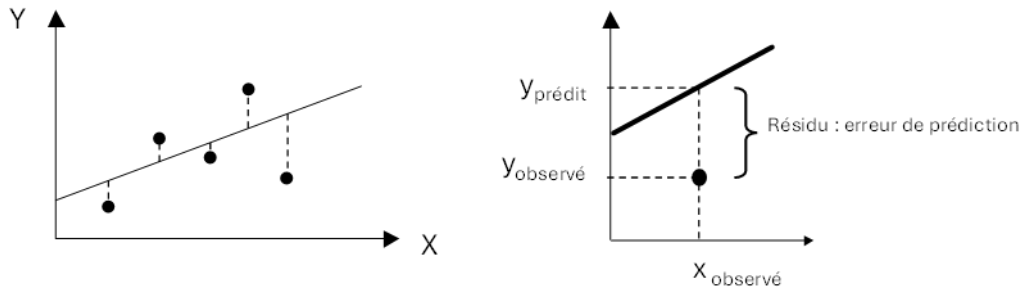


Figure 113. Régression linéaire simple : minimisation des résidus, distances verticales entre les points et la droite

Cette méthode des moindres carrés aboutit à des résultats très simples :

$$a = r.\frac{s_y}{s_x}$$
$$b = \bar{y} - a.\bar{x}$$

Équation 36. Coefficients de la droite de régression linéaire simple y=ax+b

On voit que, si la corrélation est nulle entre deux variables, alors la droite de régression est une droite horizontale dont la hauteur est la moyenne de Y.

Une fois cette équation de droite obtenue, il est possible de prédire la valeur y (notée y_0) d'un individu dont seule la valeur x est connue (notée x_0). On procède ainsi (Équation 37) :

- On calcule $\hat{y}_0$, valeur prédite par l'équation de droite (première ligne en Équation 37)
- On calcule s_{yx}, la nouvelle erreur de prédiction, qui est inférieure ou égale à l'écart type initial de Y. On voit que si la corrélation est nulle, ces deux valeurs sont identiques. On voit que si la corrélation est proche de 1 ou -1, la nouvelle erreur de prédiction est très proche de zéro (deuxième ligne en Équation 37)
- On peut ensuite calculer l'intervalle de confiance à 95% de cette prédiction, à l'aide d'une loi normale, qui nous donne le coefficient 1,96 (troisième ligne en Équation 37). La valeur inconnue de y_0 a 95% de chance de se trouver dans cet intervalle.

$$\hat{y}_0 = a.x_0 + b$$
$$s_{yx} = s_y.\sqrt{1 - r^2}$$
$$IC_{y_0;95\%} = \hat{y}_0 \pm 1{,}96.s_{yx}$$

Équation 37. Prédiction, avec intervalle de confiance, d'une valeur y inconnue avec x connu

Comme on le suppose en voyant la deuxième ligne en Équation 37, **r^2 est la part d'information** de Y prédite par le modèle de régression. On l'appelle aussi le **coefficient de détermination**[15] **R^2**.

La Figure 114 illustre comment la régression linéaire simple, lorsqu'elle est valide, permet de mieux prédire Y. Si on ne connaissait pas la taille de l'individu, on pourrait prédire son poids par l'intervalle $IC_{95} = \bar{y} \pm 1{,}96 \times s_y$, ce qui est illustré par les droites horizontales sur la Figure 114. Maintenant qu'on connait le modèle de régression appris sur les autres individus, si en plus on connait sa taille x_0, on peut désormais prédire de manière plus approprié son poids, en fonction de sa taille, grâce aux formules ci-dessus. On obtient une prédiction figurée par les droites obliques sur la Figure 114. On observe que cette prédiction est à la fois plus précise (verticalement, l'intervalle est à chaque fois plus étroit) et plus pertinente (compte tenu de la

[15] Dans le cas précis de la régression linéaire, $r^2=R^2$, mais pour d'autres régressions, la lettre R (majuscule) ne désigne pas le coefficient de corrélation linéaire r (minuscule).

taille de l'individu, le centre de l'intervalle est plus approprié). Cette prédiction étant fondamentalement linéaire, on comprend mieux les conditions de validité à venir.

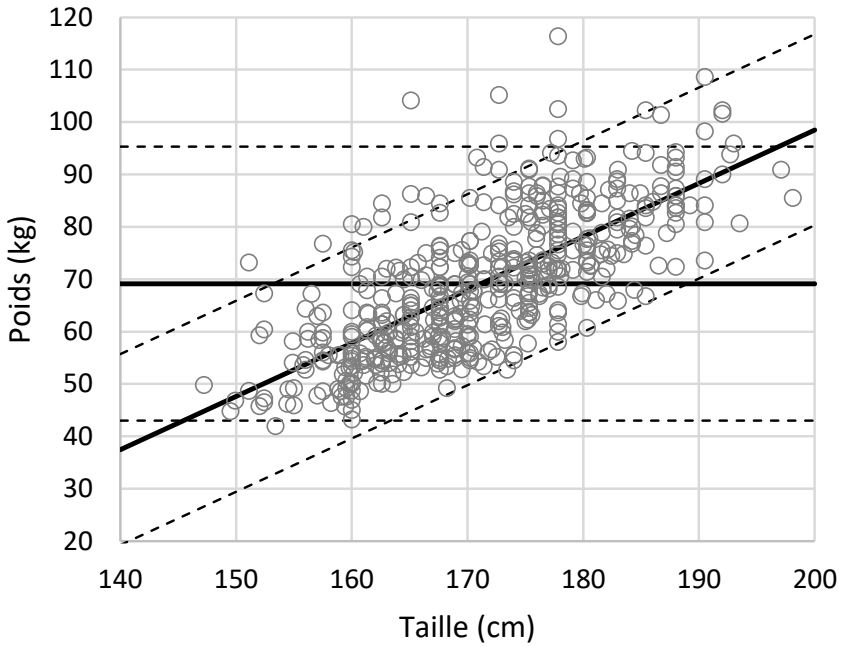


Figure 114. Prédiction du poids inconnu d'un nouvel individu.
Droites horizontales : sans connaissance a priori (avec la moyenne et l'écart type de Y).
Droites obliques : lorsqu'on connait sa taille (avec y prédit et s_{yx})

Cette prédiction est cependant assortie de conditions de validité : il faut que, au fond, le modèle linéaire soit réellement approprié. Ceci peut se vérifier ainsi, en partant de notre tableau d'individus (Figure 115) :

- Pour chaque individu *i*,
  - on calcule la valeur prédite par la droite de régression $\hat{y}_i = a.x_i + b$
  - on calcule ensuite le résidu, différence entre la valeur observée et la valeur prédite $résidu_i = \hat{y}_i - y_i$
- Ensuite, pour l'ensemble de l'échantillon,
  - On trace un histogramme des résidus : il doit avoir une allure bien symétrique
  - On trace un nuage de points des résidus en fonction de X : il doit ressembler à du bruit
    - Sa moyenne ne doit pas dépendre de x
    - Sa dispersion ne doit pas dépendre de x (homoscédasticité)

La Figure 115 illustre cette validation, en termes de résultats à obtenir ou ne pas obtenir.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

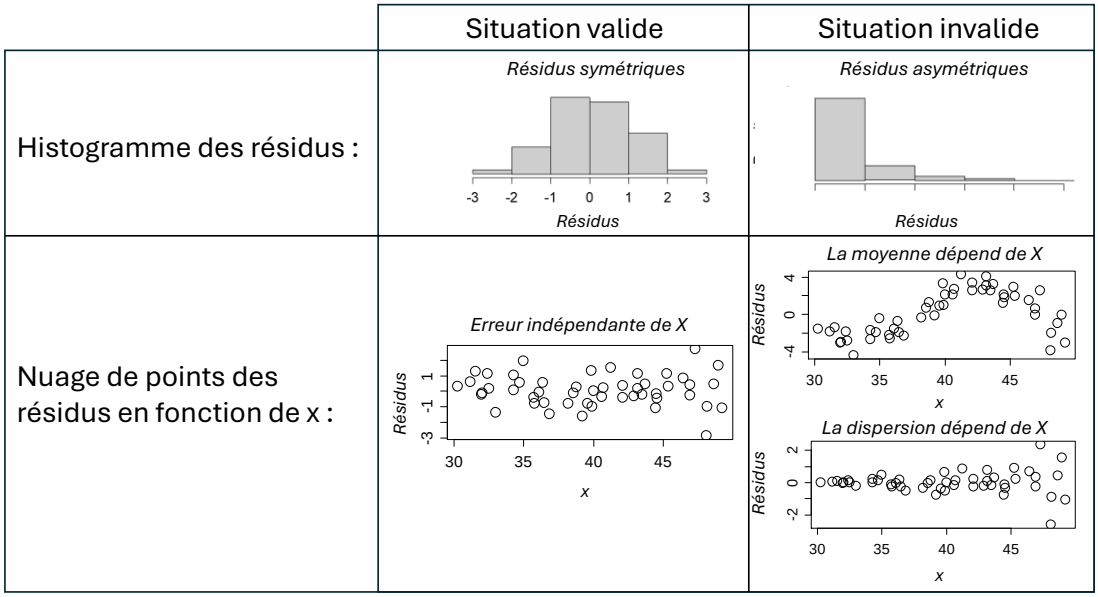


Figure 115. Analyse des résidus pour valider une régression linéaire simple

3.2.5 Une variable de survie et une variable qualitative

Comparer la survie de plusieurs groupes d'individus est une question très fréquente dans la recherche en santé. Formellement, cela revient à tester l'indépendance entre une variable de survie et une variable qualitative. La variable de groupe est alors la variable qualitative, et la variable d'intérêt est la variable de survie.

Exemple 1 : On cherche à savoir si un traitement, comparé à un placebo, est associé à une amélioration de la survie des patients.

Variable qualitative : traitement reçu {traitement ; placebo}

Variable de survie : temps jusqu'au décès (matérialisé par une colonne binaire décès 0/1, et une colonne quantitative de durée de suivi)

Exemple 2 : On cherche à savoir si un mode de vie (catégorisé) est associé à une meilleure fertilité, chez des femmes qui cherchent à tomber enceintes.

Variable qualitative : mode de vie (catégorisé selon le travail, la vie de couple, etc.)

Variable de survie : temps jusqu'au premier test de grossesse positif (matérialisé par une colonne binaire test 0/1, et une colonne quantitative de durée de suivi)

Pour répondre à ce type de question, vous pourrez simplement appliquer les outils graphiques et numériques décrits précédemment dans la section 2.4 Variables de survie en page 145, pour chaque modalité de la variable qualitative. Vous obtiendrez donc autant de courbes de Kaplan-Meier que de sous-groupe. Il sera particulièrement pertinent de les superposer.

Si vous observez une différence apparente entre plusieurs sous-groupes, il faudra réaliser un test statistique pour savoir si cette différence peut être le simple fruit du hasard. Deux tests sont couramment réalisés :

- Le **test du Log Rank**, qui est spécifiquement dédié à cette question
- Le **modèle de Cox**, qui peut être réalisé avec une seule variable explicative qualitative

Dans les deux cas, il vous sera impossible de réaliser ces tests seul, et vous aurez besoin de l'aide d'un biostatisticien.

3.2.6 Une variable de survie et une variable quantitative

Cette question, fréquente dans la recherche en santé, revient à savoir si une variable quantitative influe de manière log-linéaire la survie, en termes de *hazard ratio*, ou rapports des risque instantanés.

Exemple : On s'intéresse à des diabétiques. On mesure leur hémoglobine glyquée à l'inclusion. On cherche à savoir si cette valeur permet de prédire le délai au bout duquel ils seront traités par insuline.

Variable quantitative : hémoglobine glyquée à l'inclusion

Variable de survie : temps jusqu'au passage sous insuline (matérialisé par une colonne binaire insuline 0/1, et une colonne quantitative de durée de suivi)

Cette notion de **log-linéarité** est complexe, et liée au **modèle de Cox**, que nous utiliserons. Elle sous-entend plusieurs étapes :

- Tout d'abord, l'estimation d'un **risque instantané** d'événement, en fonction du temps
- Ensuite, l'identification d'un **effet systématique et constant** de la variable quantitative sur ce risque instantané, **log-linéaire** au sens où chaque unité supplémentaire de la variable quantitative **multiplie** le risque instantané par un même facteur

En réalité, cette notion de log-linéarité se vérifie assez peu souvent en santé. Le plus raisonnable, et qui répond déjà à la question, sera souvent de discrétiser la variable quantitative, par exemple en tertiles ou en quartiles, ou selon des seuils experts. On se retrouve alors dans le cas de la section précédente 3.2.5 Une variable de survie et une variable qualitative en page 185.

Sinon, nous n'avons pas d'outil graphique ou numérique simple à vous proposer. Si vous souhaitez réaliser une telle analyse, il vous faudra vous faire aider d'un biostatisticien qui réalisera dans votre cas un **modèle de Cox**, avec une variable explicative quantitative.

3.3 Deux variables appariées, dans plusieurs groupes

3.3.1 Préambule

Nous avions expliqué précédemment qu'il était déconseillé de procéder à des comparaisons avant-après dans un seul groupe. Nous reviendrons dessus dans le chapitre 5.2 Tests appariés dans un seul groupe, avant-après) en page 215.

En revanche, les analyses de variables appariées **dans plusieurs groupes** restent tout à fait **valides**, et correspondent même souvent à des **méthodologies de qualité**.

Dans le cas d'une variable binaire, il s'agit de savoir si des patients sont stables, s'améliorent ou s'empirent, comparativement à un autre groupe de patients (ou plusieurs) qui est suivi dans les mêmes conditions. Dans le cas d'une variable quantitative, il s'agit de savoir si des patients voient une valeur augmenter ou diminuer, comparativement à un autre groupe de patients (ou plusieurs) qui est suivi dans les mêmes conditions. Très souvent, on comparera deux groupes : un soumis à la prise en charge qu'on souhaite évaluer, l'autre soumis à une autre prise en charge (placebo ou traitement de référence). Ces études entrent généralement dans le champ des études interventionnelles comparatives, présentées dans le chapitre 6.2.2 Etudes comparatives en page 48.

On notera tout de même que, dans les recherches cliniques traditionnelles, on cherche à homogénéiser autant que possible le statut des patients à l'inclusion, si bien que ce type de méthode n'est pas forcément utilisé : le statut « avant » étant réputé identique pour tous les patients, on examine uniquement le statut « après », sans regarder l'évolution des patients.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.3.2 Deux variables binaires appariées, et une variable de groupe

Voici un exemple. On s'intéresse à une infection le plus souvent asymptomatique : l'infection à Toxoplasma Gondii. On sait que cette infection guérit spontanément dans la plupart des cas, au bout d'un délai variable. Mais on sait aussi que les individus séronégatifs peuvent contracter cette infection à tout moment, par exemple en mangeant de la salade mal lavée, ou au contact d'un chat[16]*.*

On inclut des patients, qu'on randomise dans deux groupes. Après leur randomisation, on réalise une sérologie qui permet de diagnostiquer l'infection chez certains de ces patients. Le premier groupe est soumis à un traitement antiinfectieux qu'on souhaite évaluer, le deuxième groupe est soumis à un traitement par placebo. Six mois plus tard, on réalise une deuxième sérologie chez tous les patients.

Dans chacun des deux groupes, on observe possiblement 4 statuts :

- Les patients négatifs restés négatifs : 0→0
- Les patients positifs restés positifs : 1→1
- Les patients négatifs devenus positifs : 0→1
- Les patients positifs devenus négatifs : 1→0

On souhaite savoir si, en moyenne, l'évolution est différente dans le groupe avec traitement du groupe avec placebo.

Lorsqu'on représente les données, 3 colonnes sont renseignées pour tous les individus :

- Une colonne définissant le groupe : placebo/traitement dans ce cas, mais il est possible d'avoir plus de deux groupes
- Une colonne définissant le statut à l'inclusion : 0/1
- Une colonne définissant le statut après 6 mois : 0/1

Nous créons une quatrième colonne calculée comme étant :

$$évolution = statut_{6mois} - statut_{inclusion}$$

Nous oublions les deux colonnes initiales de statut. Il nous reste ainsi deux colonnes :

- Une colonne définissant le groupe : placebo/traitement dans ce cas, mais il est possible d'avoir plus de deux groupes
- Une colonne définissant l'évolution (3 modalités) :
  - Vaut -1 pour les patients positifs qui deviennent négatifs
  - Vaut +1 pour les patients négatifs qui deviennent positifs
  - Vaut 0 pour les patients qui restent stables, positifs ou négatifs

Pour répondre à la question posée, nous posons l'hypothèse nulle H_0 : l'évolution est indépendante du groupe de traitement. Il suffit alors de réaliser un test du Khi² d'indépendance entre les deux colonnes ci-dessus, après éviction des effectifs des sujets stables, comme défini dans le chapitre 3.2.2.5 Test du Khi² d'indépendance en page 155.

3.3.3 Deux variables quantitatives appariées, et une variable de groupe

Voici un exemple. On s'intéresse à des patients diabétiques de type 2, asymptomatiques. On s'intéresse en particulier à l'hémoglobine glyquée, HbA1c. L'HbA1c doit valoir 5% chez les individus sains, et s'élève dans le diabète type 2. Par exemple, un diabétique équilibré peut avoir une HbA1c à 6% (l'objectif varie selon l'âge), et un diabétique déséquilibré peut avoir une HbA1c à 10%.

On inclut des patients, qu'on randomise dans deux groupes. Après leur randomisation, on réalise une première mesure de l'HbA1c. Le premier groupe est soumis à un nouveau

[16] Le chat se lèche l'anus puis les poils, l'humain caresse le chat puis se touche la bouche. Sinon, l'ami d'un ami d'un ami m'a aussi raconté que… non, finalement, oubliez cette histoire.

traitement qu'on souhaite évaluer, le deuxième groupe est soumis à un traitement de référence. Six mois plus tard, on réalise un deuxième dosage chez tous les patients.

Dans chacun des deux groupes, chaque patient a deux mesures de l'HbA1c. On souhaite savoir si, en moyenne, l'évolution est différente dans le groupe avec le nouveau traitement, par rapport au groupe avec le traitement de référence.

Lorsqu'on représente les données, 3 colonnes sont renseignées pour tous les individus :

- Une colonne définissant le groupe : placebo/traitement dans ce cas, mais il est possible d'avoir plus de deux groupes
- Une colonne définissant le dosage à l'inclusion : un nombre réel compris généralement entre 4 et 15
- Une colonne définissant le dosage après 6 mois : un nombre réel compris généralement entre 4 et 15

Nous créons une quatrième colonne calculée comme étant :

$$évolution = dosage_{6mois} - dosage_{inclusion}$$

Nous oublions les deux colonnes initiales de dosage. Il nous reste ainsi deux colonnes :

- Une colonne définissant le groupe : placebo/traitement dans ce cas, mais il est possible d'avoir plus de 2 groupes
- Une colonne définissant l'évolution : nombre réel compris entre -11 et +11

Pour répondre à la question posée, nous posons l'hypothèse nulle H_0 : la moyenne de l'évolution est indépendante du groupe de traitement. Sous réserve des conditions de validité, il suffit alors de réaliser un test de Student, comme défini dans le chapitre 3.2.3.5 Test de Student pour échantillons indépendants, avec ou sans correction de Welch en page 164. S'il y avait plus de deux groupes, on réaliserait une ANOVA, analyse de la variance.

3.4 Cas particuliers d'analyses bivariées

Dans cette section, nous verrons des cas particuliers d'analyses bivariées. Pour ces cas particuliers, les outils définis précédemment restent tout à fait valides, et permettraient de tester l'indépendance (ou presque) entre les deux variables. Pourtant, ce ne sont pas les plus utilisés, essentiellement pour une raison : au fond, dans les cas dont il est question, **il est déjà évident que les deux variables étudiées ne sont pas indépendantes** (ex : un test diagnostique et la maladie qu'il cherche à détecter). On cherche plutôt à **quantifier la force de leur association**. Cette manière de quantifier la force de l'association dépend de la configuration et du champ disciplinaire de la question posée, d'où les situations énumérées ci-dessous :

- **Facteurs de risque** ou protecteurs en épidémiologie (voir chapitre 3.4.1 en page 188)
- **Tests diagnostiques** à réponse binaire (voir chapitre 3.4.2 en page 195)
- **Outils de détection** d'un nombre indéterminé d'événements (voir chapitre 3.4.3 en page 198)
- **Tests diagnostiques** à réponse quantitative (voir chapitre 3.4.4 en page 199)
- **Accord entre deux juges**, sans gold standard, coefficient **Kappa** (voir chapitre 3.4.5 en page 205)

3.4.1 Facteurs de risque ou protecteurs en épidémiologie

3.4.1.1 Préambule

Dans ces études, nous nous intéresserons typiquement à un facteur binaire, et à une pathologie binaire. Nous chercherons à savoir si ce facteur est indépendant, ou est associé à une augmentation du risque de maladie (**facteur de risque**), ou est associé à une diminution du risque de maladie (**facteur protecteur**).

Le design associé à cette problématique a été évoqué en section 6.1.2 Etudes analytiques en page 43 : il s'agit principalement, par tradition, des études observationnelles analytiques réalisées en épidémiologie. Cela dit, les indicateurs que nous verrons peuvent très bien être utilisés dans d'autres études, ce n'est simplement pas traditionnel.

Les deux indicateurs que nous calculerons, l'**odds ratio (OR)** et le **risque relatif (RR)**, visent surtout à quantifier la force de l'association et préciser son sens (risque ou protection). On les utilise généralement lorsque l'association est déjà connue. Si l'association était totalement nouvelle, on commencerait par affirmer sa non-indépendance, à l'aide d'un test statistique bivarié comme le Khi² (voir section 3.2.2 Deux variables qualitatives en page 153).

3.4.1.2 Exemple de calcul du risque relatif et de l'odds ratio en pratique

Dans l'exemple ci-dessous, nous nous intéressons à une étude exposé-non-exposé incluant 1000 individus alcooliques et 1000 individus non-alcooliques. Nous noterons « M » et « NM » les malades et non-malades respectivement, et « E+ » et « E- » les exposés et non-exposés respectivement.

Nous partirons d'un tableau de contingence, qui peut aisément être obtenu d'un tableau croisé dynamique impliquant les deux variables binaires (exposition et maladie).

	A	B	C	D	E	F	G	H	I
1		M	NM	P(M) dans la ligne			Indicateurs		
2	E+	73	927	0.073	=B2/(B2+C2)		RR=	1.97	=D2/D3
3	E-	37	963	0.037	=B3/(B3+C3)		OR=	2.05	=(B2*C3)/(B3*C2)

Figure 116. Calcul du risque relatif et de l'odds ratio avec un tableur

Pour calculer le risque relatif, nous calculons d'abord la proportion de malades parmi les individus exposés d'une part, et non-exposés d'autre part (cellules D2 et D3 en Figure 116, les formules sont recopiées sur la droite). Le risque relatif est le rapport entre ce nombre calculé chez les exposés, et ce nombre calculé chez les non-exposés (cellule H2 en Figure 116, la formule est recopiée sur la droite). Il répond à la question : par combien le risque de maladie est-il multiplié, lorsqu'on est exposé, par rapport aux non-exposés ?

Pour calculer l'odds ratio, nous calculons le produit des effectifs favorables à une relation de risque (E+ & M, et E- et NM), divisé par le produit des effectifs favorables à une relation protectrice (E+ & NM, et E- et M) (cellule H3 en Figure 116, la formule est recopiée sur la droite). Cet odds ratio ne semble pas répondre à une question aisément formulable, mais retourne un nombre du même ordre de grandeur que le risque relatif dans les situations les plus fréquentes. Nous verrons par la suite que, pour de nombreuses raisons, **l'odds ratio est souvent préféré au risque relatif** (voir chapitre 3.4.1.5 Mieux comprendre le risque relatif et l'odds ratio en page 191).

Comme nous l'avions énoncé dans la section 6.1.2 Etudes analytiques en page 43 et comme nous l'expliquerons après, il est **interdit de calculer le risque relatif dans les études de cas-témoin**.

3.4.1.3 Calcul du risque relatif et de l'odds ratio, en général

Définissons à présent le risque relatif et l'odds ratio de manière plus systématique. Nous n'évoquerons pas ces quantités en population, mais directement dans un échantillon.

Tableau 18. Présentation typique d'un tableau croisé de contingence en épidémiologie analytique

	M	**NM**
E+	a	b
E-	c	d

Le risque relatif divise l'estimation de la probabilité d'être malade sachant qu'on est exposé, par l'estimation de la probabilité d'être malade sachant qu'on n'est pas exposé (Équation 38).

$$RR = \frac{\hat{P}(M/E+)}{\hat{P}(M/E-)} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

Équation 38. Calcul du risque relatif dans l'échantillon

L'odds ratio divise la cote de la maladie chez les exposés (nombre de malades divisé par le nombre de non-malades), par la cote de la maladie chez les non-exposés (Équation 39). C'est un rapport de cotes[17].

$$OR = \frac{a/b}{c/d} = \frac{a.d}{b.c}$$

Équation 39. Calcul de l'odds ratio dans l'échantillon

Ces deux nombres s'interprètent ainsi :

- $[0;1[$ facteur protecteur
- 0 facteur indépendant
- $]1;+\infty[$ facteur de risque

On notera que l'interprétation de la taille d'effet est asymétrique. Ainsi, une valeur de 0,5 (le risque est divisé par 2) est aussi importante en termes de taille d'effet qu'une valeur de 2 (le risque est multiplié par 2). Plus généralement, une valeur de 1/k est aussi importante en termes de taille d'effet qu'une valeur de k. Les facteurs protecteurs varient sur un intervalle de largeur 1, tandis que les facteurs de risque varient sur un intervalle de largeur infinie. Mentalement, il faudrait se plutôt se représenter une échelle équilibrée de Log(RR) ou Log(OR).

3.4.1.4 Tester la significativité de l'association

Si vous le souhaitez, vous pourrez tester la significativité statistique d'une association entre un facteur et une maladie. Votre observation étant réalisée dans un échantillon, si vous observez l'absence d'indépendance entre le facteur et la maladie, la question est d'extrapoler cette absence d'indépendance à la population. Un arbre décisionnel est proposé en Figure 117.

Si vous devez réaliser cela seul, le plus simple reste de réaliser un test du Khi² d'indépendance, comme nous l'avons vu précédemment dans le chapitre 3.2.2 Deux variables qualitatives en page 153.

Si vous demandez l'aide d'un statisticien, traditionnellement, on calcule l'intervalle de confiance à 95% du RR ou de l'OR. Sous l'hypothèse nulle d'indépendance, en population le RR ou l'OR devrait valoir 1. Si cet intervalle ne contient pas la valeur 1, on peut rejeter l'hypothèse nulle et conclure à la non-indépendance. Certains, plus rarement, réalisent un test statistique dont l'hypothèse nulle est la même, et rejettent cette hypothèse si la p valeur est inférieure à 5%.

[17] Le mot « cote » s'écrit sans accent et désigne, dans les paris, le rapport entre les chances de perdre et celles de gagner (« odds » en Anglais, au pluriel).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

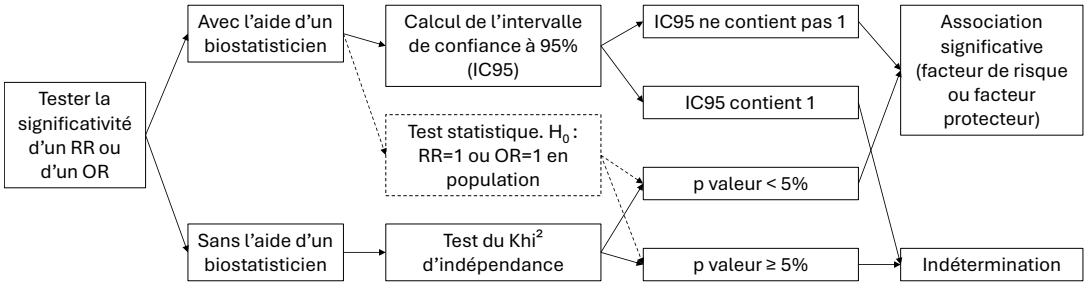


Figure 117. Arbre décisionnel : tester la significativité d'un RR ou d'un OR

Véritablement, la question n'est en général plus de savoir si un facteur est associé à une maladie, mais plutôt de quantifier la **taille de l'effet**, estimée par le RR ou l'OR. Pour ce faire, un test statistique n'est pas vraiment nécessaire, mais la formulation de l'intervalle de confiance à 95% du RR ou de l'OR, apporte une notion de précision de l'estimation qui sera appréciée par le lecteur.

3.4.1.5 Mieux comprendre le risque relatif et l'odds ratio

3.4.1.5.1 Etudes compatibles avec le RR et l'OR

Tout d'abord, discutons des conditions dans lesquelles le risque relatif et l'odds ratio peuvent être calculés. Nous reprendrons le tableau de contingence typique présenté dans le Tableau 18 en page 189, dans le cadre des études épidémiologiques présentées en section 6.1.2 Etudes analytiques en page 43.

Dans les **cohortes** prospectives ou historiques, le calcul du RR et de l'OR ne pose aucun problème.

Dans une étude de type **exposé-non-exposé**, en comparaison avec une étude de cohorte, généralement la proportion d'exposés est artificiellement sur-représentée afin de rendre l'étude réalisable. Cela correspond, dans un tableau de contingence, à multiplier par un facteur k supérieur à 1 les effectifs a et b. On constate (Équation 40) que cela n'altère pas les estimations du RR et de l'OR.

$$RR: \quad \frac{k.a(c+d)}{c(k.a+k.b)} = \frac{k.a(c+d)}{k.c(a+b)} = \frac{a(c+d)}{c(a+b)}$$

$$OR: \quad \frac{(a.k).d}{(b.k).c} = \frac{a.d}{b.c}$$

Équation 40. Comportement du RR et de l'OR dans les études exposé-non-exposé

Dans une étude de type **cas-témoin**, en comparaison avec une étude de cohorte, généralement la proportion de malades est artificiellement sur-représentée afin de rendre l'étude réalisable. Cela correspond, dans un tableau de contingence, à multiplier par un facteur k supérieur à 1 les effectifs a et c. On constate (Équation 41) que cela n'altère pas l'estimation de l'OR, mais rend fausse l'estimation du RR. Pour cette raison, **seul l'odds ratio peut être calculé dans les études de cas-témoin**.

$$RR: \quad \frac{k.a(k.c+d)}{k.c(k.a+b)} \neq \frac{a(c+d)}{c(a+b)}$$

$$OR: \quad \frac{(k.a).d}{b.(k.c)} = \frac{a.d}{b.c}$$

Équation 41. Comportement du RR et de l'OR dans les études cas-témoin

3.4.1.5.2 Que signifient le RR et l'OR ? Comment les interpréter ?

En apparence, le risque relatif est plus interprétable car il répond directement à la question « par combien le risque est-il multiplié, lorsqu'on est exposé, par rapport aux non-exposés ? ». Cette facilité n'est qu'apparente, comme l'illustre l'exemple qui suit.

Soit un risque relatif de 2 :

Appliqué à une prévalence chez les non-exposés de 1/10 000, il a un effet mineur : on passe à 2/10 000, autrement dit la prévalence de la non-maladie passe de 99,99% à 99,98%, ce qui est négligeable.

Appliqué à une prévalence de 50%, il a un effet majeur : la prévalence passe de 50% à 100%, autrement dit la prévalence de la non-maladie est divisée par l'infini.

Appliqué à une prévalence de 75%... ce n'est pas possible.

Ce risque relatif, en apparence simple, ne peut en réalité pas être interprété sans connaître la prévalence. Ce n'est donc pas un indicateur simple.

L'odds ratio, en apparence, est un indicateur moins naturel. Pourtant, on le comprend mieux si on se réfère à la fonction sigmoïde (Équation 42 et Figure 118), qui est un cas particulier de fonction logistique.

$$sigmoide(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Équation 42. Fonction Sigmoïde

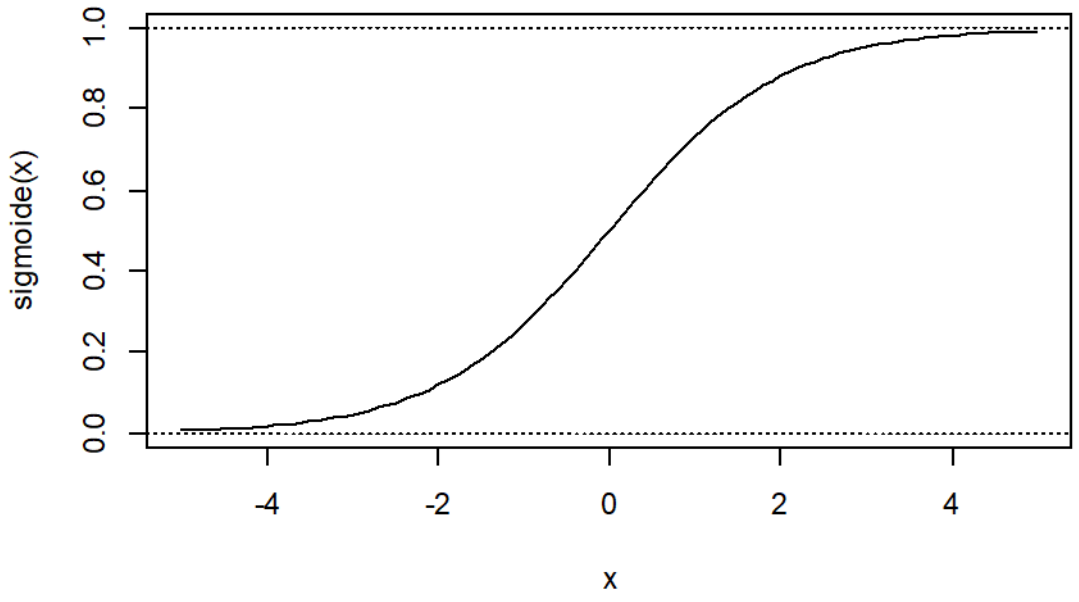


Figure 118. Fonction Sigmoïde

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

En Figure 118, l'ordonnée représente également la prévalence, et l'abscisse correspond à une situation donnée. Lorsqu'on applique un odds ratio de valeur OR, on se décale sur la courbe d'une valeur $\ln(\mathrm{OR})$ vers la droite, et on obtient ainsi la nouvelle prévalence[18].

Nous illustrons cet effet sur la Figure 119. Nous partons d'une prévalence de 1,54% (point le plus à gauche) et appliquons successivement des odds ratios de 2 : l'abscisse se décale à chaque fois de $\ln(2)$, et la prévalence (en Y) augmente en suivant une sigmoïde. On observe que l'effet augmente progressivement, puis « se tasse ». De l'expérience des épidémiologistes, cette manière de décrire un effet constant sur une prévalence (ex : progression d'une pandémie, ou inversement effet d'une politique publique sur le taux d'alphabétisation, etc.) est assez naturelle, c'est une des raisons pour laquelle les OR sont assez appréciés.

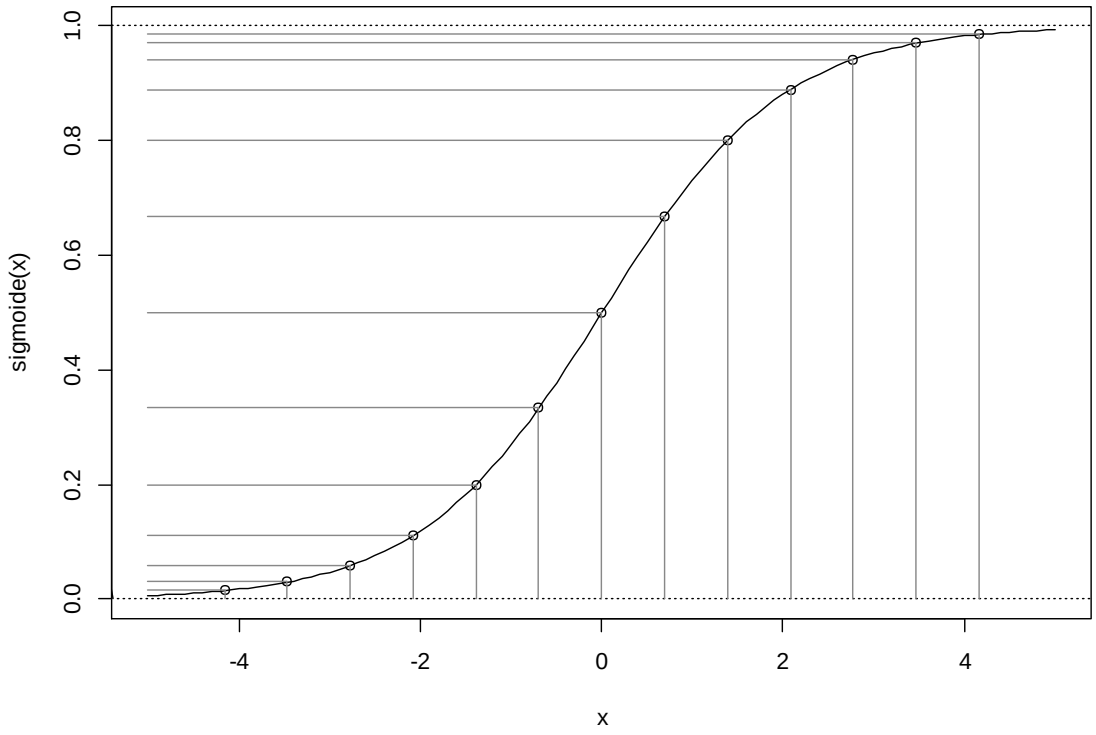


Figure 119. Effet d'un odds ratio de 2 (décalages successifs vers la droite en abscisse) sur la prévalence (décalages successifs vers le haut en ordonnée)

On peut aller plus loin, et observer que la fonction sigmoïde est assez proche de trois fonctions connues sur certains intervalles (Figure 120). L'interprétation de ces approximations est la suivante :

Pour des **états peu fréquents** (ex : de 0 à 10%, ce qui est le plus fréquent en santé ; partie gauche de la Figure 120), la fonction sigmoïde se confond avec la fonction exponentielle $y = e^x$: **l'OR multiplie la fréquence de l'état par *OR***, comme le faisait le risque relatif.

