Pr Emmanuel Chazard

Objectif Thèse niveau 1 : *Poussin pressé*

Votre mémoire académique de A à Z (M1, M2, thèse d'exercice, thèse d'université) : 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.

Avec le soutien du CIMES
Collège national des enseignants d'informatique médicale, biomathématiques, méthodes en épidémiologie, statistique

Objectif Thèse

http://editions.chazard.org

Objectif thèse niveau 1 : « Poussin pressé »

Vous devez mener une étude quantitative en santé sans logiciel de statistique ? Vous voulez aller à l'essentiel, sans vous perdre dans les explications ? Sur moins de 100 pages, nous suivrons toutes les étapes pour mener un mémoire académique (M1, M2, thèse d'exercice, thèse d'université) : autorisations, 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, couverture : Emmanuel Chazard

Relecture et corrections : Clémence Duriez

Illustration sur la couverture : Frédérique Chazard

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.

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. 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 la **version simplifiée et condensée** du livre **« Objectif thèse niveau 2 : Poulet consciencieux »**. Il est édité par http://editions.chazard.org. Différentes versions, certaines gratuites, sont proposées sur le site.

Du même auteur, sur http://editions.chazard.org :

- Objectif Thèse niveau 2 : Poulet Consciencieux (2025)
- Objectif Thèse niveau 3 : Coq méthodique (à venir en 2026)
- Navigation en catamaran de sport, pratique et théorie (2024)

Dépôt légal auprès de la BnF en avril 2025, sous le numéro 10000001154471.

ISBN 978-2-9579934-2-0

9 782957 993420

Sommaire

Sigles, acronymes et abréviations

AUC	*Area under the Curve*
BDD	Base de données
Cim10	Classification Internationale des Maladies, 10ème version
Cnil	Commission Nationale de l'Informatique et des Libertés
Consort	*Consolidated Standards of Reporting Trials*
CPP	Comité de Protection des Personnes
CSP	Catégorie Socio-Professionnelle
DFG	Débit de Filtration Glomérulaire
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*
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
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
MR*xxx*	Ixième méthodologie de référence de la CNIL (ex : MR005)
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
SD	*Standard deviation* (déviation standard)
Strobe	*STrengthening the Reporting of OBservational Studies*
VPN	Valeur Prédictive Négative
VPP	Valeur Prédictive Positive

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 !

Etudiants en fin d'études en santé, en Master santé, en Thèse d'exercice ou en Thèse d'université en santé : cet ouvrage correspond 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. Nous réaliserons les analyses statistiques avec un tableur (Excel ou Calc), et proposerons des conduites à tenir simples et opérationnelles.

Internes en santé publique, en master de statistique ou en thèse d'université méthodologique : tournez-vous plutôt vers un ouvrage plus détaillé de la collection « Objectif Thèse ».

Sur le site http://editions.chazard.org, vous trouverez trois ouvrages :

- Niveau 1 : poussin pressé (celui-ci !)
- Niveau 2 : poulet consciencieux (avec de nombreuses explications, mais toujours avec un tableur et sans logiciel de statistique)
- Niveau 3 : coq méthodique (avec les modèles multivariés, la gestion avancée de données, et la programmation en statistique avec R)

Pour ces ouvrages, vous trouverez généralement une **version PDF gratuite**, une version **papier broché**, et une version **eBook**. La présente version, « poussin pressé », est volontairement beaucoup plus courte, et permet une mise en pratique immédiate. Si vous souhaitez mieux comprendre les fondements des conseils qui vous sont prodigués, n'hésitez pas à vous procurer la version « poulet consciencieux » (dont le PDF est gratuit), qui est nettement plus détaillée, mais s'adresse au même public que le présent ouvrage.

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

A Clémence, Frédérique et Louis.

	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

Concevoir l'étude

Nous aborderons les étapes suivantes :

1 Mener une recherche bibliographique

1.1 Définitions, environnement de publication

La notion de littérature blanche ou grise n'est pas très importante :

- **littérature blanche** : documents publiés par des éditeurs professionnels, traçables avec numéro ISBN ou ISSN
- **littérature grise** : les autres documents

Ce qui compte pour vous :

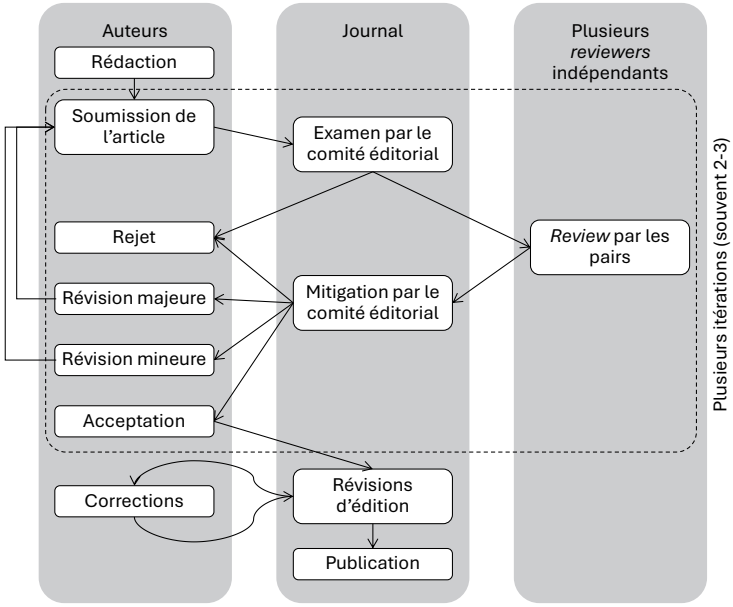
- ☺ **Sources de qualité**, à citer : **articles scientifiques** publiés dans des journaux scientifiques, **rapports scientifiques** publiés par certaines autorités (HAS, FDA, EMA…), **certains livres** publiés par des **auteurs reconnus**
- 😐 **Sources de qualité intermédiaire**, à lire et +/- à citer : mémoires d'étudiants, rapports techniques, sociétés savantes et collèges d'enseignants
- ☹ **Sources non-fiables**, à ne pas citer : tutoriels, sites grand public, presse scientifique grand public, presse non-scientifique

Articles scientifiques revus par les pairs :

- référencés dans une **base de données bibliographique** comme http://pubmed.gov , parce que l'ensemble du journal l'est
- garantie de processus éditorial : **revue par les pairs**
- garantie de qualité et indépendance

Processus de r**evue par les pairs**, ou **peer review** :

- s'appuie sur des chercheurs concurrents, généralement bénévoles (reviewers)
- processus permettant de filtrer les articles et d'améliorer leur qualité (ci-dessous)



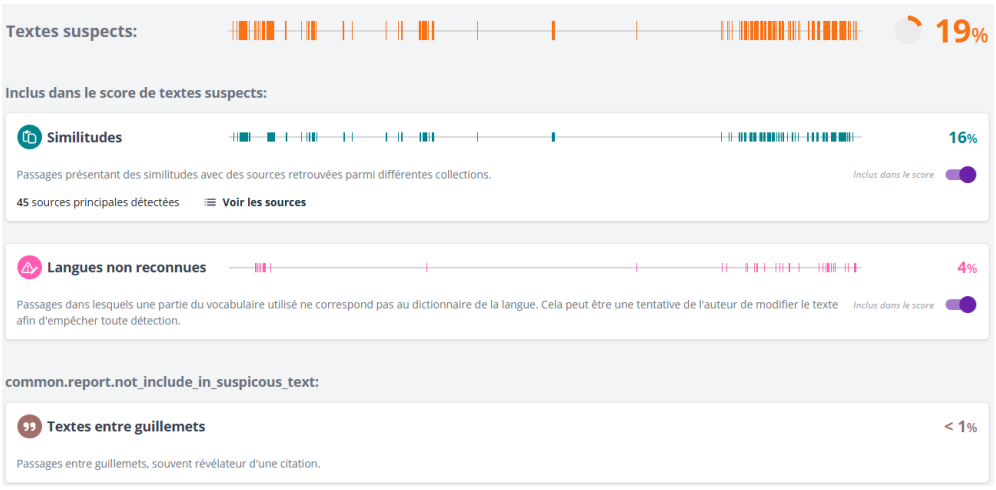
Bases de données bibliographiques :

- La référence pour vous : **Pubmed** http://pubmed.gov moteur sur la base **Medline** de la **NLM** (*National Library of Medicine*)
- Les autres : attention, certaines sont affiliées à des éditeurs => qualité variable

Droits d'auteur et copyright :

- Notion complexe et variable selon les pays
- **Idées et concepts** : réutilisation illimitée, en citant la source
- **Texte** : réutilisation limitée, entre guillemets, en citant la source
- **Images** : réutilisation limitée tolérée, en citant la source. Possibilité de refaire vous-même le dessin, ou d'utiliser une image en licence ouverte

Exemple de rapport de détection de plagiat par Compilatio Magister® :



Modèles économiques de publication d'articles :

- **modèle traditionnel par abonnement** : publication gratuite, lecture payante
- **modèle *open access*** : publication payante, lecture gratuite
- **conférences scientifiques avec actes** : publication réservée aux participants au congrès (qui ont payé l'inscription), lecture souvent gratuite
- **revues prédatrices** : détournent le modèle ***open access*** à des fins mercantiles. Mauvaise qualité

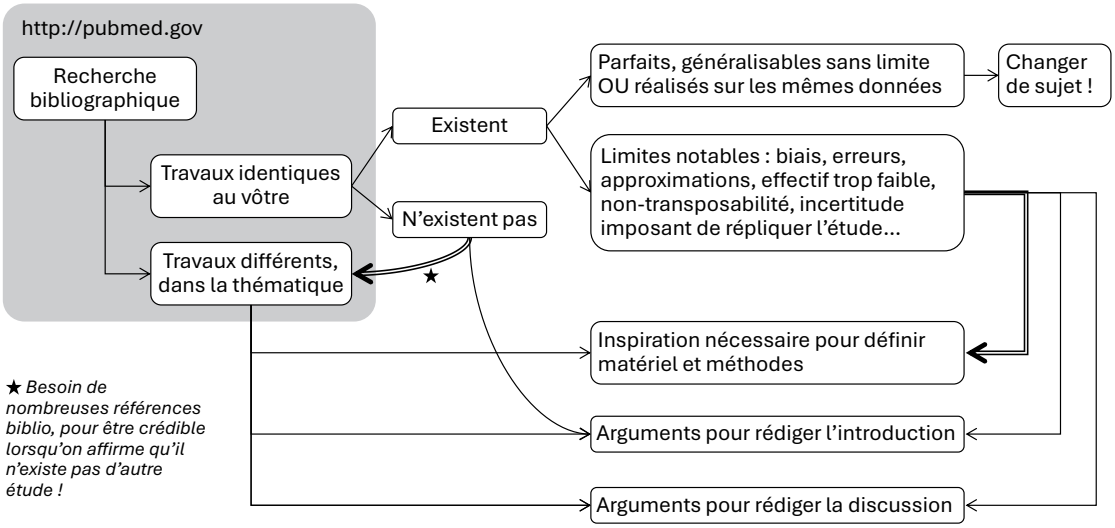
Bibliométrie :

- désigne la **quantification automatisée** de l'activité de publication (nombre d'articles, nombre de citation de ces articles, etc.)
- évaluation des journaux : notamment l'***impact factor*** **(IF)**
- évaluation des établissements d'affiliation : notamment **Sampra** et **Sigaps**
- évaluation des chercheurs : notamment **Sampra** et **Sigaps**

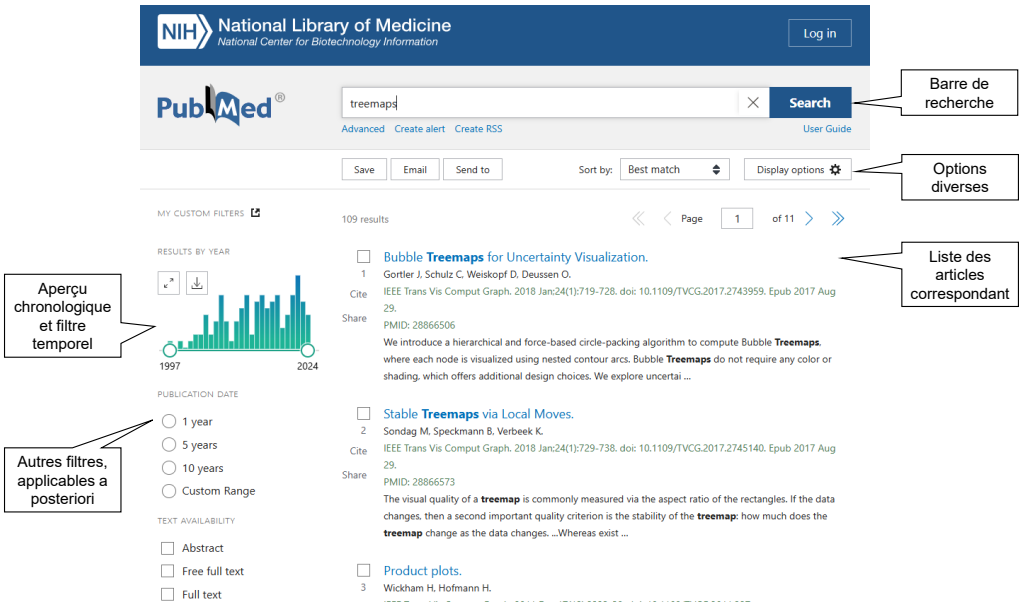
1.2 Pourquoi et comment ?