Pour les **états équilibrés** (ex : de 30 à 70%, partie médiane de la Figure 120), la fonction sigmoïde se confond avec la droite d'équation $y = \frac{x}{4} + 0{,}5$: **l'OR ajoute** $0{,}25 \times ln(OR)$ **à la fréquence de l'état**.

Pour des **états très fréquents** (ex : de 90 à 100%, ce qui correspond souvent au statut « non-malade » en santé, partie droite de la Figure 120), la fonction sigmoïde se confond avec la

[18] « ln » désigne le logarithme népérien, ou logarithme naturel. Exemples de valeurs :
ln(0)= -Inf ln(1)=0 ln(2)=0,693 ln(10)=2,303

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

fonction $y = 1 - e^{-x}$: **l'OR divise la fréquence de l'état complémentaire par *OR***, comme le ferait le risque relatif de ne pas être malade.

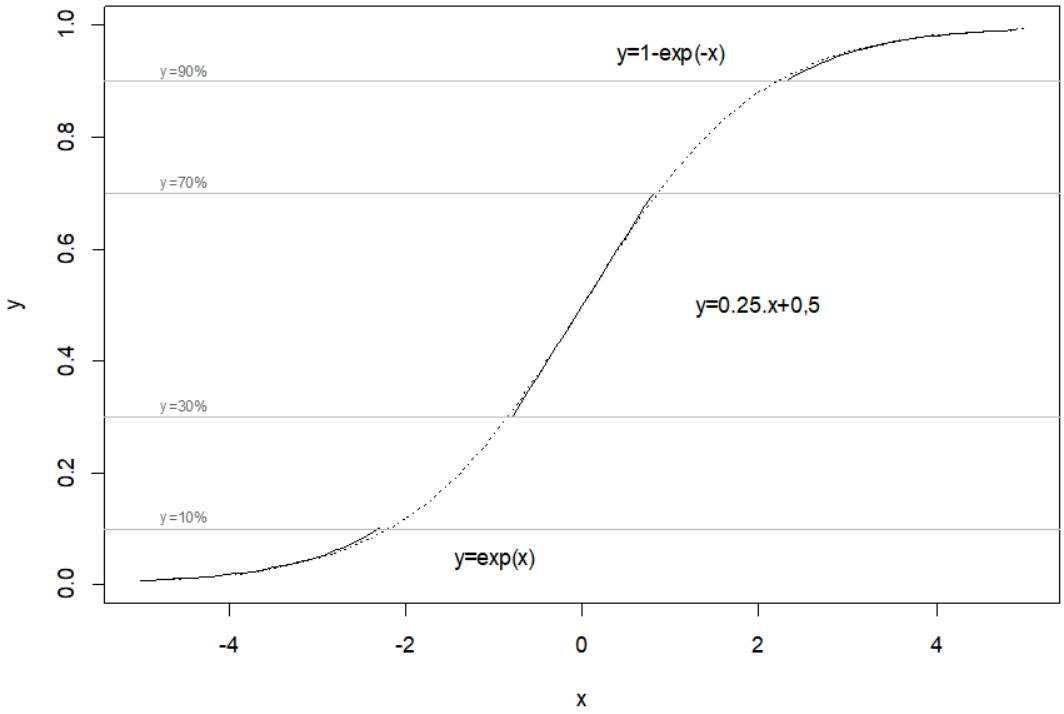


Figure 120. Approximation de la fonction sigmoïde par trois fonctions, sur certains intervalles de Y

3.4.1.5.3 Autres éléments de comparaison RR et OR, synthèse

Un autre argument en faveur de l'OR est qu'on pourrait imaginer lier aisément une variable X à la fréquence d'une maladie, avec une relation numérique proposée par la Figure 119 : c'est justement ce que fait la **régression logistique**. Cette régression logistique a l'avantage de permettre des analyses multivariées.

Le risque relatif peut également être calculé dans une approche multivariée, avec la **régression de Poisson**, qui est cependant moins utilisée en santé. Alors que la régression logistique est fortement utilisée pour les analyses portant sur des individus tous connus à l'avance, la régression de Poisson sera volontiers utilisée pour des analyses portant sur des secteurs géographiques ou des intervalles de temps, dans lesquels on comptera des événements ou individus malades, sans nécessairement connaître le nombre total d'individus disponibles.

Dans les méta-analyses, qui consistent à agréger les résultats de plusieurs études, il arrive fréquemment qu'on ait accès seulement aux résultats publiés par ces études mais pas aux données source. Alors, le risque relatif est apprécié car il est aisément **agrégeable**, au sens où le risque relatif total est une moyenne pondérée des risques relatifs des deux études. Ce n'est pas le cas de l'odds ratio.

L'ensemble de ce chapitre nous permet de dresser un tableau comparatif du RR et de l'OR (Tableau 19). Il en résulte de que, de manière générale, **les chercheurs en santé préfèrent s'intéresser à l'odds ratio**.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 19. Comparaison du RR et de l'OR

Critère	Risque relatif	Odds ratio
Utilisable dans les cohortes et études exposé-non-exposé	Oui	Oui
Utilisable dans les cas-témoins	Non ☹	Oui
Interprétation intuitive pour les états rares	Oui	Oui
Interprétation intuitive pour les états fréquents	Non ☹	Une fois habitué…
Symétrie	Non ☹	Oui
Agrégeable	Oui	Non ☹
Analyse multivariée	En écologique (régression de Poisson)	Sur les individus (régression logistique)

3.4.2 Tests diagnostiques à réponse binaire

3.4.2.1 Préambule

De nombreuses études en santé visent à évaluer l'aptitude d'un test à diagnostiquer une maladie. Dans ces études, nous nous intéresserons typiquement à un test diagnostique binaire, et à une pathologie binaire. Ce test diagnostique peut être de différentes natures : examen biologique, examen radiologique, test clinique, question posée au patient, etc.

Evidemment, nous savons déjà que ce test diagnostique n'est pas indépendant de la pathologie, c'est pourquoi, bien que ce ne soit pas faux, il serait vain de réaliser un test statistique d'indépendance comme le Khi² (voir section 3.2.2 Deux variables qualitatives en page 153).

Nous ne cherchons pas non plus à quantifier l'association entre le test et la pathologie en termes de risque relatif ou d'odds ratio, car ces mesures n'ont pas trop d'utilité dans ce contexte.

La notion de malade ou non-malade est identifiée au préalable avec des moyens déterminés dans l'expérience : nous appellerons cette classification le « **gold standard** » : elle sera considérée comme exacte.

Nous considèrerons que le test classe correctement les individus s'il revient négatif chez les non-malades, et positif chez les malades. Ainsi, 4 quantités rapportant chacune la proportion de bien-classés au sein d'un sous-groupe, seront rapportées. On espère que chacune de ces quantités soit la plus proche possible de 100% :

- La **sensibilité (Se)** est la proportion de bien classés parmi les malades
- La **spécificité (Sp)** est la proportion de bien classés parmi les non-malades
- La **valeur prédictive positive (VPP)** est la proportion de bien classés parmi les tests positifs
- La **valeur prédictive négative (VPN)** est la proportion de bien classés parmi les tests négatifs

Nous nous autorisons ici un rappel de Français. La diagnostique (nom commun féminin) désigne l'art de poser les diagnostics ; ce nom est peu utilisé. Le diagnostic (nom commun masculin) désigne l'avis qu'on souhaite se faire sur une pathologie, ou l'avis qu'on porte effectivement sur cette pathologie. L'adjectif qualificatif correspondant (diagnostique(s)) s'accorde en genre et en nombre avec le nom qu'il qualifie, mais se termine par « -que » au masculin comme au féminin : on parle de **test diagnostique** autant que de **procédure diagnostique**.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.4.2.2 Exemple de calcul de Se, Sp, VPP et VPN, en pratique

Nous étudions la relation entre un test (douleur thoracique à l'interrogatoire) et une maladie (lésions coronaires à la coronarographie, la coronarographie est donc le gold standard dans cette étude). Nous traçons (par exemple avec un tableau croisé dynamique) un tableau de contingence, et ajoutons les totaux des lignes et des colonnes (Figure 121). Nous calculons les quatre quantités d'intérêt à l'aide de simples ratios, dont les formules sont reproduites sur la droite (colonne i en Figure 121).

	A	B	C	D	E	F	G	H	I
1			*malades*	*non-malades*					
2			M	NM			Se :	0.950	*=C3/C5*
3	*tests positifs*	T+	950	250	1200		Sp :	0.444	*=D4/D5*
4	*tests négatifs*	T-	50	200	250		VPP :	0.792	*=C3/E3*
5			1000	450	1450		VPN :	0.800	*=D4/E4*

Figure 121. Calcul de Se Sp VPP et VPN avec un tableur

Remarquez qu'au numérateur de ces quatre quantités, on ne retrouve que des effectifs d'individus bien classés : les malades ayant un test positif en C3, et les non-malades ayant un test négatif en D4. S'agissant d'une proportion à chaque fois, le dénominateur est la somme du numérateur et d'un autre effectif.

3.4.2.3 Calcul de Se, Sp, VPP et VPN, en général

On définit quatre sous-groupes d'individus :

- Les vrais positifs (VP) sont les individus malades qui ont un test positif (à juste titre)
- Les vrais négatifs (VN) sont les individus non-malades qui ont un test négatif (à juste titre)
- Les faux positifs (FP) sont les individus non-malades qui, pourtant, ont un test positif
- Les faux négatifs (FN) sont les individus malades qui, pourtant, ont un test négatif

Ces effectifs peuvent se représenter comme suit dans un tableau de contingence (Tableau 20).

Tableau 20. Tableau de contingence d'un test diagnostique binaire

	M	**NM**
T+	VP	FP
T-	FN	VN

La **sensibilité** estime la probabilité d'obtenir un test positif sachant qu'on est malade. La **spécificité** estime la probabilité d'obtenir un test négatif sachant qu'on est non-malade (Équation 43). Ces deux quantités décrivent la **validité intrinsèque du test** : elles sont mesurables dans des **conditions expérimentales**, lorsqu'on connait déjà le statut malade ou non du sujet étudié. On parle de validité intrinsèque car elle n'est pas influencée par la prévalence de la pathologie étudiée : la sensibilité n'est calculée que chez les malades, et la spécificité n'est calculée que chez les non-malades, rendant ces quantités insensibles à toute modification de la proportion de malades dans l'échantillon (Figure 122).

La **valeur prédictive positive** estime la probabilité qu'un sujet soit malade, sachant qu'il a un test positif. La **valeur prédictive négative** estime la probabilité qu'un sujet soit non-malade, sachant qu'il a un test négatif (Équation 43). Ces deux quantités décrivent la **validité extrinsèque du test** : elles seront utiles à la décision en pratique clinique : le clinicien réalise un test sur un patient pour se faire une idée de son statut malade ou non-malade. On parle de

validité extrinsèque car elle est fortement influencée par la prévalence de la pathologie étudiée (Figure 122).

$$Se = \hat{P}(T^+/M) = \frac{VP}{VP + FN}$$
$$Sp = \hat{P}(T^-/NM) = \frac{VN}{VN + FP}$$
$$VPP = \hat{P}(M/T^+) = \frac{VP}{VP + FP}$$
$$VPN = \hat{P}(NM/T^-) = \frac{VN}{VN + FN}$$

Équation 43. Calcul de Se, Sp, VPP et VPN dans un échantillon

Le théorème de Bayes permet de calculer VPP et VPN en fonction de Se Sp et de la prévalence (Équation 44). Ce calcul théorique ne vous sera d'aucune utilité, car en pratique vous disposerez de vos effectifs et pourrez calculer directement ces quatre quantités dans votre échantillon.

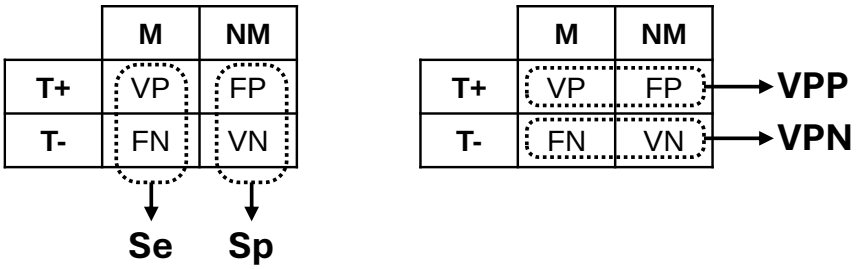


Figure 122. Données utiles au calcul de Se, Sp, VPP et VPN

$$VPP = \frac{Se \times P}{Se \times P + (1 - Sp)(1 - P)}$$
$$VPN = \frac{Sp \times (1 - P)}{Sp \times (1 - P) + (1 - Se) \times P}$$

Équation 44. Calcul de la VPP et de la VPN en fonction de Se, Sp et P (la prévalence de la maladie)

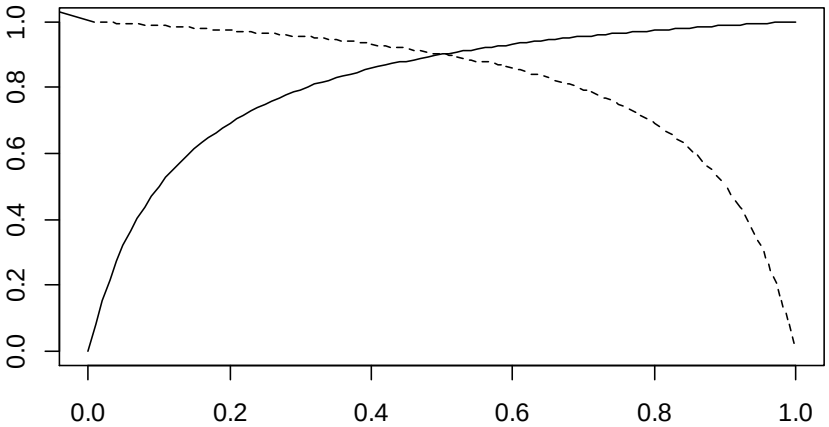


Figure 123. Exemple d'un test avec Sp=Se=90% : évolution en ordonnée de la VPP (trait plein) et de la VPN (trait pointillé) en fonction de la prévalence (abscisse)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

La Figure 123 illustre le comportement de la VPP et de la VPN en fonction de la prévalence de la pathologie, pour le cas d'un test qui est théoriquement bon (Sp=Se=0,9). On constate un effondrement catastrophique de la VPP pour les prévalences faibles, et de la VPN pour les prévalences élevées.

En pratique, en santé, **toutes les maladies sont rares** (fréquence inférieure à 5%). Il en résulte que, très souvent, les VPP des tests diagnostiques sont mauvaises, c'est pourquoi ils ne doivent être appliqués que dans les conditions recommandées. Ainsi, par exemple, le PSA ne doit pas être dosé en dépistage systématique du cancer de la prostate, mais seulement en cas de symptôme évocateur. De telles précautions permettent d'utiliser les tests dans des sous-groupes dans lesquels la prévalence de la pathologie est augmentée, ce qui permet d'améliorer la VPP de ces tests.

Nous **déconseillons fortement** de calculer la proportion de cas bien classés (Équation 45). Nous expliquerons pourquoi dans le chapitre dédié au coefficient kappa (voir chapitre 3.4.5.1 Préambule en page 205).

$$accuracy = \frac{VP + VN}{n}$$

Équation 45. Calcul de l'accuracy, ou proportion de concordance observée, fortement déconseillé

3.4.3 Outils de détection d'un nombre indéterminé d'événements

Dans certaines situations, on cherche à détecter des événements, mais il est difficile voire impossible de dire combien d'éléments sont analysés. En voici deux exemples.

Exemple 1 : On analyse un millier de courriers de sortie rédigés par des médecins. Chaque courrier est susceptible de contenir des concepts médicaux en nombre variable (ex : « appendicite », « anesthésie locale », etc.). On demande à des médecins d'annoter ces courriers, c'est-à-dire identifier et coder les concepts présents. On utilise un logiciel d'intelligence artificielle pour réaliser la même tâche. On considère que les médecins sont le gold standard, et on souhaite évaluer le logiciel au regard de ce gold standard.

Dans notre tableau de contingence (Figure 124 en page 199), le nombre de concepts détectés par les médecins devient le total de la première colonne (GS+). Le nombre de concepts détectés par le logiciel devient le total de la première ligne (T+). En examinant ces deux listes, on peut déduire le nombre de concepts détectés de manière consensuelle par l'un et par l'autre, qui sont des vrais positifs (VP). Les autres nombres (FP et FN) en découlent par soustraction. Il sera donc aisé de calculer Se et VPP. Cependant, nous n'avons aucun moyen de déterminer les nombres restants (total de GS-, total de T-, et donc VN) de manière univoque : faut-il prendre chaque mot du courrier ? faut-il prendre chaque groupe nominal ? Faut-il plafonner ce nombre à 20 concepts par courrier ? etc.

Exemple 2 : On analyse des suivis électrocardiographiques de patients sur 24h. L'objectif est de détecter des passages en fibrillation atriale (entre 0 et plusieurs dizaines par patient, de durées très variables). Le gold standard est un cardiologue. L'outil de détection est un outil d'IA qui analyse le tracé électrocardiographique.

De même, on saura dénombrer aisément les totaux de GS+ et T+ ainsi que leur intersection VP, mais on sera en difficulté pour déterminer les totaux de GS- et T-, ainsi que leur intersection VN.

Dans ces cas typiques de détection d'événements sur un nombre indéfini d'éléments, **soit le nombre de VN est inconnu, soit il est très élevé**. Par conséquent, la Sp et la VPN seront soit incalculables, soit toujours très élevées, et donc sans intérêt pour évaluer l'outil. **On calculera donc uniquement Se et VPP** (Figure 124) :

- La **Se** répond à la question : si un événement existe, quelle est la probabilité qu'il soit détecté ? Dans ce contexte, on l'appelle également **rappel** (*recall* en Anglais).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- La **VPP** répond à la question : si l'outil détecte un événement, quelle est la probabilité que cet événement soit réel ? Dans ce contexte, on l'appelle également **précision** (*precision* en Anglais)[19].

Par habitude en informatique de santé, cette attitude est également suivie pour toutes les détections informatisées, même lorsque le nombre de VN est calculable et modéré.

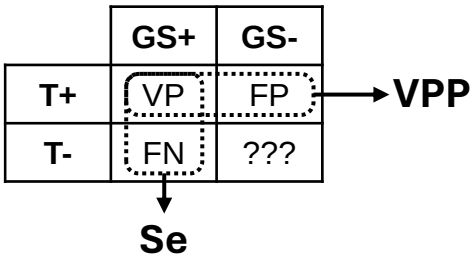


Figure 124. Tableau de contingence d'un outil de détection d'un nombre indéterminé d'événements

On notera que la VPP (précision) varie dans le même sens que la Sp (Équation 44), même si elle présente la particularité d'être négativement pénalisée lorsque la prévalence est faible (contrairement à la Sp).

Enfin, on produira fréquemment un indicateur synthétique : la **F-mesure**, ou **F-Score**, ou **mesure F1**, qui est une **moyenne harmonique de Se (rappel) et VPP (précision)**. Il s'agit d'une forme de moyenne, mais pas de la moyenne arithmétique : c'est l'inverse de la moyenne des inverses de Se et VPP (Équation 46). Pour prendre un exemple simple, si l'une vaut 1/2 et l'autre vaut 1/4, alors la F-mesure vaudra 1/3.

$$F-score = \left(\frac{Se^{-1} + VPP^{-1}}{2}\right)^{-1} = 2 \times \frac{Se \times VPP}{Se + VPP}$$

Équation 46. Calcul de la F-mesure

Dans un tableur, le calcul sera très simple (Figure 125) : on calcule Se et VPP comme étant de simples rapports, puis la fonction **moyenne.harmonique()** d'Excel ou Calc permet de calculer la F-mesure comme étant la moyenne harmonique de Se et VPP.

	A	B	C	D	E	F	G	H	I	J
13			Gold standard							
14			GS+	Gs-			Se = rappel =	0.950	=C15/C17	
15	Outil de	T+	950	250	1200		VPP = précision =	0.792	=C15/E15	
16	détection	T-	50	???						
17			1000				F-score =	0.864	=MOYENNE.HARMONIQUE(H14;H15)	

Figure 125. Calcul de la F-mesure avec un tableur

3.4.4 Tests diagnostiques à réponse quantitative

3.4.4.1 Préambule

Il arrive qu'on souhaite prédire une variable binaire (par exemple, le statut malade/non-malade), mais que le test qu'on utilise fournisse une réponse quantitative. C'est le cas par exemple du fibroscan, ou élastométrie hépatique, qui fournit une mesure en kilopascals, comprise entre 2 et 75. Une mesure élevée évoque une cirrhose hépatique.

[19] Ne pas confondre ce terme avec l'« accuracy », qui est un très mauvais indicateur, que nous définirons plus bas comme étant la « proportion de concordance observée ».

Pour binariser cette réponse, on pourra utiliser un seuil. Cependant, le choix de ce seuil devra faire l'objet d'un compromis. Par exemple, si le test est positif au-delà du seuil fixé, si on élève ce seuil on augmentera la spécificité, mais on diminuera la sensibilité. Réciproquement, si on abaisse ce seuil on augmentera la sensibilité, mais on diminuera la spécificité.

Nous exposerons ici la **courbe ROC** (*Receiver Operating Characteristic*) et son aire sous la courbe AUC (*Area under the Curve*), qui permettent de répondre simultanément à plusieurs questions :

- Visualiser le comportement du test en fonction du seuil choisi
- Fournir une aide pour choisir un bon seuil
- Savoir si, globalement, le test a une bonne capacité de discrimination, et ce sans avoir à choisir un seuil en particulier

Exemple : On mesure le taux de Bêta-HCG chez 500 femmes enceintes. On observe ultérieurement la présence d'une trisomie 21. Ce taux est un nombre réel positif. Le tableau de contingence suivant présente ce taux discrétisé en 3 catégories, à l'aide de 2 seuils (Tableau 21).

Tableau 21. Exemple : taux de Bêta-HCG et risque de trisomie 21 (T21)

Taux Bêta-HCG	Nb fœtus T21	Nb fœtus normal
x≥2	65	41
1,5≤x<2	15	62
x<1,5	20	297
Total	**100**	**400**

Cela nous laisse la possibilité d'utiliser deux seuils différents, chacun aboutissant à un calcul différent de Se et de Sp (Figure 126).

Seuil 1,5 :	M	NM
T+	80	103
T-	20	297
	Se =80/100 =0,8	Sp =297/400 =0,74

Seuil 2 :	M	NM
T+	65	41
T-	35	359
	Se =65/100 =0,65	Sp =359/400 =0,90

Figure 126. Choix de deux seuils différents

La courbe ROC représente chaque couple {Se ; Sp} sur un graphique carré de côté 1. Nous verrons dans la section suivante sa mise en œuvre, puis nous reviendrons sur l'interprétation détaillée par la suite.

3.4.4.2 Tracer et exploiter une courbe ROC, avec un tableur

Avec cet exemple, on atteint un peu la limite de ce qu'il est courant de réaliser avec un tableur.

La première étape consiste à choisir les seuils que nous utiliserons. Idéalement, et si vous souhaitez être certain de ne pas vous tromper, il faut utiliser comme seuils[20] :

- un nombre quelconque inférieur à toutes les valeurs rencontrées
- puis la liste dédoublonnée de toutes les valeurs rencontrées
- puis un nombre quelconque supérieur à toutes les valeurs rencontrées

Il est moins souhaitable, mais toujours possible, de choisir arbitrairement une séquence régulière de seuils (ex : les nombres compris entre 50 et 120, par pas de 10).

[20] En vous proposant cette méthode, vous aurez un seuil de trop, mais cela n'a aucune importance

Ensuite, nous procédons pour le premier seuil comme en Figure 127 (les formules sont ici affichées, mais ce seront bien sûr les valeurs qui seront visibles). Nous écrivons ici le seuil « en dur » en cellule D15, puis pour chaque individu nous calculons la sensibilité pour le seul individu (colonne D). Elle vaut :

- 1 si l'individu est malade et détecté (lorsque, dans notre exemple, la réponse est supérieure ou égale au seuil)
- 0 si l'individu est malade et non-détecté (lorsque, dans notre exemple, la réponse est inférieure au seuil)
- [vide] si l'individu est non-malade

Ensuite, pour chaque individu nous calculons la spécificité pour le seul individu (colonne E). Elle vaut :

- 1 si l'individu est non-malade et non-détecté (lorsque, dans notre exemple, la réponse est inférieure au seuil)
- 0 si l'individu est non-malade et détecté (lorsque, dans notre exemple, la réponse est supérieure ou égale au seuil)
- [vide] si l'individu est malade

Les formules des colonnes C et D sont écrites une seule fois, puis étendues sur tout le tableau de données. Faites bien attention aux dollars, qui permettent de figer certaines références : leur emploi approprié permet d'étendre les formules sans difficulté.

La sensibilité dans l'échantillon est la moyenne des sensibilités individuelles (cellule D16). La spécificité dans l'échantillon est la moyenne des spécificités individuelles (cellule D17).

	A	B	C	D	E
1	id	statut (GS)	reponse	Se	Sp
2	1	1	83	=SI(ET($B2=1;$C2>=D$15);1; SI(ET($B2=1;$C2<D$15);0; ""))	=SI(ET($B2=0;$C2<D$15);1; SI(ET($B2=0;$C2>=D$15);0; ""))
3	2	0	25	=SI(ET($B3=1;$C3>=D$15);1; SI(ET($B3=1;$C3<D$15);0; ""))	=SI(ET($B3=0;$C3<D$15);1; SI(ET($B3=0;$C3>=D$15);0; ""))
4	3	0	72	=SI(ET($B4=1;$C4>=D$15);1; SI(ET($B4=1;$C4<D$15);0; ""))	=SI(ET($B4=0;$C4<D$15);1; SI(ET($B4=0;$C4>=D$15);0; ""))
5	4	1	72	=SI(ET($B5=1;$C5>=D$15);1; SI(ET($B5=1;$C5<D$15);0; ""))	=SI(ET($B5=0;$C5<D$15);1; SI(ET($B5=0;$C5>=D$15);0; ""))
6	5	0	14	=SI(ET($B6=1;$C6>=D$15);1; SI(ET($B6=1;$C6<D$15);0; ""))	=SI(ET($B6=0;$C6<D$15);1; SI(ET($B6=0;$C6>=D$15);0; ""))
7	6	1	14	=SI(ET($B7=1;$C7>=D$15);1; SI(ET($B7=1;$C7<D$15);0; ""))	=SI(ET($B7=0;$C7<D$15);1; SI(ET($B7=0;$C7>=D$15);0; ""))
8	7	1	53	=SI(ET($B8=1;$C8>=D$15);1; SI(ET($B8=1;$C8<D$15);0; ""))	=SI(ET($B8=0;$C8<D$15);1; SI(ET($B8=0;$C8>=D$15);0; ""))
9	8	1	81	=SI(ET($B9=1;$C9>=D$15);1; SI(ET($B9=1;$C9<D$15);0; ""))	=SI(ET($B9=0;$C9<D$15);1; SI(ET($B9=0;$C9>=D$15);0; ""))
10	9	0	51	=SI(ET($B10=1;$C10>=D$15);1; SI(ET($B10=1;$C10<D$15);0; ""))	=SI(ET($B10=0;$C10<D$15);1; SI(ET($B10=0;$C10>=D$15);0; ""))
11	10	1	38	=SI(ET($B11=1;$C11>=D$15);1; SI(ET($B11=1;$C11<D$15);0; ""))	=SI(ET($B11=0;$C11<D$15);1; SI(ET($B11=0;$C11>=D$15);0; ""))
12	11	0	10	=SI(ET($B12=1;$C12>=D$15);1; SI(ET($B12=1;$C12<D$15);0; ""))	=SI(ET($B12=0;$C12<D$15);1; SI(ET($B12=0;$C12>=D$15);0; ""))
13	12	0	8	=SI(ET($B13=1;$C13>=D$15);1; SI(ET($B13=1;$C13<D$15);0; ""))	=SI(ET($B13=0;$C13<D$15);1; SI(ET($B13=0;$C13>=D$15);0; ""))
14					
15			Seuil	0	
16			Se	=MOYENNE(D2:D13)	
17			Sp	=MOYENNE(E2:E13)	

Figure 127. Calcul de Se et de Sp pour un seuil donné, sur deux colonnes du tableau de données

Une fois les deux colonnes (D et E) et les 3 cellules (D15:D17) correctement définies, nous réalisons simplement un copier-coller autant de fois que nécessaire vers la droite, puis nous écrivons les seuils qui nous intéressent dans la ligne 15 : tous les calculs se mettent à jour automatiquement (Figure 128).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

	id	statut (GS)	reponse	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp	Se	Sp
2	1	1	83	1		1		1		1		1		1		1		1		1		1		1		0	
3	2	0	25		0		0		0		0		0		1		1		1		1		1		1		1
4	3	0	72		0		0		0		0		0		0		0		0		1		1		1		1
5	4	1	72	1		1		1		1		1		1		1		1		1		0		0		0	
6	5	0	14		0		0		0		0		1		1		1		1		1		1		1		1
7	6	1	14	1		1		1		1		0		0		0		0		0		0		0		0	
8	7	1	53	1		1		1		1		1		1		1		1		0		0		0		0	
9	8	1	81	1		1		1		1		1		1		1		1		1		1		0		0	
10	9	0	51		0		0		0		0		0		0		1		1		1		1		1		1
11	10	1	38	1		1		1		1		1		1		0		0		0		0		0		0	
12	11	0	10		0		0		0		1		1		1		1		1		1		1		1		1
13	12	0	8		0		0		1		1		1		1		1		1		1		1		1		1

Seuil	0	8	10	14	25	38	51	53	72	81	83	90
Se	1	1	1	1	0.83	0.83	0.67	0.67	0.5	0.33	0.17	0
Sp	0	0	0.17	0.33	0.5	0.67	0.67	0.83	0.83	1	1	1

Figure 128. Calcul de Se et de Sp pour tous les seuils souhaités

Enfin, maintenant que nous disposons de Se et Sp pour chaque seuil, nous pouvons les remettre dans un tableau vertical, « proprement » (Figure 129). Ce tableau comporte une ligne par seuil, triés par ordre croissant. Cela peut se faire avec un copier-coller en valeur en supprimant les lignes vides, ou pour les plus expérimentés à l'aide d'une recherche automatique à l'aide de la fonction **rechercheh()**.

Dans ce tableau, nous ajoutons une colonne **1-Sp** (colonne AG en Figure 129), qui donnera l'abscisse du graphique. Nous traçons la courbe ROC à l'aide d'un nuage de points avec lignes droites, les points étant masqués. Nous définissons :

- En ordonnée, la sensibilité (Se)
- En abscisse, 1 moins la spécificité (1-Sp)

Une fois le graphique réalisé, il faudra veiller à le redimensionner pour qu'il ait une forme carrée, faute de quoi la lecture du graphique serait biaisée (voir à droite sur Figure 129).

Seuil	Se	Sp	1-Sp	$(1-Se)^2+(1-Sp)^2$	Surface
0	=RECI	=REC	=1-AF22	=(1-AE22)^2+(1-AF22)^2	=(AE22-AE23)*(AF22+AF23)/2

…

Seuil	Se	Sp	1-Sp	$(1-Se)^2+(1-Sp)^2$	Surface
0	1.00	0.00	1.00	1.000	0
8	1.00	0.00	1.00	1.000	0
10	1.00	0.17	0.83	0.694	0
14	1.00	0.33	0.67	0.444	0.069
25	0.83	0.50	0.50	0.278	0
38	0.83	0.67	0.33	**0.139**	0.111
51	0.67	0.67	0.33	0.222	0
53	0.67	0.83	0.17	**0.139**	0.139
72	0.50	0.83	0.17	0.278	0.153
81	0.33	1.00	0.00	0.444	0.167
83	0.17	1.00	0.00	0.694	0.167
90	0.00	1.00	0.00	1.000	XXXXX
					0.806

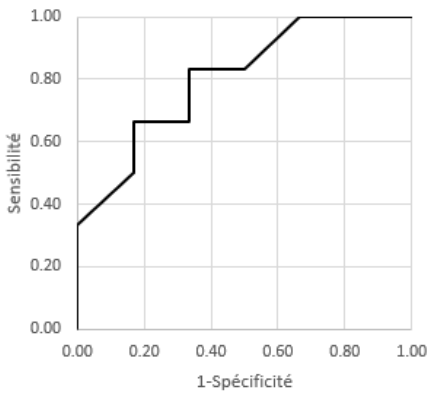


Figure 129. Tableau des seuils, tracé de la courbe ROC, choix d'un bon seuil, calcul de l'AUC (formules reportées en haut de la figure)

Sur cette courbe, nous pourrons également identifier le point le plus proche du coin en haut à gauche, c'est-à-dire correspondant à $Sp = Se = 1$. Ceci pourrait être réalisé graphiquement. Pour chaque seuil, nous pouvons également calculer la quantité $(1 - Se)^2 + (1 - Sp)^2$ qui est

le carré de la distance entre chaque point et le coin en haut à gauche (colonne AH en Figure 129). Le (ou les seuils) qui minimise cette distance est un seuil qui peut s'avérer utile en pratique. Dans notre exemple, les seuils 38 et 53 sont ceux qui minimisent cette distance.

Enfin, il est possible de calculer l'aire sous la courbe ROC (nous verrons plus bas comment l'interpréter). Il s'agit de l'aire en bas à droite de la courbe. Pour ce faire, pour chaque seuil à l'exception du dernier, nous écrivons la formule présentée en cellule AI22, puis nous la prolongeons vers le bas (hormis la dernière ligne). L'aire sous la courbe est la somme de ces quantités (voir cellule AI34 en Figure 129). Ici elle vaut 0,806 ou 80,6%.

3.4.4.3 La courbe ROC, en général

La courbe ROC est tracée dans un carré, chaque axe allant de 0 à 1. Elle relie un point par seuil. Chaque point correspond aux coordonnées suivantes :

- En ordonnée, la sensibilité (Se)
- En abscisse, 1 moins la spécificité (1-Sp)

Si la courbe était tracée d'une autre manière, cela ne changerait rien au fond, à condition de modifier en conséquence l'interprétation.

On notera que cette courbe devrait toujours être reliée à deux coins du graphique. Nous l'illustrons dans le cas où le test est considéré comme positif au-dessus du seuil choisi :

- Si le seuil est infiniment bas, alors tous les individus sont positifs, et donc Se=1 et Sp=0, donc Y=1 et X=1.
- Si le seuil est infiniment élevé, alors tous les individus sont négatifs, et donc Se=0 et Sp=1, donc Y=0 et X=0.

Naturellement, si le test est considéré comme positif en-deçà du seuil choisi, il suffit d'intervertir « infiniment bas » et « infiniment haut » dans les phrases ci-dessus.

Pour une raison similaire, la courbe progresse (ou stagne) sur chacun des deux axes au fur et à mesure que le seuil s'élève, mais ne revient jamais en arrière.

Si la série de contient pas d'ex-aequo, ou si les ex-aequo ont toujours le même statut (malade ou non-malade), alors la courbe aura un aspect de marches d'escalier : il n'y aura pas de segment oblique. Cet aspect, qui peut troubler, n'est pas du tout synonyme d'erreur.

On appelle **point parfait** le coin en haut à gauche, correspondant à $Sp = Se = 1$. On appelle **diagonale de la chance** la droite d'équation $y = x$, qui correspond à une réponse aléatoire, autrement dit un test sans intérêt diagnostique (Figure 130).

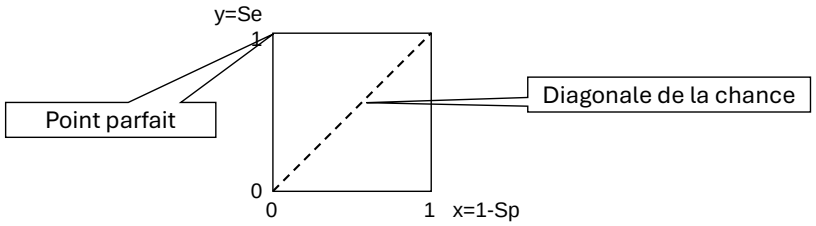


Figure 130. Zone graphique d'une courbe ROC

Pour évaluer la qualité diagnostique d'un test, sans avoir à choisir un seuil en particulier, on pourra mesurer la surface située en bas et à droite de la courbe : on l'appelle **l'aire sous la courbe, AUC** (Area Under the Curve). Cette aire est généralement comprise entre 0,5 et 1 :

- 0,5 correspond à un test répondant aléatoirement, donc dénué d'intérêt diagnostique
- 1 correspondant à un test parfait, ne se trompant jamais (Figure 131)

Il ne faut surtout pas interpréter cette AUC dans l'intervalle [0 ; 1], mais bien dans l'intervalle [0,5 ; 1]. Pour clarifier cela, certains proposent de calculer la **capacité prédictive (CP)** qui, elle, varie bien sur l'intervalle [0 ; 1] (Équation 47).