Recherche bibliographique :

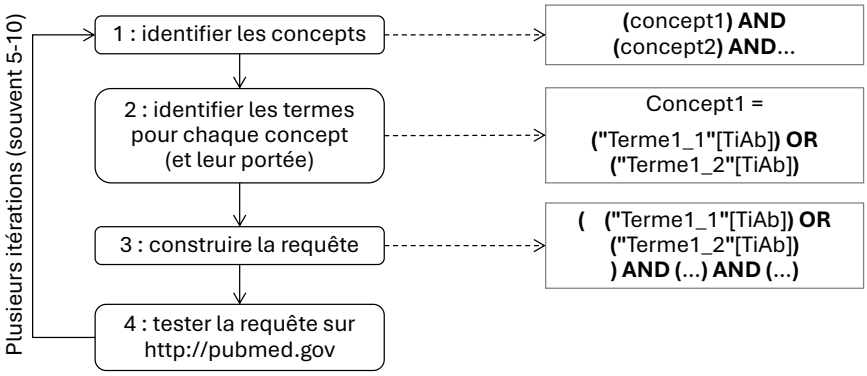
- Indispensable pour savoir si on réalise un travail
- Indispensable pour mieux définir le travail
- Indispensable pour rédiger (voir page 73)



Mener une recherche bibliographique sur http://pubmed.gov :



La démarche de construction d'une requête est **itérative** :



Construire une chaîne de requête structurée en 3 étapes :

- **#1 : identifier les concepts**
  Ex : *effets indésirables liés à l'interruption d'un traitement par statine*
  Devient : *interruption & statine & effet_indésirable*
- **#2 : identifier les termes dans chaque concept**
  Ex : *« statine » → statin, statins, fluvastatin, lovastatin, mevastatin, pitavastatin, pravastatin, rosuvastatin, simvastatin*
- **#3 : agencer la requête** avec des parenthèses, des opérateurs (AND, OR) et des indicateurs de portée ([Title], [Title/Abstract], etc.). Utiliser le fichier automatisé sur http://objectifthese.org
- **#4 : itérations.** Lire quelques articles, améliorer la requête, et recommencer

Exemple de construction d'une requête avec le fichier d'Objectif Thèse :

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			

Le fichier produit automatiquement la requête suivante :

*(("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])) **AND** (("stop"[Title]) OR ("discontinuation"[Title]) OR ("interruption"[Title]) OR ("deprescription"[Title])) **AND** (("adverse"[Title])) **AND** (("2010"[Date - Publication] : "2024"[Date - Publication]))*

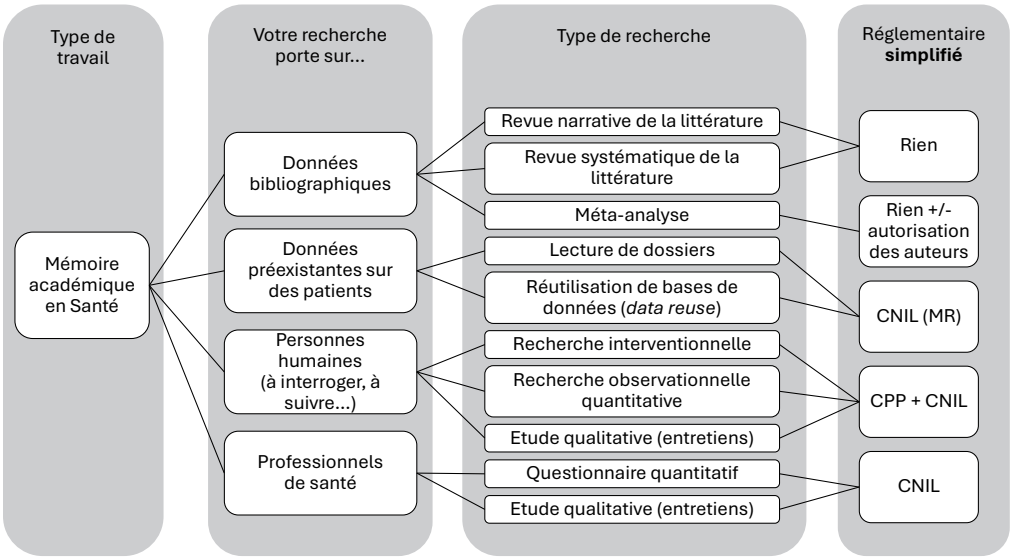
Nous verrons plus tard comment collecter les articles avec **Zotero** (voir page 75).

2 Quelles autorisations pour quelle étude

2.1 Types d'études et autorisations

Les autorisations à obtenir dépendent de l'objet étudié :

- **Bibliographie** (articles, etc.) : **revues de la littérature**, ou certaines **méta-analyses** → aucune autorisation
- **Données** sur des patients (dossier), sans jamais les rencontrer : **recherche sur les données**, ou recherche n'impliquant pas la personne humaine **RNIPH**, ou **réutilisation de données → CNIL**
- **Patients** : recherches impliquant la personne humaine **RIPH**, interventionnelles ou observationnelles, y compris les questionnaires[1] → **CNIL et CPP**
- **Professionnels de santé** : évaluations de pratiques professionnelles **EPP → CNIL**



2.2 Protection des données et CNIL

Impacts d'une éventuelle divulgation de données personnelles de santé :

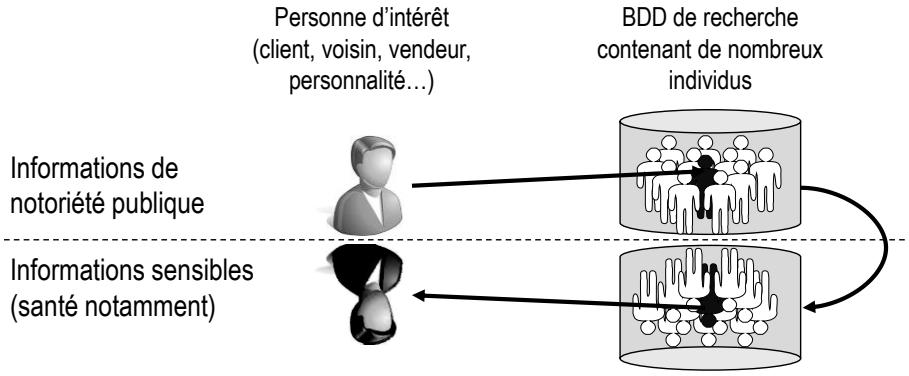
- Nuire à la réputation d'une personne, ragots, médisance
- Empêcher l'accès au crédit bancaire, augmenter le coût des assurances
- Empêcher un recrutement, une promotion, une nouvelle mission
- Léser les intérêts immobiliers et financiers (viager)

Types de bases de données (BDD) :

- **BDD nominatives** : contenant tout ou partie du nom et du prénom, numéro de sécurité sociale, numéro de titre d'identité, numéro de compte bancaire, téléphone, courriel, photo du visage, etc.
- **BDD anonymes** : selon la loi française, uniquement les données agrégées (plusieurs individus dans chaque ligne de données)
- **BDD indirectement nominatives** : potentiellement toute BDD contenant des informations individuelles, même en partie cachées

[1] Même les **questionnaires peuvent être dangereux** : si on pose trop de questions sur le suicide à certains patients psychiatriques, cela peut augmenter le risque de suicide.

Procédé de réidentification d'une personne dans une BDD indirectement nominative :



Exemple de réidentification d'une personne dans une BDD nationale de recherche :

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

La **CNIL**, Commission Nationale de l'Informatique et des Libertés :

- Fait appliquer le **RGPD**, règlement général de protection des données
- Délivre les autorisations pour les RIPH et les RNIPH
- Contrôle notamment :
  - Information et recueil du consentement des sujets
  - Absence de collecte de données **sensibles** (ethnie, religion, politique…)
  - **Protection** des identités et des données
  - **Parcimonie** des données collectées
  - Limitation des **traitements**
  - **Sécurité** du stockage, absence de **partage**, **destruction** planifiée
- Certaines **méthodologies de référence** peuvent simplifier la procédure

2.3 Protection des personnes et CPP

Le **CPP**, Comité de Protection des Personnes :

- Autorisation indispensable si des chercheurs rencontrent des patients (RIPH)
- Même pour les simples questionnaires
- Autorisation porte sur :
  - Caractère **éthique** du protocole
  - Evaluation correcte du **nombre de sujets nécessaires** (NSN) (page 19)
  - **Rigueur méthodologique** du protocole

3 Enquête quantitative ou qualitative ?

Etudes quantitatives :

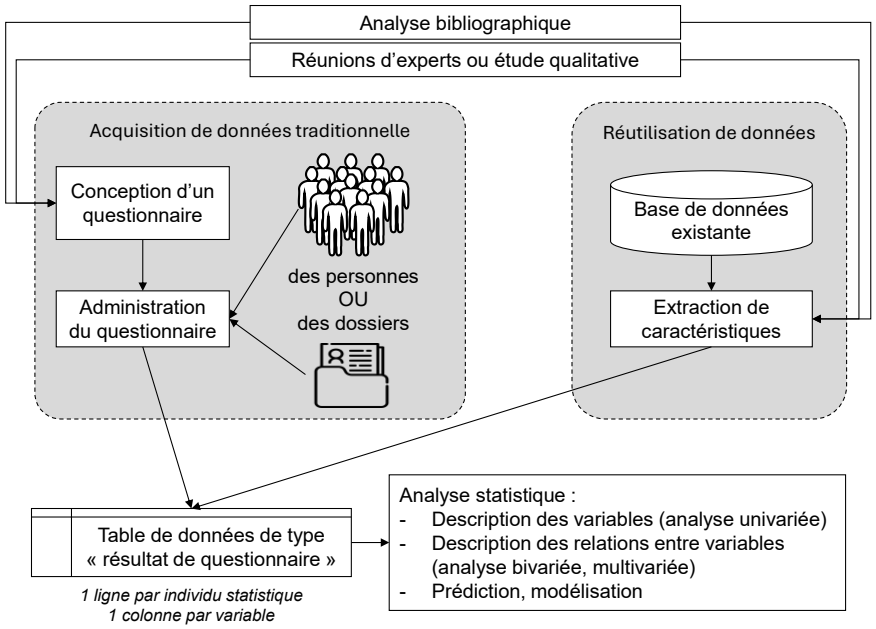
- Calculent des **quantités** (ex : proportions, moyennes)
- Dites « enquêtes » quantitatives si utilisent des **questionnaires**

Enquêtes qualitatives :

- **Listent des options**, sans calculer de quantité
- Utiles principalement pour préparer une enquête quantitative, sinon insuffisantes

	Enquête quantitative par questionnaire	**Enquête qualitative par entretien semi-dirigé**
Effectif	Des centaines d'individus	6-12 personnes
Support	Questionnaire fermé	Guide d'entretien, ouvert
Qui mène ?	Auto-administré (rempli par le sujet) ou hétéro-administré (rempli par l'enquêteur)	Entretien mené par enquêteur de manière à « faire parler » le sujet
Durée	5 minutes par sujet	1/2h ou 1h par sujet
Objectif stratégique	Répondre directement à une question scientifique	Préparer ou interpréter une enquête quantitative
Objectif opérationnel	Estimer des moyennes, des proportions…	Lister toutes les options possibles, comprendre

Déroulement d'une étude quantitative :



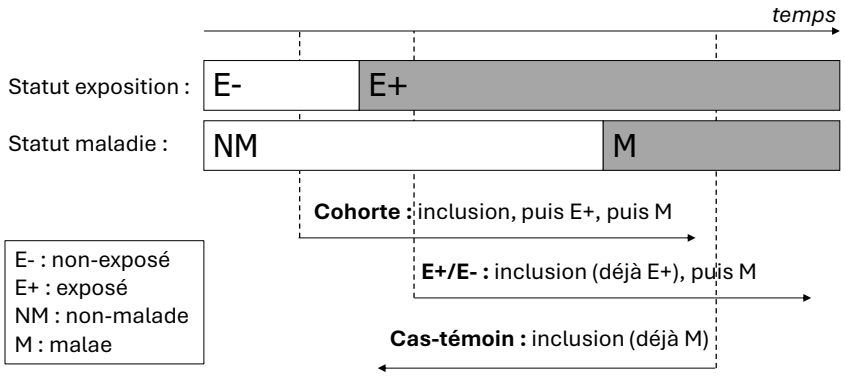
4 *Designs* les plus fréquents en étude quantitative

4.1 Etudes observationnelles

Etudes **observationnelles** :

- Pas de modification du cours du soin, hormis examens complémentaires
- Etudes **descriptives** (ne croisent pas les variables) : transversales (à un temps donné) ou longitudinales (suivi)
- Etudes **analytiques** : lien entre une exposition et une maladie
    - **Cohorte** : les sujets inclus sont encore non-exposés et non-malades
        - **cohorte prospective** : calcul possible de l'incidence, du risque relatif et de l'odds ratio
        - **cohorte historique** : émulation de cohorte sur des dossiers (recherche sur les données ; incidence, risque relatif, odds ratio)
        - **études pronostiques** : cohorte ne portant que sur des personnes exposées à une intervention

- **Exposé-non-exposé** : sujets inclus dans deux groupes, exposés ou non. Possibilité de calculer le risque relatif et l'odds ratio
- **Cas-témoin** : sujets inclus dans deux groupes, malades et non-malades. Possibilité de calculer l'odds ratio



		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

4.2 Etudes interventionnelles

Etudes **interventionnelles** :

- La prise en charge du patient est modifiée pour l'étude
- **Déconseillées** pour les mémoires académiques : besoin d'une autorisation CPP
- Etudes **non-comparatives** : un seul bras de traitement
- Etudes **comparatives** : plusieurs traitements, pour plusieurs groupes

Classification des études **comparatives** :

- Selon le **type d'intervention** :
  - **essai thérapeutique** si prestation ou produit de santé
  - **intervention de santé publique** dans les autres cas (sport, régime, etc.)
- Selon le **bras contrôle** :
  - **absence d'intervention** : **déconseillé** car l'**effet placebo** et l'effet cocooning peuvent déjà améliorer les résultats de santé
  - **prise en charge similaire, sans l'intervention**
  - **autre intervention déjà validée**
- Selon l'**affectation au bras** de l'étude :
  - essai non-randomisé (non souhaitable)
  - **essai randomisé contrôlé** (RCT, *randomized controlled trial*) si l'affectation dans un bras est aléatoire (c'est la meilleure méthode)
- Selon la **connaissance** du bras :
  - **sans aveugle**
  - **simple aveugle** lorsque le patient ignore dans quel bras il se trouve
  - **double aveugle** lorsque, en plus, les investigateurs ignorent cela
  - **triple aveugle** lorsque, en plus, les statisticiens ignorent cela

4.3 Etudes comparatives quasi-expérimentales

Les études **quasi-expérimentales** sont des études observationnelles, comparatives, dénuées de biais d'indication :

- Design **ici-ailleurs** : compare un lieu de soin à un autre, chacun son protocole
- Design **avant-après** : compare deux périodes, chacune son protocole

5 Questionnaires : taux de sondage, taux de réponse

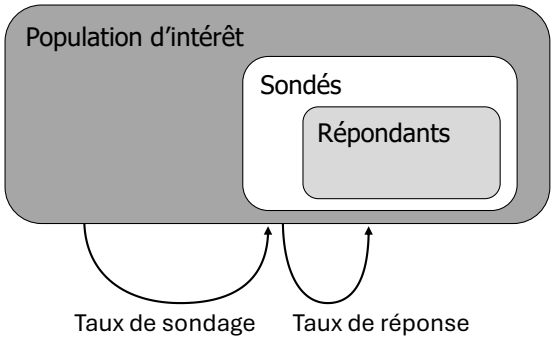
Concernant les enquêtes quantitatives par questionnaire :

- **Taux de sondage** = proportion de personnes interrogées parmi la population. Il est toujours faible en pratique.
- **Taux de réponse** = proportion de répondants parmi les personnes interrogées. Il doit être le plus élevé possible.

Deux conditions pour pouvoir interpréter les résultats :

- Une **sélection aléatoire ou pseudo-aléatoire** des sondés
- Un **taux de réponse entre 67% et 100%**, sinon très probable biais de sélection (voir comment l'atteindre en page 21)

Les méthodes des instituts de sondage (ex : échantillons représentatifs, méthode des quotas) sont trop faibles méthodologiquement, et inutilisables en recherche en santé.



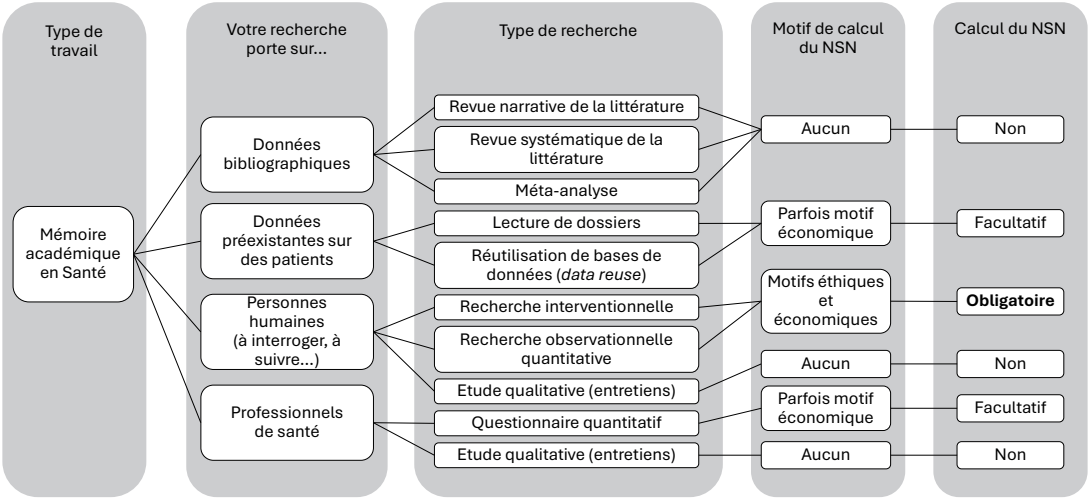
Exemple :
Population d'intérêt = $1{,}4.10^6$ patients BPCO
Sondés = 200 patients tirés au sort
Répondants = 170 patients

Taux de sondage = $200/1{,}4.10^6 = 0{,}0143\%$
Taux de réponse = 170/200 = 85%

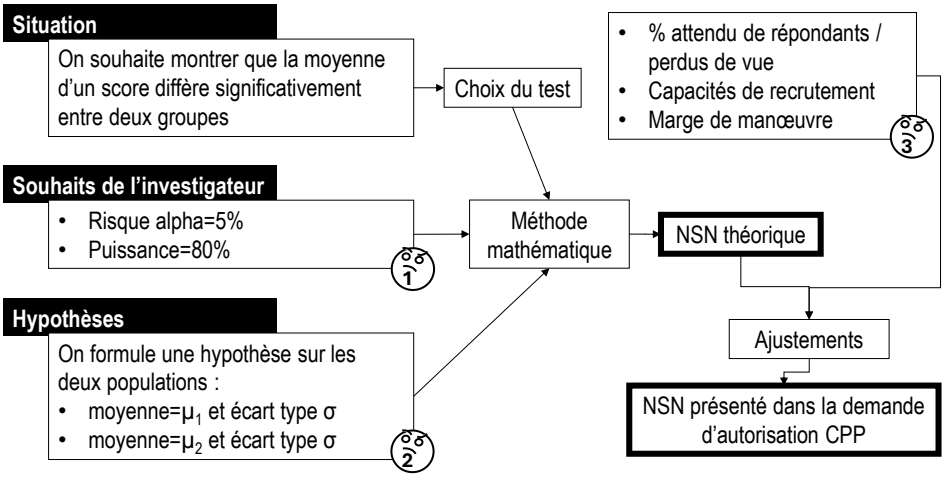
6 Calculer le nombre de sujets nécessaires

Le **nombre de sujets nécessaires (NSN)** :

- Calcul obligatoire uniquement dans les RIPH
- Pour un seul test statistique : celui de l'objectif principal de l'étude
- Proposé par les investigateurs au CPP, qui l'accepte ou non
- Motifs éthiques et économiques : ni trop, ni trop peu de patients dans une étude



Le schéma suivant montre comment calculer le NSN pour **comparer deux moyennes** : les entrées (1), (2) et (3) sont spéculatives et il est possible de fortement influencer le résultat en fonction du paramétrage initial.



Conduite à tenir pour un mémoire académique hors RIPH : **ne pas calculer le NSN** car ce n'est pas obligatoire ! Le bon sens prime, comme dans l'exemple suivant :

Exemple : on souhaite obtenir 80 réponses à un questionnaire. Méthode : envoi postal, avec enveloppe retour, et relances téléphoniques. On espère 90% de courriers bien adressés et 70% de réponses. Calcul : 80 / 0,7 / 0,9 = 126,98. On veut envoyer 127 courriers à 3€ soit un budget de 381€. Arrondi à 400€, cela fera 133 courriers.

Recueillir les données

Nous verrons dans ce chapitre comment :

- Concevoir un questionnaire (chapitre 1 en page 21)
- Augmenter le taux de réponse (chapitre 2 en page 23)
- Sélectionner les personnes à interroger (chapitre 3 en page 24)
- Saisir les données (chapitre 4 en page 25)
- Corriger les données (chapitre 5 en page 28)
- Gérer les données manquantes (chapitre 6 en page 31).

1 Concevoir un questionnaire

1.1 Composants de formulaires

Rappel sur les types de variables :

- **quantitatives discrètes** (ex : nombre d'enfants), ou **continues** (ex : taille, poids)
- **qualitatives monovaluées**, ordonnées ou non (le stade d'un cancer 1, 2a, etc.).
- **binaires** : oui/non, 0/1, vrai/faux, etc.
- **qualitatives multivaluées** : à plusieurs réponses possibles (ex : les différents moyens de transport utilisés par le sondé).

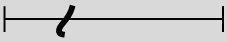
Composants de formulaire à saisie libre (surtout variables quantitatives) :

- ***textbox*** (1, 2, 3 ci-dessous) : saisie mono-ligne
- ***textarea*** (4 ci-dessous) : saisie multiligne, plutôt pour les commentaires

1	Dans quelle ville exercez-vous ? *Indiquez la commune, et non le lieu-dit*	
2	Quel âge avez-vous ? *Âge en années entières révolues.*	ans
3	Quel jour sommes-nous ? *Format jj/mm/aaaa*	JJ / MM / AAAA
4	Commentaires *Votre réponse ne sera pas analysée, c'est juste pour vous permettre de vous défouler. Vous perdez votre temps ;-)*	

Composants de formulaire à saisie contrainte, une seule réponse possible :

- ***radiobox*** (1, 2, 5 ci-dessous) : multiligne ou monoligne, idéale pour les variables qualitatives, ou quantitatives discrétisées (ex : âge en classes)
- **échelle visuelle analogique**, **EVA** (3 ci-dessous) : on mesure la distance entre le bord gauche et la coche pour obtenir une variable quantitative continue
- ***checkbox unique*** (4 ci-dessous), essentiellement pour le consentement

1	Quel est votre mode d'exercice ? *Une seule réponse possible. En cas d'exercices multiples, cochez la bulle correspondant à votre mode d'exercice principal.*	◯ Urbain ◯ Semi-rural ◯ Rural
2	Quel est votre sexe ? *Une seule réponse possible.*	◯ Femme ◯ Homme
3	Quel est votre niveau de douleur ? *Réalisez une seule coche comme ci-dessous :*  *gauche : aucune douleur, droite : la douleur la plus forte que vous puissiez imaginer.*	0% ⊢——⊣ 100%
4	Pouvons-nous publier vos réponses ? *Cochez cette case si vous acceptez que l'ensemble de votre questionnaire soit publié.*	☐ Oui, j'y consens
5	Vous souhaitez fermer votre cabinet *Indiquez votre niveau d'accord avec cette proposition.*	◯ Pas du tout d'accord ◯ Plutôt pas d'accord ◯ Ni d'accord, ni pas d'accord ◯ Plutôt d'accord ◯ Tout à fait d'accord

Echelle de Likert (5 ci-dessus ; détaillée par la suite) :

- cas particulier de *radiobox*, idéale pour une information subjective
- une proposition radicale et simple suivie de 3, 5 ou 7 propositions

Saisie contrainte à plusieurs réponses possibles : les ***checkbox*** **multiples**

Quels modes de transport utilisez-vous pour aller travailler ? *Plusieurs réponses possibles. Indiquez les modes de transport que vous utilisez au moins une fois, lors d'une semaine type.*	☐ Voiture individuelle ☐ Covoiturage ☐ Bicyclette ☐ Transports en commun ☐ Autre

Terminologies :

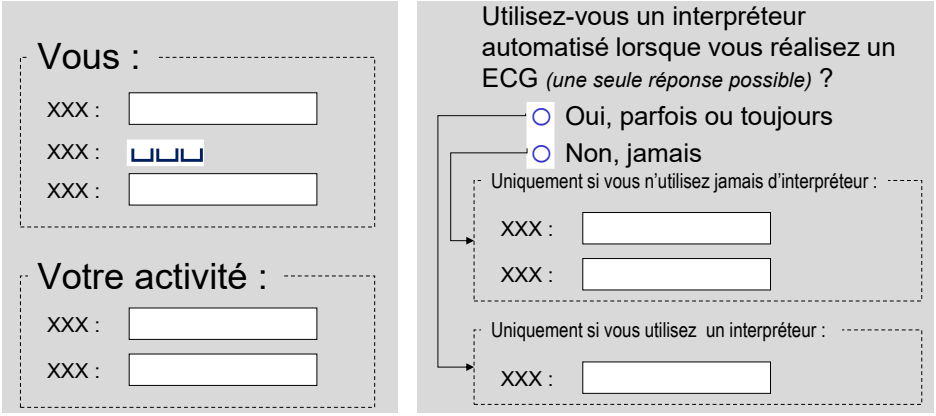
- Dictionnaires qui listent des libellés, ou associent des codes à des libellés
- A utiliser préférentiellement si existent déjà : rigueur et comparabilité
- Exemple ci-après : catégories socio-professionnelles (CSP) de l'Insee

Code	Intitulé
AZ	Agriculture, sylviculture et pêche
BE	Industrie manufacturière, industries extractives et autres
FZ	Construction
GI	Commerce de gros et de détail, transports, hébergement et restauration
…	…

1.2 Conseils pour la présentation sur papier

Pour améliorer la mise en page, le taux de réponse et la qualité des réponses :

- Faire apparaître des **zones sémantiques** (de quoi parle-t-on ?)
- Expliciter les **dépendances fonctionnelles** (répondez seulement si…)
- Faire apparaître en **blanc les zones** attendant une réponse



Il est souhaitable de regrouper et aligner les **échelles de Likert** :

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 d'opinions antivaccins est motivée par le profit financier	◯	◯	◯	◯	◯

1.3 Autres considérations

Adaptez la mise en page aux techniques d'impressions et aux circonstances :

- **Photocopieuses mécaniques** : strict noir et blanc, sans niveaux de gris
  => éviter les fonds et les trames
- **Imprimantes à jet d'encre** : encre hydrosoluble
  => éviter de colorer les zones de réponses, éviter la pluie
- **Imprimantes laser** : déposent une couche d'enrobage
  => résiste à l'eau, mais difficile d'écrire sur les zones imprimées

Quelle que soit la technique, **laisser un fond blanc dans les zones de réponses**.