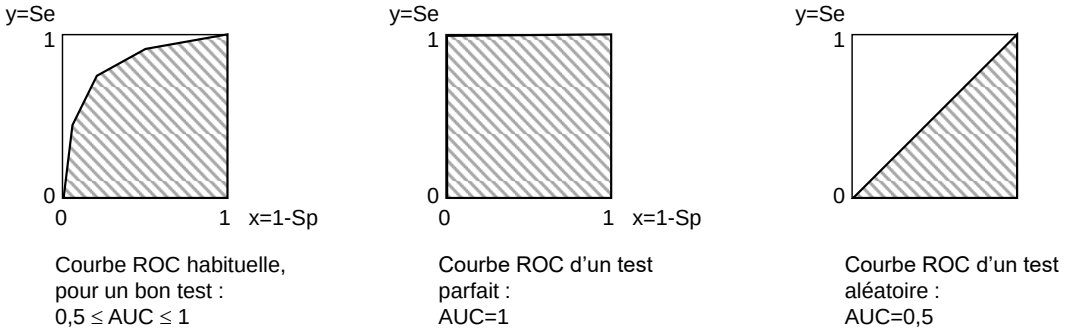


Figure 131. Interprétation de l'AUC d'une courbe ROC

Le calcul de l'aire sous la courbe ROC n'est pas complexe : on peut simplement décomposer la surface en un ensemble de trapèzes rectangles, décrivant la bande horizontale entre deux seuils (Figure 132). Si *Sp* ne change pas entre deux seuils il s'agit d'un rectangle, et si *Se* ne change pas entre deux seuils, il s'agit d'une bande de hauteur nulle : dans les deux cas la description par un trapèze rectangle reste correcte. La surface d'un tel trapèze est $hauteur \times (grandBase + petiteBase)/2$. L'AUC est la somme de toutes les surfaces ainsi calculées, comme nous l'avions fait en Figure 129, en page 202. L'Équation 47 formalise ce calcul.

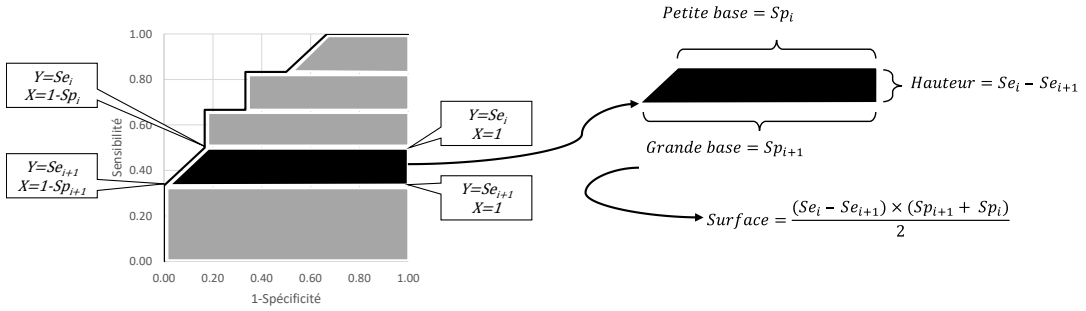


Figure 132. Principe du calcul de l'aire sous la courbe ROC

$$AUC = 0{,}5 \times \sum_{i=1}^{k-1} (Se_i - Se_{i+1}) \times (Sp_i + Sp_{i+1})$$

$$i = numéro\ du\ seuil, \quad k = nombre\ de\ seuils$$

$$CP = 2 \times AUC - 1$$

Équation 47. Calcul de l'aire sous la courbe et de la capacité prédictive d'une courbe ROC

Une méthode permet de calculer **l'intervalle de confiance à 95% de l'AUC**. Sa variation est complexe, car elle dépend certes des qualités du test, mais également du nombre de seuils et des éléments géométriques qui découlent de tout cela. Pour cette raison, elle est calculée par Bootstrap et non en utilisant une méthode asymptotique.

Enfin, la courbe ROC peut être utile pour déterminer un seuil. En l'absence de tout autre argument, on peut choisir le seuil correspondant au **point le plus proche du point parfait**. Pour ce faire, nous calculons pour chaque seuil la quantité $(1 - Se)^2 + (1 - Sp)^2$ qui est le

carré de la distance entre le point correspondant et le point parfait. Le (ou les seuils) qui minimise cette distance est un seuil qui peut s'avérer utile en pratique.

Cette approche suppose que la sensibilité et la spécificité ont le même poids dans le choix du seuil. En pratique, il n'en est rien, car dans la pratique clinique ce sont plutôt la VPP et la VPN qui comptent, or elles dépendent de la prévalence mais, surtout, le poids qu'on souhaite leur donner dépendra des implications pratiques, cliniques, éthiques et médico-économiques. Pour un premier test de dépistage, on souhaitera souvent privilégier la sensibilité. Pour un test de confirmation, on souhaitera souvent privilégier la spécificité. En conclusion, même s'il est fréquent de présenter ce seuil choisi sur critère géométrique, il faut garder à l'esprit que **les arguments métier sont toujours plus forts que cet argument purement géométrique**.

Enfin, tout ce qui a été exposé pour une courbe ROC présentant Se et Sp, est également applicable à une courbe présentant d'autres métriques. Ainsi, en informatique de santé, il est fréquent de présenter des **courbes Se et VPP**, car dans de nombreux cas, la spécificité n'est pas calculable, ou sans intérêt, comme nous l'avons vu au chapitre 3.4.3 Outils de détection d'un nombre indéterminé d'événements en page 198.

3.4.5 Accord entre deux juges, sans gold standard

3.4.5.1 Préambule

Nous nous intéresserons ici à l'analyse de deux variables binaires (ou qualitatives) qui représentent chacune un jugement visant à classer les individus dans des catégories, comme « malade » versus « non-malade ». A la différence des chapitres précédents (tests diagnostiques binaires, outils de détection, tests diagnostiques quantitatifs), la différence majeure est qu'**on ne sait pas quel est le véritable statut** des individus : aucune des deux informations n'est le gold standard. On s'intéresse seulement à **quantifier la concordance** entre les deux jugements.

La **concordance** est la conformité, la similitude entre deux (ou plus) informations se rapportant au même objet. Ce terme couvre les informations qualitatives ou quantitatives, nous nous intéresserons ici uniquement aux informations qualitatives.

Attention : la notion de concordance diffère de celle de liaison statistique ou de corrélation, car la concordance sous-tend un véritable appariement : il s'agit de deux évaluations d'un même caractère.

L'existence d'une concordance implique l'existence d'une corrélation.
Exemple : le diagnostic d'infarctus du myocarde réalisé par deux méthodes

L'existence d'une corrélation n'implique pas forcément l'existence d'une concordance.
Exemple : le diagnostic d'infarctus du myocarde en fonction du sexe

Les cadres d'utilisation de la concordance sont nombreux :

- Concordance **entre deux méthodes**, deux appareils, deux tests, deux réactifs, etc.
- Concordance **entre deux observateurs**, utilisant la même méthode de mesure : **reproductibilité inter-opérateurs**
- Concordance **entre deux itérations** de la même méthode de mesure : **reproductibilité intra-opérateur, répétabilité de la mesure**

Il faut prendre conscience que la simple **proportion d'accord** (ou proportion de concordance observée) entre deux jugements **n'est pas un critère acceptable**, car elle n'est interprétable que lorsque les deux statuts sont bien équilibrés. En voici un contrexemple.

Pour dépister l'infection au VIH, on utilise deux tests diagnostiques : le test A, qui est un test de référence, et le test B, qui est défectueux et est toujours négatif. Calculons la proportion de concordance observée entre ces deux tests.

	A+	A-	*total*
B+	0	0	*0*
B-	4	996	*1000*
total	*4*	*996*	*1000*

Ces deux tests sont en accord parfait dans $(0+996)/1000 = 99{,}6\%$ *des cas : ce nombre parait excellent, et pourtant un des deux tests est défectueux.*

L'exemple qui précède illustre pourquoi la proportion de concordance observée ne doit jamais être utilisée. Elle est généralement utilisée par les auteurs qui souhaitent berner les lecteurs, et les impressionner par des nombres flatteurs, qui servent souvent à dissimuler de piètres résultats scientifiques.

Nous nous intéresserons ici uniquement au **coefficient Kappa de Cohen**[38], qui permet de mesurer la concordance entre **deux jugements catégoriels** (souvent binaires).

3.4.5.2 Exemple de calcul du coefficient Kappa, en pratique

Nous représentons tout d'abord un tableau de contingence présentant les cas étudiés, classés en colonnes selon un premier jugement, et en lignes selon un deuxième jugement (colonnes ABCD en Figure 133). Ce tableau pourra être généré à l'aide d'un tableau croisé dynamique.

Nous créons ensuite un deuxième tableau de même structure, dont les totaux de lignes et de colonnes sont les mêmes que le tableau précédent (colonnes FGHI en Figure 133). Dans ce tableau, nous complétons les cases de la diagonale d'accord (A+ & B+ d'une part, A- & B- d'autre part) par les effectifs théoriques, que nous aurions observés si les deux jugements avaient été indépendants. Chaque effectif, potentiellement non-entier, est le produit du total de la ligne par le total de la colonne, divisé par le total général. On pourrait compléter tout le tableau mais ce ne sera pas nécessaire.

Nous calculons ensuite la proportion de concordance observée P_0, comme étant la proportion représentée par la diagonale des cas concordants dans le tableau d'effectifs observés (voir cellules K3 et L3 en Figure 133).

Nous calculons ensuite la proportion de concordance théorique P_c, comme étant la proportion représentée par la diagonale des cas concordants dans le tableau d'effectifs théoriques (voir cellules K4 et L4 en Figure 133).

Nous calculons enfin le coefficient Kappa comme étant $K = (P_o - P_c)/(1 - P_c)$ (voir cellules K5 et L5 en Figure 133).

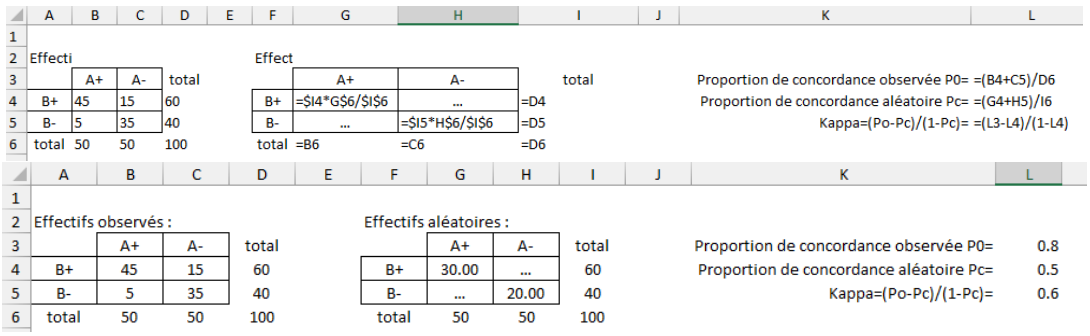


Figure 133. Calcul du coefficient Kappa de Cohen avec un tableur (haut : formules ; bas : résultats)

Ce coefficient s'interprète comme une proportion d'accord entre les deux juges. En cas de valeur négative, ce qui peut accidentellement survenir, il suffit de l'interpréter comme $K = 0$. On notera que l'intervalle d'interprétation est [0 ; 1], mais l'intervalle de variation théorique est plutôt]-inf ; 1].

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.4.5.3 Calcul du coefficient Kappa de Cohen, en général

La Figure 134 schématise les étapes de calcul du coefficient Kappa. Nous reprenons ici les étapes numérotées sur cette figure.

La première étape consiste à représenter les effectifs observés (1) sous forme d'un tableau de contingence croisé. Nous nous intéresserons uniquement à la diagonale formée par les effectifs concordants. On peut ensuite calculer la proportion de concordance observée P_o (2), qui ne doit jamais être publiée ni commentée. Comme nous l'avons vu en introduction, elle peut être élevée dès lors que les effectifs sont déséquilibrés, ce qui ne signifie aucunement que les tests sont concordants. Nous calculons ensuite les effectifs théoriques, que nous aurions dû observer sous l'hypothèse nulle, qui est que les deux jugements sont indépendants (3). Ces effectifs théoriques ne sont pas arrondis à l'entier. Ils permettent de calculer la proportion de concordance théorique (4). Le **coefficient Kappa de Cohen** peut enfin être calculé (5). La plupart du temps, le calcul s'arrête là. Ce coefficient **s'interprète comme un indicateur d'accord entre les deux juges, sur l'intervalle [0 ; 1]**. Une valeur nulle (ou négative, ce qui peut arriver) indique une absence totale d'accord entre les deux jugements (ex : au moins un des deux répond aléatoirement), tandis qu'une valeur de 1 indique un accord parfait entre les deux jugements.

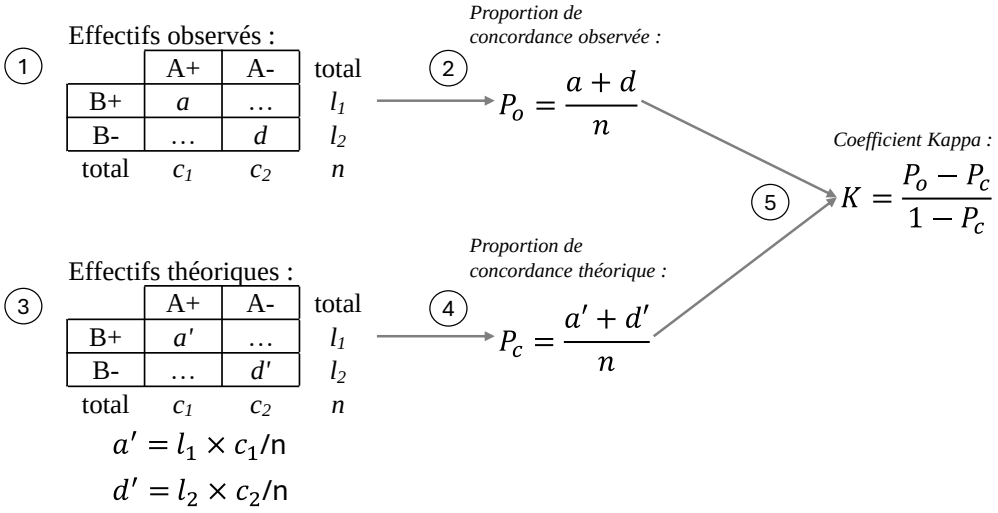
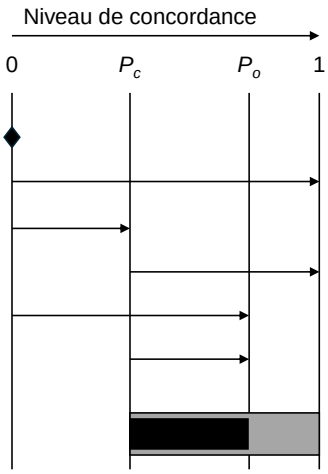


Figure 134. Calcul du coefficient Kappa

3.4.5.4 Mieux comprendre le coefficient Kappa de Cohen

3.4.5.4.1 Interprétation détaillée de la valeur du Kappa

L'interprétation du coefficient Kappa s'explique selon la Figure 135. On souhaiterait, en partant de zéro, obtenir une concordance totale, c'est-à-dire une proportion de concordance égale à 1. Or le simple hasard explique que la concordance soit déjà égale à P_c. Il nous reste donc un chemin à parcourir de 1-P_c. Nous observons une concordance observée P_o, qui finalement constitue un gain par rapport au hasard de (seulement) P_o-P_c. La proportion de chemin ainsi parcouru est de (P_o-P_c)/(1-P_c) : c'est la valeur du coefficient Kappa.



Valeur	Interprétation
0	Niveau zéro : absence de concordance
1	Concordance totale, qu'on souhaiterait atteindre
P_c	Valeur obtenue par le seul fait du hasard
$1 - P_c$	Chemin restant à parcourir
P_o	Valeur obtenue dans l'échantillon
$P_o - P_c$	Gain par rapport au hasard
$K = \frac{P_o - P_c}{1 - P_c}$	Chemin ainsi parcouru, parmi le chemin qu'il restait à parcourir

Figure 135. Compréhension du calcul du coefficient Kappa

L'auteur de ce coefficient a proposé les intervalles suivants pour son **interprétation** (après arrondi au dixième) :

- Concordance excellente si $0{,}8 < K \leq 1$
- Concordance bonne si $0{,}6 < K \leq 0{,}8$
- Concordance moyenne si $0{,}4 < K \leq 0{,}6$
- Concordance faible si $0{,}2 < K \leq 0{,}4$
- Concordance négligeable si $0{,}0 \leq K \leq 0{,}2$

On observe que les proportions de concordance sont **doublement symétriques**, et il en est donc ainsi du coefficient Kappa lui-même :

- Il est possible d'intervertir les deux jugements, sans modifier la valeur du coefficient Kappa
- Il est possible d'inverser les réponses pour chaque jugement (ex : malade versus non-malade), sans modifier la valeur du coefficient Kappa, à condition de faire de même pour les deux jugements

3.4.5.4.2 Plage de valeurs du coefficient Kappa

Intéressons-nous tout d'abord aux valeurs théoriques du coefficient Kappa. Il ne peut excéder 1, mais il peut théoriquement atteindre une **valeur minimale négative**, et potentiellement assez éloignée de 0 si la proportion de concordance observée est nulle (Équation 48). Concrètement, toute valeur négative doit être interprétée **comme une valeur nulle**, à moins qu'il ne s'agisse d'une erreur d'encodage.

$$K_{min} = \frac{-P_c}{1 - P_c}$$

Équation 48. Valeur minimale du coefficient Kappa

3.4.5.4.3 Inférence statistique

L'interprétation du coefficient Kappa se fera sur une valeur estimée dans un échantillon, elle sera donc plus prudente si l'échantillon est de petite taille, et plus confiante dans le cas contraire. Il est également possible de réaliser plusieurs types d'**inférence statistique** :

- calculer l'IC95 du coefficient Kappa en population
- tester la nullité du coefficient Kappa en population (ce qui, au fond, n'est pas très utile car on sait déjà que les deux jugements ne sont pas indépendants)
- tester si en population deux coefficients Kappa sont significativement différents

Ces inférences statistiques ne sont pas très fréquentes dans la littérature, et ne seront pas décrites ici. Vous n'aurez pas besoin d'y recourir.

3.4.5.4.4 En cas de mauvaise concordance : vrai désaccord ou mauvaise calibration ?

Cependant, si la valeur du coefficient Kappa est faible, certains voudront savoir si c'est le fait d'un vrai désaccord entre les deux jugements, ou si c'est simplement parce que l'un est moins souvent positif que l'autre. Autrement dit, si on recalibrait les deux jugements à l'identique, le coefficient Kappa serait-il meilleur ?

Pour ce faire, reprenons la Figure 134, qui montrait comment le coefficient Kappa est calculé. Nous la complétons vers le bas, et obtenons ainsi la Figure 136.

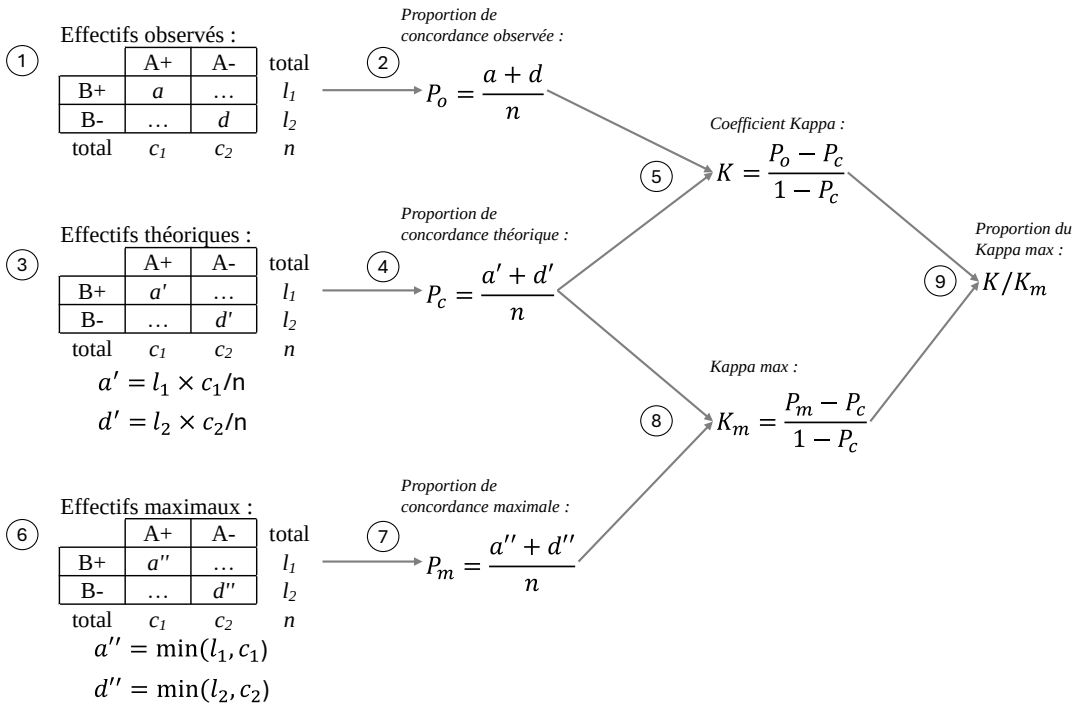


Figure 136. Calcul du coefficient Kappa (1 à 5) puis de la proportion du Kappa max (6 à 9)

Nous commentons ci-dessous la Figure 136 à l'aide des numéros qui s'y trouvent. On peut calculer les effectifs maximaux qu'on aurait pu obtenir sur la diagonale de concordance (6), qui sont à chaque fois la plus petite valeur parmi le total de la ligne et le total de la colonne. On aurait, au mieux, pu obtenir ces effectifs, et donc obtenir la proportion de concordance maximale P_m (7), et donc au mieux le coefficient Kappa max (valeur maximale du Kappa compte tenu de la calibration actuelle des deux jugements) K_m (8). On pourra ainsi commenter le coefficient Kappa en **proportion du Kappa max** (9).

L'idée sous-jacente est la suivante. Si le coefficient Kappa n'est pas satisfaisant, deux cas de figure sont possibles :

- Si la proportion du Kappa max est néanmoins élevée, cela signifie que le problème vient principalement d'un **problème de calibration** : un jugement est beaucoup plus souvent positif que l'autre.
- Si la proportion du Kappa max est faible, cela signifie que les tests, au fond, ne sont pas d'accord, et qu'une révision par un quelconque moyen de leur calibration ne sera pas bénéfique. Le problème est alors bien un **problème de désaccord**.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3.4.5.4.5 Effet de la rareté des caractères étudiés

On pourra également noter que le coefficient Kappa est pénalisé par les prévalences extrêmes, et notamment par les prévalences faibles. Nous rappellerons que, en santé, toutes les maladies sont rares.

Comme préalable, nous rappellerons que si le coefficient Kappa peut mesurer la concordance entre deux jugements en ignorant la réalité, il peut tout aussi bien être utilisé pour **mesurer l'accord global entre un test et un gold standard**, comme le fait la F-mesure par exemple, car « qui peut le plus peut le moins ».

La Figure 137 représente le cas où un des deux jugements est le gold standard, et où l'autre jugement lui est lié par une sensibilité et une spécificité de 90%. Nous représentons le coefficient Kappa en trait plein, et la F-mesure (moyenne harmonique de la sensibilité et de la valeur prédictive positive) en pointillés. En se rappelant que la plupart des pathologies en santé sont rares et ont une fréquence inférieure à 5%, on notera que le Kappa et la F-mesure sont **fortement pénalisées par les prévalences faibles**, et atteignent des valeurs relativement proches. Ce comportement est dénoncé par certains comme une faiblesse du Kappa. Nous le considérons en réalité **comme une qualité en pratique**, car selon nous les valeurs obtenues rendent réellement compte de l'**utilisabilité clinique** des jugements dans de tels contextes.

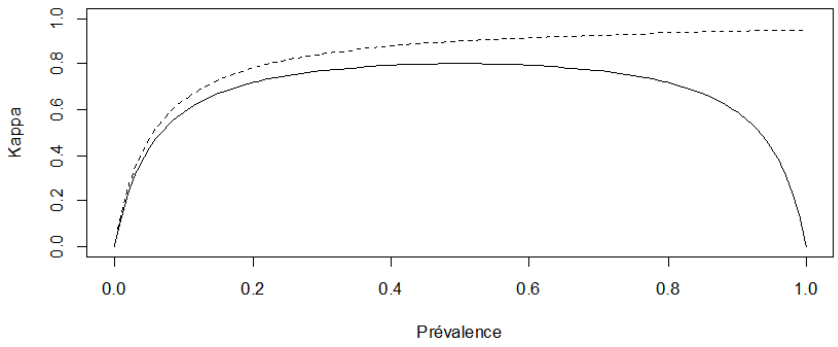


Figure 137. Valeur du coefficient Kappa (ou de la F-mesure en pointillés) en fonction de la prévalence, dans le cas où Se=Sp=0,9

3.4.5.4.6 Jugements comportant plus de deux modalités

Il est possible de calculer un coefficient Kappa entre deux jugements comportant 3 ou plus modalités. Le fondement en est simple : **il suffit que P_o et P_c soient calculés selon les mêmes règles**, le coefficient Kappa sera alors interprétable.

Voyons un premier exemple **sans pondération** (Figure 138) : nous nous intéressons uniquement aux cas d'accord parfait entre les deux jugements, cas qui sont situés sur la diagonale. Nous calculons P_o comme étant la proportion d'accords parfaits (2), puis calculons les effectifs théoriques sous l'hypothèse d'indépendance entre les deux jugements (3), puis P_c de la même manière que P_o (4), et enfin le coefficient Kappa comme précédemment (5).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

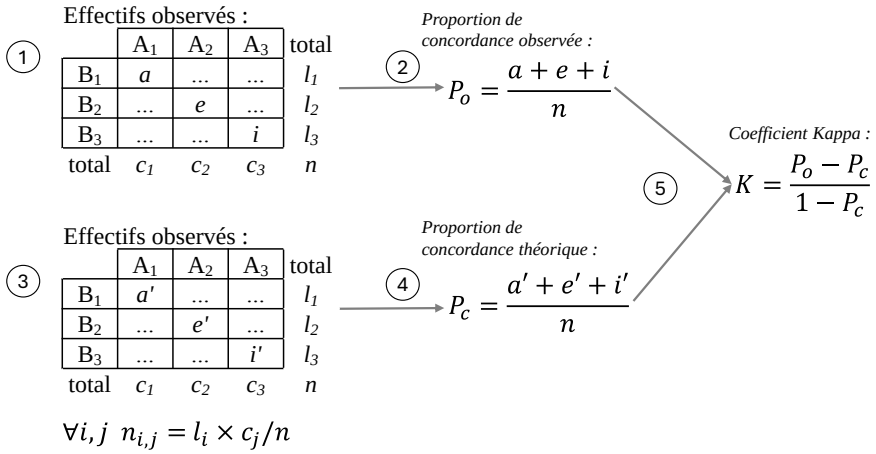


Figure 138. Calcul du coefficient Kappa entre deux jugements à 3 modalités, sans pondération

Il est également possible de calculer un coefficient Kappa **avec pondération**[39]. Dans l'exemple de la Figure 139, nous considérons que les effectifs de la diagonale constituent un accord parfait, pondéré à 1, mais que les effectifs proches de la diagonale correspondent à un accord partiel, pondéré à 0,5, tandis que les derniers effectifs correspondent à un désaccord. Il nous suffit de les prendre en compte avec ces pondérations de la même manière dans P_o et P_c, puis la formule du coefficient Kappa reste inchangée.

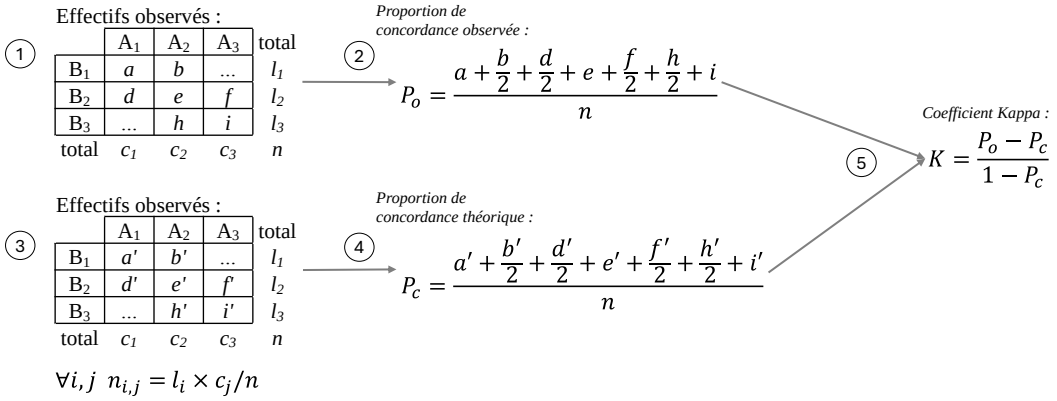


Figure 139. Exemple de calcul du coefficient Kappa entre deux jugements à 3 modalités, avec pondération

Au final, tout ce qui est cohérent est possible, tant que le même calcul est utilisé tant pour P_o que pour P_c.

3.4.5.4.7 Extension à plus de deux jugements

Enfin, si vous devez évaluer plus de deux jugements simultanément, l'attitude la plus fréquente consiste à calculer **un coefficient Kappa par paire** de jugements. Par exemple, pour 3 jugements A, B et C, on calculera 3 coefficients Kappa : A&B, B&C, C&A.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

4 Analyses statistiques multivariées, en bref

Les analyses multivariées sont des méthodes qui permettent d'analyser plusieurs variables simultanément. On sépare habituellement ces méthodes en deux familles.

Les **méthodes non-supervisées** considèrent toutes ces variables en même temps et au même plan, et visent à identifier des sous-groupes d'individus plus ou moins semblables, au vu de ces variables. On citera par exemple la méthode des K plus proches voisins, les nuées dynamiques, les classifications hiérarchiques ascendantes, les analyses en composantes principales, etc. Ces méthodes sont souvent utilisées comme méthodes exploratoires, ou pour la recherche d'individus similaires dans une perspective d'intelligence artificielle. Elles sont assez peu utilisées dans les publications de recherche clinique ou épidémiologique. Leur interprétation n'est pas aisée et, d'ordinaire, elles ne sont pas utilisées dans les mémoires académiques en santé.

Les **méthodes supervisées, très utilisées en recherche biomédicale,** s'intéressent à **une variable à expliquer**, et tentent de l'expliquer ou de la prédire avec l'ensemble des autres variables. Les méthodes les plus fréquentes en recherche sont les **régressions**, on peut également citer les **arbres**. Les régressions sont extrêmement souvent rencontrées dans les publications de recherche clinique ou épidémiologique, en particulier dans les recherches suivant un **design observationnel**[21]. Lorsqu'on s'intéresse à l'effet d'un facteur sur une pathologie, afin d'identifier l'effet de ce facteur ajusté sur d'autres facteurs (l'âge, le sexe, les pathologies chroniques, etc.), on lancera généralement une régression pour prédire ou expliquer la pathologie, en fonction d'un ensemble de variables, qui mélangeront le facteur étudié et tous les **facteurs d'ajustement**. Le plus souvent, du point de vue mathématique, aucune différence ne sera faite entre le facteur étudié et les facteurs d'ajustement : ils sont mis dans le même sac des « variables explicatives ».

Parmi les régressions, le choix de la méthode dépend uniquement du **type de la variable à expliquer**, comme illustré en Figure 140. De manière simplifiée, toutes ces méthodes utilisent une combinaison linéaire des variables explicatives, pour tenter de prédire, directement ou indirectement, la variable à expliquer.

De manière comparable, parmi les arbres, le choix de la méthode dépend presqu'uniquement du type de la variable à expliquer, comme illustré en Figure 141. De manière simplifiée, toutes ces méthodes utilisent les variables explicatives pour identifier des sous-groupes d'individus, qui soient relativement homogènes au sein du groupe et suffisamment différents entre les groupes, pour tenter de prédire, directement ou indirectement, la variable à expliquer.

[21] Inversement, dans les essais randomisés contrôlés, qui sont les études les plus « haut de gamme », l'affectation aléatoire des individus est censée gommer les différences entre individus, au point que l'analyse répondant à l'objectif principal est généralement une simple analyse bivariée (le plus souvent, un test du Khi² ou un test de Student).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

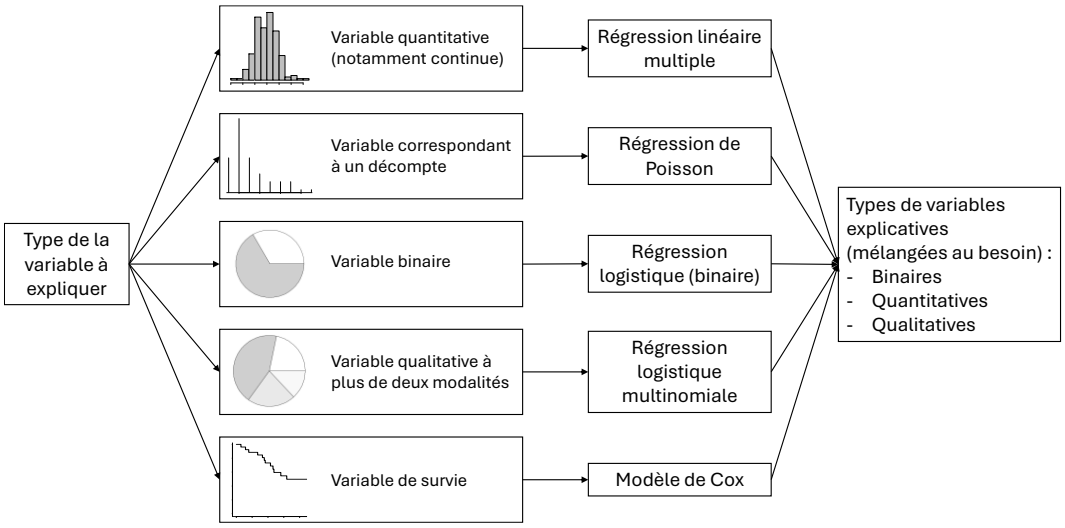


Figure 140. Différents types de régressions, en fonction de la variable à expliquer

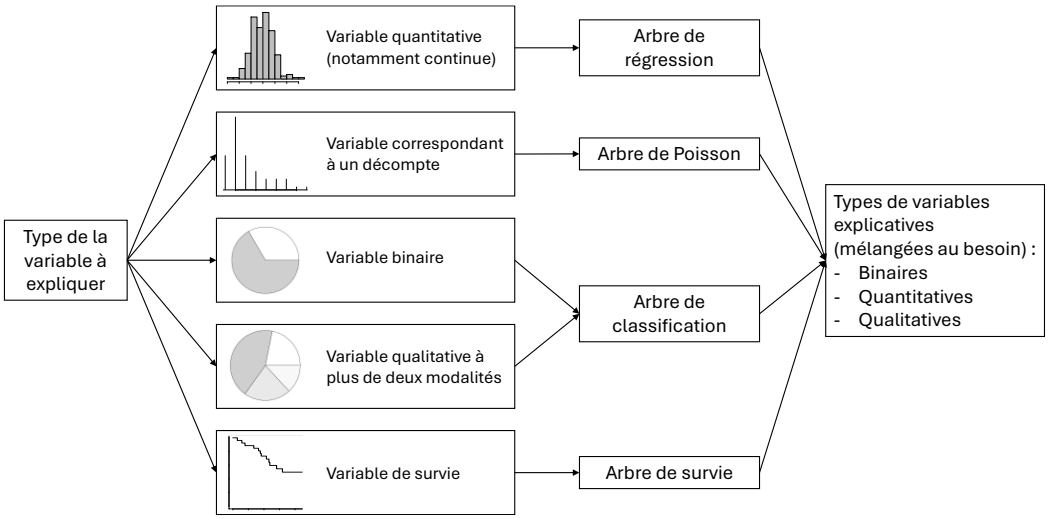


Figure 141. Différents types d'arbres, en fonction de la variable à expliquer

Toutes ces méthodes sont relativement complexes notamment en ce qui concerne leur préparation, leur mise en œuvre, leur validation et leur interprétation. Si vous souhaitez y recourir, vous devrez faire appel à un biostatisticien.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

5 Réflexions sur certains tests statistiques ou leur paramétrage

Dans ce chapitre, nous aborderons certains éléments de discussion sur les tests statistiques.

Nous verrons que, bien qu'ils soient hautement valides, certains tests ne sont pas recommandés, soit parce qu'ils s'utilisent dans des études de piètre qualité méthodologique (5.1 Tests de comparaison à une norme (ici-ailleurs) en page 214 et 5.2 Tests appariés dans un seul groupe, avant-après en page 215), soit parce que leur utilisation est impropre (5.3 Tests qu'on réalise en espérant ne pas rejeter H0 en page 216).

Nous évoquerons les classifications des tests (5.4 Test paramétrique ou non-paramétrique ? Asymptotique ou exact ? en page 217).

Puis nous aborderons un problème essentiel de paramétrage des tests statistiques : 5.5 Test unilatéral ou bilatéral ? Et pourquoi 5% ? en page 219.

Enfin, nous expliquerons ce qu'est la correction de Bonferroni et quand elle doit être appliquée : 5.6 Correction de Bonferroni en page 222.

5.1 Tests de comparaison à une norme (ici-ailleurs)

Nous avons vu trois tests qui permettent de comparer un paramètre observé (la moyenne ou la proportion) à une valeur attendue :

- Le test binomial
- Le test du Khi² d'adéquation
- Le test de Student de comparaison d'une moyenne observée à une moyenne attendue

D'un point de vue mathématique, ces tests sont hautement valides. D'un point de vue pédagogique, ils permettent une bonne introduction aux tests statistiques.

Cependant, d'un **point de vue méthodologique**, les recherches qui les utilisent sont **généralement de piètre qualité**.

En voici une illustration :

Nous disposons d'un médicament homéopathique miraculeux, le Phallus Elongator 22CH (PE22CH), qui est administré pour rallonger le pénis, y compris en état flaccide (c'est-à-dire au repos).

Nous mesurons le pénis de 50 patients qui ont pris du PE22CH. Cette étude est réalisée au mois de juin. Les participants sont installés dans une cabine en début d'après-midi et sont invités à mesurer eux-mêmes leur membre.

La taille moyenne obtenue (12,5 cm) est comparée à la taille moyenne publiée dans une autre étude de référence (8,7 cm). Dans cette étude de référence, 50 participants ont été inclus le matin en février lors du service militaire, et leur pénis a été mesuré par un infirmier du Service de Santé des Armées.

Le test statistique réalisé confirme que la taille moyenne des patients sous PE22CH est significativement différente de la taille moyenne de l'étude de référence (p<0,001).

On comprend immédiatement que, si les conditions de mesure ne sont pas les mêmes dans les deux études, il n'est pas étonnant que les résultats observés soient différents. Les différences sont nombreuses : la température, l'heure de la journée, les conditions de stress de la mesure, le protocole même de mesure, l'objectivité ou subjectivité de la mesure, etc. Il serait évidemment erroné d'en conclure que cette différence de moyenne est imputable au traitement, le PE22CH, alors que les conditions de mesure ne sont pas les mêmes.

Le principal problème des études comparant un paramètre observé à une valeur issue d'une autre étude, est que le protocole de mesure n'est jamais strictement identique dans les deux bras. Les différences observées sont probablement plus le fait des **différences de protocole**, que du phénomène sous-jacent. **Il n'est donc pas recommandé d'utiliser de tels tests**.

5.2 Tests appariés dans un seul groupe, avant-après

Nous avons évoqué trois tests de comparaison de paramètres appariés au sein d'un unique groupe :

- Le test de Student apparié (comparaison de deux moyennes appariées)
- Le test de Wilcoxon apparié (comparaison de deux moyennes appariées)
- Le test de McNemar (comparaison de deux proportions appariées)

D'un point de vue mathématique, ces tests sont hautement valides. Cependant, d'un **point de vue méthodologique**, les recherches qui les utilisent pour réaliser une **comparaison avant-après dans un seul groupe** sont **généralement de piètre qualité**. Cette critique ne s'adresse pas aux comparaisons gauche-droite par exemple, qui sont tout à fait valides.

En voici une illustration :

Nous disposons d'un médicament homéopathique miraculeux, le Phallus Elongator 22CH (PE22CH), qui est administré pour rallonger le pénis, y compris en état flaccide (c'est-à-dire au repos).

Nous incluons 100 patients. Nous mesurons leur pénis le matin. Puis nous leur administrons le PE22CH. Nous leur servons un repas le midi, puis évaluons leur longueur pénienne en début d'après-midi (temps requis pour que le PE22CH « actionne les ondes et les bienfaits, et élimine les toxines », tout ça tout ça, vous savez bien). Le protocole de mesure est, autant que possible, identique le matin et l'après-midi. Il est réalisé par les mêmes personnes, qui ont d'ailleurs participé au repas le midi. La taille mesurée est en moyenne plus élevée qu'avant la prise du PE22CH, ce qui est confirmé par un test statistique ($p<0,001$).

On comprend immédiatement que, même si la mesure a été réalisée selon le même protocole, de nombreux facteurs ont changé :

- la mesure a lieu en début d'après-midi, instant de la journée où le pénis a tendance à être le plus long hors stimulation, du fait de l'activation du système nerveux parasympathique durant la digestion
- les participants, stressés au début de la journée, se sont détendus
- les participants ont pu discuter avec les personnes chargées de les mesurer
- la mesure étant répétée, les participants sont moins stressés (banalisation)

L'administration d'un produit, aussi inactif soit-il, semble donc être associé à une efficacité mesurable. Il serait évidemment erroné d'en conclure que cette différence de moyenne est imputable au traitement, le PE22CH, alors que les circonstances sont très différentes, malgré l'application d'un protocole de mesure identique. On obtiendrait exactement le même résultat avec un test de quotient intellectuel également.

Deux exemples de comparaison avant-après sont bien connus en population, et aboutissent à des conclusions erronées.

*Le premier exemple concerne **les rhumes et l'homéopathie**. Lorsqu'un patient présente un rhume, il peut décider d'acheter un médicament homéopathique. Il lui faut généralement 1 à 2 jours pour acheter ce médicament. C'est également la durée de la phase la plus désagréable du rhume. Ainsi, dès la première prise du médicament, le rhume semble s'améliorer.*

*Un deuxième exemple concerne le fameux **bracelet du surfeur**, bracelet plastique qui prétend appuyer sur un point d'acupuncture et améliorer l'équilibre. Le vendeur demande au surfeur de se tenir debout pieds joints, le bouscule, et le surfeur perd l'équilibre. Il lui propose d'enfiler*

le bracelet magique, reproduit la même manœuvre et, surprise, le surfeur qui sait exactement ce qui va se passer, est nettement plus stable. C'est ici la répétition de l'expérience qui semble confirmer l'effet magique du bracelet : un homme averti en vaut deux.

Le principal problème des **études avant-après dans un seul groupe**, est que, même si le protocole de mesure se veut inchangé, il existe :

- un **effet propre du temps qui passe** (certaines maladies guérissent spontanément, comme les rhumes). Cet effet est encore plus marqué lorsque le fait de présenter une maladie aiguë est le critère d'inclusion
- un **effet placebo**, dans lequel le conditionnement du sujet fait que l'administration du traitement diminue réellement les symptômes
- un **effet « cocooning »** lié à la prise en charge dans une étude (certains symptômes réels diminuent du simple fait de l'écoute : anxiété, tremblement, douleur, insomnie, constipation, etc.)
- un **effet lié à la répétition** : **banalisation** des mesures, **gain d'expérience** par le sujet

Les différences avant-après observées sont probablement plus le fait d'un ou plusieurs de ces facteurs, que du traitement évalué. Les tests de comparaison appariés dans un seul groupe sont principalement utilisés dans ces études avant-après. **Il n'est donc pas recommandé de les utiliser**.

Encore une fois, ces critiques ne s'appliquent pas aux **comparaisons gauche-droite, qui restent valides**. Dans ces études, si nécessaire, on veillera à **équilibrer ou randomiser** l'ordre des mesures : tantôt gauche puis droite, tantôt droite puis gauche.

Ces commentaires ne discréditent pas non plus les études avant-après **comparatives**, c'est-à-dire réalisées dans deux groupes (ou plus) de patients, par exemple un groupe sous traitement et un groupe sous placebo (comparaison de 4 moyennes appariées, ou de 4 proportions appariées).

5.3 Tests qu'on réalise en espérant ne pas rejeter H_0

L'ensemble des tests que nous avons vus (comparaison d'un paramètre à une norme, analyses bivariées) étaient exécutés en espérant rejeter l'hypothèse nulle H_0. La Figure 142 rappelle le fonctionnement de ces tests. Nous formulions tout d'abord deux hypothèses : l'hypothèse H_0 décrit une égalité simple, suffisamment simple pour permettre un calcul de probabilités. L'hypothèse H_1 est l'hypothèse alternative, qui décrit tout sauf H_0. Elle est trop complexe pour être exploitée. Ensuite, nous supposons que H_0 est vraie, et calculons la p valeur, qui est un indicateur de plausibilité de l'observation faite dans l'échantillon, en supposant H_0 vraie.

Si cette p valeur est inférieure à 5%, cela signifie que, en supposant H_0 vraie, nous avions moins de 5% de chances d'observer cet échantillon ou d'autres échantillons plus inattendus encore. Cette situation est jugée comme incompatible avec H_0, que nous rejetons. Nous pouvons donc accepter l'hypothèse alternative H_1.

Si cette p valeur est supérieure à 5%, cela signifie que, en supposant H_0 vraie, nous avions au moins 5% de chances d'observer cet échantillon ou d'autres échantillons plus inattendus encore. L'hypothèse H_0 est alors jugée plausible. Cette situation ne permet pas de rejeter H_0, mais ne permet pas non plus de la confirmer : on se trouve dans une situation d'indécision.

Il faut donc retenir de cela que les p valeurs supérieures à 5% ne permettent jamais de conclure quoi que ce soit d'utile pour la suite. En particulier, ce ne sont pas des arguments suffisants pour prouver que l'hypothèse H_0 est vraie. **On ne peut jamais prouver que H_0 est vraie**.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

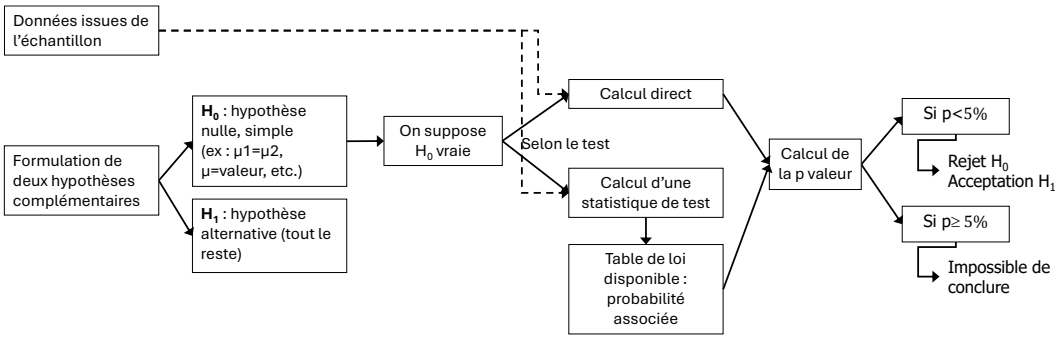


Figure 142. Fonctionnement général des tests statistiques d'hypothèse

Le problème est le suivant : certains tests statistiques sont réalisés en espérant ne pas rejeter H_0, autrement ils sont utilisés pour prouver que H_0 est vraie. En voici quelques exemples :

- Le test de **Kolmogorov Smirnov** et le test de **Shapiro-Wilk** sont souvent utilisés pour montrer qu'une distribution observée dans un échantillon suit une distribution normale
- Le test de **Fisher-Snedecor** est souvent utilisé pour montrer que deux variances observées ne sont pas trop différentes
- Le test de calibration de **Hosmer-Lemeshow** est utilisé pour montrer qu'une régression logistique fournit un résultat conforme en termes de proportion au résultat attendu
- Etc.

Ces utilisations de tests sont **incorrectes**, c'est pourquoi nous ne vous les avons pas proposées dans cet ouvrage, et vous avons proposé des critères décisionnels plus simples. De plus, le résultat est trop lié à la taille d'échantillon, et ne tient pas compte de la seule taille d'effet : sur un échantillon de petite taille, comme tous les tests, ils peineront à rejeter l'hypothèse nulle. Sur les très grands échantillons, ils rejetteront l'hypothèse nulle, et ce pour des déviances qui n'ont aucune significativité réelle. **Nous vous conseillons donc de ne pas utiliser ces tests statistiques.**

Un cas particulier toutefois est représenté par les **essais de non-infériorité**, que nous ne détaillons pas dans cet ouvrage. Le raisonnement mené dans ces essais pose moins de problème, car ces études s'appuient au préalable sur le calcul d'un nombre de sujets nécessaire. Si ce calcul est honnête et que l'échantillon atteint l'effectif souhaité, alors le raisonnement mené est considéré comme valide.

5.4 Test paramétrique ou non-paramétrique ? Asymptotique ou exact ?

Il est fréquent d'opposer les tests paramétriques aux tests non-paramétriques, avec l'idée que les premiers seraient réservés aux grands échantillons et aux distributions normales, et les deuxièmes aux petits échantillons et aux distributions bâtardes, et utiliseraient les rangs. Cette classification est **inexacte**. Cela ne revêt pas une importance capitale, mais nous préciserons les choses pour votre culture.

Un **test paramétrique** est un test qui s'appuie sur l'estimation d'un paramètre : moyenne, proportion, écart-type, etc.

Un **test non-paramétrique** est un test qui ne s'appuie pas sur l'estimation d'un paramètre : certains s'intéressent directement aux effectifs, d'autres s'intéressent aux rangs, etc.

Ces termes sont souvent utilisés pour désigner, à tort, respectivement les tests asymptotiques et les tests exacts. Cette dichotomie est plus intéressante que la précédente.

Un **test exact** est un test au cours duquel la p valeur est calculée de manière exacte. Elle peut être calculée par la personne qui exécute le test (ex : test binomial), ou elle peut avoir été

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

calculée par l'auteur du test, qui met à disposition une table de correspondance permettant de trouver cette p valeur sans la recalculer (ex : test de Wilcoxon).

Un **test asymptotique** est un test dans lequel la p valeur est exacte lorsque l'échantillon est infiniment grand, et d'une valeur approchée mais acceptable lorsque l'échantillon est suffisamment grand (ex : n≥30) ou que certaines conditions de validité sont remplies (ex : la variable étudiée suit une loi normale).

Nous vous proposons une classification des différents tests statistiques développés dans cet ouvrage, dans le Tableau 22.

Tableau 22. Classification des tests statistiques développés dans cet ouvrage

Cadre	Test	Paramétrique ?	Exact ?
Comparer une proportion à une norme	Test binomial	non-paramétrique	exact
	Test du Khi² d'adéquation	non-paramétrique	asymptotique
Comparer deux proportions appariées (un seul groupe)	Test de McNemar	non-paramétrique	asymptotique
Comparer une moyenne à une norme	Test de Student de comparaison d'une moyenne observée à une moyenne attendue	paramétrique	asymptotique
Comparer deux moyennes appariées (un seul groupe)	Test de Student de comparaison de deux moyennes appariées	paramétrique	asymptotique
	Test des rangs signés de Wilcoxon pour séries appariées	non-paramétrique	exact
Analyse bivariée qualitatif-qualitatif	Test du Khi² d'indépendance (+/-Yates)	non-paramétrique	asymptotique
Analyse bivariée qualitatif-quantitatif	Test de Student pour échantillons indépendants (+/- Welch)	paramétrique	asymptotique
Analyse bivariée quantitatif-quantitatif	Test de nullité du coefficient de corrélation de Pearson	paramétrique	asymptotique
	Test de nullité du coefficient de corrélation de Spearman	non-paramétrique	exact

Il existe dans ces classifications des cas trompeurs ou ambigus.

Le **test binomial** est un test **non-paramétrique**, car il ne s'intéresse pas aux proportions mais aux effectifs (c'est presque pareil, mais bon…). Il s'agit d'un test **exact** (ce qui illustre que tous les tests exacts ne s'appuient pas forcément sur les rangs).

Le **test du Khi²** (dans toutes ses variantes, y compris le test de McNemar) est un test **non-paramétrique**, car il ne s'intéresse pas aux proportions mais aux effectifs (c'est presque pareil, mais bon…). Il s'agit d'un test **asymptotique**, car il requiert des effectifs théoriques minimaux.

Le **test de Wilcoxon** pour séries appariées, et le **test de Wilcoxon-Mann-Whitney** sont des tests **non-paramétriques** et **exacts** portant sur les rangs mais, pour des effectifs élevés, une approximation par la loi normale permet d'accélérer les calculs. Ils ont donc une facette asymptotique.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Le coefficient de corrélation de Spearman n'est pas considéré comme un paramètre, car les rangs n'ont de sens que dans un échantillon et non dans une population de taille infinie. Ce coefficient n'a donc pas la prétention d'estimer un paramètre existant en population. Le **test de nullité du coefficient de corrélation de Spearman** est donc habituellement classé comme **non-paramétrique**. Il utilise une loi de Student, mais elle permet d'obtenir une p valeur exacte du fait des propriétés mathématiques des rangs : c'est donc un **test exact** malgré l'utilisation d'une loi elle-même utilisée par de nombreux tests asymptotiques.

5.5 Test unilatéral ou bilatéral ? Et pourquoi 5% ?

Lorsque nous calculons directement la p valeur, comme nous l'avons détaillé dans le test binomial (voir 2.2.4.2 Test binomial en page 113), nous supposons que l'hypothèse H_0 est vraie, calculons les probabilités associées à différentes observations possibles, et additionnons la probabilité de l'observation de l'échantillon avec d'autres probabilités :

- Si le test est « unilatéral », nous prenons en compte toutes les observations plus extrêmes que notre échantillon, du même côté
- Si le test est « bilatéral », nous prenons en compte toutes les observations moins probables que notre échantillon, des deux côtés

Reprenons l'exemple de la page 115 : nous avions un échantillon de 16 individus, une probabilité attendue de 49% d'être un homme, et observions un nombre x d'hommes dans l'échantillon. En Figure 143, nous représentons la p valeur obtenue en fonction du nombre x d'hommes observé. La courbe noire, au-dessus, représente la p valeur bilatérale, tandis que la courbe grise, en-dessous, représente la p valeur unilatérale. La partie droite propose un zoom sur les ordonnées proches de 5%, seuil de rejet de l'hypothèse nulle. On observe que **les p valeurs unilatérales sont plus faibles que les p valeurs bilatérales**. Dans le cas précis où nous observons 4 hommes sur 16 individus, le test unilatéral amène à rejeter H_0, mais pas le test bilatéral.

La Figure 144 rappelle comment l'une et l'autre de ces p valeurs sont calculées selon le cas, pour l'exemple où l'on observe 4 hommes parmi les 16 individus.

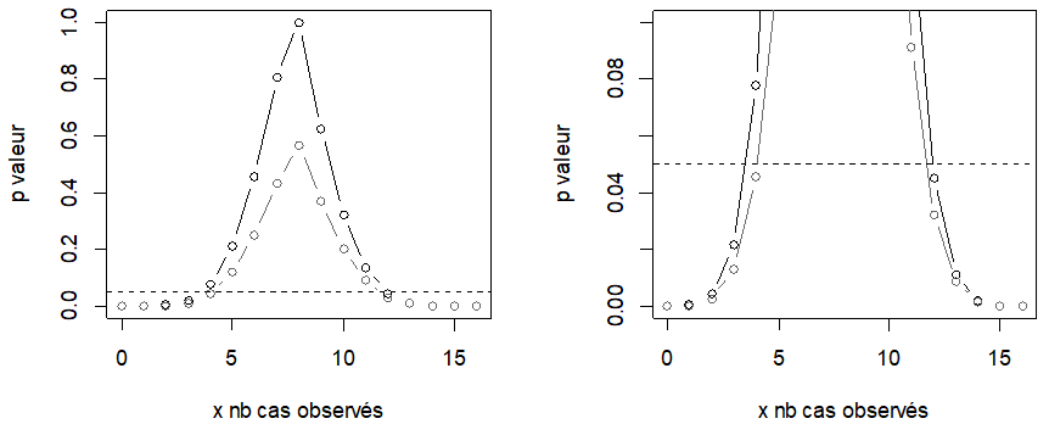


Figure 143. Exemple d'un test binomial, n=16, p_0=0,49, p valeur bilatérale (noir) ou unilatérale (gris) en fonction du nombre de cas observé. Pointillés : seuil d'interprétation de 5%.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

n=	16
p0=	0.49
x=	4

Effecfif observé	Proba sous H0	p valeur BILAT.	p valeur UNILAT.
0	0.00002	0.00002	0.00002
1	0.00032	0.00032	0.00032
2	0.00232	0.00232	0.00232
3	0.01040	0.01040	0.01040
4	0.03249	0.03249	0.03249
5	0.07491	-	-
6	0.13195	-	-
7	0.18110	-	-
8	0.19575	-	-
9	0.16718	-	-
10	0.11244	-	-
11	0.05892	-	-
12	0.02359	0.02359	-
13	0.00697	0.00697	-
14	0.00144	0.00144	-
15	0.00018	0.00018	-
16	0.00001	0.00001	-
Somme	**100.00%**	**7.77%**	**4.56%**

Figure 144. Exemple d'un test binomial, n=16, p_0=0,49, et x=4. Calcul de la p valeur bilatérale (7,77%) ou unilatérale (4,56%)

L'effet est le même lorsque le raisonnement s'appuie sur une loi symétrique (Student, Normale, etc.), et qu'on définit une zone de rejet. Comme l'illustre la Figure 145, pour un test bilatéral cette zone de rejet est partagée entre les deux extrémités (en-dessous de -1,96 et au-dessus de +1.96 pour une loi normale). Pour un test unilatéral de supériorité, cette zone de rejet se trouve au-dessus de 1,645. Pour un test unilatéral d'infériorité, cette zone de rejet se trouve en-dessous de -1,645. Ces seuils, en valeur absolue, sont plus permissifs que le seuil de 1,96, et permettrons plus facilement de rejeter l'hypothèse nulle H_0.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

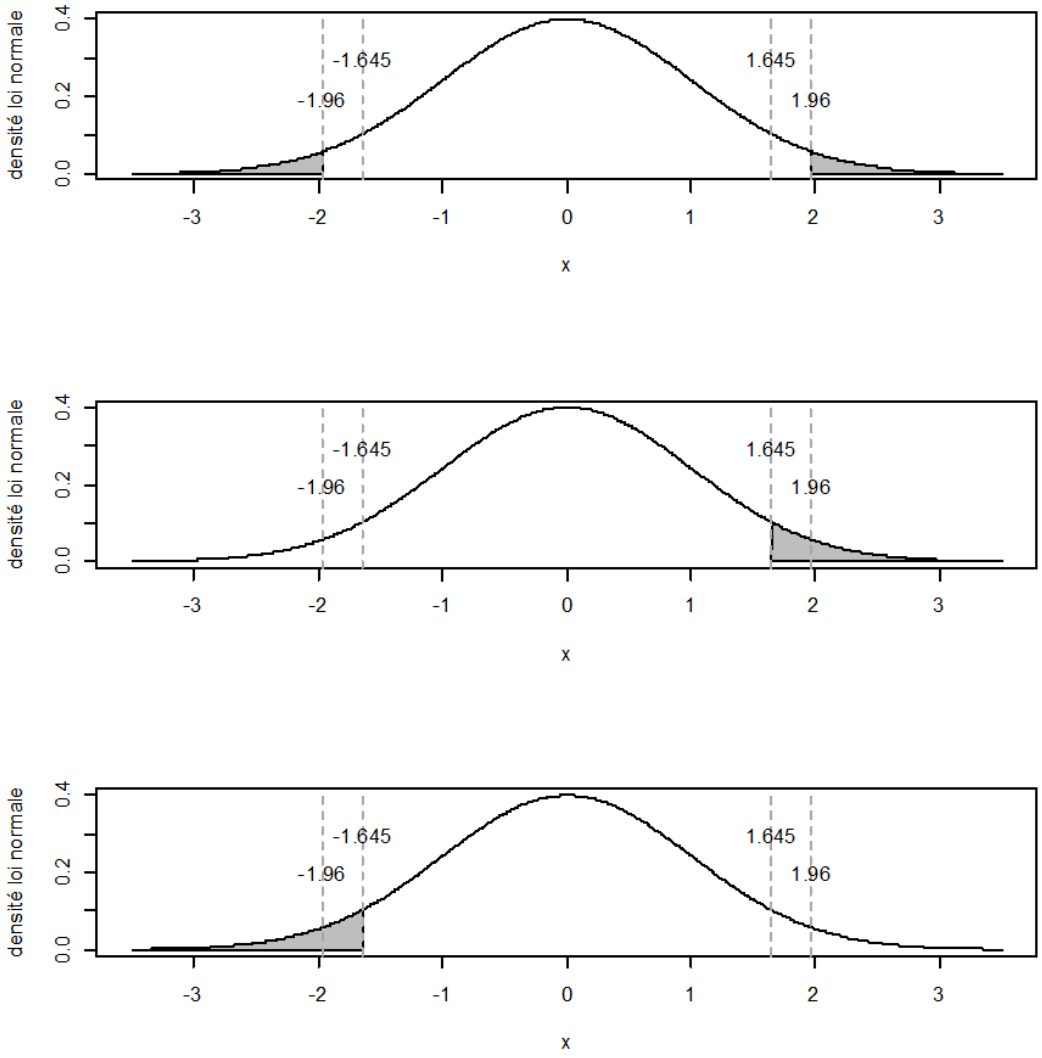


Figure 145. Exemple des zones de rejet à 5% de H_0 (fond gris) sur une loi normale (de haut en bas) : test bilatéral, test unilatéral de supériorité, test unilatéral d'infériorité

Alors, si les tests unilatéraux sont plus puissants, pourquoi ne pas les utiliser ? Plusieurs raisons peuvent être avancées de manière consensuelle.

Certains tests, comme tous les tests utilisant la loi du Khi², s'appuient sur une statistique de test qui est toujours positive (à cause de la puissance 2 utilisée dans son calcul), et donc sont toujours bilatéraux. Autoriser les tests unilatéraux reviendrait à **accepter un jeu malsain**, qui serait de choisir un test non pas parce qu'il est le plus valide, mais parce qu'il peut être utilisé en unilatéral, et permet de sauver des situations limites.

Deuxièmement, en recherche en santé, le risque alpha ou **risque de première espèce** (risque de rejeter à tort l'hypothèse nulle) est considéré comme **très grave** : une fausse découverte pourrait amener à soumettre des patients à des traitements dangereux, sans preuve scientifique robuste. Or de nombreuses raisons non-statistiques amènent à ce que les associations fortuites soient trop représentées dans la littérature scientifique, au-delà du risque alpha (analyses statistiques « data-driven », biais de non-publication, etc.). Imposer arbitrairement que tous les tests soient bilatéraux est une manière simple et efficace de limiter ce risque de première espèce.

Troisièmement, la réalisation de tests unilatéraux est un peu hypocrite : au fond, on sait déjà dans quel sens sera la différence, soit parce qu'on évalue un traitement qu'on espère être meilleur que le placebo, soit parce qu'on a observé les données (ex : deux moyennes) avant de décider de lancer un test statistique.

Enfin, si la FDA et l'EMA autorisent des tests unilatéraux, conscientes de leur hypocrisie, elles imposent alors que leur seuil d'interprétation soit de 2,5% et plus de 5%, ce qui amène à choisir exactement les mêmes seuils de rejet de H_0 (mais d'un seul côté)[40].

Toutes ces raisons amènent à une conclusion simple et opérationnelle : en-dehors des tests de non-infériorité, **les tests statistiques doivent être bilatéraux et interprétés au seuil de 5%**.

Le raisonnement est exactement le même pour les **intervalles de confiance**, car ils peuvent être utilisés pour réaliser, implicitement, un test de comparaison d'un paramètre observé à une valeur attendue : ils doivent être **calculés à 95%, avec une probabilité de rejet de 2,5% à gauche et autant à droite**.

5.6 Correction de Bonferroni

5.6.1 Exposé sur problème

Imaginons un jeu de données **généré aléatoirement**, et donc ne comportant normalement aucune association statistique, hormis celles purement issues du hasard. On cherche à réaliser des tests statistiques d'indépendance entre les différentes variables.

Intéressons-nous à un couple quelconque de variables. L'hypothèse nulle, qui est que ces variables sont indépendantes, est vraie. Donc, mécaniquement, la probabilité d'observer une association statistique avec le test *i* (sachant que H_{0i} est vraie), est de :

$$P(t_i^+) = P(t_i^+|H_{0i}) = \alpha = 5\%$$

Si nous réalisons ce genre de test à *k* reprises, entre *k* couples différents de variables, la probabilité qu'au moins un des tests soit significatifs est :

$$P(t_1^+ \cup t_2^+ \cup \ldots \cup t_k^+) = 1 - P(t_1^- \cap t_2^- \cap \ldots \cap t_k^-)$$

$$= 1 - (1-\alpha) \times (1-\alpha) \times \ldots \times (1-\alpha) = 1 - (1-\alpha)^k = 1 - 0{,}95^k$$

La Figure 146 montre ce qu'on appelle l'inflation du risque de première espèce, en fonction du nombre de tests réalisés.

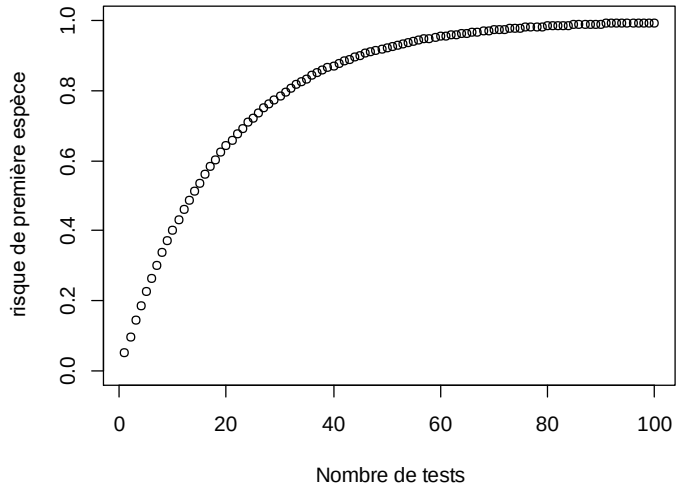


Figure 146. En supposant qu'il n'existe aucune association statistique, probabilité qu'au moins un test statistique trouve une association, en fonction du nombre de tests réalisés

En bref, plus on réalise de test, plus on de chances de découvrir à tort des associations statistiques fortuites. On appelle cela l'**inflation du risque alpha**, ou inflation du risque de première espèce.

5.6.2 Corrections possibles

La formule précédente peut aisément être retournée, pour identifier le seuil d'interprétation à utiliser pour interpréter chacun des *k* tests statistiques, de telle manière que le risque total reste de 5%, malgré l'inflation du risque alpha. Selon Šidák, ce seuil est :

$$\propto_{test\ indiv} = 1 - (1 - \propto_{total})^{1/k} = 1 - 0{,}95^{1/k}$$

Bonferroni a proposé une approximation légèrement conservatrice mais très populaire :

$$\propto_{test\ indiv} \approx \frac{\propto_{total}}{k} = \frac{5\%}{k}$$

On observe que les valeurs des deux corrections sont très proches (Tableau 23).

Tableau 23. Seuils de significativité corrigés (Šidák et Bonferroni)

Nombre de tests	Correction de Sidak	Correction de Bonferroni
1	0.05	0.05
2	0.0253205	0.025
3	0.0169524	0.0166667
4	0.0127415	0.0125
5	0.0102062	0.01
10	0.0051162	0.005
20	0.0025614	0.0025
30	0.0017083	0.0016667
50	0.0010253	0.001
75	0.0006837	0.0006667
100	0.0005128	0.0005

5.6.3 Quand, théoriquement, utiliser ces corrections ?

En soi, il n'est pas anormal de réaliser plusieurs tests statistiques sans mettre en place de correction.

La correction de Bonferroni devrait être utilisée chaque fois qu'on réalise plusieurs tests, en espérant qu'au moins un des tests soit significatif. Concrètement, cela a du sens lorsque tous

les tests réalisés se rapportent **au même objet** et peuvent, en cas de positivité, amener **la même conclusion**. En voici ceux exemples.

Premier exemple (mauvais) : Pour montrer la supériorité d'un somnifère A sur un somnifère B, on les administre à deux groupes de patients. On compare chacune des caractéristiques suivantes entre les deux groupes, à l'aide d'un test de Student : délai d'endormissement, durée de sommeil profond, délai avant le premier réveil, délai jusqu'au lever effectif, nombre de cauchemars par semaine, etc. Si sur au moins une de ces caractéristiques A est meilleur que B, on conclura que, globalement, A est meilleur que B, sans tenir compte des autres tests.

Deuxième exemple (mauvais) : Pour prouver l'origine génétique d'une maladie, on séquence le génome entier de patients. On identifie 200 allèles candidats. On réalise 200 tests du Khi², entre chacun des allèles examinés et la maladie. On conclura que la maladie est d'origine génétique si au moins un des tests revient significatif.

Dans les deux cas précédents, si on souhaite absolument réaliser ces études, il faudra impérativement appliquer une correction de Bonferroni.

5.6.4 Conduite à tenir en pratique

La conduite à tenir en pratique est présentée en Figure 147.

Le cas le plus simple est si vous réalisez **une étude non-soumise à une déclaration** très précise des objectifs principaux (exemple : étude par réutilisation de données, évaluation des pratiques professionnelles, étude portant sur des personnes qui ne sont pas des patients, etc.). Dans ce cas, en règle générale, vous n'appliquerez **pas de correction** de Bonferroni, à moins que, comme dans les exemples précédents, plusieurs tests portent clairement sur le même objet afin d'aboutir à une même conclusion. La contrepartie de cette liberté est que vos résultats auront **une portée moindre** que dans les études protocolisées, car vous aurez eu la possibilité de multiplier les tests et, plus généralement, de réaliser des tests qui n'étaient initialement pas prévus, au fur et à mesure que vous avez découvert les données puis les résultats préliminaires (*data-driven statistical analyses*). L'interprétation des p valeurs sera assez souple : on ne criera pas victoire pour des p valeurs de 4,5%, et les résultats que vous mettrez en évidence mériteront peut-être **d'être confirmés** par des protocoles plus solides. Et ce pour une raison très simple : personne ne pourra vérifier si vous n'avez pas réalisé de nombreux tests puis publié seulement les résultats significatifs.

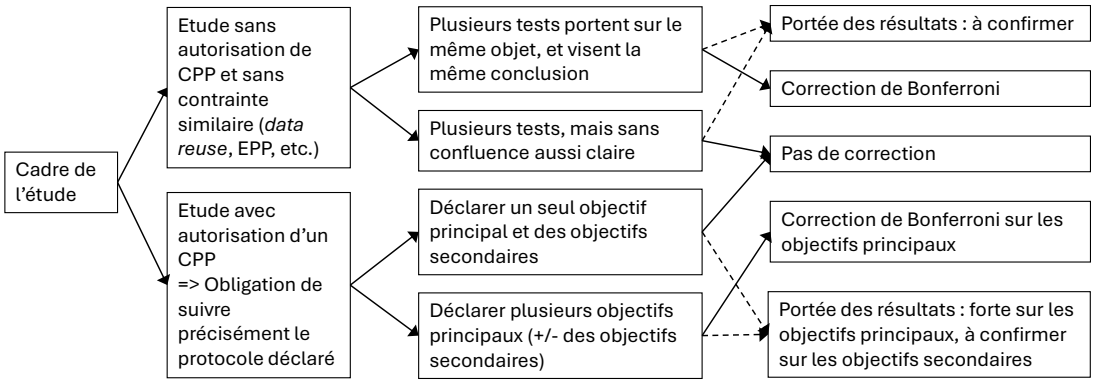


Figure 147. Conduite à tenir : appliquer une correction de Bonferroni

Si vous réalisez une **étude soumise à déclaration** à un CPP, les choses sont différentes : la demande d'autorisation vous contraint à déclarer un objectif principal, et en théorie vous n'aurez pas le droit de publier vos résultats en modifiant cet objectif. Les choses sont plus vraies encore pour les essais thérapeutiques. La plupart du temps, vous choisirez **un seul objectif principal** et n'aurez donc **aucune correction** à appliquer. D'autres analyses pourront être définies en objectif secondaire mais, comme les objectifs secondaires n'ont pas de portée

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

décisionnelle, il ne sera pas non plus nécessaire d'appliquer de correction de Bonferroni sur ces objectifs secondaires. Si vraiment vous souhaitez déclarer **plusieurs objectifs principaux**, il sera indispensable d'appliquer une **correction de Bonferroni sur ces objectifs principaux**. Les objectifs principaux ayant été définis bien avant d'accéder aux données, étant traçables et publics, la portée des résultats que vous obtiendrez sera forte concernant l'objectif principal. Les p valeurs inférieures à 5%, même de très peu (4,9% par exemple), seront interprétées comme un rejet clair de l'hypothèse nulle, et ce d'autant plus que le nombre de sujets nécessaire aura été calculé et enregistré dans votre déclaration de protocole, rendant impossible d'inclure plus de sujets que nécessaire pour atteindre l'objectif principal. En revanche, concernant les objectifs secondaires, la portée des résultats sera comparable à ceux obtenus dans une recherche non soumise à l'avis d'un CPP.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6 Interpréter une association statistique en général

6.1 Discuter la significativité statistique

Lorsque leur p-valeur est **inférieure à 5%**, les tests statistiques bivariés permettent seulement, de rejeter l'hypothèse nulle et, généralement, de conclure que les deux variables **ne sont pas indépendantes**. Cela ne confirme jamais, en tant que telle, la taille d'effet de l'association que vous aurez constatée (ex : en moyenne, le poids augmente de 340g par centimètre de taille corporelle). Cela ne confirme jamais non plus que cette taille d'effet est importante : **la p valeur rend compte du niveau de confiance que vous avez lorsque vous rejetez l'hypothèse nulle**, et non du niveau auquel la réalité s'écarte de l'hypothèse nulle. Si vous vous intéressez à la **taille d'effet**, vous devrez recourir à des indicateurs plus simples (ex : différence de deux moyennes, risque relatif, etc.), éventuellement assortis d'un intervalle de confiance, qui donnera une idée de la précision de l'estimation réalisée dans votre étude.