Mode d'administration du questionnaire :

- Questionnaire **hétéro-administré** : complété par l'investigateur face au sujet
- Questionnaire **auto-administrés** : rempli par le sujet, seul
- Utilisé par l'investigateur seul : papier ou **e-CRF** (*electronic clinical research form*)

Attention à respecter l'**autorisation CNIL** obtenue, et donc l'anonymat.

2 Augmenter le taux de réponse des professionnels de santé aux questionnaires auto-administrés

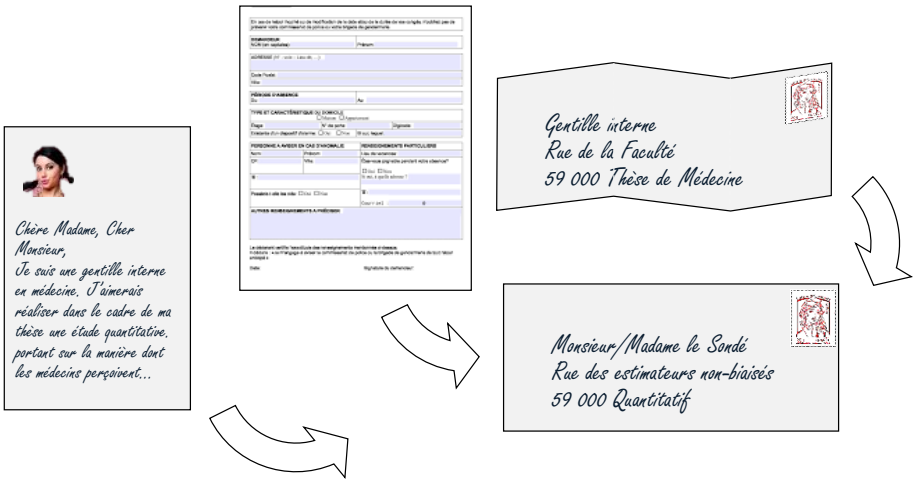
Questionnaires anonymes auto-administrés soumis à des professionnels de santé :

- **Electroniques** :
  - rapides à faire, économiques, fournissent des données propres
  - taux de réponse faible ou inconnu (<10%) => biais de sélection majeur !

- **Sur papier** : à privilégier !
  - le seul moyen d'obtenir un taux de réponse connu et élevé (~70%)
  - permettent de maîtriser l'inclusion et la sélection aléatoire (voir page 24)
  - garantissent l'unicité des réponses
  - limitent les fuites de données

Conseils pour **améliorer le taux de réponse** d'un questionnaire postal (3€/sondé) :

- Enveloppe principale DL 11x22 affranchie au tarif 20 grammes
- Questionnaire soigné, **une seule feuille** A4 recto-verso imprimée en **couleurs**
- **Lettre d'accompagnement** A5 soignée avec **photographie** de l'expéditeur
- **Enveloppe de retour** DL 11x22 comportant l'adresse et **affranchie** à 20g



-

Chronologie de la campagne :

- Préparer et envoyer tous les courriers (ex : 200 courriers)
- Attendre 2 semaines
- Réaliser une relance téléphonique systématique, proposer de renvoyer le courrier
- Attendre 2 semaines puis clôturer la campagne
- Saisir les réponses dans un tableur au fur et à mesure de leurs arrivées

3 Sélectionner les sondés par tirage au sort

Inférence statistique :

- Le fait de réaliser des tests statistiques, calculer des intervalles de confiance, etc.
- Est théoriquement valide uniquement si l'échantillon est **tiré au sort** de la population d'intérêt (voir page 18).

Tirage au sort sur liste finie en pratique avec un tableur :

(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

Tirage au sort sur des pages (annuaire, site web, etc.) :

- Créer un tableau comportant des numéros allant de 1 au nombre total d'individu
- Procéder de même que précédemment
- Retrouver les individus sur les pages. *Ex : sur un annuaire présenté avec 10 individus par page, le 26ème individu se trouve sur la 3ème page, en 6ème position.*

Tirage au sort prospectif : pour sélectionner un à un des individus au fil de l'eau, choisir un dispositif physique permettant d'obtenir les probabilités souhaitées. Exemples :

- Pour 50% : jet de pièce, inclus si {Face}
- Pour 25% : tirer une carte dans un jeu complet, inclus si {Cœur}
- Pour 3/13 : tirer une carte dans un jeu complet, inclus si {2 ; 3 ; 4}

4 Saisir des données dans un tableur

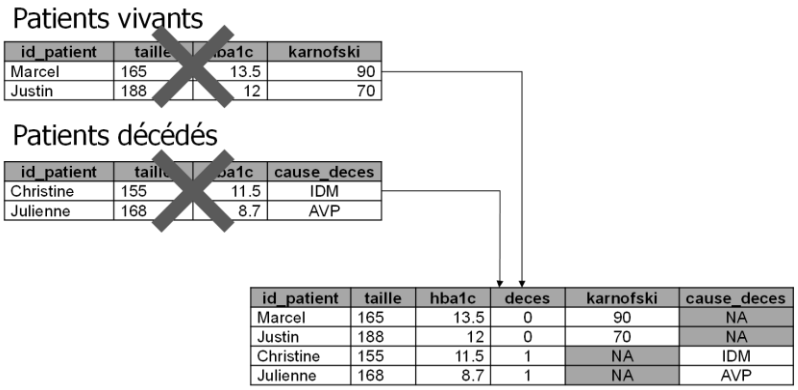
4.1 Principes généraux

Le recueil est généralement réalisé dans une **unique table** :

- **une ligne par individu** statistique
  - le plus souvent une personne physique
  - assez souvent, un séjour hospitalier, une consultation
  - parfois, un œil, une dent, etc.
- **une colonne par variable**
- **noms de variables** en première ligne, sans fusion ni fractionnement. Exemple :

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

Dans l'exemple suivant, nous regroupons les patients dans une seule table :



Dans l'exemple suivant, nous analysons des consultations et non des personnes :

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

(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

Dans l'exemple suivant, nous analysons des yeux et non des personnes :

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

(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

Les **noms de variables** :

- Figurent en **première ligne**, sans fusion ni fractionnement
- Pas de « chapeau » : utilisez plutôt des **préfixes** (exemple ci-dessous)
- Chaque nom est **unique** et débute par une lettre
- Idéalement, en **minuscules, sans espace, sans accent ni caractère spécial**

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

↓ devient

vanesth_date	vanesth_nom	vanesth_duree	vchir_date	vchir_nom	vchir_duree
…	…	…	…	…	…

Respect de l'**anonymat** :

- **Licéïté** : ne saisir que les variables désignées dans l'**autorisation CNIL**
- **Parcimonie** : si des données nominatives ou indirectement nominatives sont autorisées, les saisir dans un tableau séparé, confidentiel
- **Traçabilité :** donner un **numéro** à chaque questionnaire, écrit en gros sur la feuille, et reporté dans le tableau : il permettra de revenir au papier si besoin
- **Sécurité** : stocker les questionnaires **sous clef et les détruire** après l'étude

4.2 Détails en fonction du type de variable

Identifiants :

- Nombre entier ou chaîne de caractères courte
- Valeur unique si c'est l'identifiant de l'individu statistique
- Aucune valeur manquante autorisée

Variables quantitatives, discrètes ou continues :

- S'alignent **automatiquement à droite** si elles sont bien saisies
- Ne pas écrire l'unité dans la case ; garder la même pour toute la colonne
- Pas de **séparateur de milliers**
- Utiliser le **séparateur décimal** du système d'exploitation (virgule ou point)
- **Durées** : saisir un nombre décimal dans une seule unité
  (ex : un mois dure 30,44 jours, une année dure 365,25 jours)
- Ne pas utiliser le symbole « % »

- Indiquer toute la **précision** disponible, sans arrondi
- Données manquantes : laisser la case vide ou saisir « NA » si destiné à un statisticien

Variables binaires :

- Saisir **0** (zéro) pour non/faux/absent/jamais (etc.)
- Saisir **1** (un) pour oui/vrai/présent/toujours (etc.)
- Données manquantes : laisser la case vide ou saisir « NA » si destiné à un statisticien

Dates :

- Attention : souvent indirectement nominatives !
- Les valeurs s'alignent **automatiquement à droite** si elles sont bien saisies
- Choisir un format sans ambiguïté et utiliser le même dans toute la colonne
- Données manquantes : laisser la case vide ou saisir « NA » si destiné à un statisticien

Variables qualitatives non-ordonnées :

- Saisir directement des mots.
- Privilégier du **texte simple et court** sans caractères spécial (ex : préférer « SF » à « sage-femme » ou « maïeuticien » ou « maïeuticienne »)
- **Ne pas remplacer ces valeurs par des nombres**, c'est inutile et source d'erreur !
- Vérifier l'absence de variation typographique (ex : Bleu, bleu, [espace]bleu, etc.)
- Données manquantes : laisser la case vide ou saisir « NA » si destiné à un statisticien

Variables qualitatives ordonnées (ou « ordinales ») :

- Respecter les consignes précédentes
- Choisir des étiquettes de texte dont l'**ordre alphabétique** est naturel.
  Ex : 0_aucun 1_brevet 2_bac 3_licence 4_master 5_these
- Pour les **échelles de Likert**, saisir des nombres de -2 à +2 :
  - 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**

Variables qualitatives multivaluées :

- Les saisir comme autant de **variables binaires** que de réponses possibles
- Prendre des décisions cohérentes en cas de réponses incohérentes
- Utiliser un **préfixe commun** à toutes ces variables. Exemple :

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

Variables décrivant un événement temps-dépendant (survie) :

- Saisir une première colonne binaire : l'événement a-t-il été observé ?
- Saisir une deuxième colonne quantitative positive indiquant (même unité) :
  - Soit le délai de survenue (si l'événement s'est produit)
  - Soit la durée du suivi (s'il ne s'est pas produit)
- Utiliser un préfixe commun à ces noms de variables (cf. ci-dessous)
- Pas de valeur manquante autorisée : alors saisir 0 et 0

Avez-vous été opéré du deuxième œil ?

Suite à votre première chirurgie de la cataracte, avez-vous subi la même chirurgie sur l'autre œil ?