Comme nous l'avons déjà discuté plus haut, une p valeur inférieure à 5% a une valeur décisionnelle très forte dans une étude où le protocole a été écrit en avance et enregistré, contient un seul objectif principal qui a permis d'obtenir cette p valeur, et inclut un nombre de sujets limité par un calcul du NSN et l'autorisation obtenue. Ainsi par exemple, dans les essais thérapeutiques de phase 3, une seule étude suffit à autoriser la mise sur le marché d'un médicament, même si la p valeur obtenue est de 4,9%. Dans les autres cas, pour tout un tas de raison plus liées au protocole qu'aux mathématiques, les p valeurs proches de 5% doivent être interprétées avec prudence, et généralement une seule étude ne sera pas suffisante pour entraîner des actions de santé publique.

Lorsque leur p-valeur est **supérieure à 5%**, les tests statistiques ne permettent absolument **aucune conclusion** :

- Il se peut que, réellement, il n'y ait **rien de spécial** à constater (ex : les variables sont bien indépendantes)
- Il se peut que **l'effectif de l'échantillon soit trop faible**, conclusion qu'on ne pourra jamais porter
- Il se peut que le **modèle choisi soit inapproprié** (ex : on choisit un modèle linéaire pour décrire une relation en U)
- Dans une analyse multivariée, il se peut que la variable étudiée soit éclipsée par d'autres variables (**multi-colinéarité**)
- Etc.

Nous insistons sur le fait que, si la p valeur est supérieure à 5%, il faut se garder d'interpréter les résultats. Il est particulièrement malvenu de parler de « tendance » ou de « différence sensible mais non-significative » ou encore de « manque de puissance de l'étude ». **On ne sait pas, un point c'est tout !**

6.2 De la significativité statistique à la causalité et à l'explication

Disons-le clairement : il n'existe **aucune méthode statistique pour prouver une causalité**. Prouver la causalité relève non pas des statistiques, mais de la **méthodologie de l'étude**, nous reviendrons plus tard dessus. Les méthodes statistiques seules permettent d'affirmer qu'il existe une association statistique (avec un risque d'erreur), et c'est tout. Cette association statistique peut avoir différentes positions vis-à-vis de la relation de causalité :

- Elle peut être la conséquence d'une vraie causalité : le tabagisme augmente le risque de cancer du poumon.
- Elle peut être la conséquence d'une causalité de sens inverse : le fait d'avoir un cancer du poumon augmente la probabilité d'être fumeur, mais la causalité est inverse.

- Elle peut être la conséquence d'une cause commune : le fait d'avoir un briquet dans la poche est associé à une augmentation de probabilité d'avoir un cancer du poumon, mais le tabagisme est la cause commune de ces deux états.
- Elle peut être la conséquence d'une association fortuite, liée à une multitude de facteurs historiques, comportementaux, ou autres : le fait d'être Hawaïen était autrefois associé au risque d'avoir le SIDA[22]. Il ne faut pas y voir de causalité, mais simplement l'évolution de l'épidémie dans les années 1980.
- Etc.

L'essai randomisé contrôlé contre placebo ou traitement de référence, et en double ou triple aveugle, qui est la forme la plus aboutie d'étude comparative, est la seule méthodologie d'étude qui permette de démontrer la causalité, car l'exposition est modifiée par l'expérimentation, et que les patients sont affectés aléatoirement au sous-groupe soumis à l'intervention à évaluer ou au sous-groupe de comparaison, ce qui devrait neutraliser l'effet des facteurs de confusion (voir 6.2.2 Etudes comparatives en page 48).

Les **études observationnelles**, inversement, souffrent d'un biais majeur, qui est le **biais d'indication**. Si on veut comparer deux traitements par exemple, dans une étude observationnelle, les soignants choisissent l'un ou l'autre des traitements, mais pas de manière aléatoire : ils les choisissent parce qu'ils pensent que c'est la meilleure option pour le patient, compte tenu de nombreux facteurs, tels l'état de la maladie, les attentes du patient, les recommandations de bonne pratique, la disponibilité des traitements, etc. Il est possible que la différence d'effet observée entre les deux groupes soit plus le fait de ces facteurs qui motivent la décision, que de la décision elle-même. Nous reviendrons sur le biais d'indication plus bas. Il existe des méthodes statistiques pour diminuer l'impact de ce biais d'indication : il s'agit des **scores de propension**, qui peuvent être utilisés pour apparier les individus, ou ajuster l'analyse multivariée. Les études qui utilisent ces méthodes se présentent comme réalisant de « l'**inférence causale** » ou étant des « **émulations d'essais cliniques** ». Ces termes ne devraient pas être propagés, car ils sont excessifs et trompeurs : ces méthodes ne sont pas capables de prouver la causalité, mais simplement de diminuer, lorsqu'on dispose des données nécessaires, l'influence du biais d'indication. Elles sont très utiles et généralement assez efficaces, mais ne sont pas à mettre au même niveau que les RCT pour autant.

La **notion de causalité** peut être obtenue de plusieurs manières :

- Directement, en obtenant une association statistique en objectif principal d'un **essai randomisé contrôlé**
- Ou, directement mais approximativement, dans une étude observationnelle utilisant un **score de propension**
- Ou, indirectement et sans certitude, par un **faisceau d'arguments** :
    - Une association statistique dans une étude observationnelle (ou en objectif secondaire d'un essai randomisé contrôlé)
    - Une confirmation par la littérature
    - Des arguments physiopathologiques forts
    - Eventuellement, une relation dose-effet (l'effet est d'autant plus fort que l'exposition mesurée est forte)
    - Eventuellement, des expériences de challenge, dé-challenge, re-challenge (l'effet apparaît suite à l'exposition, disparaît après la fin de l'exposition, réapparaît à une nouvelle exposition)
    - Etc.

[22] On parlait à l'époque des quatre « H », qui étaient les principaux facteurs de risque de SIDA au début de l'épidémie : héroïnomane, homosexuel, hémophile, hawaïen. Ce moyen mnémotechnique est aujourd'hui dépassé.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

L'étape suivante est de passer de la causalité à l'explication. Là encore**, il n'existe aucune méthode pour affirmer avec certitude une explication**. L'explication nécessite à la fois :

- Une causalité, montrée comme précédemment
- Et de nombreux éléments liés aux connaissances valides du moment. En pratique, l'expérience montre que lorsque ces connaissances évoluent, cela remet en question très souvent l'explication, même lorsque la causalité reste acquise

D'un point de vue plus conceptuel, l'opinion de l'auteur de cet ouvrage est que **la recherche d'une explication est un peu éculée**. Elle était très présente il y a plusieurs siècles à l'époque où la médecine prenait plus de vie qu'elle n'en sauvait, et ne devrait plus être une priorité. Deux notions sont clairement prioritaires aujourd'hui :

- La recherche d'une **association statistique sans explication** (notion de facteur de risque ou protecteur, sans plus) : elle est déjà utile simplement pour identifier les patients à risque, et adapter ainsi la prise en charge (ex : chez un patient héroïnomane, il peut être utile de proposer une sérologie VIH et hépatites B et C).
- La quantification de l'**effet d'une intervention, sans chercher à l'expliquer** : la seule chose qui compte réellement, c'est de savoir, pour un patient donné, si une intervention sera bénéfique, neutre ou préjudiciable. La question du pourquoi et du comment est, reconnaissons-le, très secondaire.

Mon opinion : « Très souvent, on se moque de l'explication ! ». Nous illustrerons ce propos radical par un exemple bien connu.

Exemple de la mort subite du nourrisson :

Alertés par l'incidence des morts subites du nourrisson, les scientifiques ont tenté d'expliquer cela, et ont mis au point un raisonnement cohérent : couché sur le dos, en cas de régurgitation, le nouveau-né inhalerait ses vomissements, et décèderait. S'en sont suivies des recommandations généralisées, appuyées par des campagnes d'information financées par les contribuables dans les années 1960 et suivantes : il fallait coucher les nouveau-nés sur le ventre. Cette recommandation suivait un raisonnement de recherche d'explication, mais ignorait la réalité des associations statistiques.

En 1985, des chercheurs ont observé que, plus un état avait investi dans ces campagnes d'information du grand public, plus les décès étaient nombreux. D'autres arguments, ignorant totalement la possible explication, mais constatant simplement l'association statistique, ont permis de promouvoir exactement l'inverse : en 1987 aux Pays-Bas, puis en 1992 aux Etats-Unis, des campagnes promouvant le couchage sur le dos ont permis de diminuer de plus de 50% la mortalité. En France, entre 1991 et 1997, il a été possible de diviser par 4 la mortalité !

Dans cet exemple, le raisonnement qui cherchait une explication a tué des milliers d'humains, tandis que le raisonnement qui ignorait l'explication a sauvé des milliers de vies dans le monde. Au fond, **savoir « pourquoi » n'a aucune importance en termes populationnels**. Prédire l'impact d'une intervention, même si on ne l'explique pas, est essentiel. Pour modérer ce qui précède, on pourra dire que les recommandations en population doivent s'appuyer sur les preuves statistiques en ignorant l'explication mais que, lorsqu'un soignant décide d'une stratégie diagnostique ou thérapeutique pour un patient particulier, il est important que ce soignant puisse mener un raisonnement déductif, non pas pour contester la recommandation de manière populationnelle, mais pour l'appliquer avec discernement en tenant compte des particularités d'un patient spécifique. Il est possible que, même si une attitude est recommandée à l'échelle populationnelle, elle puisse être discutée sur un cas clinique très précis, avec de solides arguments tenant compte de toute la complexité qu'on connait du cas à traiter. Nous en donnons un exemple.

Exemple : un médecin reçoit un patient pour une ulcération chronique de l'oropharynx. Les recommandations indiquent que toute ulcération chronique de l'oropharynx doit être

considérée comme un carcinome, et doit être biopsiée. Ce médecin sait que ce patient a passé une semaine de vacances sexuellement actives à Mykonos en Grèce. Il n'applique pas la recommandation, et réalise en première intention une sérologie syphilitique, qui revient positive. Il a bien fait : il s'agissait d'un chancre syphilitique, acquis par contamination lors d'une fellation. Le traitement antibiotique a permis de guérir ce patient, et son conjoint a pu être soigné à son tour alors qu'il n'avait pas remarqué sa contamination.

Dans cet exemple, la bonne attitude diagnostique a été choisie d'une part en connaissant les recommandations, mais d'autre part en gardant le sens critique nécessaire et en intégrant toutes les connaissances que le praticien avait sur son patient. Cependant, la plupart du temps, la recommandation devra être appliquée, car la plupart du temps elle est valide, étant elle-même fondée sur des preuves statistiques.

6.3 Principaux biais en épidémiologie et en recherche clinique

Dans toute étude, il importe de rechercher des biais pour s'assurer que le résultat peut être interprété correctement et généralisé au-delà de l'échantillon observé.

6.3.1 Définitions : erreur, biais, biais différentiel, biais conservateur

6.3.1.1 Erreur ou biais ?

Une **erreur** est un processus qui fait que la mesure que nous en registrons n'est pas exactement la vraie valeur. Dans le cas des variables quantitatives, cette erreur est le fait d'un **bruit**, qui est la différence entre la valeur réelle et la valeur mesurée. Un bruit est une quantité qui suit une **loi normale de moyenne nulle**. Ce bruit peut être liée à l'appareil de mesure, à la précision de la mesure par l'opérateur, etc. Dans le cas d'un échantillon issu d'une population, une erreur est liée à l'échantillonnage, s'il est vraiment aléatoire.

Exemple 1 : on s'intéresse à la longueur d'un seul morceau de fer de 10cm. On le mesure plusieurs fois : on obtient une fois 10,056cm, une fois 9,983cm, etc.

Exemple 2 : on s'intéresse au poids moyen des humains adultes, et on suppose qu'en population il vaille 71kg. Dans un premier échantillon de 30 individus on mesure une moyenne de 72,35kg, dans un deuxième échantillon de 30 individus on mesure une moyenne de 69,84kg, etc.

L'erreur est, par définition, équilibrée et aléatoire : elle ne peut pas être prédite, elle est autant négative que positive, et elle ne dépend pas de facteurs connus. L'erreur n'est pas un problème pour nous : toutes les méthodes statistiques sont capables de la gérer, si elle est équilibrée et aléatoire.

Un **biais** est une erreur qui n'est pas équilibrée, ou qu'on peut prédire en partie du fait d'autres facteurs.

Exemple 1 : on s'intéresse à la longueur d'un seul morceau de fer de 10cm. On utilise un double décimètre, mais l'opérateur ne regarde pas la graduation orthogonalement, mais un peu « de biais », dans son cas il regarde de l'intérieur vers l'extérieur. Il mesure plusieurs fois le morceau : il obtient une fois 10,202cm, une fois 9,953cm, etc. La moyenne des mesures, sur un très grand nombre, est plus élevée que 10cm. La mesure est biaisée.

Exemple 2 : on s'intéresse au poids moyen des humains adultes, et on suppose qu'en population il vaille 71kg. On sélectionne des individus qui prennent le métro. Dans un premier échantillon de 30 individus on mesure une moyenne de 71,05kg, dans un deuxième échantillon de 30 individus on mesure une moyenne de 68,64kg, etc. Le mode de recrutement des participants exclut de fait les personnes obèses, qui ne prennent plus le métro, et les personnes qui utilisent tout le temps la voiture, qui en moyenne marchent moins que celles qui prennent le métro. La moyenne des mesures, sur un très grand nombre, est plus faible que 71kg. Ces moyennes sont donc biaisées.

6.3.1.2 Biais : différentiel ou non ?

En **épidémiologie descriptive**, l'existence d'un biais est problématique, car il empêche d'estimer les quantités qui nous intéressent (moyennes, prévalences, incidences, etc.). L'estimation devient alors une « borne basse » ou une « borne haute », qui peut éclairer la décision, mais de manière partielle seulement.

En **épidémiologie analytique**, où l'on réalise des analyses univariées ou bivariées, il est possible qu'un biais ne pose pas trop de problème, s'il est **non-différentiel**, c'est-à-dire s'il ne perturbe pas l'estimation des quantités qui nous intéressent, par exemple un risque relatif ou un *odds ratio*. Nous l'illustrons ici dans le cas bivarié binaire-binaire classique, mais cette notion peut aisément être généralisée.

Exemple de biais non-différentiel :

On réutilise les données du PMSI, pour savoir si un tabagisme codé lors d'une hospitalisation en 2010 est un facteur de risque de maladie d'Alzheimer lors d'une hospitalisation codée en 2020.

On sait que le tabagisme, en 2010 notamment, est très largement sous-codé (moins de 5% des séjours, alors qu'il concerne 30% de la population). On ne peut donc pas utiliser cette information en épidémiologie descriptive, car on sous-estimerait sa prévalence. En revanche, il n'y a pas de raison particulière qu'il soit plus ou moins bien codé en 2010 selon que les patients qui présenteront ou non une maladie d'Alzheimer 10 ans plus tard. L'erreur de codage est un biais car elle n'est pas équilibrée (c'est plutôt chez les fumeurs qu'on code mal), mais elle est indépendante de la pathologie étudiée, donc elle n'affectera pas l'odds ratio : ce biais de codage est donc non-différentiel.

Exemple de biais différentiel :

On réutilise les données du PMSI, à savoir un groupe de séjours de 2020 qui présentent un code de cancer du poumon, et un autre groupe de séjours de 2020 qui présentent un cancer de l'ovaire. On souhaite calculer un odds ratio entre le tabagisme (codé dans ces mêmes séjours) et le type de cancer.

Là encore, on sait que le tabagisme est très largement sous-codé, ce qui empêche d'estimer sa prévalence sur cette base de données. Cependant, on sait également que les pneumologues sont sans doute plus sensibles aux pathologies respiratoires et coderont mieux le tabagisme, que les chirurgiens gynécologiques, surtout que ces chirurgiens pensent (car c'est l'état actuel des connaissances) que le tabagisme n'a d'impact ni sur la pathologie, ni sur sa prise en charge. L'erreur de codage est un biais différentiel, car le tabagisme est mieux codé dans un groupe que dans l'autre, ce qui augmentera automatiquement la valeur de l'odds ratio dans ce cas.

6.3.1.3 Biais différentiel : conservateur ou non ?

Enfin, tout n'est pas perdu. Il se peut qu'un biais soit différentiel, mais qu'il atténue la taille d'effet qu'on est capable de mesurer. Dans ce cas, il peut être qualifié de **biais différentiel conservateur** : c'est plutôt une bonne nouvelle, car si on observe une différence, la différence réelle est encore plus forte que la différence observée, ce qui nous laisse en confiance quand nous affirmons que cette différence existe.

Cette distinction, entre biais conservateur et biais non-conservateur, relève de l'interprétation fine du cas étudié. Elle suppose de disposer d'éléments externes valides et de bien connaître le domaine d'application de l'étude.

6.3.1.4 Synthèse

La Figure 148 présente un arbre décisionnel, de ce qu'il est possible d'interpréter correctement en cas d'erreur, de biais, différentiel ou non, conservateur ou non.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

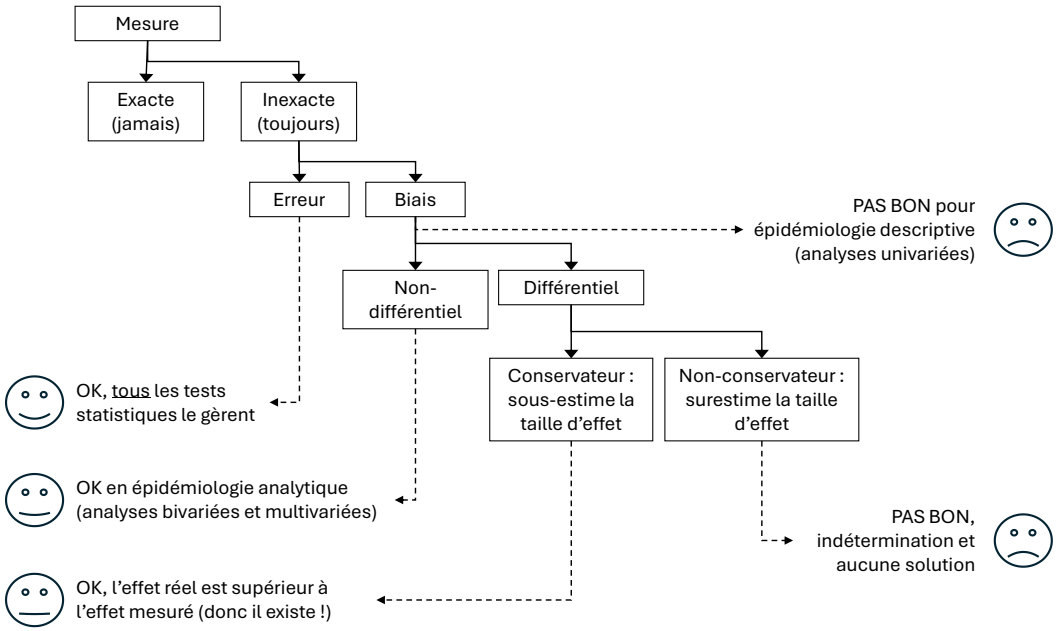


Figure 148. Arbre décisionnel : interpréter un résultat avec une erreur, un biais, etc.

Dans les parties suivantes, nous détaillerons les biais les plus fréquents.

6.3.2 Biais de sélection

Les biais de sélection sont les biais qui **résultent d'une anomalie de sélection** des participants, laquelle anomalie fait que l'échantillon n'est plus aléatoirement issu de la population d'intérêt. Selon leur nature, ces biais peuvent être différentiels ou non, conservateurs ou non, et donc impacter ou non l'interprétation des résultats. En voici quelques exemples, non-limitatifs :

- **Biais de recrutement** : par exemple, lorsqu'on recrute des patients parmi les patients hospitalisés, alors que notre étude prétend s'intéresser à tous les patients atteints d'une maladie. On imagine volontiers que les patients hospitalisés ne sont pas dans le même état clinique que les patients vivant à domicile. On imagine aussi que les cas les plus sévère sont peut-être déjà décédés, et ne sont pas pris en compte.
- **Biais de non-réponse** : lorsqu'on envoie un questionnaire et que la proportion de répondants est relativement faible, on peut supposer que les personnes qui n'ont pas répondu sont différentes de celles qui ont répondu (voir 7 Questionnaires : taux de sondage, taux de réponse en page 51)
- **Biais de censure informative** : dans les études prospectives, il est possible que certains patients quittent l'étude de manière non-aléatoire, par exemple parce que le traitement qu'ils ont reçu présente trop d'effets indésirables, ou les tue. Ce phénomène est généralement différent dans les groupes étudiés, et a un impact majeur. Pour y remédier, il est important de réaliser non pas des analyses **per protocole**, c'est-à-dire avec les patients qui terminent l'étude, mais des analyses **en intention de traiter**, c'est-à-dire en tenant compte de tous les patients initialement inclus, mais en prévoyant dès le début comment imputer les données manquantes pour ces patients. L'imputation devra être pessimiste (ex : considérer qu'ils ont tous eu des effets indésirables, et qu'aucun n'a eu d'effet bénéfique du traitement).
- Etc.

6.3.3 Biais d'information

Les biais d'information sont les biais qui **résultent d'une anomalie dans le recueil d'information** sur les participants. Ces biais peuvent porter sur toutes les variables, y compris

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

le groupe d'exposition, ou l'événement à prédire. Selon leur nature, ces biais peuvent être différentiels ou non, conservateurs ou non, et donc impacter ou non l'interprétation des résultats. En voici quelques exemples, non-limitatifs :

- **Biais de classement** : lorsque l'erreur impacte directement le groupe dans lequel le patient est affecté (ex : fumeur versus non-fumeur)
- **Biais de désirabilité sociale** : dans les questionnaires, il peut arriver que les personnes interrogées donnent des réponses inexactes, en fonction de l'image qu'ils souhaitent renvoyer (ex : avez-vous déjà eu des relations extraconjugales ?). Cela peut également s'appliquer aux faits illégaux, immoraux, aux opinions politiques ou religieuses extrêmes, etc.
- **Biais de mémorisation** : dans les enquêtes rétrospectives, les participants peuvent mal répondre à des questions, car ils ont oublié les faits anciens. Par exemple (c'est prouvé !), les femmes qui ont accouché, quelques années plus tard, ont tendance à sous-estimer la douleur qu'elles ont ressentie. Le biais de mémorisation est évident dans les études comparant les patients déments aux autres patients
- **Biais de codage** : en réutilisation de données, on sait que certaines pathologies ou prises en charge sont bien codées, tandis que d'autres sont sous-codées (ex : états n'ayant aucun impact tarifaire, tel le tabagisme), et d'autres pourraient être sur-codés (ex : codes dont la définition est ambiguë, et qui ont un impact positif sur les tarifs facturables)
- **Effet nocebo, effet placebo, effet cocooning** : la seule inclusion du patient dans une étude ou dans un bras de l'enquête, peut le pousser à ressentir et sur-déclarer certains effets positifs ou négatifs
- **Biais liés à l'enquêteur** : si l'enquêteur connait le traitement auquel le patient a été exposé, il peut être plus à l'affut de certaines complications que dans l'autre groupe, et laisser croire que ces complications sont plus fréquentes dans un groupe que dans l'autre. Le double aveugle est très important lorsqu'il est possible dans les RCT
- Etc.

6.3.4 Biais de confusion

Les facteurs de confusion n'altèrent pas directement les estimations faites dans l'étude, mais plutôt leur interprétation en termes de causalité.

Nous avons déjà évoqué ce problème dans la section 6.2 De la significativité statistique à la causalité et à l'explication en page 226. Nous pourrons citer deux autres exemples :

- **Biais protopathique** : par exemple, une maladie silencieuse provoque des symptômes non-étiquetés, et donc la prise d'un traitement, finalement la maladie émerge, et est diagnostiquée. Le traitement symptomatique se retrouve être un facteur de risque (statistique) de la maladie, sans bien sûr en être la cause.
- **Biais d'immortalité** : ce biais survient lorsqu'on veut comparer des patients qui présentent un événement (qui survient tard) à des patients qui ne l'ont pas présenté. Ceux qui ont eu l'événement sont nécessairement plus âgés que les autres, et forcément ils n'ont pas pu décéder avant cet événement (sinon ils auraient été classés dans l'autre groupe), faisant croire à tort que cet événement les a rendus immortels avant sa survenue. Il s'agit, au fond, d'un problème de design, qui ne pourrait pas survenir dans une véritable cohorte prospective, dans laquelle on ne connait pas l'avenir avant de l'avoir vécu.
- **Biais d'indication** : ce biais est très important, car il survient dans la quasi-totalité des études observationnelles comparatives. Nous lui dédions une partie séparée, ci-dessous.
- Etc.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

6.3.5 Un des biais de confusion, le biais d'indication

Les **études observationnelles comparatives** souffrent d'un biais majeur, qui est le **biais d'indication**. Prenons l'exemple d'une étude qui cherche à comparer deux traitements, indiqués pour la même pathologie. Chez les patients suivis, les soignants choisissent l'un ou l'autre des traitements, mais pas de manière aléatoire : ils les choisissent parce qu'ils pensent que c'est la meilleure option pour le patient, compte tenu de nombreux facteurs, tels l'état de la maladie, les attentes du patient, les recommandations de bonne pratique, la disponibilité des traitements, etc. Il en résulte que l'impact mesuré (ex : mortalité, score fonctionnel, morbidité, etc.) est la conséquence non seulement de l'intervention choisie, mais probablement plus encore des **facteurs qui ont amené les soignants à choisir l'intervention**, puisque ce choix n'est pas aléatoire ! Ce biais est considéré comme **majeur**. Vous devez systématiquement le rechercher si votre étude n'est pas un RCT.

Exemple (réel mais simplifié) : Dans le traitement des AOMI compliquées (artériopathie oblitérante des membres inférieurs), deux options sont possibles : la revascularisation par intervention chirurgicale sur les artères (déboucher ou remplacer l'artère bouchée), ou l'amputation de tout ou partie d'un membre (couper le pied ou la jambe qui est mal vascularisé).

Une étude montre que les patients qui subissent une intervention de revascularisation survivent plus longtemps que ceux qui subissent une amputation (après ajustement sur l'âge, le sexe et les comorbidités codées dans le PMSI). La conclusion de l'étude est qu'il faudrait privilégier la revascularisation à l'amputation.

L'analyse du cas est simple : le chirurgien a choisi pour chaque patient d'appliquer l'une ou l'autre des options en intégrant différents facteurs :

- Des facteurs visibles dans les données analysées pour l'étude : âge, sexe, diabète, description sommaire de la nécrose
- Des facteurs invisibles dans les données analysées :
    - La sévérité de l'obstruction artérielle
    - La sévérité des lésions de nécroses, au-delà de la piètre finesse du codage des diagnostics en CIM10
    - L'état fonctionnel du patient (chez un patient qui marche encore, l'amputation est moins souhaitable que chez un patient grabataire)
    - Les antécédents du patients et autres considérations cliniques, qui laissent présager son potentiel de cicatrisation, etc.
    - Le souhait du patient
    - Le contexte social, de mode de vie, d'hygiène, etc.
    - Etc.

De manière évidente, pour deux patients qui ont les mêmes valeurs dans les données accessibles aux chercheurs, les patients auxquels on propose l'amputation sont en réalité en état clinique nettement plus lourd, et avec une espérance de vie nettement plus faible, que ceux auxquels on propose la revascularisation. C'est donc pour cette raison qu'on croira, à tort, que l'amputation est responsable d'une morbi-mortalité supérieure.

Le biais d'indication peut être amoindri de plusieurs manières :

- On peut l'amoindrir par des **méthodes traditionnelles** :
    - **Apparier** les patients pour que les deux groupes soient équilibrés et comparables sur les variables choisies (ex : âge, sexe, sévérité des lésions…)
    - **Stratifier** les analyses, c'est-à-dire réaliser plusieurs comparaisons au sein de sous-groupes relativement homogènes (ex : une première analyse chez les hommes de 20 à 50 ans, etc.)
    - **Ajuster** les analyses, c'est-à-dire réaliser des analyses multivariées qui prennent en compte plusieurs variables explicatives en même temps, et tentent d'exprimer l'effet propre de chaque variable

- On peut l'amoindrir plus fortement à l'aide de **scores de propension**, visant une prétendue « **inférence causale** » ou « **émulation d'essai clinique** » (ces termes sont excessifs comme nous l'avons discuté) :
    - **Apparier** les patients sur ce fameux score de propension, pour que les deux groupes soient équilibrés en termes d'arguments pour une intervention ou l'autre, de manière que ce choix d'intervention, au sein d'une paire, paraisse quasiment aléatoire
    - **Ajuster** les analyses sur ce score de propension (on parle également de surajustement, car les variables constituant le score de propension sont présentées une deuxième fois dans le modèle multivarié)

Vous noterez que, dans tous les cas, ces méthodes ne peuvent s'appliquer **que sur les données disponibles**. Or, comme nous l'avons vu dans notre exemple des amputations de membres, les données réelles qui motivent l'indication thérapeutique sont, pour la plupart, **absentes du jeu de données : il n'existe alors aucune méthode statistique pour annuler totalement ce biais d'indication**, on ne pourra que l'atténuer, plus ou moins !

6.4 Analyses de sensibilité

Lorsqu'on réalise une étude dans laquelle on suppose que des biais peuvent entacher le résultat d'erreur, il est possible d'y adjoindre des analyses de sensibilité. Le terme « sensibilité » doit s'entendre au sens « dans quelle mesure les résultats sont-ils sensibles à (influencés par) certains biais ? ».

Une **analyse de sensibilité** est une analyse supplémentaire qui est réalisée en modifiant le protocole, pour voir quel impact cela a sur les résultats. Souvent, cela consiste à sélectionner un sous-ensemble des patients de l'étude, sur lesquels les données sont présumées plus fiables. On peut par exemple utiliser un critère d'inclusion plus précis (donc plus spécifique mais moins sensible). On peut également appliquer d'autres filtres, comme ajouter un délai de washout (période avant l'inclusion, dans laquelle on assure n'observer aucun événement). On peut également augmenter la spécificité des événements pris en comptes, en utilisant des critères de jugement plus restrictifs. Si, en reproduisant la même analyse sur ce sous-ensemble, les résultats sont proches, ou diffèrent d'une manière prévisible, ce sera un bon argument soit pour conforter les résultats obtenus dans l'ensemble de l'échantillon, soit pour prévoir quel impact a un biais sur les résultats.

Il n'existe **pas de conduite à tenir** claire et standardisée sur quelles analyses de sensibilité réaliser, ni comment les interpréter. Les analyses de sensibilité relèvent purement de l'expertise du chercheur sur son sujet. Il sera utile également de se référer à la littérature scientifique publiée dans votre domaine d'étude. Les analyses de sensibilité sont généralement présentées en résultat secondaire de la recherche. En général, on ne fixe pas de seuil de différence acceptable entre le résultat principal et le résultat de l'analyse de sensibilité. En voici trois exemples.

Exemple 1 : on cherche à évaluer l'incidence du cancer du sein dans une grande base de données médico-administrative comportant des séjours hospitaliers de 2010 à 2020. On considère comme cas incidents toutes les patientes qui ont un premier séjour à partir de 2013, mais n'ont aucun séjour avec cancer du sein en 2010 et 2011. On décrit ce procédé comme le fait s'imposer une période de washout de deux ans.

L'inquiétude est la suivante : il se peut que des patientes aient eu un cancer du sein, puis une rémission de deux ans, puis une récidive. Elles sont considérées à tort comme de nouveaux cas. On réalisera alors plusieurs analyses de sensibilité, avec des périodes de washout de 3, 4, 5, 6, 7 et 8 ans. Dans chaque analyse de sensibilité l'effectif sera plus faible, mais si les taux d'incidence calculés ne diminuent pas, ce sera un bon argument pour conforter l'analyse principale.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Exemple 2 : on s'intéresse au risque de fracture chez les patients qui ont une ostéoporose. On inclut ces patients, et on les suit dans le temps, en notant les épisodes de fractures identifiés dans les dossiers.

L'inquiétude vient du fait qu'il n'est pas exclu que certaines fractures résultent non pas de l'ostéoporose, mais de métastases osseuses de cancers. On réalisera une analyse de sensibilité en excluant tous les patients qui ont un antécédent de cancer. Cette exclusion n'est pas réalisée dans l'analyse principale, car on sait que chez les hommes âgés, la prévalence du cancer de la prostate est très élevée. Si l'incidence de fracture n'est pas diminuée dans cette analyse de sensibilité, ce sera un argument en faveur du résultat initial.

Exemple 3 : on s'intéresse au devenir de patients ayant une maladie rare. Ils sont identifiés parce qu'ils ont un séjour hospitalier dans lequel on retrouve un code de cette maladie.

L'inquiétude vient du fait que le diagnostic de cette maladie pourrait être posé à tort par des non-spécialistes. Une première analyse de sensibilité consistera à n'inclure que les patients pour lesquels ce code apparait en position de diagnostic principal ou relié (par opposition aux diagnostics associés significatifs). Une deuxième analyse de sensibilité consistera à n'inclure que les patients pour lesquels ce code apparait dans deux séjours hospitaliers différents.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Rédiger et présenter le document

Les conseils que nous présenterons ici sont autant valables pour la rédaction de votre mémoire académique que pour un article scientifique.

Nous verrons tout d'abord comment utiliser un logiciel de traitement de texte, tel Microsoft Word ou LibreOffice Writer, dans la partie 1 Utiliser un traitement de texte de manière appropriée en page 236.

La partie suivante sera dédiée au logiciel de gestion de la bibliographie : 2 Installer et utiliser Zotero, logiciel de bibliographie en page 243.

Nous parlerons à proprement parler de la rédaction scientifique en partie 3 Rédiger les différentes parties du mémoire en page 250.

Plus prosaïquement, le passage au papier sera traité en partie 4 Imprimer et diffuser le document en page 264.

Enfin, nous évoquerons simultanément la préparation du diaporama et la soutenance orale en partie 5 Utiliser un logiciel de présentation pour la soutenance orale en page 269.

1 Utiliser un traitement de texte de manière appropriée

L'ensemble des choses qui vous sont présentées dans cette section le sont de manière plus complète et visuelle dans la vidéo d'Objectif Thèse dédiée au traitement de texte (accès libre et gratuit) : http://www.objectifthese.org

Ce même site web vous propose également un modèle de document Word spécialement optimisé pour les mémoires académiques. De nombreuses fonctionnalités ont déjà été paramétrées dans ce document.

Nous verrons dans ce chapitre comment utiliser **Microsoft Word** et **LibreOffice Writer** (ou OpenOffice Writer) de manière appropriée pour réaliser un mémoire académique. Nous supposons que vous savez déjà utiliser ces logiciels pour des tâches simples comme saisir et imprimer un courrier. Nous verrons ici les fonctions indispensables pour un document structuré. Ce que nous verrons vous permettra à la fois d'augmenter la qualité du document et son homogénéité, mais surtout de gagner énormément de temps, en automatisant certaines tâches.

1.1 Généralités sur les styles

Vous savez certainement déjà utiliser un traitement de texte pour saisir et formater un document. Il est très important que vous ne réalisiez aucune mise en forme locale (ex : modifier la taille d'un paragraphe à la main), mais plutôt que vous utilisiez les styles. Lorsque vous affectez un style à un paragraphe (par exemple), ce paragraphe prend automatiquement toutes les propriétés de mise en forme décrites dans le style. Si vous modifiez la définition du style, même dans un document de 300 pages, tous les paragraphes portant ce style verront instantanément et automatiquement leur mise en forme mise à jour. L'utilisation des styles permet :

- Une énorme économie de temps dans la mise en forme du document

- Une homogénéité de présentation garantie, contribuant au sentiment d'une mise en forme soignée
- Pour certains styles de titre, des fonctionnalités supplémentaires indispensables (voir paragraphe suivant)

Le style par défaut d'un paragraphe quelconque est le style « Normal ».

Si vous copiez des paragraphes d'un document vers un autre document :

- Les styles qui n'existent pas dans le document cible seront créés avec la même définition que dans le document source
- Si des styles existent sous le même nom dans les deux documents, le nouveau contenu sera automatiquement reformaté avec les styles du document cible

Il existe également des fonctionnalités pour appliquer les styles d'un document à un autre document, sans modifier le contenu de cet autre document.

Le site http://objectifthese.org vous propose, en plus des vidéos et tutoriels, un modèle de document parfaitement adapté pour les mémoires académiques. Ce document contient déjà des styles spécialement redéfinis pour les mémoires académiques.

1.2 Le cas particulier des styles « Titre X »

Les styles nommés « Titre », « Titre 1 », « Titre 2 », etc. existent déjà dans le modèle de document par défaut, et ont été améliorés dans le modèle de document proposé sur le site http://objectifthese.org.

Si vous utilisez correctement ces styles, vous gagnerez du temps et garantirez une stricte homogénéité de tous vos titres. Il sera notamment possible de définir une fois pour toute :

- L'indentation (décalage vers la droite), en fonction du niveau de titre (1 à 9)
- La police, le gras, l'italique et la couleur de vos titres
- Les règles de numération automatique de vos titres, en respectant la hiérarchie, y compris avec répétition (ex : 1.3.1, 1/C/3/, etc.)
- L'espacement au-dessus et au-dessous de chaque titre, sans insérer de paragraphe vide

Comme expliqué dans la vidéo sur http://objectifthese.org, les styles de type « Titre X » permettent en particulier :

- Une **numérotation hiérarchique** automatique cohérente dans tout le document
- La mise en place, en début de document, d'un **sommaire** avec les numéros de pages, automatique et dynamique (qui se met à jour)
- L'utilisation du **mode plan** pour réorganiser rapidement et de manière sécurisée tout le document
- L'utilisation du **volet de navigation**, sur la gauche, pour naviguer rapidement dans le document
- Des **renvois automatiques** et dynamiques à des titres dans le document

1.3 Afficher les caractères non-imprimables

Afin de tirer pleinement parti de votre traitement de texte, il est important d'afficher les caractères non-imprimables.

Pour ce faire, vous pouvez cliquer sur le bouton reproduit à gauche. Cela permet d'afficher explicitement plusieurs caractères non-imprimables, qui sont présentés en Tableau 24.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tableau 24. Principaux caractères non-imprimables mais affichables

Caractère	Représentation	Raccourci clavier PC
Espace	·	[espace]
Espace insécable	°	[ctrl]+[maj]+[espace]
Marque de fin de paragraphe	¶	[entrée]
Retour de chariot	↵	[maj]+[entrée]
Tabulation	→	[tab] ou [ctrl]+[tab]
Saut de page	..... Saut de page	[ctrl]+[entrée]
Saut de colonne	..... Saut de colonne	[ctrl]+[maj]+[entrée]

Vous connaissez bien l'**espace**. Il sera désormais représenté sous la forme d'un point suspendu. Il sera surtout important de distinguer l'espace des autres caractères non-imprimables.

L'**espace insécable** (Tableau 24), présenté comme une bulle, est un espace qui ne sera pas coupé par le passage à la ligne, et qui ne pourra pas être étiré en cas d'alignement justifié. Il garantit ainsi la solidarité entre les termes qu'il sépare. Il est utilisé en particulier devant les ponctuations doubles (nous y reviendrons), comme séparateur de milliers dans les nombres, ou comme séparateur entre un nombre et son unité.

La **marque de fin de paragraphe** (Tableau 24) permet de passer au paragraphe suivant. C'est ce qu'on fait en pressant [entrée] sur le clavier. Les propriétés de paragraphes sont donc constantes entre deux marques de fin de paragraphe (au sein d'un même paragraphe), mais peuvent différer uniquement après une telle marque.

Le **retour de chariot** (Tableau 24) peut être utilisé pour passer à la ligne sans changer de paragraphe. Le plus souvent, nous l'utilisons au sein d'une légende de figure ou de tableau, dans un titre, ou encore dans une liste à puces. Les propriétés de paragraphe sont ainsi ininterrompues.

La **tabulation** (Tableau 24) est un espace de largeur variable. Sa largeur est autant étendue que nécessaire, pour que le bloc de caractères suivant respecte les conditions imposées par les taquets de tabulation. Elle est également utilisée dans les listes à puces. Nous ne détaillerons pas trop cela, mais le tutoriel vidéo d'Objectif Thèse en fait une démonstration plus complète. Lorsqu'on souhaite ajouter une tabulation au sein d'un tableau ou au tout début d'un paragraphe, il faut y adjoindre la touche « contrôle » pour éviter l'activation de commandes non-désirées.

Le **saut de page** (Tableau 24) permet de passer à la page suivante. Le **saut de colonne** permet de passer à la colonne suivante au sein de la même page. Il ne faut surtout pas insérer de paragraphes vides pour atteindre cet objectif.

1.4 Afficher les champs dynamiques sur trame grise

Il est également très important de bien distinguer le contenu statique (saisi tel quel par l'auteur) du contenu dynamique, qui est censé se mettre à jour automatiquement. Le contenu dynamique inclut notamment :

- La numérotation des titres*
- La numérotation des listes à puces*
- La numérotation des pages
- Le début (mot clef et numéro) des légendes de figures et tableaux (ex : « Figure 3 »)
- Les renvois vers chacun des éléments précédents

- La table des matières
- Les tables des illustrations (figures, tableaux, équations)
- Le glossaire
- Les citations Zotero, puis la bibliographie en fin de document
- Etc.

Hormis les éléments marqués d'un astérisque dans la liste ci-dessus, tous ces éléments peuvent être affichés **sur fond gris** (non imprimable), ce qui permet de mieux comprendre le document et d'éviter des altérations fortuites du contenu dynamique. Il faut pour ce faire modifier l'affichage.

Dans Microsoft Word 365 :
Fichier > options > avancées > Champs avec trame : toujours

Dans LibreOffice :
Affichage > trame de fonds de champ (ctrl+F8)

La mise à jour de certains de ces éléments n'est pas instantanée, et nécessite de demander une **mise à jour des champs**, ce qui peut être fait de deux manières. :

- Sélectionner l'élément à mettre à jour ou tout le document ([contrôle]+[a]), et appuyez sur la touche F9
- Faites un clic droit sur l'élément à mettre à jour puis cliquez sur « mettre à jour les champs » ou « actualiser le champ sélectionné »

1.5 Figures et légendes

1.5.1 Alignement

Dans un mémoire académique, les figures doivent être « alignées sur le texte », et présentées dans un paragraphe unique. Elles ne doivent pas être encadrées par le texte, ce qui deviendrait vite ingérable (contrairement à ce que l'on ferait pour un article de journal, dont la mise en page doit rester compacte). Pour aligner une figure sur le texte :

Dans Microsoft Word : *clic droit sur l'image > taille et position > habillage du texte > Aligné sur le texte*

Dans LibreOffice : *clic droit sur l'image > ancre > comme caractère*

1.5.2 Type d'image

Lorsque la figure provient d'un logiciel qui génère du dessin vectoriel (ex : logiciel de présentation Powerpoint / Impress, tableur Excel / Calc), il est très important de coller l'image **en tant que métafichier**, et non en tant qu'image bitmap ou, pire, en tant qu'objet incorporé. Pour ce faire :

Dans Microsoft Word : *Coller > collage spécial > Image (métafichier amélioré)*

Dans LibreOffice : *Edition > collage spécial > collage spécial > métafichier (…)*

Ceci permet plusieurs choses :

- Eviter la pixélisation
- Permettre, lorsqu'on le souhaite, de redimensionner l'image sans problème, y compris en rompant le ration hauteur-largeur
- Eviter le recalcul des objets incorporés (ex : graphiques Excel), ce qui ralentirait l'ordinateur et affecterait la mise en forme des images

1.5.3 Légende d'une figure

On doit ensuite **générer une légende** avec numérotation automatique. Pour ce faire,

dans Microsoft Word ou dans LibreOffice : *Clic droit sur l'image > Insérer une légende.*

Puis dans les deux cas, choisir le type d'élément approprié (Figure, Tableau, Equation, etc.) et compléter la légende avec le texte de votre choix. Après validation, la légende est créée. Il est usuel de faire figurer les légendes en-dessous des figures qu'elles décrivent. N'hésitez pas, si vous souhaitez passer à la ligne dans une légende, à utiliser le **retour de chariot** (ctrl+Entrée) plutôt que la fin de paragraphe (Entrée), afin de ne pas interrompre la légende.

Pour les manipulations suivantes, il faut bien comprendre que, en dépit des apparences, la légende n'est aucunement liée à l'image. Il s'agit d'un **paragraphe autonome**, doté par défaut du style Légende, et qui comporte en son début un signet un peu particulier. Ce signet permettra aux renvois de connaître sa position, son type (Figure, etc.), son numéro séquentiel, et le texte situé à sa droite. Ce paragraphe peut être à son tour déplacé, modifié, remis en forme, etc. Si ce paragraphe est copié, un nouveau numéro de référence sera créé, si bien que les deux légendes seront indépendantes l'une de l'autre.

Les légendes ainsi créées présentent plusieurs particularités :

- Elles **se renumérotent automatiquement** en cas de déplacement, insertion de nouvelle figure, ou suppression d'une figure existante
- Elles peuvent faire l'objet de **renvois dynamiques** insérés par l'auteur du document
- Elles sont **automatiquement référencées** dans une « table des illustrations » si l'auteur en crée une en fin de document.

Si vous n'êtes pas l'auteur d'une des images, pensez à citer l'auteur (entre parenthèses par exemple) dans la légende de l'image.

1.5.4 Renvois dans le texte

Les figures ne doivent pas rester orphelines : chaque figure **devra donc être citée** au moins une fois dans le texte, à l'aide d'un renvoi.

Dans Microsoft Word ou dans LibreOffice : *Menu insertion > Renvoi*

Il faudra ensuite sélectionner le grand type d'élément (ici, une figure), choisir l'élément dans la liste, puis choisir le type de renvoi à afficher. Le plus souvent, on affiche le mot clef (« Figure ») suivi du numéro séquentiel, ce qui peut être fait ainsi :

Dans Microsoft Word : *texte et numéro uniquement*

Dans LibreOffice : *catégorie et numéro*

On peut également faire appel à d'autres éléments, tel le numéro de page, le texte complet de la légende (c'est déconseillé), le positionnement « ci-avant/ci-après », etc. Il est possible de juxtaposer les éléments, comme en Figure 149 :

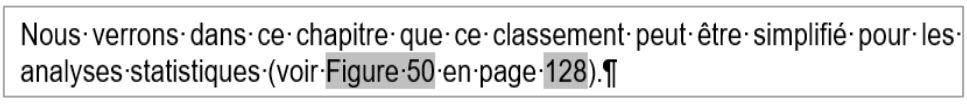


Figure 149. Exemple de renvoi composite : texte et numéro, puis numéro de page

1.5.5 Table des illustrations

Quelque part en fin de document, vous pourrez insérer une fois pour toutes une table des illustrations, qui listera toutes les figures du document, avec leur numéro, leur légende et le numéro de la page.

Dans Microsoft Word : *Références > insérer une table des illustrations*  *(choisir « figure » et paramétrer)*

Dans LibreOffice : *Insertion > Table des matières ou index > table des matières, index ou bibliographie (choisir type="index des figures")*

Notez que les tables des illustrations sont déjà présentes dans le modèle de document proposé sur le site http://objectifthese.org. Ces tables peuvent être insérées précocement : elles seront mises à jour par la suite sans difficulté.

1.6 Tableaux et légendes

Des tableaux peuvent être insérés dans le document, d'une manière similaire aux figures. L'alignement d'un tableau est accessible dans ses propriétés. Là aussi, nous vous conseillons de ne pas « habiller » le tableau par du texte. Il s'agit du positionnement par défaut. Tout comme pour les figures, vous devrez utiliser :

- Une légende « tableau » généralement positionnée au-dessus du tableau
- Des renvois dans le texte
- Une table des illustrations « tableaux », en fin de document

Ces fonctionnalités sont très proches de celles utilisées pour les figures, seul le paramétrage final les distingue. Vous pourrez aisément adapter les paragraphes précédents, dédiés aux figures, aux tableurs.

Si vous comptez réaliser une version anglophone de votre document, pour gagner un peu de temps, utilisez le libellé « table » au lieu du libellé « tableau » : il est acceptable en Français, et déjà approprié en Anglais.

1.7 Rappels sur la typographique et la ponctuation

1.7.1 Ponctuation et espacements

Les règles d'espacement à proximité des marques de ponctuations n'ont réellement été précisées en Français qu'avec l'avènement de l'imprimerie. Ces règles sont les suivantes :

- Les ponctuations simples (. , …) ne sont pas précédées d'un espace ; elles sont suivies d'un espace (que nous représentons par un carré ci-dessous).
    - *Ex : Voici,□et□voilà.□Encore□voilà.*
- Les ponctuations doubles (; : ! ?) sont précédées par un espace insécable (représenté par une bulle ci-dessous), et sont suivies d'un espace normal. L'espace insécable a été présenté dans le chapitre 1.3 Afficher les caractères non-imprimables en page 237.
    - *Ex : Je□demande○:□Venez-vous○?□Non○!*
- Les marques d'encadrement ([] () -- "" ‘’) ne présentent aucun espace intérieur, et un espace extérieur, à moins qu'elles soient suivies d'une ponctuation simple.
    - *Ex : Je□"vois"□-difficilement-□deux□couleurs□(rouge,□jaune).*
- Par dérogation à ce qui précèdent, les guillemets français nécessitent de plus un espace insécable à l'intérieur
    - *Ex : Je□«○devine○»□une□certaine□frustration…*

En langue anglaise, la principale différence est que les ponctuations doubles s'écrivent comme les ponctuations simples, et requièrent un espace après, mais aucun espace avant. Les guillemets français n'existent pas.

1.7.2 Typographie des nombres dans le texte

En Français, le séparateur décimal est la virgule. Pour les nombres supérieurs à mille, il est possible d'utiliser le point ou l'espace comme séparateur de milliers. Ce n'est pas obligatoire. Le point étant ambigu, nous vous conseillons d'utiliser l'espace insécable.

Ex : Cette maison coûte 1 500 000,01€.

En Anglais, le séparateur décimal est le point. La virgule est utilisée comme séparateur des milliers.

Ex : This house costs 1,500,000.01€.

Il importe de se poser la question des arrondis, et des chiffres significatifs. On pourrait recommander, notamment dans des articles de sciences dures, de présenter tous les nombres avec 3 chiffres significatifs :

1,123456 devient 1,12
1,99987 devient 2,00
1 234 456 devient 1 230 000

En pratique en Santé, l'usage est plutôt de **conserver toute la précision entière** des nombres, et de **réaliser un arrondi à 1 ou 2 chiffres** après la virgule. Cependant, on choisira l'unité ou la présentation (pourcentage par exemple) qui rendra cet arrondi tolérable :

Un âge moyen de « 1,08333 ans » sera converti et noté « 13,0 mois »
Une proportion de « 0,123456 » sera convertie et notée « 12,3% »

/! Lorsque le nombre produit est « plus rond » que la précision retenue, il importe de faire figurer des **zéros à droite** de la virgule pour **rappeler la précision** choisie (d'où le « 13,0 » et non « 13 » ci-dessus).

Pour les **p valeurs** (probabilités rendant compte de la vraisemblance de l'observation sous l'hypothèse nulle), il vaut mieux éviter de raisonner en termes de précision (centième, millième, etc.), et plutôt raisonner en nombre de chiffres significatifs. Bien qu'il n'y ait pas de règle consensuelle, on proposera ce qui suit :

- Indiquer 2 chiffres significatifs pour les p valeurs supérieures à 1%
- Indiquer 1 chiffre significatif pour les p valeurs inférieures à 1%
- Noter plus simplement « p=0 » pour les p valeurs inférieures à 10^{-10} (ce type de choix sera indiqué dans la section « méthodes » du document, nous y reviendrons)

Voici quelques exemples de retranscriptions de p valeurs :

p valeurs : 0,82 0,045 0,002 2e-5 0

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2 Installer et utiliser Zotero, logiciel de bibliographie

2.1 Difficultés liées à l'affichage de la bibliographie

Un mémoire académique, tout comme un article scientifique, doit s'appuyer sur des références bibliographiques. Les **références** sont citées en fin de document (zone grisée à droite, Figure 150). Dans le texte, ces références sont citées au fil du texte : on appelle cela des **citations** (inserts grisés à gauche, Figure 150).

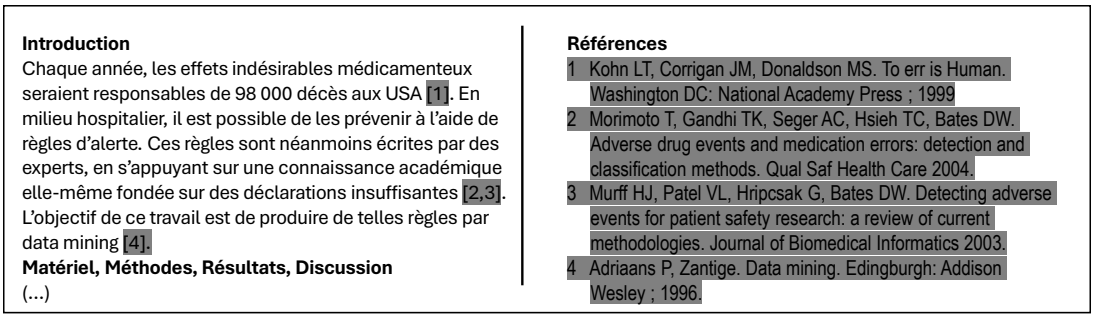


Figure 150. Bibliographie dans un mémoire : citations (gauche) et références (droite)

Différentes normes existent pour présenter les citations et les références. Dans la **norme Vancouver**, qui est utilisée dans la recherche en santé, les citations sont numérotées de manière séquentielle (1, 2, 3, etc.). L'ordre des numéros est la conséquence directe de l'ordre de citation des références. En fin de document, les références doivent être classées dans ce même ordre.

Vous entrevoyez déjà une première difficulté : dès que les citations changent, il faut modifier tous les numéros de renvois, et la liste de références en fin de document.

En outre, selon la norme de citation utilisée, les citations doivent être condensées :

[1][2] devient [1, 2]
[1][2][3] devient [1-3]
[1][2][3][8] devient [1-3, 8]

Il s'agit d'une deuxième difficulté lorsqu'on doit renuméroter toutes les références.

Enfin, la présentation des références est propre à chaque journal. Elle résulte de la fusion des métadonnées de chaque référence, avec des règles très précises. Un exemple en est donné en Figure 151.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

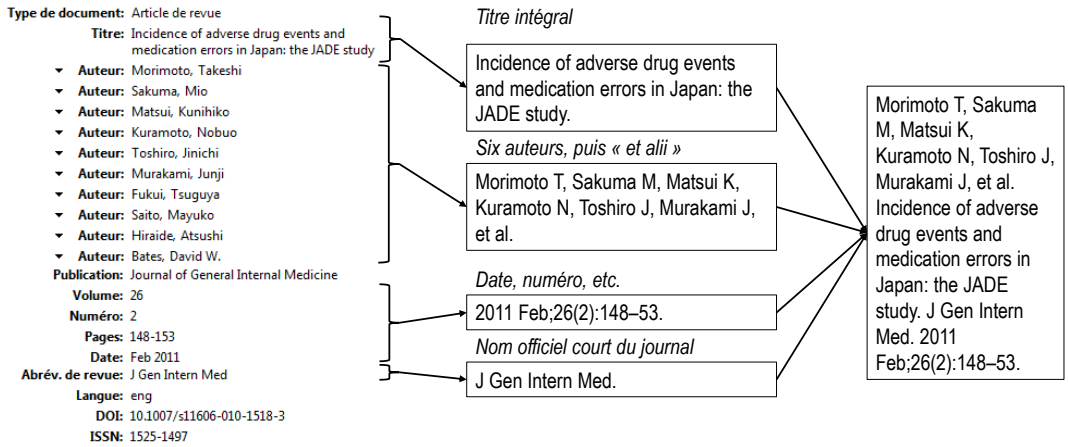


Figure 151. Des métadonnées à la référence mise en forme. Chaque journal diffère…

Vous voyez là la troisième difficulté : mettre en forme les références en fonction des exigences de chaque journal, en restant strictement homogène entre les différentes références du document.

Toutes ces difficultés, qui peuvent paraître insurmontables si gérées à la main, sont aisément surmontées par les utilisateurs de Zotero, logiciel gratuit et open source de gestion de la bibliographie.

2.2 Installer Zotero

Zotero est un logiciel multiplateforme, c'est à dire qu'il fonctionne sur tous les systèmes d'exploitation (Windows, Linux, Mac OS) et permet de synchroniser les références entre les différents appareils. Il est en outre interfaçable avec les navigateurs web (pour enregistrer rapidement des références) et avec les traitements de texte (pour insérer ces références dans vos documents).

La vidéo d'Objectif Thèse propose une démonstration simple et rapide de l'utilisation de Zotero : http://www.objectifthese.org

Avec un navigateur web quelconque, rendez-vous sur la page https://www.zotero.org/download/ puis téléchargez et installez Zotero.

Figure 152. Différents logos de Zotero

Cette installation comportera :

- Le logiciel principal de Zotero
- Le connecteur vers votre traitement de texte : Microsoft Word, LibreOffice Writer, OpenOffice Writer
- Le connecteur vers votre navigateur web : Chrome, Edge, Firefox, Safari
- Des modules permettant la lecture directe des documents PDF

Les logiciels déjà installés sur votre poste sont détectés, ce qui permet aux plugins de liaison d'être automatiquement installés. Si vous installez un autre logiciel par la suite, il est possible de réinstaller les plugins nécessaires via le menu du logiciel Zotero : *Edition* > *Paramètres*.

Figure 153. Navigateurs web gérés par les connecteurs de Zotero

Comme de nombreux logiciels, Zotero sera capable de vous prévenir de mises à jour, qui seront installées rapidement.

2.3 Créer un compte (facultatif et gratuit)

La création d'un compte sur le site https://www.zotero.org/user/register est facultative. Cela permet de sauvegarder automatiquement votre bibliographie en ligne. Le compte est gratuit pour une quantité de fichiers inférieure ou égale à 300Mo (c'est déjà énorme). Vous pourrez ensuite accéder à votre bibliographie :

- sur l'interface en ligne si vous travaillez sur un poste sur lequel Zotero n'est pas installé
- depuis n'importe quel ordinateur sur lequel vous aurez installé Zotero : votre bibliographie locale se met automatiquement à jour, sans effort !

Nous vous conseillons de créer le compte si vous êtes dans un des cas suivants :

- Situation 1 : vous utilisez Zotero fréquemment dans au moins un projet sur une longue durée (mémoire ou thèse) et vous ne souhaitez pas perdre tout votre travail en cas de plantage de votre ordinateur
- Situation 2 : vous travaillez simultanément sur plusieurs postes informatiques, et souhaitez synchroniser votre travail entre les différents postes sans effort
- Situation 3 : vous devez partager votre bibliographie, par exemple avec votre encadrant, ou dans le cadre d'un travail collectif
- Situation 4 : vous souhaitez créer une bibliothèque publique, par exemple pour alimenter automatiquement un site web avec vos publications

Si vous avez un doute, vous pourrez créer ce compte plus tard.

Pour **créer un compte**, créez tout d'abord votre compte sur la page https://www.zotero.org/user/register. Définissez un mot de passe spécifique à Zotero car il faudra le ressaisir dans l'interface : ne mettez pas le même mot de passe que pour votre messagerie personnelle par exemple. Puis lancez Zotero, et cliquez sur le menu *Edition* > *Paramètres* > *Synchronisation*. Saisissez alors le login et le mot de passe que vous avez définis lors de l'enregistrement.

Si vous souhaitez **partager votre bibliographie** avec vos collègues ou encadrants (ou même avec le public si vous le souhaitez), cliquez sur le menu : *Fichier* > *Nouvelle bibliothèque* > *Nouveau groupe…* Vous pourrez alors paramétrer cette bibliothèque partagée et inviter d'autres personnes à y contribuer.

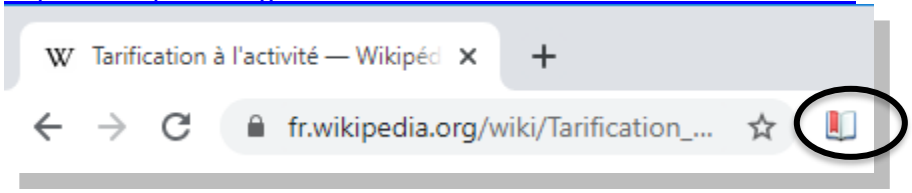
PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.4 Utiliser Zotero pour créer et maintenir votre bibliothèque personnelle

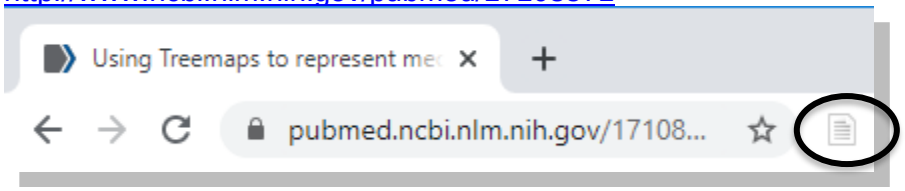
2.4.1 Insérer automatiquement un document dans Zotero

La manière la plus simple pour insérer un document dans Zotero est de le trouver sur Internet en utilisant le navigateur pour lequel vous avez intégré le plugin de Zotero :

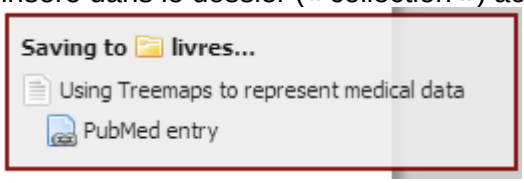
1. Lancez tout d'abord Zotero, positionnez-vous dans le dossier (« collection ») dans lequel vous souhaitez insérer le document (par défaut, restez dans la racine)
2. Naviguez sur internet avec votre navigateur. Lorsque des pages visitées sont prêtes à être référencées, une icône apparait dans la barre d'adresse du navigateur juste à droite de l'adresse. L'icône change selon la nature du document visité. Voici quelques exemples :
   a. La page « T2A » de Wikipedia :
   http://fr.wikipedia.org/wiki/Tarification_%C3%A0_l'activit%C3%A9



   b. Un article référencé Pubmed :
   http://www.ncbi.nlm.nlm.nih.gov/pubmed/17108571



3. Cliquez sur l'icône à droite de la barre d'adresse : le document est alors automatiquement inséré dans Zotero. Une discrète fenêtre s'affiche en bas à droite du navigateur, elle disparait en quelques secondes sans demander d'action de votre part. Elle vous informe que le clic a été pris en compte, et que le document a été inséré dans le dossier (« collection ») actuellement ouvert sous Zotero :



2.4.2 Enrichir un document déjà enregistré

Cette section concerne l'interface de Zotero.

Comme vous l'avez déjà vu dans l'aperçu rapide, cliquer sur un document permet d'afficher à droite les informations qui servent à indexer le document.

En cliquant sur le petit triangle à gauche du nom du document, on peut voir la liste des items disponibles dans le document. Il s'agit généralement de l'entrée « Pubmed » par exemple si le document provient de cette base de données, et du document PDF s'il est automatiquement rapatrié.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

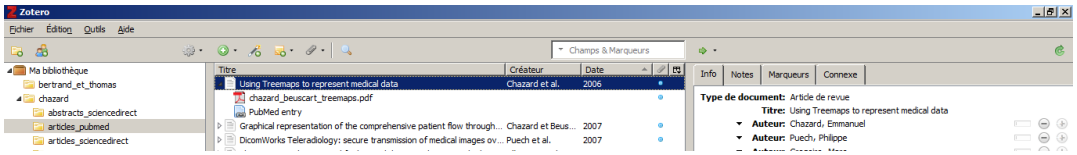


Figure 154. Visualisation d'une entrée dans Zotero

Il est possible d'ajouter à la main le document PDF d'un article, si vous vous l'êtes procuré d'une autre manière. Pour ce faire :
clic droit > ajouter une pièce jointe > joindre une copie enregistrée du fichier.

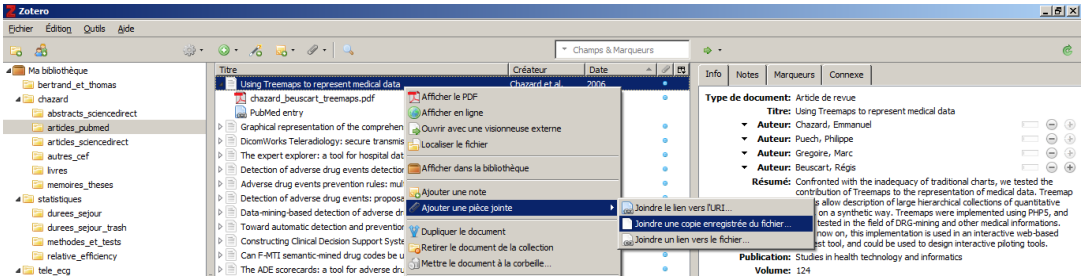


Figure 155. Joindre un fichier PDF à une entrée dans Zotero

Il est possible d'ajouter des annotations personnelles directement dans la fenêtre qui s'affiche aussi à droite lorsqu'on clique sur un document PDF.

L'ensemble de ces modifications (annotation, PDF joint, etc.) se synchronisent automatiquement et sans effort avec votre compte Zotero si vous en avez créé un, ou si vous décidez de le créer a posteriori.

2.4.3 Gérer les doublons

Pour détecter les doublons, il suffit de se rendre dans la collection « doublons », qui existe toujours en bas du volet de gauche. Les articles en doublons s'affichent automatiquement. Vous pouvez les sélectionner par groupe (ou un par un) pour les supprimer à l'aide du clic droit ou, mieux, les fusionner à l'aide du gros bouton qui s'affiche sur le volet de droite. S'il s'agit de vrais doublons, nous vous conseillons de ne jamais supprimer de document, mais plutôt de les fusionner. Ainsi, dans les documents qui utilisent déjà ces références, la correction sera faite automatiquement.

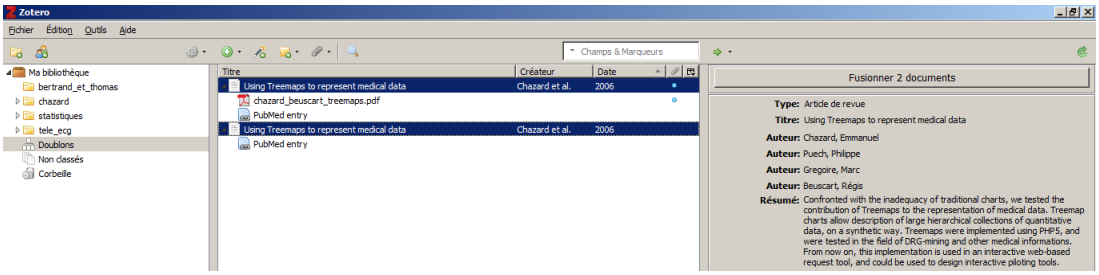


Figure 156. Fusion des doublons avec Zotero

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

2.5 Utiliser Zotero pour citer les références dans un traitement de texte

2.5.1 Intégration de Zotero dans le traitement de texte

L'installation de Zotero a permis l'ajout de boutons de commande dans votre logiciel de traitement de texte (Microsoft Word, LibreOffice Writer, OpenOffice Writer).

Dans Microsoft Word, un ensemble de boutons devient désormais accessible dans le ruban, sous l'intitulé « Zotero » (Figure 157).

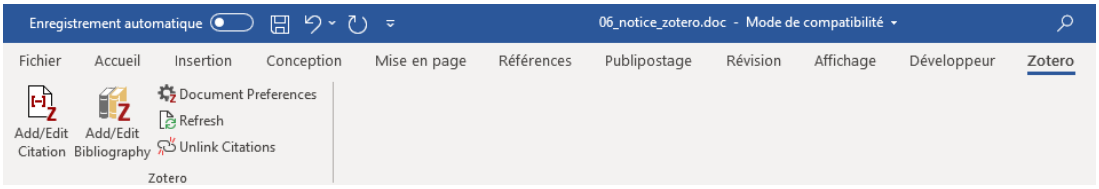


Figure 157. Boutons supplémentaires sous Microsoft Word

Dans LibreOffice Writer, il s'agit d'un jeu de boutons affichés dans le ruban (s'il ne s'affiche pas, invoquez la commande « Affichage > Barres d'outils > Zotero »), comme en haut à gauche en Figure 158.

Figure 158. Boutons supplémentaires sous LibreOffice Writer

Ces boutons correspondent aux actions suivantes (des étiquettes de légende apparaissent lorsqu'on survole les boutons) :

Insérer ou éditer une citation dans le texte

Insérer ou éditer une bibliographie (en fin de document)

Synchroniser avec Zotero (en cas de modification d'une référence dans Zotero alors qu'elle a déjà été citée dans le document)

Préférences de Zotero pour ce document texte

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Remplacer tous les liens dynamiques par du texte figé (il ne sera plus possible de mettre à jour la bibliographie, mais cette fonction peut être utile sur une copie d'un document)

2.5.2 Insérer ou modifier une citation dans le texte

Pour citer un élément de votre bibliothèque (article scientifique par exemple) dans votre traitement de texte, positionnez-vous là où vous souhaitez citer la référence, et cliquez sur le bouton ci-contre. La première fois, Zotero vous demande de choisir un style de citation, choisissez « Vancouver ».

Puis recherchez le document qui vous intéresse, de trois manières :

- Le dernier item incorporé s'affiche et peut être sélectionné
- Ou, commencez à saisir un mot du titre ou un nom d'auteur, puis cliquez sur la proposition qui vous convient
- Ou, si vous ne trouvez pas votre document de la sorte, cliquez sur le « Z » rouge à gauche puis sur « vue classique ». Double-cliquez sur le document qui vous intéresse

Quand vous avez fini de trouver le ou les documents, validez par « entrée » dans le cadre rouge.

Il vous est possible de modifier une citation, si par exemple vous souhaitez citer un deuxième document au même endroit, alors qu'il y en a déjà un. Pour ce faire, cliquez avec le bouton gauche sur la citation, puis de nouveau sur le bouton ci-contre. Ne rajoutez pas une nouvelle citation à l'aide du premier bouton : les deux zones ne seront pas fusionnées.

Actuellement, il n'existe pas de procédure automatique dans Zotero pour fusionner deux citations adjacentes. Vous pouvez utiliser une macro disponible sur Internet, ou réintégrer les citations manuellement.

2.5.3 Insérer une bibliographie en fin de document

Une fois votre document peuplé de citations, la création de la bibliographie (avec les bons numéros) est automatique. Pour ce faire, rendez-vous à l'endroit où vous souhaitez la créer, puis appuyez sur le bouton ci-contre : la création est automatique et instantanée, il n'y a rien d'autre à faire. La bibliographie se mettra à jour automatiquement chaque fois que nécessaire, sans nécessiter d'action de votre part.

Si vous utilisez le modèle de document proposé sur Objectif Thèse, les références sont déjà créées à la fin du document.

Vous pourrez très aisément modifier le style de citation via le menu de Zotero. Pour un mémoire académique, je vous conseille le style « *Vancouver (superscript, brackets, only year in date)* » qui est celui utilisé dans le présent document. Le style « *Elsevier Vancouver* » est également apprécié. Pour un article scientifique, utilisez simplement le style portant le nom du journal.

2.5.4 Autres fonctionnalités

De très nombreuses fonctionnalités sont disponibles. Elles sont sans cesse plus nombreuses (intégration dans Google Drive, génération automatique de pages web présentant une bibliographie, etc.).

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3 Rédiger les différentes parties du mémoire

3.1 Organisation selon le plan IMMRaD

Depuis une trentaine d'années, les articles scientifiques suivent presque tous le plan **IMMRaD**, pour *introduction material methods results and discussion*. Vous devez également suivre strictement ce plan pour la rédaction de votre mémoire académique. De prime abord, on pourrait penser que les meilleurs scientifiques sont les personnes les plus créatives, et que cette créativité devrait se ressentir dans la rédaction scientifique. Il n'en est rien. Les meilleurs scientifiques sont surtout les personnes les plus rigoureuses, celles qui maîtrisent parfaitement l'état de l'art (ce qui existait avant leur travail), qui mettent à jour des connaissances en s'intégrant parfaitement dans les connaissances existantes et l'état de l'art des méthodes utilisables, et relatent leurs travaux sans jamais flouer le lecteur.

Le fait de suivre un plan imposé identique pour tous les articles scientifiques et mémoires académiques présente plusieurs avantages :

- En cas de lecture non-linéaire (ce qui est presque toujours le cas), on sait parfaitement où trouver l'information, en fonction du type d'information qu'on cherche. Il n'est pas nécessaire de lire un article en entier pour trouver les informations utiles
- Cette rédaction sépare clairement les éléments objectifs (pour lesquels le lecteur n'a pas besoin d'être sur ces gardes) des éléments interprétatifs ou subjectifs (pour lesquels le lecteur pourra avoir un avis différent)
- Cette rédaction sépare clairement les connaissances issues du travail (internes), des connaissances issues d'autres travaux (externes), qui doivent naturellement être cités

Le plan IMMRaD, que nous développerons plus bas, correspond à celui d'une **recette de cuisine** :

- **Introduction** : contexte de réalisation, avantages et défauts connus d'autres recettes, objectif de la recette
- **Matériel** (au sens « matériaux ») : ingrédients nécessaires à la recette (œufs, farine, beurre, etc.)
- **Méthodes** : actions réalisées sur les ingrédients (couper, hacher, mélanger, cuire, laisser reposer, etc.)
- **Résultats** : résultat brut (un cake de 500g prêt en 30 minutes de préparation et 1h30 de cuisson)
- **Discussion** : commentaire des résultats au regard des objectifs et de ce qui était connu, autres possibilités de procédé, etc.

3.2 Rédaction de l'introduction

3.2.1 Objet de l'introduction

L'introduction **pose le contexte** en exposant les connaissances utiles à prendre en compte avant de réaliser le travail. Sa rédaction nécessite une **recherche bibliographique** approfondie et préalable à la réalisation du travail. L'introduction sera ainsi caractérisée par la présence de très nombreuses références bibliographiques (c'est aussi le cas de la discussion, mais pas des autres parties). Les éléments cités dans l'introduction sont donc principalement **externes et antérieurs** à l'étude, mais s'y mêlent également des éléments **subjectifs** qui critiquent utilement l'existant, et justifient ainsi l'existence du travail qui sera ensuite réalisé. Néanmoins, aucun élément propre à l'étude ne doit encore être mentionné, cela viendra plus tard.