◯ Oui — Si oui, combien de temps après ? ___ ___ ___ mois

◯ Non — Si non, depuis combien de temps avez-vous été opéré ? ___ ___ ___ mois

Le formulaire précédent pourra être saisi comme suit :

id	chir_oeil2_evt	chir_oeil2_delai	commentaire_didactique
1	1	4	Evénement au bout de 4 mois
2	1	5	Evénement au bout de 5 mois
3	0	8	Pas d'événement pendant 8 mois
4	0	10	Pas d'événement pendant 10 mois

4.3 Saisie numérique de certaines variables

En général, il ne faut pas saisir en numérique des variables qualitatives (ex : couleur des cheveux). Cela reste possible dans certains cas cependant. Exemples :

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

Pour les variables quantitatives discrétisées (pour des raisons d'anonymat) :

- On peut saisir le centre de classe
- Cela permettra de réaliser certains calculs par la suite si les classes sont assez nombreuses (moyenne, médiane, corrélation linéaire, régression linéaire).
- Exemple :

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
…	…

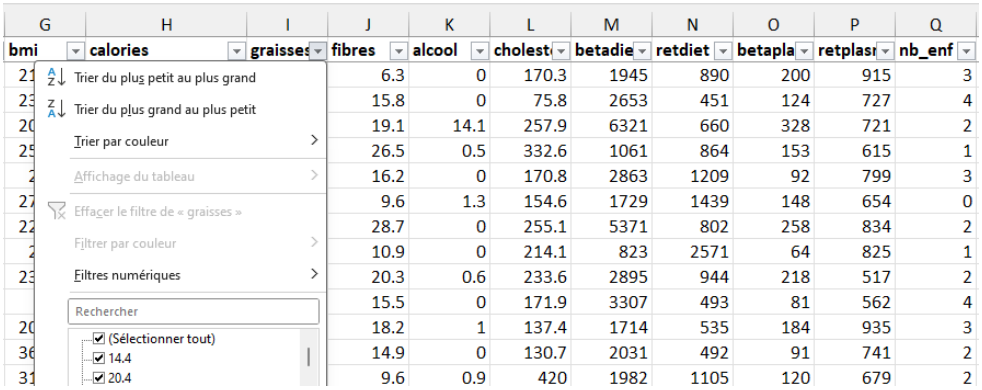
5 Vérifier, corriger et recoder des données

5.1 Détecter et corriger les erreurs de saisie

Activez le filtre automatique sur votre tableau de données :

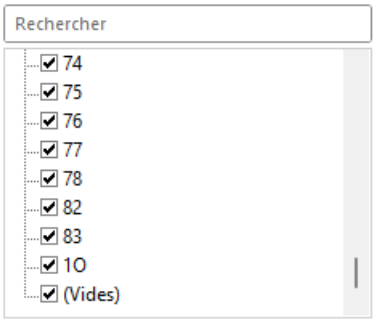
- Avec Microsoft Excel : *menu Données > Filtre*
- Avec LibreOffice Calc : *menu Données > Autofiltre*

- Ceci ajoute une liste déroulante des valeurs rencontrées à chaque colonne :

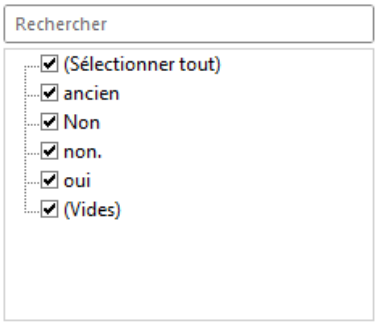


G	H	I	J	K	L	M	N	O	P	Q
bmi	calories	graisses	fibres	alcool	cholest	betadie	retdiet	betapla	retplasr	nb_enf
21			6.3	0	170.3	1945	890	200	915	3
23			15.8	0	75.8	2653	451	124	727	4
20			19.1	14.1	257.9	6321	660	328	721	2
25			26.5	0.5	332.6	1061	864	153	615	1
2			16.2	0	170.8	2863	1209	92	799	3
27			9.6	1.3	154.6	1729	1439	148	654	0
22			28.7	0	255.1	5371	802	258	834	2
2			10.9	0	214.1	823	2571	64	825	1
23			20.3	0.6	233.6	2895	944	218	517	2
			15.5	0	171.9	3307	493	81	562	4
20			18.2	1	137.4	1714	535	184	935	3
36			14.9	0	130.7	2031	492	91	741	2
31			9.6	0.9	420	1982	1105	120	679	2

Pour les **variables quantitatives**, les valeurs anormales (hors limites), illégales (mauvais types) et manquantes s'affichent aux extrémités de la liste des valeurs.



Pour les **variables qualitatives ou binaires**, on peut aisément détecter les variations typographiques non-souhaitées. Il faudra les corriger. **/!\ Attention :** sur les versions récentes d'Excel, le tableau croisé dynamique est plus sensible que le filtre automatique.



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

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

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

Selon la version du tableur, les **dates** s'affichent parfois de manière hiérarchique. Les valeurs anormales apparaissent aux extrémités de la liste.

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

Parfois, il est nécessaire de recoder certaines variables :

- Ex : regrouper les modalités des variables qualitatives
- Ex : discrétiser les variables quantitatives (les mettre en classes)
- /!\ Ne pas remplacer les valeurs, mais créer **de nouvelles colonnes**.

Discrétisation d'une variable quantitative :

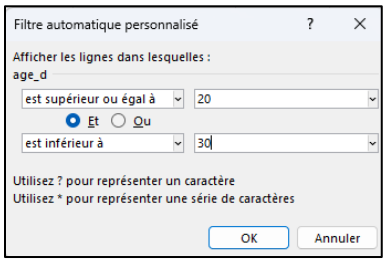
- Désigne le fait de la « mettre en classe » : elle devient qualitative ordonnée
- Il n'y a pas de règle de fixation de seuils. Ils peuvent être :
  - des seuils retrouvés dans la littérature
  - des seuils intuitifs pour les experts du domaine
  - les arrondis que tout le monde attend, par habitude
  - les quartiles observés, etc.

Option 1 - recodage manuel :

Accessible aux moins expérimentés :

- Recopier la colonne à recoder
- Appliquer un filtre automatique sur tout le tableau
- Sur la nouvelle colonne uniquement, sélectionner les valeurs (ou plages de valeurs) à modifier, et les remplacer
- Pour utiliser le tiret débuter la saisie par une apostrophe : « '20-30 »

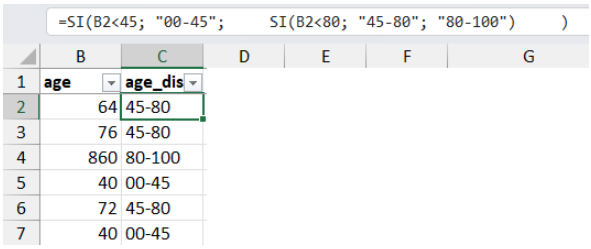
id	age	age_d
1	64	64
2	76	76
3	860	860
4	40	40
5	72	72
6	40	40
7	65	65
8	58	58
9	35	35
…	…	…



id	age	age_d
56	29	20-30
97	22	20-30
101	25	20-30
128	26	20-30
151	25	20-30
164	23	20-30
180	27	20-30
181	27	20-30
222	26	20-30
…	…	…

Option 2 - recodage par formule :

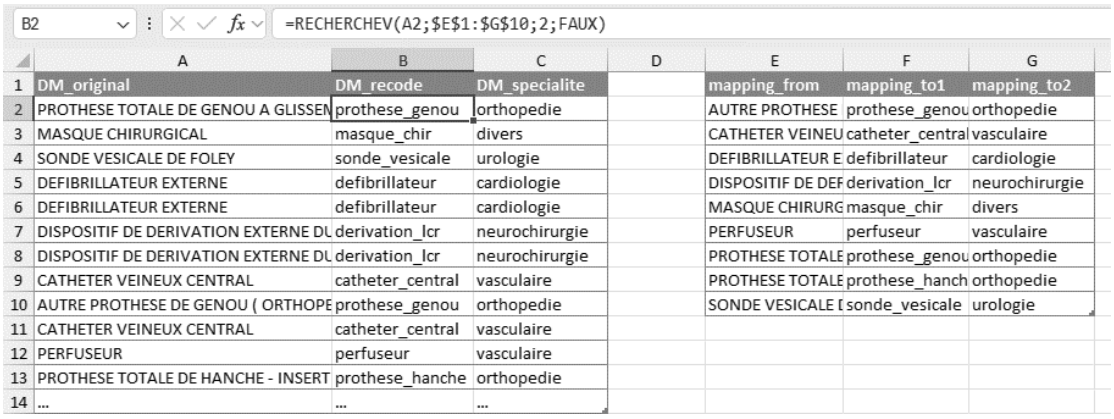
Il est possible de recoder les variables en insérant une formule :



	B	C	D	E	F	G
1	age	age_dis				
2	64	45-80				
3	76	45-80				
4	860	80-100				
5	40	00-45				
6	72	45-80				
7	40	00-45				

Option 3 : recodage par table de correspondance

Cette méthode permet de définir les substitutions dans un autre tableau. Elles sont ainsi traçables, réversibles et ré-exécutables en cas de mise à jour des données (exemple ci-dessous).



	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 I	sonde_vesicale	urologie
11	CATHETER VEINEUX CENTRAL	catheter_central	vasculaire				
12	PERFUSEUR	perfuseur	vasculaire				
13	PROTHESE TOTALE DE HANCHE - INSERT	prothese_hanche	orthopedie				
14	...	...	...				

6 Gérer les données manquantes

Données manquantes :

- Cases vides ou valeurs « NA » dans un tableau de données
- Parfois imputables à dire d'expert
  ex : score Glasgow =15 pour les patients qui rentrent à domicile
- Tout à fait normales en cas de dépendance fonctionnelle (exemple ci-dessous)

Avez-vous été opéré du deuxième œil ?

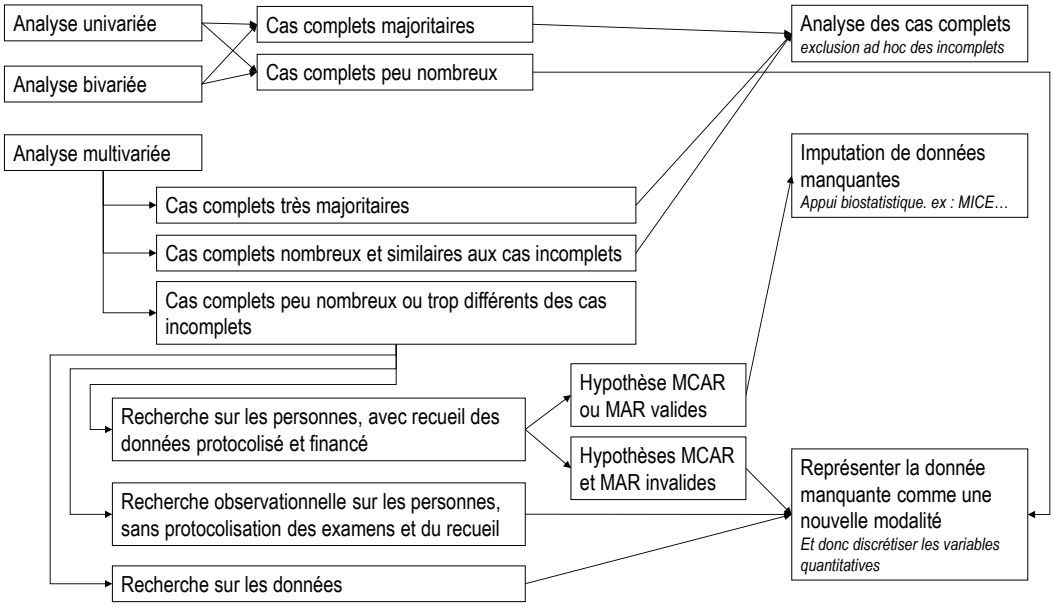
◯ Oui - - - Si oui, combien de temps après ? ___ ___ ___ jours

-

◯ Non

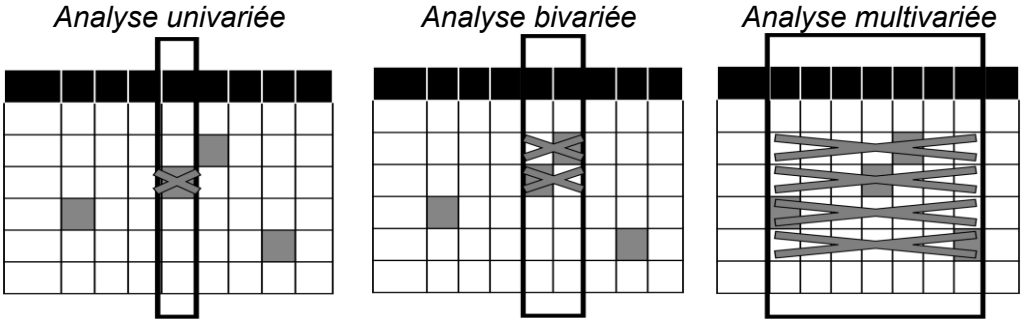
Trois **attitudes** (voir arbre décisionnel ci-après) :

- Ne rien faire et **analyser les cas complets** pour les variables considérées
- **Imputer les données manquantes** (statisticien), uniquement si ces données sont manquantes aléatoirement, indépendamment de leur valeur (MAR ou MCAR)
- Traiter les données manquantes comme une **nouvelle modalité** non-manquante



Analyse des **cas complets** :

- Approche par défaut de toutes les fonctions de tableur ou logiciels statistiques
- Pour chaque analyse, exclut temporairement les individus ayant des valeurs manquantes sur les variables considérées
- Parfois problématique en multivarié. Illustration ci-dessous :



Comment améliorer l'analyse multivariée des cas complets ?

- Exclure certaines variables pour augmenter la proportion de cas complets
- Construire des variables composites. *Ex : « insuffisance rénale », fausse en l'absence d'information, vraie si on trouve un DFG faible, ou un code CIM10 d'insuffisance rénale, ou un acte de dialyse rénale*

Décrire le NA comme une **nouvelle modalité** :

- Simple, efficace, et toujours valide
- Utile si les données manquantes sont assez nombreuses
- Variables qualitatives ou binaires : fonctionne toujours
- Variables quantitatives : nécessite une **discrétisation** préalable. Exemple :

id_patient	age	age_d
1	63	"25-64"
2	59	"25-64"
3	*NA*	"inconnu"
4	45	"25-64"
5	86	"75-109"
6	*NA*	"inconnu"

Réaliser les analyses statistiques

1 Préambule

Les analyses statistiques peuvent être classées en :

- **Analyses univariées** (chapitre 2 page 33) :
  - Chaque colonne est analysée seule
  - Systématiquement fait dans toutes les études
  - Ce cadre inclut :
    - La **description** des variables (graphiques, paramètres)
    - Le calcul d'**intervalles de confiances** des moyennes et des proportions
    - Les **tests** de comparaison d'un paramètre observé à une valeur **attendue**
    - Les **tests** de comparaison « de deux moyennes / proportions **appariées** »
- **Analyses bivariées** (chapitre 3 page 47) : liens entre deux variables
  - Le plus souvent tests d'**indépendance statistique**
  - Parfois quantification de la **force de l'association** (facteurs de risque et protecteurs, tests diagnostiques, courbe ROC, coefficient Kappa)
- **Analyses multivariées** (chapitre 4 page 64) : notamment pour expliquer ou prédire une variable en particulier, à partir de plusieurs variables

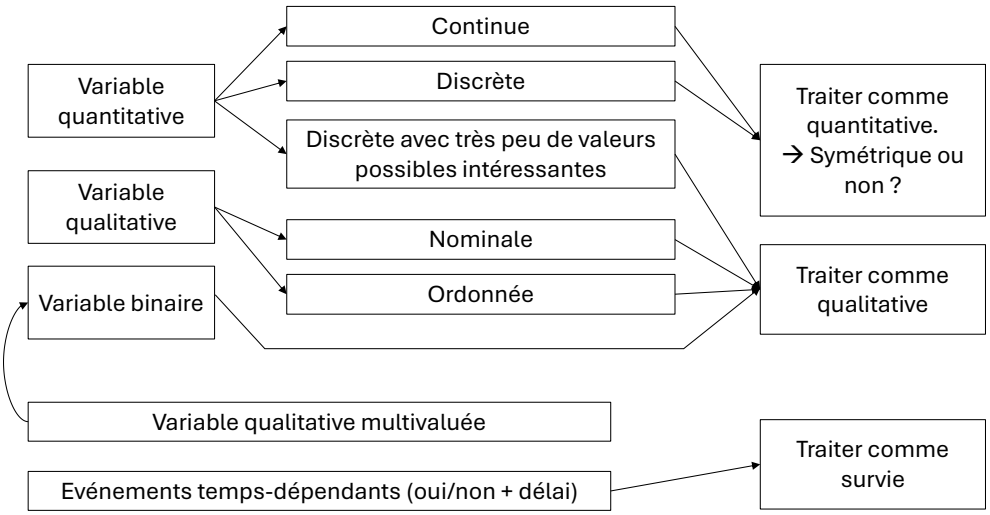
Nous livrerons enfin quelques réflexions sur les tests (chapitre 5 page 65), et verrons comment interpréter les p valeurs obtenues (chapitre 6 page 68).

2 Analyses statistiques univariées

2.1 Définir le type de variable

Types de variables :

- Présentés dans le schéma ci-dessous
- Concrètement, 3 grands types de variables *in fine* : **quantitative, qualitative ou de survie**.



Discrétisation d'une variable quantitative :

- Transforme des nombres en modalités qualitatives ordonnées pertinentes

- Guidée avant tout par l'expertise et le bon sens, au vu des valeurs observées
- *Exemple : le nombre d'enfants d'un foyer, « aucun », « un », « plus d'un »*
- *Exemple : l'âge, "[15;25[", "[25;65[", "[65;75[", "[75;110["*
- Idéalement, recueillir la variable avec toute la précision disponible, puis discrétiser si besoin lors de l'analyse statistique

Quels **paramètres ou indicateurs** calculer pour décrire une variable :

- En théorie, une foule de paramètres
- En pratique, normes internationales https://www.equator-network.org/ :
  - **Strobe** pour les études **observationnelles**, dont les questionnaires
  - **Consort** pour les **essais randomisés**
  - **Prisma** pour les **revues systématiques** de la littérature

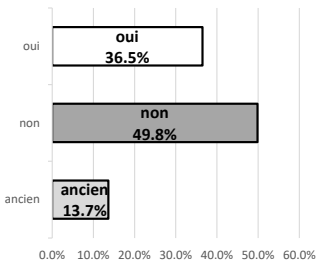
2.2 Variables qualitatives

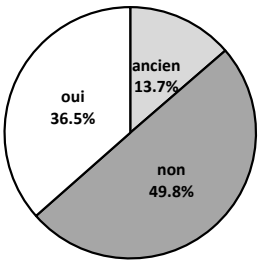
2.2.1 Description et présentation

Description d'une variable qualitative :

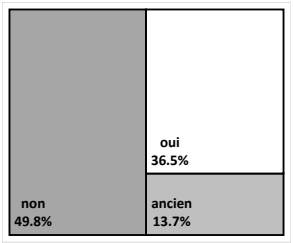
- Toujours rapporter **l'effectif et la proportion** de chaque modalité rencontrée, dans une phrase ou un tableau, éventuellement après regroupements pertinents
- Avec un tableur : utiliser un **tableau croisé dynamique**, ou **table dynamique**, ou **pilote de données**, ou ***pivot table*** (selon votre logiciel)
- Graphiques (facultatifs) : représenter ces proportions, généralement en surface

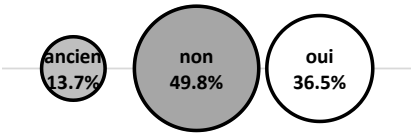
Exemples de graphiques faisant varier une dimension : diagramme en barres et diagramme en secteurs (camembert, *pie chart*)





Exemples de graphiques faisant varier deux dimensions : treemap et diagramme en bulles





La présentation en phrase s'adapte au sous-type de variable :

- **Variable qualitative quelconque** : par fréquence décroissante, en homogénéisant la précision (ex : un chiffre après la virgule) :
  Ex : Parmi les patients, 22 (44,0%) sont bruns, 16 (32,0%) sont blonds etc.
- **Variable qualitative ordonnée** : conserver l'ordre des modalités
  Ex : Treize patients (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.
- **Variable binaire** : rapporter la modalité la plus marquante, ou la plus fréquente
  Ex : Deux patients (4,0%) sont décédés.
  Ex : L'échantillon comporte 36 (72,0%) femmes.
- **Variable qualitative multivaluée** :
  Ex : 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.

2.2.2 Calcul de l'intervalle de confiance d'une proportion

Calcul de l'**intervalle de confiance à 95%** d'une proportion :

- Intervalle dans lequel la vraie proportion en population a 95% de chances d'être
- **Ne doit pas être calculé, sauf** si c'est votre objectif principal
- Calculable avec la **méthode de Wald**, qui utilise la **loi normale**
- Prérequis : observer **au moins 5 individus** dans chaque modalité de la variable

Mise en œuvre dans Microsoft Excel ou LibreOffice Calc :

- Disposer l'effectif total de l'échantillon (P1)
- Disposer la fréquence du caractère étudié (nombre compris entre 0 et 1) (P2)
- Calculer la demi-largeur de l'intervalle de confiance (P5)
- Borne basse = proportion observée diminuée de ce nombre (P6)
- Borne haute = proportion observée augmentée de ce nombre (P7)

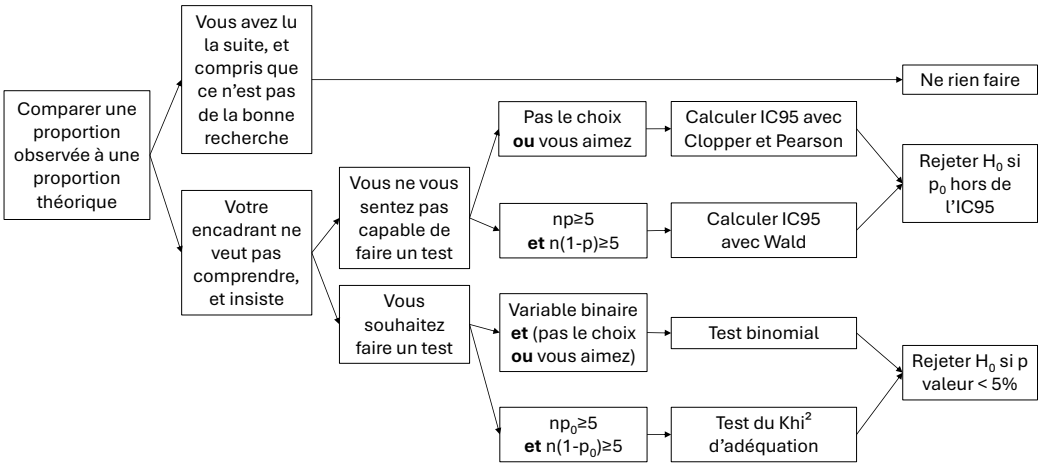
	O	P	Q
1	Effectif total :	315	
2	Fréquence :	13.70%	
3			
4	IC à 95% :		
5	demi-intervalle :	3.80%	=1.96*RACINE(P2*(1-P2)/P1)
6	borne basse :	9.90%	=P2-P5
7	borne haute :	17.50%	=P2+P5

2.2.3 Tests de comparaison d'une proportion observée à une proportion attendue

2.2.3.1 Introduction

Principes de ces tests :

- On **observe** dans l'échantillon une **proportion** *p*
- On souhaite savoir si cette observation est **compatible avec l'hypothèse nulle H_0 :** « L'échantillon est issu aléatoirement d'une population caractérisée par une certaine proportion p_0 » (hypothèse externe)
- On calcule la **p valeur** (ou petit p, ou p) :
  - **En supposant H_0 vraie**, c'est la **plausibilité** de cette observation précise
  - Probabilité d'observer cette observation, ou une observation plus éloignée encore de l'observation attendue sous H_0 (calcul bilatéral ou unilatéral, cf. post)
- Conclusion :
  - Si p valeur < 5% : « *p* est significativement différente de p_0 ».
  - Si p valeur > 5% : « on ne met pas en évidence de différence statistiquement significative entre *p* et p_0 ».
  - Ce seuil de 5% est une convention en santé : c'est le **risque α** (alpha) ou **risque de première espèce**



2.2.3.2 Test binomial

Le test binomial :

- **test exact** et **non-paramétrique**
- permet de comprendre ce qu'est la p valeur en général
- s'appuie sur la **loi binomiale**, que vous avez peut-être étudiée au lycée

Exemple : Nous nous intéressons au cancer du larynx. Dans un échantillon de 16 malades, 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 cette hypothèse (49%) ?

Mise en œuvre avec un tableur (voir image ci-après) :

- Disposer l'effectif de l'échantillon (16 en B1)
- Disposer la fréquence attendue du sexe masculin (0,49, en B2 ; hypothèse H_0)
- Tracer un tableau listant le nombre d'hommes que nous aurions théoriquement pu observer parmi 16 individus : 0, 1, 2, …, 15, 16 (colonne A).
- Pour chaque ligne du tableau, calculer la probabilité que nous aurions eue d'observer ce nombre, en supposant H_0 vraie, en utilisant la formule Excel **loi.binomiale()** ou Calc **loibinomiale()**. La somme de ces probabilités vaut bien 1 (colonne B)
- Pour conclure :

  - L'échantillon observé était associé à une probabilité de 2,36%
  - Identifier les lignes associées à une probabilité inférieure ou égale à 2,36% : on recherche des deux côtés (cela caractérise un **test bilatéral**)
  - Calculer la somme de ces probabilités : **c'est la p valeur**, ici 4,53%
- Conclusion statistique : p valeur < 5%, nous rejetons l'hypothèse nulle et concluons que **la proportion observée d'hommes est significativement différente de 49%**
- Interprétation médicale : le sexe masculin est probablement un facteur de risque de cette maladie

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%

Le calcul de la p valeur procède d'un **raisonnement par l'absurde** : nous supposons H_0 vraie, et calculons la plausibilité de l'observation : c'est la p valeur.