3.2.2 Sous-partie « contexte »

La majeure partie de l'introduction présente donc l'état de l'art (les connaissances actuelles et les méthodes de recherche) et le contexte du travail. Dans un article scientifique, sans le dire clairement, cette partie devra principalement répondre à quatre critiques majeures :

Critique n°1 : Ce travail ne sert à rien

Pour répondre à cette critique, il vous faudra dresser l'état de l'art de ce qui est connu dans le domaine, pour faire apparaître un manque, et convaincre le lecteur que ce travail était indispensable.

Critique n°2 : Ce travail a déjà été fait

Si un travail est nécessaire, le lecteur bougon pensera que, bien entendu, ce travail a déjà été réalisé par des personnes plus brillantes que vous. Votre difficulté sera la suivante : il est possible de montrer qu'un travail a déjà été fait, simplement en le citant, mais il est impossible de montrer avec certitude qu'il n'a pas été réalisé. Cependant, lorsqu'on affirme qu'un travail n'a pas été fait, on est plus crédible si on peut montrer qu'on a lu de nombreux travaux dans le même domaine, et qu'aucun d'eux n'est similaire au travail en cours. Il s'agit donc ici plus de crédibilité que de preuve.

Critique n°3 : Ce travail n'est pas réalisable ou n'apportera pas la réponse attendue

Vous aurez besoin là aussi de montrer qu'un faisceau d'éléments, de préférence externes, laisse penser que le travail est réalisable et apportera la réponse attendue.

Critique n°4 : vous n'êtes pas la bonne personne ou pas la bonne équipe

On revient là à un problème de crédibilité. Cependant, si vous avez répondu aux trois premières critiques, vous êtes déjà crédible. En particulier, grâce à la bibliographie réalisée et exposée pour réfuter les trois premières critiques, vous serez capable de produire des résultats, éventuellement peu ambitieux, mais parfaitement intégrables dans le mur des connaissances collectives, telle une brique de petite taille, mais solide et des mêmes dimensions que les autres briques.

Dans un article scientifique, l'introduction tiendra généralement sur une page ou moins. Dans un mémoire académique, cette introduction occupera entre 10 et 30 pages selon le cas. Dans ce cas, en plus de l'introduction qu'on rédigerait pour un article, on complètera l'introduction avec des rappels des connaissances académiques sur le sujet, qu'on pourrait lire par exemple dans un ouvrage didactique.

Par exemple, si le travail porte sur une pathologie particulière, on pourra suivre le plan suivant :

- Rappels anatomiques
- Rappels physiologiques
- La pathologie en question : physiopathologie, épidémiologie, facteurs de risque, etc.
- La question plus précise qui est posée…

3.2.3 Sous-partie « objectif »

L'introduction se conclut toujours par un **court** paragraphe qui énonce l'**objectif** du travail. Il peut s'agir d'une seule phrase, concise et précise.

En recherche clinique, il est fréquent de distinguer un objectif principal et des objectifs secondaires. Cette dichotomie est rendue nécessaire par le calcul du nombre de sujets nécessaire, et le raisonnement très particulier lié au risque de première espèce, puisqu'à cet objectif principal correspondra un seul test statistique. Dans les recherches sur les données, ou les recherches en santé dont la portée est moins ambitieuse que conclure sur la supériorité d'un traitement sur un autre, cette distinction principal/secondaire n'est pas nécessaire.

Il est également possible d'énoncer un objectif opérationnel (qu'allons-nous faire ici) et un objectif stratégique (au fond, à quoi cela servira-t-il). En voici un exemple :

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

L'objectif stratégique est de contribuer à réduire la morbimortalité liée aux effets indésirables du médicament. Pour ce faire, l'objectif opérationnel est d'évaluer l'incidence des effets indésirables hémorragiques liés aux anticoagulants durant les hospitalisations tout venant.

3.3 Rédaction de la partie Matériel et méthodes

3.3.1 Exemple de structuration

En rédaction scientifique, le titre « matériel » est en fait un anglicisme : c'est la traduction impropre de « *material* », qui désigne le « matériau » et non le « matériel ». Dans certains cas, cette section « matériel » est séparée de la section « méthodes » : on y met alors la base de données utilisée, les patients, etc.

Dans une **recherche sur des patients**, le plan pourra être le suivant :

- *Design* de l'étude
    - Forme générale (cohorte, exposé/non-exposé, cas/témoin, etc.)
    - Critères d'inclusion / exclusion (âge, sexe, se présentant à, tirage au sort, patients consécutifs, etc.)
    - Critères d'exposition (variable définissant les groupes notamment)
    - Critère principal de jugement (survie, score, etc.)
    - Critères secondaires de jugement
    - …
- Patients et établissements (personnes incluses, services, hôpitaux)
- Données (données recueillies sur les patients)
- Analyse de données (ce qu'on souhaite calculer ou tester, sans citer les méthodes statistiques)
- Analyse statistique (méthodes employées, sans citer les variables)
- Cadre réglementaire (financement, CPP, CNIL)

Dans une **recherche sur des données**, le plan pourra être le suivant :

- Base de données source / ensemble de dossiers patients / etc.
- *Design* de l'étude
    - Forme générale (le plus souvent cohorte historique)
    - Critères d'inclusion / exclusion (âge, sexe, critères, etc.)
- Données disponibles (données natives)
- Extraction de caractéristiques (construction de variables depuis les données, notamment par des regroupements de codes)
- Analyse de données (ce qu'on souhaite calculer ou tester, sans citer les méthodes statistiques)
- Analyse statistique (méthodes employées, sans citer les variables)
- Cadre réglementaire (financement, CNIL)

On notera que la partie « analyse de données » indique ce qu'on souhaite analyser ou comparer, en citant clairement les variables impliquées, mais sans citer les méthodes statistiques.

Ex : Nous comparerons les survies sans réhospitalisation pour reprise chirurgicale entre les groupes définis par le type de chirurgie, et chercherons à identifier les facteurs influents. Nous ferons de même avec la survie jusqu'au décès intra-établissement, quelle que soit la cause du décès.

Inversement, la partie « analyse statistique » indique quelles méthodes seront utilisées en fonction du cas de figure, sans jamais citer les noms des variables.

Ex : Les survies seront décrites avec l'estimateur de Kaplan-Meier. Les facteurs de risque d'événement temps-dépendant seront identifiés et caractérisés à l'aide d'un modèle de Cox. Les résultats seront exprimés en termes de hazard ratios avec intervalles de confiance à 95%.

Dans l'exemple ci-dessus, on comprend qu'il y aura deux analyses multivariées de survie, une pour prédire la réhospitalisation et l'autre pour prédire le décès. Pourtant, la méthode statistique n'est décrite qu'une seule fois, de manière générique. Ceci permet d'éviter les redites. De même, les noms des méthodes statistiques ne seront pas cités dans les résultats, par la suite.

Nous reviendrons immédiatement après sur la partie dédiée aux analyses statistiques.

Une toute dernière partie, **cadre réglementaire**, précise, le cas échéant :

- le financement éventuel de l'étude
- les autorisations CNIL, CPP, etc.
- l'information des sujets et le recueil de leur consentement

3.3.2 Chapitre « Analyse statistique » en particulier

Le paragraphe « analyses statistiques » est habituellement un des derniers paragraphes de la section « matériel et méthodes ». Vous pourrez reprendre directement le texte qui suit. Une fois les analyses terminées, il sera important de supprimer toutes les informations qui ne sont pas nécessaires.

Notez que le modèle de documents fourni sur le site http://objectifthese.org contient déjà un texte prêt à l'emploi.

3.3.2.1 Analyses univariées

Les variables qualitatives, binaires, ou discrètes avec très peu de modalités sont exprimées en effectif et pourcentage.

Les variables quantitatives sont exprimées en moyenne et écart type (SD) si l'histogramme révèle une distribution d'allure symétrique, et médiane premier et troisième quartile (Q1, Q3) dans le cas contraire.

Les survies sont étudiées avec l'estimateur de Kaplan-Meier. Les intervalles de confiance des survies à 95% (IC95) sont calculés à l'aide d'une loi normale.

[souvent inutile] *Les intervalles de confiance des proportions à 95% (IC95) sont calculés à l'aide d'une loi [choisir] binomiale / normale.*

[souvent inutile] *Les intervalles de confiance des moyennes à 95% (IC95) sont calculés à l'aide d'une loi de Student.*

3.3.2.2 Analyses bivariées

La relation entre deux variables qualitatives est analysée à l'aide d'un test [choisir] *exact de Fisher / du Khi².*

La relation entre une variable qualitative et une variable quantitative est analysée à l'aide [choisir] *d'un test de Student / d'une analyse de la variance ANOVA / d'un test de Wilcoxon-Mann-Whitney / d'un test de Kruskal-Wallis.*

La relation entre deux variables quantitatives est analysée à l'aide [choisir] *du test de nullité du coefficient de corrélation de Pearson / du test de nullité du coefficient de corrélation de Spearman / du test de nullité de la pente d'une régression linéaire simple.*

La relation entre une variable de survie et une variable qualitative est analysée à l'aide d'un test du Log Rank.

3.3.2.3 Analyses multivariées

Les relations entre les covariables candidates et une variable quantitative sont modélisées et testées à l'aide d'une régression linéaire multiple. Les résultats sont exprimés en termes de coefficient assorti d'un intervalle de confiance à 95%.

Les relations entre les covariables candidates et une variable binaire sont modélisées et testées à l'aide d'une régression logistique. Les résultats sont exprimés en termes d'odds ratio (OR) assorti d'un intervalle de confiance à 95%.

Les relations entre les covariables candidates et une variable de survie sont modélisées et testées à l'aide d'un modèle de Cox. Les résultats sont exprimés en termes de hazard ratio (HR) assorti d'un intervalle de confiance à 95%.

[Préciser dans tous les cas une des options suivantes] :

- *Seules les covariables retrouvées dans la littérature sont incluses dans l'analyse*
- *Seules les covariables associées en bivarié à la variable d'intérêt avec une p valeur inférieure à 20% sont incluses dans l'analyse*
- *Les covariables disponibles sont toutes incluses dans l'analyse, et sont sélectionnées automatiquement à l'aide d'une procédure pas-à-pas* [choisir] *ascendante/descendante/bidirectionnelle. Seul le modèle final est présenté.*
- *Les covariables disponibles sont toutes incluses dans l'analyse, et sont filtrées itérativement à dire d'expert. Seul le modèle final est présenté.*

3.3.2.4 Significativité

Les tests statistiques sont bilatéraux. Les p valeurs sont considérées comme significatives au seuil de 5%. Les intervalles de confiance sont calculés à 95%.

3.3.3 Non pas l'historique, mais le procédé garanti *a posteriori*

Lorsque vous lisez un livre de recettes de cuisine, l'auteur vous indique ce qu'il faut faire pour obtenir le résultat décrit. L'auteur ne vous raconte pas comment lui-même a tenté diverses options, essuyé des échecs, pour finalement arriver à ce résultat.

De la même manière, il ne s'agit pas ici de décrire **ce que vous avez réellement fait**, mais plutôt **ce qu'il faudrait faire pour atteindre ce résultat**, autrement dit, ce que vous feriez si vous deviez retrouver les mêmes résultats mais plus rapidement, à partir des mêmes données. Il ne s'agit bien sûr pas de mentir, car vous devez garantir que l'application des méthodes à votre matériel donne exactement le résultat présenté. Mais si, historiquement, vous avez appliqué les procédés A, B et C, et que vous êtes absolument certain que le procédé D donne exactement le même résultat, vous pouvez présenter uniquement le procédé D.

Exemple : « On a d'abord inclus tous les patients, puis on a refait l'analyse en excluant les patients de plus de 70 ans, puis on a refait l'analyse en excluant les patients de moins de 15 ans. »

Deviendra : « Les patients âgés de moins de 15 ans ou de plus de 70 ans sont exclus. »

3.4 Rédaction de la partie résultats

3.4.1 Généralités

La partie résultats produit les résultats de vos analyses, de la manière la plus **froide et neutre** possible.

Dans un livre de recettes de cuisine, l'auteur garantit au lecteur que, s'il applique les méthodes décrites aux matières premières décrites, il obtiendra strictement le résultat décrit.

De la même manière, une bonne rédaction scientifique garantit au lecteur que, s'il applique les méthodes décrites aux matériaux décrits, il obtiendra strictement le résultat décrit.

Ceci implique plusieurs choses importantes :

- Les résultats sont **neutres**, et ne présentent **aucun élément d'interprétation** des résultats : ces éléments seront déplacés dans la discussion
- Les résultats ne rappellent **aucun élément de contexte**, **aucune comparaison** des méthodes à la littérature, **aucune comparaison** des résultats à la littérature : tous ces éléments seront déplacés dans l'introduction ou la discussion
- Les résultats ne présentent pas l'ensemble des résultats historiquement obtenus par l'auteur, y compris ses tâtonnements, hésitations, revirements, résultats inutiles, etc. Si l'auteur devait reprendre l'ensemble de l'étude en un temps limité, voici ce qu'il obtiendrait.
- Les résultats ne rappellent **pas les noms des méthodes** utilisées, en particulier les méthodes statistiques. Elles sont présentées dans la section méthodes.
- Généralement, la section résultats ne contient **pas ou peu de référence bibliographique**.

3.4.2 Exemples de plans de résultats

Le modèle de document Word proposé sur le site Objectif Thèse contient déjà une suggestion de plan générique : http://www.objectifthese.org

Nous vous proposons ci-dessous un exemple de plan qui conviendra très bien pour la plupart des études cliniques s'intéressant à des **patients**, ou réutilisant des **données** portant sur des personnes :

- *Flowchart* (diagramme de flux d'inclusion des patients : cette partie sera détaillée dans le chapitre suivant)
- Descriptif des patients à l'inclusion
- Suivi des patients dans le temps
- [puis réponse aux questions posées en objectif]
- …

Pour un **questionnaire**, le plan sera un peu plus simple :

- *Flowchart* (allant des personnes initialement sondées, aux questionnaires finalement analysés ; cf. chapitre suivant)
- Caractéristiques des participants
- [puis descriptif des réponses, dans un ordre thématique]
- …

Pour une **revue de la littérature**, le plan peut également être le suivant :

- *Flowchart* (allant des articles initialement inclus, aux articles finalement retenus ; cf. chapitre suivant)
- Processus d'annotation (si plusieurs personnes étaient chargées de l'annotation, préciser la concordance inter-juges, le processus de mitigation, etc.)
- Caractéristiques des articles retenus
- Critères de qualité des articles retenus
- [puis réponse aux questions posées en objectif]
- …

Au sein de chaque partie, l'organisation doit sembler naturelle au lecteur, de manière qu'il puisse trouver l'information qu'il cherche sans lire le paragraphe en entier, de manière intuitive. Par exemple, lorsqu'on souhaite décrire des patients, on peut utiliser un ordre qui est chronologique dans le cours de la vie et de la maladie de ces patients :

- âge, sexe (car liés à la naissance)
- maladies chroniques préexistantes, facteurs de risque de la maladie d'intérêt

- informations au diagnostic de la maladie d'intérêt (ex : âge, symptômes, mode de découverte, etc.)
- prise en charge initiale (ex : admission aux urgences, etc.)
- prise en charge principale, dans un ordre naturel (ex : chirurgie, traitement)
- suites immédiates (ex : réanimation, décès, sortie de l'hôpital, etc.)
- suites à plus long terme (ex : récidives, complications, réinterventions, rémission, guérison, etc.)

3.4.3 Flowchart

3.4.3.1 Principe général

Le ***flowchart***, ou **diagramme de flux**, constitue très souvent la première partie d'une section « résultats ».

Le principe général d'un *flowchart* est présenté en Figure 159. Le nombre total d'individus est présent en haut du diagramme. Puis, par une succession d'étapes, certains individus sont **exclus** et figurent **sur la droite** du graphique. **En bas** du graphique, on trouve les individus **finalement inclus** dans l'étude. Chaque rectangle décrit un sous-ensemble d'individus et documente son effectif avec « (n=…) ». Il est indispensable que le lecteur puisse contrôler les effectifs et calculer lui-même des proportions lorsqu'il le souhaite. Le contrôle des effectifs ne doit pas poser de problème. Dans l'exemple de la Figure 159 :

- A chaque nœud, le nombre au-dessus est égal à la somme des exclus (à droite) et des restants (en-dessous) :
  dans l'exemple de la Figure 159, $a = b + c$ et $c = e + d$
- La différence entre le début est la fin du flowchart se retrouve dans l'ensemble des exclusions :
  dans l'exemple de la Figure 159, $a - e = b + d$

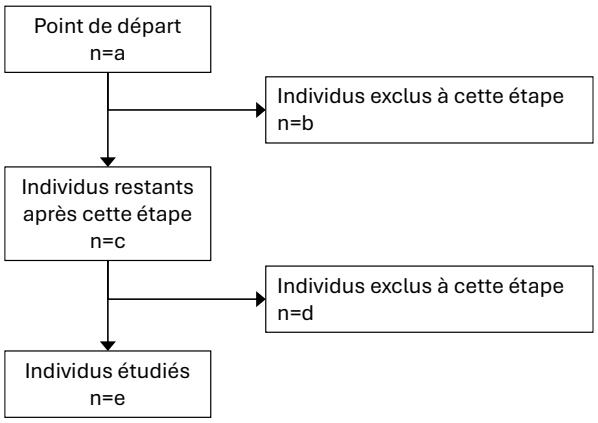


Figure 159. Structure générale d'un flowchart

Ce principe général se décline aisément selon le type d'étude. Passé ce principe général, il n'y a aucune règle, et chaque auteur est libre de présenter le diagramme qui lui parait le plus pertinent. Vous pourrez vous inspirer des trois exemples qui suivent, ou des exemples déjà publiés dans votre discipline. Dans ce cas, privilégiez les articles récents, car la tradition des *flowcharts* est actuellement en train de se construire, et les exemples anciens ne sont pas forcément les meilleurs.

3.4.3.2 Trois exemples de flowchart

La Figure 160 montre un exemple de *flowchart* correspondant à une revue de la littérature. Il devra être complété par une équation de recherche, présentée dans la partie « méthodes ».

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

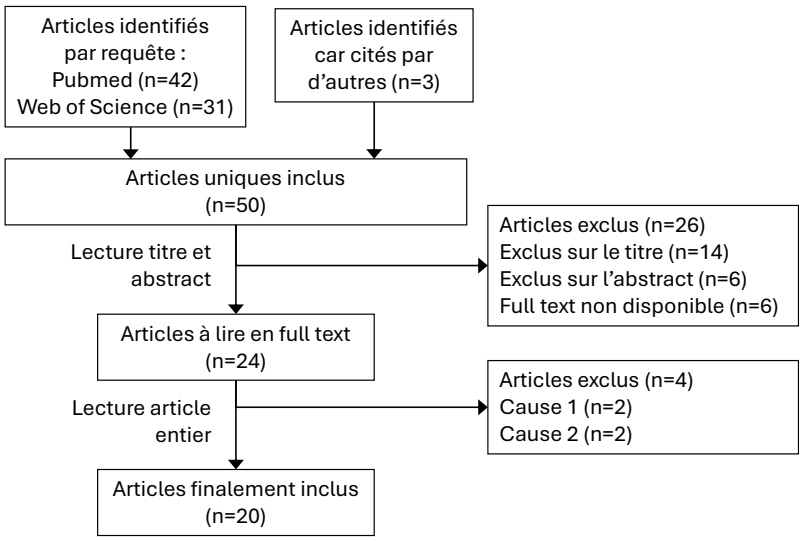


Figure 160. Exemple de flowchart pour une revue de la littérature

La Figure 161 présente un *flowchart* pour un questionnaire papier envoyé à des médecins. Les effectifs présentés doivent permettre au lecteur de calculer le taux de sondage et, surtout, le taux de réponse (ces notions ont été évoquées dans le chapitre 7 Questionnaires : taux de sondage, taux de réponse en page 51). On rappellera que les questionnaires numériques ne permettent généralement pas de réaliser un *flowchart* exhaustif, ce qui est regrettable.

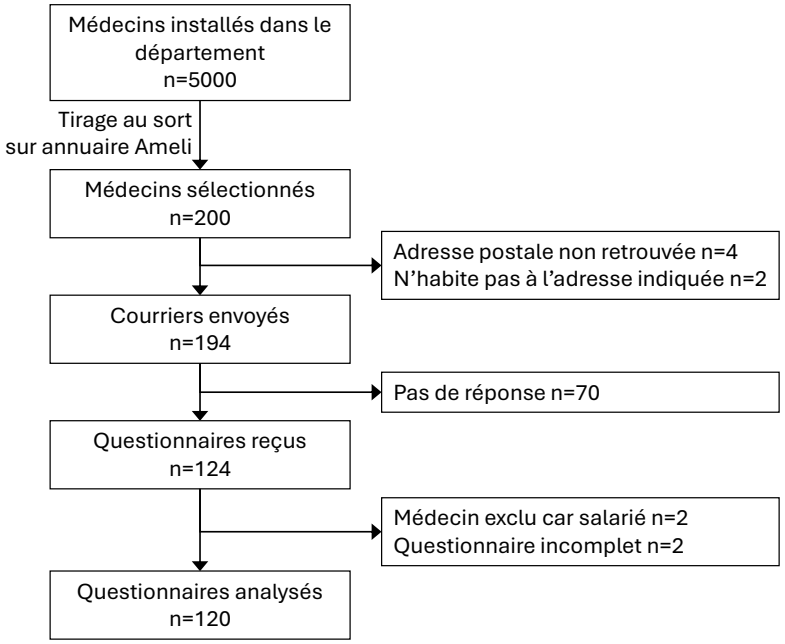


Figure 161. Exemple de flowchart pour un questionnaire papier

La Figure 162 présente un *flowchart* pour une étude réalisée sur des patients, ou sur des dossiers de patients. Ce *flowchart* détaille les critères d'exclusion rencontrés. Dans cet exemple, le *flowchart* indique également la constitution des deux groupes de patients qui seront comparés.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

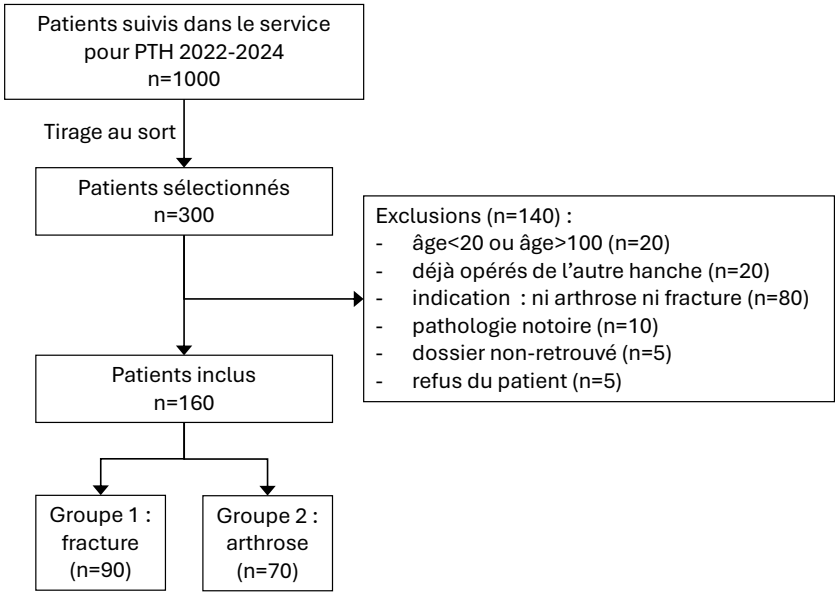


Figure 162. Exemple de flowchart pour une étude sur des patients ou des dossiers

3.4.3.3 Conseils techniques pour la réalisation du flowchart

Concernant la réalisation pratique du *flowchart*, nous vous conseillons d'utiliser un logiciel de présentation tel **Microsoft Powerpoint** (qui est très agréable à utiliser) ou **LibreOffice Impress** (qui est bon, gratuit et open source). Ces logiciels de présentation sont initialement destinés à réaliser des diapositives, ou « slides ». Cependant, ils sont très pratiques à utiliser pour réaliser des diagrammes en dessin vectoriel. Si vous préférez utiliser un logiciel de dessin vectoriel, vous atteindrez le même résultat en un temps généralement supérieur.

Le site Objectif Thèse vous propose un document Powerpoint contenant déjà des *flowcharts*, aisément réutilisables. La vidéo jointe montre certaines manipulations utiles à la construction de *flowcharts* : http://www.objectifthese.org

Nous insistons en particulier sur deux fonctions utiles à la réalisation de *flowcharts*.

La première concerne les **connecteurs**. Dans Impress ou dans les anciennes versions de Powerpoint, il s'agissait de composants spécifiques. Dans les versions actuelles de Powerpoint, il s'agit désormais de flèches, tout simplement. La Figure 163 illustre l'aspect de ces connecteurs. Le premier est une flèche rectiligne. Le deuxième est un connecteur en angle : moins esthétique, il peut s'avérer préférable dans certains diagrammes. Les deux premières flèches, lorsqu'elles sont sélectionnées, affichent des extrémités colorées différemment, qui signifient qu'elles sont « collées » aux rectangles (il suffit de déplacer leur extrémité au-dessus d'un point d'accroche, ceci est alors automatique). L'avantage est que, si les rectangles sont déplacés ou redimensionnés, ces connecteurs suivent automatiquement, ce qui laisse le diagramme cohérent. Dans le troisième exemple de la Figure 163, la flèche n'est pas connectée aux rectangles, ce qui se voit parce que ses extrémités restent de la couleur native lorsqu'elle est sélectionnée. En cas de déplacement des rectangles, la flèche ne suivra pas, ce qui est regrettable.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

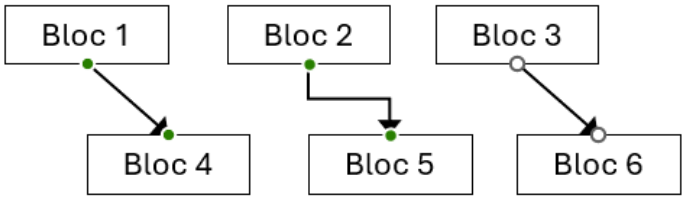


Figure 163. Flèches (aspect dans Powerpoint après sélection ; de G à D) : flèche trait, connecteur en angle, ou flèche non-reliée

La deuxième fonction très utile concerne les **alignements et répartitions**. Une fois votre *flowchart* réalisé, ne perdez pas de temps à placer précisément les rectangles à la main : utilisez les fonctions automatiques. Si vos connecteurs sont bien connectés, il suffit de sélectionner les rectangles pour les aligner ou les répartir, sans vous soucier des connecteurs, qui suivront automatiquement. Sélectionnez les rectangles, puis utilisez une des **huit fonctions** disponibles, dont trois sont présentées dans la Figure 164. La fonction « distribuer verticalement » est très utile : le logiciel identifie la forme la plus haute et la forme la plus basse, puis déplace toutes les autres de sorte que les espaces entre les formes soient homogènes. Nous vous laissons découvrir les autres fonctions : après sélection des formes :

Dans Microsoft Powerpoint : *menu Forme > Aligner > [choisir la fonction]*

Dans LibreOffice Impress : *menu Format > Aligner [ou] Répartition*

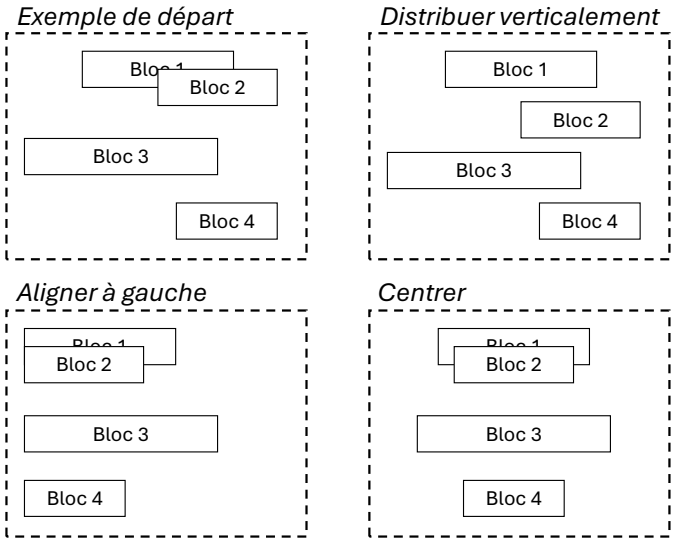


Figure 164. Trois exemples de fonctions d'alignement ou répartition automatique

Avant de positionner les éléments, il peut être utile de redimensionner certains d'entre eux non pas avec la souris, mais en entrant leur mesure exacte : c'est la manière la plus simple et rapide de s'assurer que certains auront exactement la même taille.

3.4.4 Résultats d'analyses statistiques

Pour ce qui concerne en particulier les résultats d'analyses statistiques, il est très important de suivre les recommandations et de ne pas produire trop d'indicateurs statistiques. Voir pour ce faire le cours dédié à la présentation des résultats statistiques, sur la page http://objectifthese.org (vidéo et diaporama).

Nous avons déjà proposé des modes de rédaction au fil des paragraphes de la section dédiée aux analyses statistiques, par exemple dans la section 2 Analyses statistiques univariées en

page 105. Il est très important que vous **suiviez strictement** dans les résultats les consignes que vous avez vous-mêmes énoncées dans la section « matériel et méthodes ».

Vous verrez dans les exemples ci-dessous que, de manière générale :

- Ces phrases sont **simples, prévisibles**, et permettent au lecteur de trouver très rapidement l'information, même si sa lecture est saltatoire et non continue (contrairement à la littérature)
- Ces phrases produisent le strict minimum en termes d'indicateurs numériques
- Ces phrases ne contiennent **aucune interprétation** du résultat (ce sera fait dans la section « discussion »)
- Ces phrases ne cherchent pas, contrairement à ce qu'on fait en littérature, à éviter les répétitions ou rechercher une quelconque esthétique linguistique
- Ces phrases **ne rappellent pas les méthodes statistiques** employées : cela est fait dans la section « méthodes »

Nous vous proposons ici des phrases type, en fonction du type d'analyse statistique :

- **Analyses univariées :**
  - **Variable qualitative :**
    L'échantillon comporte 54 femmes (27,0%).
    Parmi les patients, 26 (13,0%) ont un cancer de grade 1, 32 (18,0%) de grade 2, et 12 (6,0%) de grade 3.
  - **Variable quantitative symétrique :**
    L'âge moyen est de 54 ans (SD=12).
  - **Variable quantitative asymétrique :**
    La durée médiane de séjour est de 2 jours (Q1-Q3 : [0 ;5])
- **Analyses bivariées :**
  - **Qualitatif-qualitatif :**
    La proportion de malades diffère significativement selon le sexe (10,2% chez les hommes, 8,1% chez les femmes, p=0,023).
    ou
    La proportion de malades ne diffère pas significativement selon le sexe (respectivement 10,2% et 10,3% chez les hommes et les femmes, p=0,83).
  - **Qualitatif-quantitatif :**
    Les femmes sont significativement plus âgées que les hommes (respectivement 62,3 ans et 56,8 ans, p=0,013).
    ou
    L'âge ne diffère pas significativement en fonction du sexe (respectivement 62,3 ans et 31,8 ans pour les femmes et les hommes, p=0,32).
  - **Quantitatif-quantitatif :**
    L'âge et la durée de séjour sont en relation linéaire croissante (r=0,620, r^2=38,4%, p=0,003). L'équation de la droite est : (…)
    ou
    On n'observe pas de corrélation linéaire entre l'âge et la durée de séjour (p=0,31).
- **Analyses multivariées :**
  - **Régression linéaire multiple :**
    Les facteurs suivant sont associés de manière croissante à la variable Y (coefficient et IC95%) : l'âge (0,01 [0,005 ; 0,015]), le sexe masculin (1,5 [0,005 ; 0,015]), (…).
    Les facteurs suivant sont associés de manière décroissante à la variable Y : (…)
    Les facteurs suivants sont associés à des coefficients non-significativement différents de zéro : (…)
  - **Régression logistique :**
    Les facteurs suivant apparaissent comme des facteurs de risque (odds ratio

ajusté et IC95%) : l'âge (1,01 [1,005 ; 1,015]), le sexe masculin (1,50 [1,48 ; 1,52]), (…).
Les facteurs suivants apparaissent comme des facteurs protecteurs : (…)
Les facteurs suivants sont associés à des odds ratios non-significativement différents de un : (…)

- **Modèle de Cox :**
  Les facteurs suivant apparaissent comme des facteurs de risque (hazard ratio ajusté et IC95%) : l'âge (1,01 [1,005 ; 1,015]), le sexe masculin (1,50 [1,48 ; 1,52]), (…).
  Les facteurs suivants apparaissent comme des facteurs protecteurs : (…)
  Les facteurs suivants sont associés à des hazard ratios non-significativement différents de un : (…)

Certains des résultats, en particulier pour les analyses multivariées, pourront être présentés en tableau ou en graphique : dans ce cas il ne sera pas nécessaire de recopier ces résultats dans le texte. Le texte peut alors se contenter de renvoyer le lecteur aux tableaux.

3.5 Rédaction de la discussion

3.5.1 But et organisation générale

La discussion suit généralement un plan libre. Elle poursuit plusieurs objectifs officieux :

- Rappeler au lecteur que le travail répond exactement à l'objectif
- Rappeler les principaux résultats et leur interprétation
- Discuter les forces et faiblesses de l'étude, tant en termes de méthodes que de résultats
- Réintégrer les résultats obtenus dans l'ensemble de la connaissance scientifique

On constate que certains de ces objectifs font uniquement appel au travail réalisé, mais la plupart d'entre eux s'appuient largement sur les travaux publiés par d'autres chercheurs. De ce fait, il est habituel de citer de nombreuses références bibliographiques. En particulier, il est habituel de **réutiliser les références** déjà citées dans l'introduction.

Nous vous proposons le plan suivant, qui peut être utilisé dans la grande majorité des travaux :

- Principaux résultats
- Discussion de la méthode
  - Discussion interne
  - Discussion externe
- Discussion des résultats
  - Discussion interne
  - Discussion externe
- Perspectives

Nous détaillerons ces sous-parties ci-après.

Pour un mémoire académique, il est usuel de détailler le plan à l'aide de sous-parties. Dans une thèse de médecine par exemple, la discussion devrait faire une dizaine de pages au moins.

Pour un article scientifique, la discussion est beaucoup plus courte (ex : une page), dense, et le plan n'est pas explicité. Cependant, nous vous conseillons de travailler en écrivant ce plan, puis de supprimer les sous-titres au dernier moment. Il en résultera une discussion cohérente et bien structurée.

3.5.2 Principaux résultats

Dans un style « veni, vidi, vici »[23], ou à la manière d'un artisan qui reprend son devis pour décrire ses réalisations lors de la remise d'un chantier, énumérez ce que vous vouliez faire, ce que vous avez fait, et ce que vous avez obtenu. Il ne s'agit pas de refaire le catalogue des résultats, mais seulement d'insister sur la cohérence du travail et de rappeler le principal résultat.

A la lecture de cette courte partie, le lecteur doit se dire « c'est clair, c'est cohérent, c'est exactement ce qu'on voulait savoir ».

3.5.3 Discussion de la méthode

Il s'agit ici de discuter les méthodes mises en œuvre, à la fois d'un point de vue interne, et externe.

D'un point de vue interne, il faudra lister de manière honnête et constructive les **biais de l'étude** (voir chapitre 6.3 Principaux biais en épidémiologie et en recherche clinique en page 229). Naturellement, sans minimiser la réalité, il faudra faire part au lecteur de tous les **éléments rassurants** (tant sur le procédé que sur les résultats), qui pourraient amener à penser que ces biais sont conservateurs ou ont une portée limitée.

D'un point de vue externe, cette discussion **comparera la méthode** employée aux méthodes employées par les autres auteurs. Elle utilise de nombreuses références bibliographiques. Contrairement à ce qu'on pourrait penser, il sera toujours bien perçu d'avoir utilisé **les mêmes méthodes** que les autres auteurs, car c'est une manière simple et efficace de justifier de leur bienfondé, et cela permet de produire des résultats comparables. Or la section suivante porte justement sur la comparaison des résultats à ceux des autres auteurs.

3.5.4 Discussion des résultats

Il s'agit ici de discuter les résultats obtenus, à la fois d'un point de vue interne, et externe.

D'un point de vue interne, c'est l'occasion d'interpréter les résultats obtenus. C'est généralement à cette étape qu'on passe de la simple association statistique à une possible interprétation d'explication ou de causalité (voir pour ce faire la discussion du chapitre 6.2 De la significativité statistique à la causalité et à l'explication en page 226).

D'un point de vue externe, cette discussion **comparera les résultats** obtenus aux résultats obtenus par les autres auteurs. Elle utilise de nombreuses références bibliographiques. Cette partie justifie combien il est nécessaire d'utiliser des méthodes comparables aux autres auteurs, pour obtenir des résultats comparables.

3.5.5 Perspectives

Cette dernière partie est supposée orienter les autres chercheurs vers de nouveaux travaux nécessaires.
Ex : « Ces résultats devraient être confirmés par un essai randomisé contrôlé… ».