- Si p valeur < 5% : nous avons observé un fait qui serait trop peu plausible si H_0 était vraie, donc H_0 est fausse
- Si p valeur > 5% : l'échantillon observé est compatible avec H_0, cela ne prouve rien : nous faisons face à une indétermination

2.2.3.3 Test du Khi² d'adéquation

Le test du **Khi² d'adéquation** (ou Chi-deux, ou X^2, etc.) :

- Test **non-paramétrique** et **asymptotique**
- Répond à la même question que le test binomial
- Teste l'adéquation entre des proportions observées et des proportions attendues
- S'appuie sur un calcul **rapide et facile**
- Permet de traiter des variables **qualitatives à plus de 2 modalités** si nécessaire
- Impose une condition de validité : effectif théorique minimal
- Consiste à calculer la **statistique de test**, qui est ici le X^2 (khi²), puis une **p valeur**

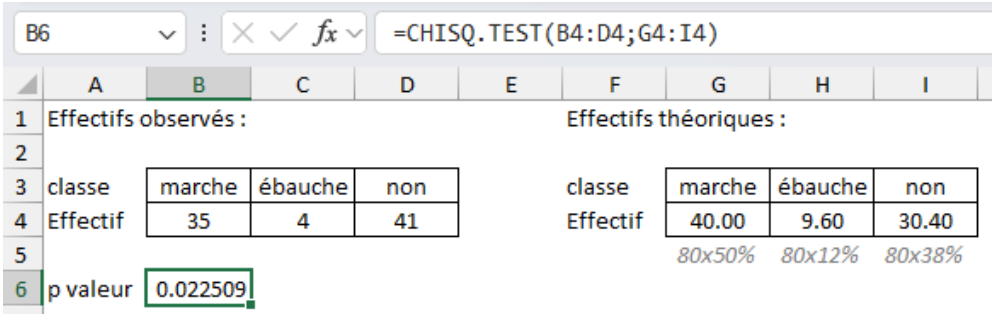
Exemple : A un âge donné, parmi les bébés normaux : 50% marchent, 12% ont une ébauche de marche, 38% ne marchent pas. Sur 80 bébés prématurés, nous observons que 35 marchent, 4 ont une ébauche de marche, 41 ne marchent pas. On souhaite savoir si cette distribution observée est significativement différente de celle attendue.

Le test se construit comme suit :

- Population étudiée : les bébés prématurés
- Variable étudiée : la marche, qualitative à 3 classes
- 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, bilatéral, risque alpha de 5%

Mise en œuvre sur une feuille de tableur (voir illustration ci-dessous) :

- Disposer un tableau des **effectifs observés**
- Disposer un tableau des **effectifs théoriques,** attendus sous H_0 (en supposant qu'elle soit vraie), en utilisant les probabilités de H_0 et l'effectif total. Conserver 1 ou 2 chiffres après la virgule
- Vérifier que chacun de ces nouveaux nombres est **supérieur ou égal à 5**
- Calculer directement la p valeur avec la formule Excel **chisq.test()** (ou **test.khideux()** anciennement, ou **test.loi.khideux()** avec Calc). Elle prend en paramètres les deux plages d'effectifs
- Conclusion statistique : p valeur<5% donc la distribution de la marche est significativement différente de celle attendue
- Conclusion médicale : les bébés prématurés présentent un retard à la marche



Le calcul de la p valeur procède d'un **raisonnement par l'absurde** : nous supposons H_0 vraie, et calculons la plausibilité de l'observation : c'est la p valeur.

- Si p valeur < 5% : l'observation est trop peu plausible sous H_0, donc H_0 est fausse
- Si p valeur > 5% : nous faisons face à une indétermination

2.2.4 Tests de comparaison « de deux proportions appariées »

Dans certaines situations, un même phénomène binaire est mesuré deux fois chez les mêmes individus : on parle de **mesures appariées avant-après**, ou **gauche-droite**.

Exemple : des participants 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.

Les questions « avant-après » ou « gauche-droite » dans un seul groupe :

- Ce ne sont pas des problèmes bivariés mais univariés
- Calculer la **variation individuelle**, par exemple : $\Delta x_i = x_{i,après} - x_{i,avant}$:
  - 0 si le statut de l'individu est stable (1→1 ou 0→0)
  - +1 si le statut de l'individu augmente (passe de 0 à 1)
  - -1 si le statut de l'individu diminue (passe de 1 à 0)
- Comparer les effectifs de « +1 » et de « -1 » à l'aide d'un **test de McNemar**

Le test de **McNemar** :

- Est un test du Khi² spécialisé
- Test **non-paramétrique** et **asymptotique**

- Hypothèse nulle H_0 : probabilité de « -1 » = probabilité de « +1 »
- Effectifs observés : nombre d'individus qui changent de statut, dans chaque cas
- Effectifs théoriques : la moitié du nombre d'individus qui changent de statut, deux fois
- Vérifier que chaque effectif théorique est bien supérieur ou égal à 5
- Calculer la p valeur avec **chisq.test()** avec Excel, ou **test.loi.khideux()** avec Calc
- Rejeter H_0 si p valeur<5%. Sinon, indétermination

J13 : =CHISQ.TEST(J3:J4;J9:J10)

	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
14	13	0	0	0

DeltaX	Effectif
-1	13
0	20
1	7
Total général	**40**

Effectifs observés

statut	effectif
"-1"	13
"+1"	7
total	20

Effectifs attendus

statut	effectif
"-1"	10.00
"+1"	10.00
total	20

p val Khi² : 0.179712495

2.3 Variables quantitatives

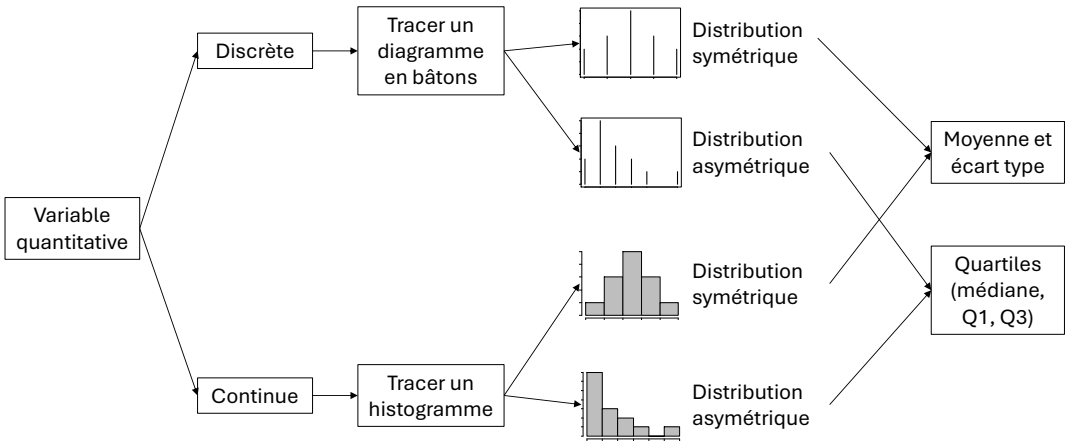
2.3.1 Description et présentation

2.3.1.1 Arbre décisionnel

La description d'une variable quantitative s'appuie sur deux indicateurs :

- un **indicateur de tendance centrale** : *les valeurs sont-elles hautes ou basses ?*
- un **indicateur de dispersion** : l*es valeurs sont-elles resserrées ou dispersées ?*

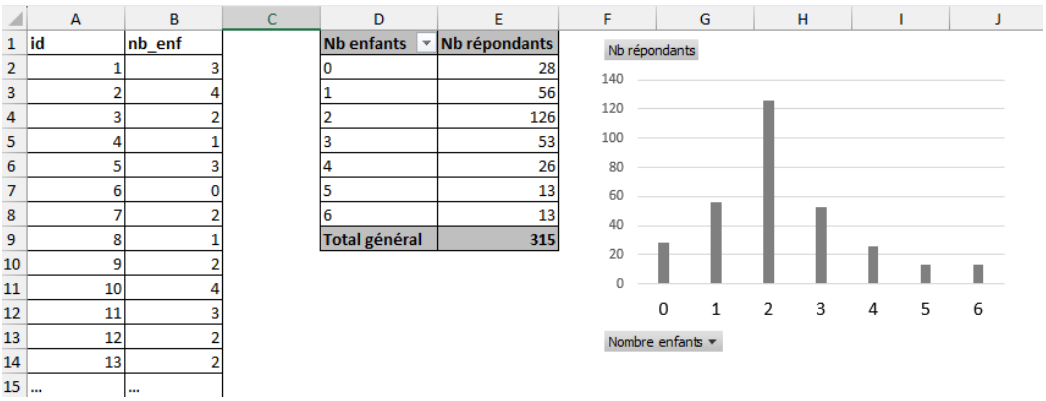
Ces indicateurs sont choisis après visualisation graphique de la distribution.



2.3.1.2 Variables quantitatives discrètes

Pour les variables quantitatives discrètes, on trace un **diagramme en bâtons** :

- Utiliser un tableau croisé dynamique pour obtenir le tableau de contingence
- Tracer ensuite un **diagramme en bâtons**. Vérifier que l'abscisse est cohérente
- Diminuer l'épaisseur des barres : ce sont des bâtons



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
...	...

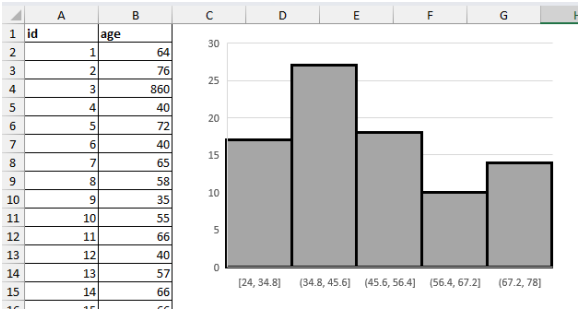
Nb enfants	Nb répondants
0	28
1	56
2	126
3	53
4	26
5	13
6	13
Total général	315

2.3.1.3 Variables quantitatives continues

Pour les variables quantitatives continues, on tracera idéalement un **histogramme en densité de fréquence** :

- Il s'appuie sur des classes contiguës de la variable
- Il est fait de rectangles dont la surface est proportionnelle à la proportion observée
- On ne lit pas les ordonnées, mais on interprète simplement la forme
- Hélas, les tableurs ne tracent pas ce type de graphique

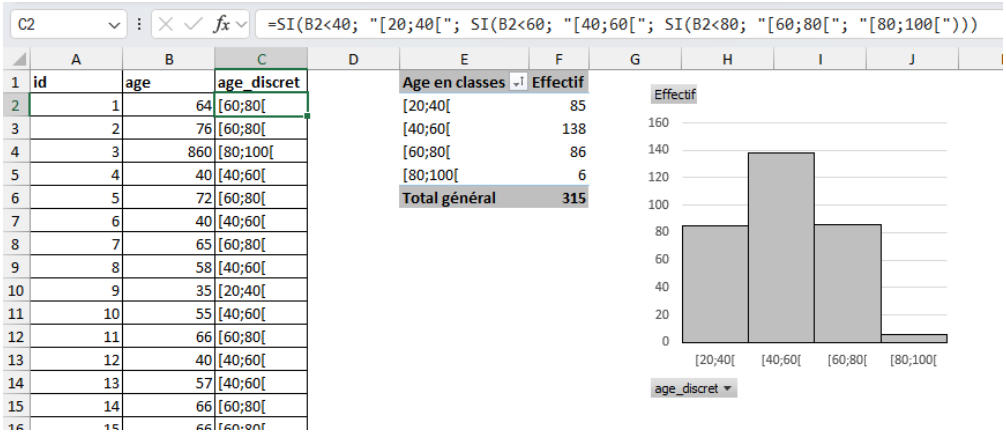
Sur des versions récentes de Microsoft Excel, demander un **histogramme en effectifs** :



id	age
1	64
2	76
3	860
4	40
5	72
6	40
7	65
8	58
9	35
10	55
11	66
12	40
13	57
14	66

Avec LibreOffice Calc, ou des versions plus anciennes de Microsoft Excel, réaliser un graphique *ressemblant à* un histogramme :

- **Discrétiser** la variable d'intérêt. Attention :
  - Les classes doivent obligatoirement être **de largeurs égales !**
  - Forcer si besoin des **zéros initiaux**, pour un tri alphabétique cohérent
- Utiliser un **tableau croisé dynamique** pour obtenir un tableau de contingence
- Tracer un diagramme en barres
- Elargir les barres pour qu'elles se touchent



id	age	age_discret
1	64	[60;80[
2	76	[60;80[
3	860	[80;100[
4	40	[40;60[
5	72	[60;80[
6	40	[40;60[
7	65	[60;80[
8	58	[40;60[
9	35	[20;40[
10	55	[40;60[
11	66	[60;80[
12	40	[40;60[
13	57	[40;60[
14	66	[60;80[

Age en classes	Effectif
[20;40[	85
[40;60[	138
[60;80[	86
[80;100[	6
Total général	315

2.3.1.4 Définition des indicateurs

Voici les indicateurs calculables (voir l'arbre décisionnel précédent) :

- **Moyenne** : somme des valeurs observées, divisée par l'effectif. Notée $\bar{x}$. Utiliser la fonction **moyenne()**.
- **Déviation standard (DS)**, ou *standard deviation* (SD), ou estimation non-biaisée de l'**écart type** : indicateur de dispersion, de même unité que la variable et sa moyenne. Utiliser les fonctions **stdeva()** ou **ecartype()** ou **ecartype.standard()**.
- **Quartiles** : ce sont les 3 valeurs seuils qui séparent l'échantillon en 4 sous-groupes d'effectifs équilibrés :
  - 25% de l'effectif est en-dessous de **Q1, premier quartile**
  - 50% de l'effectif est en-dessous de **Q2, deuxième quartile**, ou **médiane** $\tilde{x}$
  - 75% de l'effectif est en-dessous de **Q3, troisième quartile**
- **Intervalle interquartile : [Q1 ; Q3]**

2.3.1.5 Présentation des indicateurs

Distribution d'**allure symétrique** (plus ou moins) : présenter **moyenne** et **écart type**

Exemple : « Les participants sont en moyenne âgés de 32,3 ans (SD=5,8) ».