Parfois, mais plus rarement, les perspectives peuvent orienter les cliniciens ou les décideurs vers une conduite à tenir.
Ex : « Ces résultats suggèrent qu'une sérologie toxoplasmique devrait désormais faire partie du bilan somatique avant toute hospitalisation en psychiatrie ».

Il faut toujours **rester très prudent** sur le potentiel de translation en vie réelle d'un résultat de recherche : comme nous l'avons vu précédemment, de très nombreux biais, ou simplement l'erreur d'échantillonnage, peuvent amener à découvrir des résultats qui ne reflètent pas la réalité, et ne seront donc pas forcément confirmés par des études autres. La meilleure manière

[23] « je suis venu, j'ai vu, j'ai vaincu » en Latin, expression employée par Jules César en -47 pour relater une victoire militaire rapide

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

de rester prudent, est de considérer que les perspectives s'adressent **aux autres chercheurs**, et non aux cliniciens et décideurs. Sinon, il convient d'utiliser toutes les précautions nécessaires, dont le mode **conditionnel**.

3.6 Rédaction de la conclusion

Dans de nombreux journaux, la conclusion est absente. Lorsqu'elle est présente, elle prend généralement la forme d'une phrase unique, à emporter à la maison : un « *take home message* ». On peut s'imaginer expliquer l'intérêt du travail à ses grands-parents, en une seule phrase.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

4 Imprimer et diffuser le document

4.1 Eléments finaux de mise en page

Lorsque le jury souhaite disposer de place pour annoter le texte, deux options sont habituellement proposées :

- Imprimer en recto-verso mais laisser un double interligne
- Imprimer en simple interligne, mais en recto seulement

De très loin, c'est la deuxième option que nous vous recommandons : **recto simple, simple interligne**. C'est dans cette optique que le modèle de document proposé sur http://objectifthese.org a été conçu. Voici trois arguments de poids :

- Par rapport au simple interligne, le double interligne consomme nettement plus que le double de place, et perturbe fortement la mise en page et la compréhension de la structure du texte (notamment dans les tableaux et les listes à puces)
- Le double interligne est beaucoup plus long à mettre en œuvre, et nécessite de contrôler tout le document
- L'impression en recto-verso est plus complexe, car elle nécessite de contrôler précisément quelle page sera à gauche ou à droite, et d'insérer des pages blanches pour faire apparaître tous les débuts de section sur une page de droite (qui s'affiche à gauche dans Word !)

Avant l'impression d'une épreuve, parcourez le document du début jusqu'à la fin, et contrôlez, en insérant des sauts de page, que :

- les tableaux ne sont pas coupés
- les tableaux et figures ne sont pas séparées de leur légende
- aucun paragraphe trop petit ne figure seul sur une page
- les sauts de page déjà insérés soient toujours nécessaires, compte tenu de la repagination en cours

Enfin, sélectionnez tout le document ([contrôle]+[a]) puis **mettez à jour tous les champs** (F9 ou clic droit et « mettre à jour les champs »). Chaque fois que demandé, sélectionnez « mettre à jour toute la table ».

/! Si votre document comporte des entrée de glossaire, ces signets prennent plus de place lorsque les caractères non-imprimables sont affichés, que lorsqu'ils sont masqués. Il faut donc passer en mode « caractères invisibles » en cliquant sur le bouton adapté ¶ avant de mettre à jour les champs. Pour limiter leur impact, nous vous conseillons de ne pas positionner d'entrée de glossaire sur les titres et sous-titres.

Dans votre **bibliographie Zotero**, recherchez la présence de « **n.d.** » : cela indique des métadonnées manquantes dans vos références bibliographiques (très souvent, la date ou l'auteur). Complétez les champs manquants dans l'interface de Zotero (surtout pas en modifiant votre document), puis resynchronisez votre document avec Zotero en cliquant sur le bouton Zotero Refresh ci-contre dans l'interface de votre traitement de texte.

4.2 Ultimes contrôles avant impression

Si ce n'est déjà fait, définissez (sans en faire part) les **éléments de langage** : c'est la manière dont vous souhaitez nommer les concepts au fil du document. Mettez-vous d'accord avec vous-même, et corrigez les synonymes dans tout le document de manière à créer une homogénéité linguistique. Le lecteur sera ainsi rassuré, et certain de bien comprendre votre propos. Les recherches automatiques ([contrôle]+[f]) seront utiles pour identifier les

occurrences. Nous déconseillons les remplacements automatiques, car il faudra vérifier avant de remplacer.

Exemple 1 : remplacer le mot « pourcentage » par « proportion »

Exemple 2 : parler de « arthroplastie de hanche » pour l'acte, et de « prothèse de hanche » pour le matériel implanté. Ne pas parler de « pose de prothèse de hanche ».

Pour le dernier contrôle de l'orthographe, il est conseillé d'imprimer le document, puis de corriger le document **sur papier**. Ce procédé est nettement plus sensible que la lecture à l'écran, c'est un fait.

Si vous relisez vous-mêmes le document que vous avez écrit, pour rompre les automatismes, il est conseillé de lire les phrases **de la fin vers le début**, afin d'oublier le fond et de se consacrer uniquement à l'orthographe.

Profitez-en pour noter sur une feuille séparée **tous les sigles**, afin de vérifier à la fin qu'ils figurent bien sur la page dédiée aux sigles en début de document.

Enfin, de retour sur la version numérique, après mise à jour des champs, et avant impression, **recherchez le mot « erreur » ou « introuvable »** : il peut parfois témoigner d'un renvoi orphelin, dans le texte ou dans une table des matières (Figure 165).

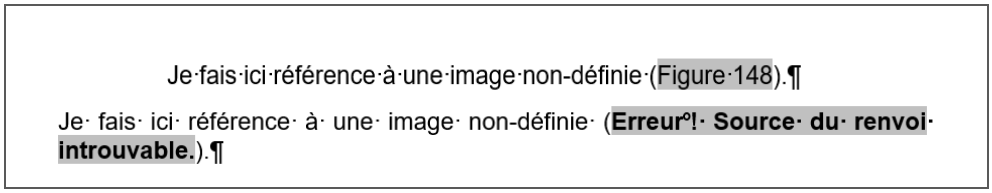


Figure 165. Exemple de renvoi orphelin, avant et après mise à jour des champs

Vous pouvez à présent enregistrer votre document au format PDF :

Microsoft Word : *Fichier > Enregistrer sous > [Choisir le format PDF]*

LibreOffice Writer : *Fichier > Exporter vers > Exporter au format PDF*

L'enregistrement en format PDF vous garantit un format d'échange qui présentera beaucoup moins de variabilité selon le poste ou le logiciel qui lit le fichier. Parcourez le document PDF une dernière fois avant de le transmettre.

4.3 Modalités d'impression non-professionnelle

Pour certains mémoires académiques de petite taille, et en fonction des exigences de votre formation, une impression maison peut être suffisante. Vérifiez alors chaque page, car il est possible que votre imprimante commette des erreurs (trainées d'encre, trainées sans encre, impression inclinée, etc.).

A la maison, on pourra **agrafer** le fascicule, ou le relier à l'aide d'une **baguette de reliure** plastique. Ces baguettes sont élégantes, faciles à poser, solides et réutilisables (Figure 166). Elles se glissent le long des feuilles, qu'elles pincent simplement. Leurs inconvénients sont qu'elles peuvent empêcher de lire le texte proche de la reliure, et qu'elles peuvent parfois se déboiter.

Dans un secrétariat standard, on pourra également perforer et relier le fascicule à l'aide d'une **baguette de reliure à anneaux en plastique** ou, mieux, d'une **baguette de reliure à anneaux en métal** (Figure 166). Il faut pour ce faire disposer d'une machine adaptée et des consommables en question. La reliure métal est parfois appelée « spirale » alors que ce n'est pas une spirale. Cette reliure est nettement plus agréable que la reliure plastique, car elle

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

nécessite des trous ronds et rectangulaires : ces trous sont plus solides, les feuilles ne s'accrochent pas entre elles et ne génèrent pas de poussière de papier.

Figure 166. Baguettes de reliure

4.4 Impression professionnelle

Pour les mémoires académiques de taille moyenne ou importante (40 pages et plus) comme les thèses de médecine, en particulier pour les ouvrages qui doivent être déposés en bibliothèque universitaire, on recourt plus facilement à une impression professionnelle. Ce type d'impression peut être réalisé par de nombreuses boutiques de photocopie à proximité des facultés, et même par des associations étudiantes. L'impression professionnelle comprend plusieurs avantages :

- Elle permet une reliure professionnelle
- Elle coûte moins cher qu'une impression jet d'encre à domicile
- Elle est généralement rapide, reliure comprise (une journée)
- La finition de l'impression est généralement excellente

On notera que, de nos jours, l'écart de prix entre la page couleur et la page en niveaux de gris est modeste. Si vous n'avez aucune figure nécessitant la couleur, l'impression en niveaux de gris reste plus économique. Mais si vous avez de-ci, de-là, des images en couleur, il reste acceptable et nettement plus aisé de tout imprimer en couleur, sans se poser de question. Enfin, comme ce sont généralement les mêmes machines, certains professionnels appliquent un tarif différent par page, même si l'impression est lancée en une seule fois.

Il vous sera toujours possible, en plus du mémoire relié, de proposer au jury un **tiré à part** de quelques figures, y compris une impression A3 si nécessaire.

Parmi les différentes options de reliure, c'est généralement le **dos carré collé** qui est proposé. Le dos carré collé consiste principalement en une bande de colle qui colle la liasse de feuilles (qui sont des feuilles simples, sans pli) à la couverture (Figure 167). Cette bande de colle est parfaite tant que l'ouvrage est plié. Lorsque l'ouvrage est ouvert, si on veut mettre les feuilles à plat il faudra plier la tranche de la couverture, car elle est solidaire de la liasse de feuilles (Figure 167 à droite).

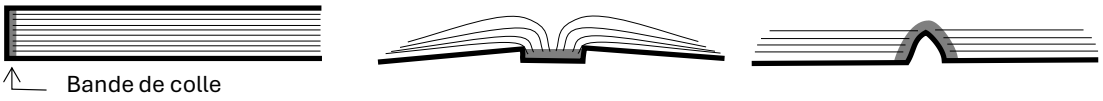


Figure 167. Aspect du dos carré collé : fermé, ouvert à plat, ouvert en pliant la tranche

Nous expliquons ici son fonctionnement, car cela impacte la mise en page pour l'auteur du document (Figure 168). Dans un premier temps (1 en Figure 168), on imprime toutes les pages

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

du manuscrit sauf la couverture. L'impression se fait sur feuilles volantes A4, qui ne sont pas pliés.

L'imprimeur récupère également de l'auteur les éléments nécessaires pour réaliser la couverture : généralement la première de couverture (qui apparaîtra au recto à l'extérieur), un texte brut pour la tranche, et la quatrième de couverture (qui apparaîtra au verso à l'extérieur). Pour les mémoires académiques, on n'imprime généralement rien à l'intérieur de la couverture (deuxième et troisième de couverture). L'imprimeur réalise alors une composition, qui tient compte de la largeur prévisible de l'ouvrage. Cette largeur dépend directement du nombre de pages. Il imprime ainsi une feuille cartonnée (2 en Figure 168) qui a pour hauteur la hauteur d'une feuille A4 (29,7cm) mais est plus large que nécessaire. Cette feuille comprend, de gauche à droite, une marge, la 4ème de couverture, la tranche, la 1ère de couverture, et une autre marge.

L'imprimeur assemble la liasse de feuilles volantes et la couverture, aligne au mieux l'ensemble, et les relie par collage à chaud (avec une colle semblable à celle des pistolets à colle, qui fond à la chaleur et durcit à température ambiante).

Une fois la reliure droite, l'ensemble passe dans un massicot électrique, qui recoupe le tout. Il en résulte un parfait alignement de tout le volume, avec une perte raisonnable de hauteur et de largeur (quelques millimètres seulement).

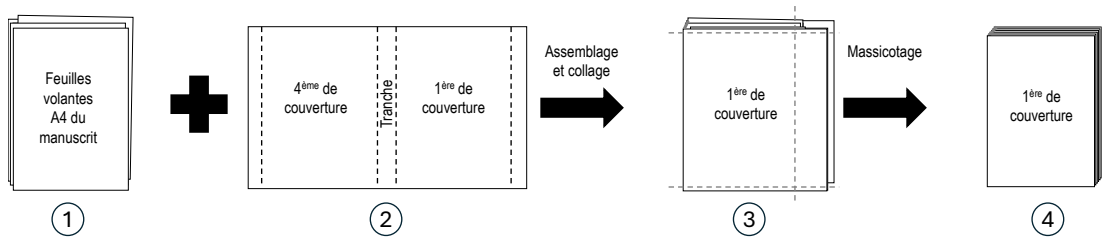


Figure 168. Réalisation d'un dos carré collé par un professionnel

Ce mode s'assemblage a quelques conséquences pratiques pour l'auteur :

- Rien ne doit être imprimé à moins d'un centimètre des bords
- La marge de gauche (en cas d'impression recto) ou la marge intérieure (en cas d'impression recto-verso) doit être de 2,5cm pour permettre une lecture aisée à l'approche de la reliure
- Il faut idéalement préparer un fichier séparé contenant 3 pages (l'imprimeur fera lui-même la mise en page de la couverture) :
    - La première de couverture
    - Le texte brut de la tranche
    - La quatrième de couverture

On notera que, comme c'est l'imprimeur qui met en page la tranche du document, rien ne garantit que les textes des tranches de différents documents soient identiquement positionnés.

4.5 Envoi par courrier postal, le cas échéant

Il n'est pas fréquent de devoir envoyer votre document par courrier postal à des membres du jury. Si tel est le cas, prenez bien note de certains conseils.

Si votre destinataire travaille dans un hôpital ou une université française, proposez-lui de lui faire livrer le document **à son domicile**, il aura plus de chances de le recevoir. De mon expérience, le quart des documents expédiés à une adresse professionnelle en établissement public n'atteint jamais le destinataire.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Evitez les accusés de réception. Préférez l'envoi par courrier postal simple. Les autres systèmes sont très contraignants pour les destinataires, car ils les obligent à faire la queue le samedi matin en bureau de poste ou, pire, dans les agences des sociétés de livraison, qui sont généralement très éloignées du centre-ville[24]. Laissez une note visible demandant aux destinataires d'**accuser réception par SMS** : ainsi, vous ne recontacterez que les personnes qui ne confirment pas.

Pour l'emballage de votre document, une **enveloppe en papier kraft** est suffisante (format 22,9 x 32,4 cm). N'ajoutez pas d'épaisseur : s'il s'agit d'une simple enveloppe elle pourra contenir dans la boîte aux lettres du destinataire, sinon il devra aller la chercher en bureau de poste aux heures d'ouverture.

[24] De plus, parmi les sociétés spécialisées dans la livraison de colis, certaines sont « spécialisées dans la distribution d'avis de passage ». Comprenez par-là que, même lorsque les destinataires sont chez eux, le livreur ne sonne jamais et laisse un avis de passage.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

5 Utiliser un logiciel de présentation pour la soutenance orale

Votre mémoire académique aura très probablement vocation à être présenté oralement. Nous verrons dans cette section comment vous y préparer.

5.1 Concevoir le diaporama, sur le fond

Concernant la taille du diaporama, on compte généralement **1 minute par diapositive**, hors titres. Pour une présentation de 15 minutes, visez en premier lieu un diaporama de 15-20 diapositives. Cependant, seule une répétition en temps réelle permettra de calibrer finement votre diaporama.

Le diaporama reprend principalement le contenu de votre mémoire académique, dans le même ordre, en suivant le même plan IMMRaD. Cependant, le diaporama pondère différemment le contenu. Globalement, il doit **mettre l'accent sur les résultats**, en passant plus vite sur les autres parties :

- L'introduction est généralement plus courte. Pour les thèses d'exercice, l'introduction se focalisera sur les aspects les plus simples, les plus visuels, les plus didactiques. Pour ainsi dire, elle est plus destinée au public, et lui permettra de comprendre le contexte dans lequel le travail est réalisé.
- La section matériel et méthodes est grandement simplifiée et raccourcie. C'est plus le design que l'analyse statistique qui est mis en avant.
- Les résultats sont assez développés, en mettant l'accent sur les éléments très graphiques et marquants. Le *flowchart* sera toujours présenté. On évitera de reproduire de grands tableaux illisibles. Inversement, les graphiques simples qui ne sont habituellement pas présents dans le document, seront produits (ex : camemberts).
- La discussion est sommaire, et ne s'étend pas sur la discussion de la méthode (« il existe un biais, mais… on aurait pu… »), mais plutôt sur l'intérêt de fond des résultats.

Les diapositives ne doivent **pas être rédigées** : leur style **télégraphique**, s'appuyant sur de nombreuses **listes à puces** hiérarchiques, doit permettre aux spectateurs distraits de rapidement se raccrocher au cours de la narration. Le fait d'avoir peu de texte permet également de l'afficher en **grande taille** sur chaque diapositive.

Pour les thèses d'exercice et les mémoires professionnels, il est d'usage de ne pas mettre de **référence bibliographique**. Pour les thèses d'université, on en met quelques-unes pour la forme, sachant qu'elles sont généralement illisibles.

5.2 Concevoir le diaporama avec un logiciel de conception

Nous vous conseillons d'utiliser **Microsoft Powerpoint** ou **LibreOffice Impress** pour concevoir votre diaporama. Le premier est excellent mais payant, le deuxième est bon, gratuit et open source. Nous vous **déconseillons formellement d'utiliser des solutions en ligne** : d'expérience elles sont plus longues à utiliser et vous font perdre toute indépendance et résilience. De plus, l'accès à votre travail n'est pas garanti dans le temps, ni même le jour de la soutenance, ce qui est un comble.

Il est impératif dès le début d'utiliser les **masques de diapositives** (Powerpoint), ou **diapos maîtresses** (Impress). Cette fonctionnalité est comparable aux styles de traitements de texte, mais concerne la dispositive dans son ensemble. Elle permet de dissocier le fond de la mise en forme, ce qui a plusieurs effets :

- Le fond est saisi sans perdre de temps à le mettre en forme. Le **gain de temps** est notable.

- La même mise en forme est appliquée identiquement à l'ensemble du diaporama, garantissant une grande **homogénéité** de forme.
- Les améliorations itératives et à peu de frais des masques permettent généralement d'atteindre une **meilleure qualité** de mise en forme.

Les masques se présentent sous la forme d'un masque principal, sur lequel il sera intéressé de passer un peu de temps pour parfaire son esthétique, et plusieurs dispositions, qui héritent des caractéristiques du premier mais proposent des mises en forme différentes (ex : texte en deux colonnes, texte à gauche et image à droite, etc.). Pour accéder aux masques d'un diaporama :

Dans Microsoft Powerpoint : *menu Affichage > Masque des dispositives*

Dans LibreOffice Impress : *menu Affichage > Diapo maîtresse*

Le site Objectif Thèse vous propose un document Powerpoint prêt à l'emploi, avec un masque de dispositives déjà défini, esthétique et fonctionnel. La vidéo jointe montre certaines manipulations utiles : http://www.objectifthese.org

Pour la plupart des blocs de texte, il pourra être intéressant d'activer l'option permettant de **rétrécir automatiquement la taille du texte** s'il est trop long. Activez tout de suite cette option pour chaque bloc de texte de votre masque de dispositives : cela vous fera gagner beaucoup de temps :

Dans Microsoft Powerpoint : *clic droit sur le cadre > Format de la forme > Options de texte > vignette texte > Réduire le texte dans la zone de débordement*

Dans LibreOffice Impress : *clic droit sur le cadre > texte > texte > adapter au cadre*

Vous pouvez ensuite **quitter le mode de masques** de dispositives pour revenir à votre diaporama :

Dans Microsoft Powerpoint : *menu Masque des dispositives > Fermer le mode masque*

Dans LibreOffice Impress : *menu Affichage > Normal*

Vous devrez également vous assurer que le **numéro de diapositive** est visible partout. Son emplacement est défini dans le masque de chaque dispositive, et n'attend qu'à être renseigné automatiquement.

Dans Microsoft Powerpoint :
menu Insertion > Entête/pied > Numéro de dispositive > Appliquer partout

Dans LibreOffice Impress :
menu Insertion > Entête et pied de page > Numéros de diapos > Appliquer partout /!\ ne pas faire insertion > numéro de diapo

Cette même fenêtre vous permettra d'afficher sur toutes les diapositives un titre générique et une date, si vous le souhaitez.

Prenez garde à l'**étalonnage des couleurs** du vidéoprojecteur : généralement les couleurs sont réétalonnées pour accroître le contraste. Ceci a pour effet que le rouge peut devenir marron. Un écrit noir sur fond rouge, très lisible sur votre l'écran, devient alors illisible en projection. Le mieux est d'utiliser des couleurs simples, d'écrire en noir sur fond blanc ou en blanc sur un fond très foncé.

Utilisez des **polices standard**, de manière à éviter les problèmes de portabilité. Nous vous conseillons les polices **sans sérif** (sans empattement), simples et disponibles sur tous les postes : Arial, Cambria, Helvetica, etc.

Privilégiez les **illustrations** (dessins, graphiques, etc.) au texte. Citez systématiquement la source si vous n'êtes par auteur d'une illustration.

Utilisez **le minimum d'animations** : elles sont souvent perçues comme peu sérieuses, peuvent perturber la narration surtout si elles sont minutées ou trop nombreuses, et ralentissent les accès rapides à une dispositive pour répondre à une question du jury.

Si vous prévoyez de faire une **démonstration** « en live » d'un logiciel, préparez tout de suite en plan B une vidéo de démonstration, et en plan C un diaporama contenant une **série de captures d'écran**. Il est très fréquent que la démonstration prévue ne puisse pas être réalisée, pour des raisons de connexion à Internet, d'équipements, de bugs, etc.

Evitez d'intégrer des modules dont la portabilité n'est pas certaine : Wooclap, vidéos, etc.

Faites plusieurs **répétitions à voix haute en temps réel**, de manière à mieux voir et corriger les erreurs, contrôler le temps de présentation, et faire baisser votre anxiété par la répétition et l'automatisation des tâches intellectuelles.

Une fois votre diaporama au point, enregistrez en plus **une version PDF** de manière à assurer une portabilité de votre document. Cette version PDF sera une version de secours, parfois salvatrice :

Dans Microsoft Powerpoint : *Fichier > Enregistrer sous > [Choisir le format PDF]*

Dans LibreOffice Impress : *Fichier > Exporter vers > Exporter au format PDF*

Notez que, actuellement, si votre document comporte des animations, la version PDF ne produira qu'une page par diapositive, et présentera cette page après exécution de toutes les animations.

5.3 Présenter le diaporama, avec le logiciel

S'il s'agit de votre présentation, il est possible que vous ayez besoin de lire un texte préparé à l'avance. Cela n'a rien de honteux, mais peut sécuriser votre présentation et faire baisser votre niveau d'anxiété. Si tel est le cas, vous utiliserez peut-être la fenêtre « commentaire » de votre logiciel de présentation.

Vous pouvez activer (c'est le cas par défaut) le **mode présentateur** de Powerpoint, ou la **console de présentation** d'Impress. Lorsque ce mode est activé et que vous présentez (Figure 169) :

- L'écran principal affiche une console destinée à l'orateur, comportant :
  - la diapositive en cours
  - la diapositive suivante
  - vos commentaires
  - le minuteur et diverses commandes
- L'écran secondaire affiche votre dispositive, telle qu'elle doit être vue par l'auditoire

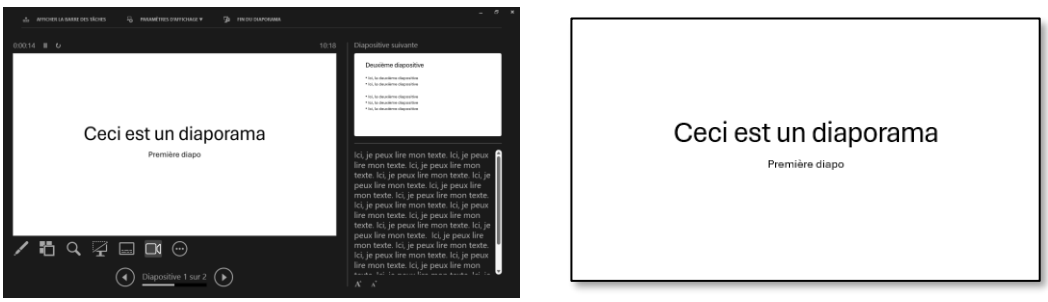


Figure 169. Mode présentateur de Microsoft Powerpoint : vue sur l'écran principal (gauche) et secondaire (droite)

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Attention : si le mode présentateur est désactivé, ou si vous n'avez qu'un seul écran, vous ne pouvez alors pas lire vos commentaires ! Il est important, pour parer toute éventualité, d'**imprimer votre texte sur du papier**.

Faites attention à un deuxième point : si vous utilisez directement Powerpoint ou Impress pour imprimer des pages comportant à la fois vos diapositives et votre texte, **si votre texte est trop long, il sera tronqué** sans message d'alerte. Vous devez contrôler que votre texte de commentaire n'atteint pas les limites de la page, avant impression (Figure 170).

Dans Microsoft Powerpoint : *menu Affichage > Pages de notes*

Dans LibreOffice Impress : *menu Affichage > Notes*

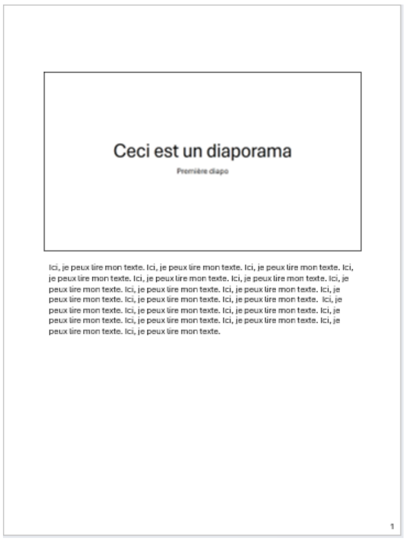


Figure 170. Affiche en "page de notes" pour voir et imprimer les commentaires

Si vous présentez sur un **logiciel de visioconférence** (Zoom, Teams, etc.), au moment où vous partagez votre écran, le logiciel vous demande de choisir ce qui sera affiché. De nombreuses personnes sont alors troublées : elles peuvent choisir l'écran entier (lequel ?) ou une application, or les deux fenêtres du logiciel de présentation (d'une part la console de présentation, d'autre par le diaporama visible par le public), apparaissent comme deux applications différentes et, très souvent, l'une d'elles n'existe pas encore. Une manière simple de s'assurer que les personnes distantes voient bien votre diaporama est alors de **désactiver le mode présentateur**, demander à votre OS de **dupliquer les deux écrans** (au lieu de les disposer en extension l'un de l'autre), puis **partager l'écran complet** en visioconférence, et non une fenêtre d'application en particulier. La conséquence, encore une fois, est que vous ne verrez plus vos commentaires.

Si vous utilisez un **pointeur laser**, pour limiter les tremblements qui révèlent votre anxiété, nous vous conseillons de tenir le pointeur **à deux mains**, en plaquant **vos coudes contre votre bassin**. Faites attention cependant : le **pointeur laser** ne devrait pas être utilisé dans les circonstances suivantes :

- Affichage sur un écran LCD (ne fonctionne pas)
- Visioconférence (invisible à distance)
- Affichage sur plusieurs projections simultanées (absent des autres écrans)
- Enregistrement de la présentation (invisible sur l'enregistrement)

Il existe d'autres manières de pointer sans utiliser de pointeur laser :

- Utiliser le pointeur commandé par la souris, rendu disponible par votre logiciel de présentation lui-même

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

- Simuler des pointages ou surlignages planifiés à l'aide d'animations (si vous savez à l'avance ce que vous allez montrer)
- Utiliser un dispositif de pointage numérique[25], qui affiche à l'écran un pointeur virtuel, ou une poursuite, ou même un zoom. Cependant, ceci ne fonctionnera que sur un ordinateur permettant l'installation du pilote.

Lors de la projection, nous vous conseillons de désactiver le minutage et le mode narrateur, de manière à mener votre diaporama à votre rythme.

[25] Exemple : l'excellent pointeur Logitech Spotlight

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Conclusion

Nous avons dans cet ouvrage envisagé toutes les étapes successives pour réaliser un mémoire académique en santé (thèse d'exercice, mémoire de master, mémoire de fin d'études, etc.).

Cet ouvrage pourra être complété par le site Objectif Thèse http://objectifthese.org, qui propose notamment des vidéos et fichiers prêts à l'emploi, le tout gratuitement et sans inscription.

Cet ouvrage est diffusé sur http://editions.chazard.org : si vous le trouvez trop long, ou au contraire pas assez complexe pour votre niveau, il est possible que vous trouviez sur ce site un ouvrage qui, dans le même esprit, correspondra mieux à votre attente.

Si cet ouvrage vous a plu, n'hésitez pas à en faire la promotion. En regard du travail fourni pour créer cet ouvrage, ma plus belle récompense sera que ce travail soit utile au plus grand nombre, même si c'est à travers sa version gratuite.

Une fois que vous aurez entièrement réalisé et soutenu votre mémoire, vous aurez peut-être déjà des regrets : « j'aurais dû faire ceci, cela… » : c'est un excellent signe ! Cela veut dire que vous avez déjà beaucoup progressé ! Si, grâce à cela, vous avez déjà obtenu votre diplôme final, il sera peut-être frustrant de ne pas mettre à profit vos nouveaux super-pouvoirs. **Et si, à votre tour, vous encadriez des étudiants ? Transmettez votre méthode de travail ! Faites progresser vos cadets ! Contribuez à améliorer les connaissances collectives ! Ce sera un honneur pour moi de vous compter parmi mes estimés collègues ;-)**

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Tables des illustrations

1 Figures

2 Equations

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

3 Tableaux

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Références

1. Dépôt légal éditeur : mode d'emploi [Internet]. BnF - Site Institutionnel [cité 2024 déc 1];Available from: https://www.bnf.fr/fr/centre-d-aide/depot-legal-editeur-mode-demploi

2. National Library of Medicine. PubMed [Internet]. 2024 [cité 2024 sept 16];Available from: https://pubmed.ncbi.nlm.nih.gov/

3. Web of Science | Clarivate [Internet]. [cité 2024 déc 1];Available from: https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/

4. Maisonneuve H. Rédaction Médicale et Scientifique [Internet]. [cité 2024 déc 1];Available from: https://www.redactionmedicale.fr

5. Virgile. La Conférence des Doyens de médecine et le CNU santé luttent contre les « revues prédatrices » [Internet]. Conférence Doyens Médecine2023 [cité 2024 déc 1];Available from: https://conferencedesdoyensdemedecine.org/la-conference-des-doyens-de-medecine-et-du-cnu-sante-luttent-contre-les-revues-predatrices/

6. Harzing AW. Publish or Perish [Internet]. [cité 2024 déc 1];Available from: https://harzing.com/

7. Chazard E. Objectif Thèse [Internet]. [cité 2024 déc 2];Available from: http://objectifthese.org

8. Zotero | Your personal research assistant [Internet]. [cité 2024 déc 2];Available from: https://www.zotero.org/

9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

10. Article L1121-2 [Internet]. [cité 2024 sept 17]. Available from: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000025457486

11. Article R1121-1 - Code de la santé publique - Légifrance [Internet]. [cité 2024 déc 6];Available from: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000043723460

12. Rights (OCR) O for C. HIPAA for Professionals [Internet]. 2015 [cité 2024 sept 17];Available from: https://www.hhs.gov/hipaa/for-professionals/index.html

13. CNIL [Internet]. [cité 2024 déc 6];Available from: https://cnil.fr/fr

14. Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés | Legifrance [Internet]. 2014 [cité 2014 sept 15];Available from: http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460

15. Loi n° 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel et modifiant la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés | Legifrance [Internet]. 2014 [cité 2014 sept 15];Available from: http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000441676

16. Directive 95/46/CE du Parlement européen et du Conseil relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données | Legifrance [Internet]. 2014 [cité 2014 sept 15];Available from: http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000697074

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

17. CNIL. MR-004. Recherches n'impliquant pas la personne humaine, études et évaluations dans le domaine de la santé [Internet]. [cité 2024 sept 18];Available from: https://www.cnil.fr/fr/declaration/methodologie-de-reference-04-recherches-nimpliquant-pas-la-personne-humaine-etudes-et-evaluations-dans-le-domaine-de-la-sante

18. CNIL. MR-005. Études nécessitant l'accès aux données du PMSI et/ou des RPU par les établissements de santé et les fédérations hospitalières [Internet]. 2018 [cité 2022 mars 1];Available from: https://www.cnil.fr/fr/declaration/mr-005-etudes-necessitant-lacces-aux-donnees-du-pmsi-etou-des-rpu-par-les-etablissements

19. Article R1123-4 - Code de la santé publique - Légifrance [Internet]. [cité 2024 déc 6];Available from: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000045303453

20. Vedel I, Kaur N, Hong QN, El Sherif R, Khanassov V, Godard-Sebillotte C, et al. Why and how to use mixed methods in primary health care research. Fam Pract 2019;36(3):365-8.

21. Kaur N, Vedel I, El Sherif R, Pluye P. Practical mixed methods strategies used to integrate qualitative and quantitative methods in community-based primary health care research. Fam Pract 2019;36(5):666-71.

22. « Taxe lapin » : y a-t-il vraiment 27 millions de rendez-vous médicaux non honorés en France chaque année ? [Internet]. 2024 [cité 2024 sept 18];Available from: https://www.lemonde.fr/les-decodeurs/article/2024/02/06/taxe-lapin-y-a-t-il-vraiment-27-millions-de-rendez-nous-medicaux-non-honores_6215039_4355770.html

23. Burns KEA, Duffett M, Kho ME, Meade MO, Adhikari NKJ, Sinuff T, et al. A guide for the design and conduct of self-administered surveys of clinicians. CMAJ Can Med Assoc J 2008;179(3):245-52.

24. French Technical Agency for Hospital Information (ATIH). ICD-10 FR 2017 for PMSI usage [Internet]. 2022 [cité 2023 janv 12];Available from: https://www.atih.sante.fr/cim-10-fr-2022-usage-pmsi

25. PMSI, T2A et facturation hospitalière en MCO, SSR, HAD et Psychiatrie [Internet]. Lille, France: 2019 [cité 2020 févr 25]. Available from: https://www.youtube.com/watch?v=LkntQ5ZLBfU

26. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet Lond Engl 1974;2(7872):81-4.

27. Newcombe RG. Confidence intervals for a binomial proportion. Stat Med 1994;13(12):1283-5.

28. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17(8):857-72.

29. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika 1934;26(4):404-13.

30. Fisher RA. Statistical methods for research workers [Internet]. Oliver and Boyd; 1925 [cité 2024 déc 18]. Available from: https://repository.rothamsted.ac.uk/item/97031/statistical-methods-for-research-workers

31. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: http://www.R-project.org/

32. Student. The Probable Error of a Mean. Biometrika 1908;6(1):1-25.

33. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc 1958;53(282):457-81.

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

34. Fisher SRA. The Design of Experiments. Oliver and Boyd; 1935.

35. Cochran WG. Some Methods for Strengthening the Common χ2 Tests. Biometrics 1954;10(4):417-51.

36. Armitage P. Tests for Linear Trends in Proportions and Frequencies. Biometrics 1955;11(3):375-86.

37. Chazard E, Ficheur G, Beuscart JB, Preda C. How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests. Value Health J Int Soc Pharmacoeconomics Outcomes Res 2017;20(7):992-8.

38. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 1960;20(1):37-46.

39. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213-20.

40. ICH Topic E 9 Statistical Principles for Clinical Trials - Note for guidance on statistical principles for clinical trials [Internet]. European Medicine Agency; 2006 [cité 2024 nov 20]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Glossaire

PDF gratuit - Version papier et autres disponible sur http://editions.chazard.org - © Emmanuel Chazard
Version A4 pour le web : la pagination est modifiée. Pour soutenir ce travail, commandez le livre.

Cet ouvrage de **338 pages**, comprenant **242 illustrations** dont **24 arbres décisionnels**, vous donnera toutes les ressources pour concevoir et réaliser votre **mémoire académique quantitatif en santé** (M1, M2, thèse d'exercice ou d'université). L'approche simple, didactique, mais rigoureuse et documentée, vous permettra notamment de réaliser toutes les **analyses statistiques avec un tableur**, sans logiciel de statistique et sans l'aide d'un biostatisticien.

Trois livres Objectif Thèse :

	Niveau 1 *Poussin pressé*	Niveau 2 *Poulet consciencieux*	Niveau 3 *Coq méthodique*
Conception, formalités, bibliographie	☑ abrégé	☑ détaillé	☑ détaillé
Recueil, correction et transformation de données	☑ abrégé, avec un tableur	☑ détaillé, avec un tableur	☑ avancé, avec R
Analyse statistique univariée et bivariée	☑ abrégée, avec un tableur	☑ détaillée, avec un tableur	☑ détaillée, avec R
Analyse statistique multivariée, rapport automatisé	-	-	☑ détaillée, avec R
Rédaction, traitement de texte, diaporama	☑ abrégée	☑ détaillée	☑ détaillée

Contenu garanti

0% intelligence artificielle
100% expérience et expertise

ISBN : 978-2-9579934-1-3 338 pages 38,90 €

9 782957 993413