Distribution d'allure **clairement asymétrique** : présenter la **médiane, le premier et le troisième quartile**.

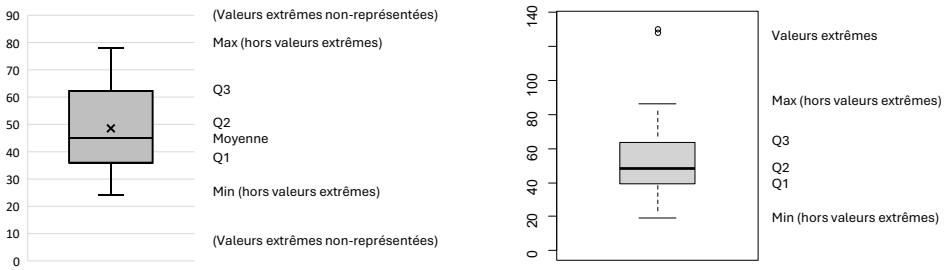
Exemple : « Les répondants ont en médiane 1 enfant (Q1-Q3 : [0 ; 3]) ».

Tous ces indicateurs ($\bar{x}$, SD, Q1, Q2, Q3) prennent l'unité de la variable d'intérêt.

2.3.1.6 Un dernier graphique utile : la boîte à moustache ou *boxplot*

La ***boxplot***, ou **boîte à moustache** :

- très utilisée, surtout lorsqu'on souhaite en juxtaposer plusieurs
- présente principalement l'étendue de la distribution et les quartiles
- son rendu diffère selon le logiciel employé (gauche : Excel ; droite : R) :



2.3.2 Calcul de l'intervalle de confiance d'une moyenne

La moyenne $\bar{x}$ est une **moyenne mesurée dans l'échantillon**, elle **estime la moyenne μ (inconnue) en population**.

L'**intervalle de confiance à 95% (IC95)** d'une moyenne :

- est l'intervalle dans lequel la vraie moyenne μ a 95% de chances de se trouver
- calculé « à 95% » par convention en recherche en santé
- **ne doit pas être calculé, sauf** si c'était votre objectif principal
- peut être calculé à l'aide de la Loi de Student

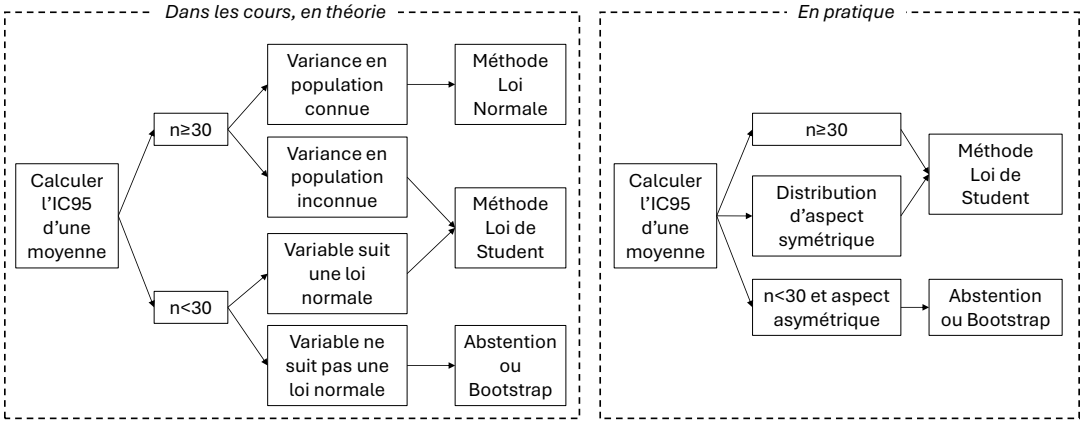
Mise en œuvre dans un tableur comme Microsoft Excel ou LibreOffice Calc :

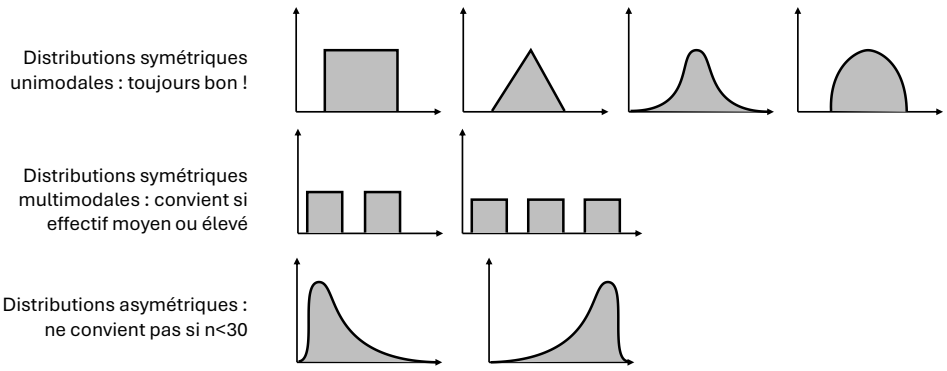
- Disposer l'effectif total avec **nb()** (cellule N1 ci-après)
- Disposer la moyenne avec **moyenne()** (cellule N2)

- Disposer la déviation standard avec **ecartype.standard()** (cellule N3)
- Calculer la demi-largeur de l'intervalle de confiance (cellule N6)
- La borne basse est la moyenne moins cette demi-largeur (cellule N7)
- La borne haute est la moyenne plus cette demi-largeur (cellule N8)
- Ici notre intervalle de confiance à 95% est [48,8 ; 52,2]

	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

Les conditions de validité sont détaillées ci-dessous. Nous vous proposons un algorithme opérationnel, plus simple et valide, sur la droite :





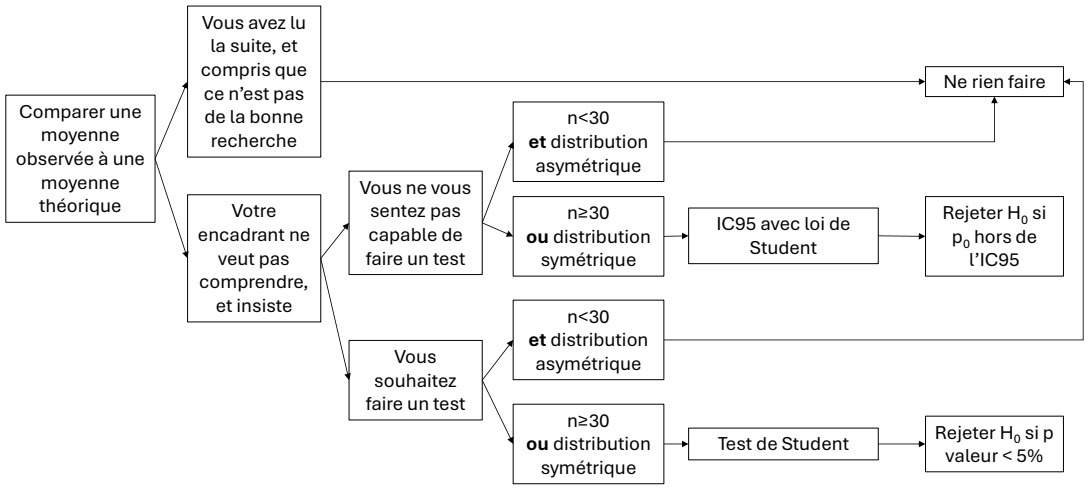
2.3.3 Tests comparant moyenne observée et moyenne attendue

2.3.3.1 Introduction

Situation typique :

- Dans un échantillon de *n* individus, on calcule une moyenne $\bar{x}$ et un écart type *DS*
- Une **hypothèse externe** indique que, en population, la moyenne *μ* inconnue vaut *m₀* ». C'est **l'hypothèse nulle** $H_0 : \mu = m_0$
- On souhaite savoir si l'observation est compatible ou non avec cette hypothèse
- On réfutera plus facilement H_0 si la différence observée $\bar{x} - m_0$ est importante par rapport à la dispersion de la variable dans l'échantillon, OU si l'effectif est élevé

Plusieurs tests sont possibles. Nous détaillerons le **test de Student** ci-après. Nous verrons cependant plus tard, dans le chapitre 5.1 Tests de comparaison à une norme en page 65, qu'il n'est pas vraiment conseillé d'utiliser ce test.



2.3.3.2 Test de Student observé-attendu

Le test « Student observé-attendu » :

- test d'adéquation entre une moyenne observée et une moyenne attendue
- il est **paramétrique** et **asymptotique**.

Exemple : 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 (en l'absence d'information, on observerait son histogramme).
- 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, 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

Dans un tableur, on procède comme suit :

- Saisir la moyenne, la déviation standard et l'effectif, ou les recalculer avec les fonctions **nb()**, **moyenne()** et **stdeva()** (ou **ecartype()** avec Calc, ou **ecartype.standard()** avec Excel)
- Calculer la statistique de test « t » (ici en cellule 10)

- Déduire la p valeur correspondante avec la formule **loi.student.bilaterale()**
- Interprétation statistique : ici p valeur <5% donc nous rejetons H_0
- Interprétation médicale : avoir un QI faible est un facteur de risque de criminalité

	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

Le calcul de la p valeur procède d'un **raisonnement par l'absurde** : nous supposons H_0 vraie, et calculons la plausibilité de l'observation : c'est la p valeur.

- Si p valeur < 5% : l'observation est trop peu plausible, donc H_0 est fausse
- Si p valeur > 5% : nous faisons face à une indétermination

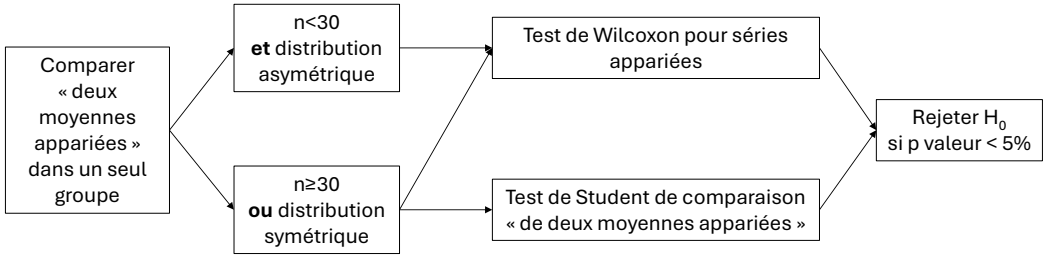
2.3.4 Tests de comparaisons « de deux moyennes appariées »

Parfois, une même variable quantitative est mesurée deux fois chez les mêmes individus : il peut s'agir de **mesures appariées avant-après**, ou **gauche-droite** par exemple.

Exemple : on mesure la taille le matin (x_{avant}, en centimètres) et le soir ($x_{après}$) de participants. 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.

Les problèmes « avant-après » ou « gauche-droite » :

- Ne relèvent pas des analyses bivariées, mais univariées
- Calculer pour chaque individu la variation. Ex : $\Delta x_i = x_{i,après} - x_{i,avant}$
- Question : la moyenne de cette variation est-elle significativement différente de zéro ?
- Solution : test de Student observé-attendu entre cette moyenne et la valeur « 0 » (voir section 2.3.3.2 page 43)
- Test **paramétrique** et **asymptotique**
- Correct mathématiquement mais douteux méthodologiquement : voir le chapitre 5.2 page 65



2.4 Variables de survie

2.4.1 Définitions

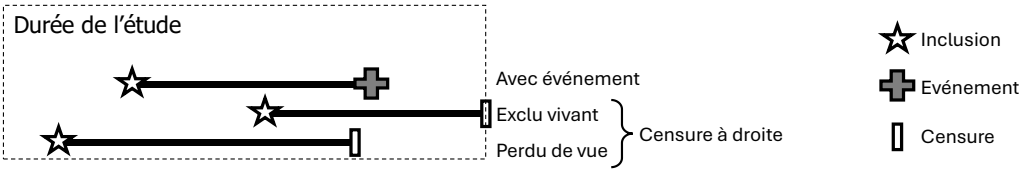
Les données de survie :

- Décrivent la survenue (inconstante) d'un événement dans le temps
- Evénement péjoratif (décès, début d'une maladie, accident, rechute, etc.) ou mélioratif (retour à domicile, rémission, reprise de la marche, fin d'un arrêt de travail, etc.)

Trois cas de figure se présentent :

- L'événement est observé précisément
- Des individus ne présentent toujours pas l'événement au terme de l'étude : on parle d'**exclus vivants** (mais ces individus ne seront pas exclus !)
- Des individus quittent l'étude en cours, mais n'ont pas présenté l'événement durant leur suivi : on parle de **perdus de vue** (eux aussi feront partie de l'analyse)

Les « perdus de vue » et les « exclus vivant » sont traités identiquement : on parle de **censure à droite** car on sait seulement que le délai est supérieur à une durée.



2.4.2 Présentation de données

Rappel sur la saisie des données (voir page 27) :

- Première colonne : binaire, indique si l'événement a été observé
- Deuxième colonne : indique le délai sans événements (nombre réel positif) :
  - Si l'événement a été observé : délai au bout duquel l'événement a eu lieu
  - Si l'événement n'a pas été observé : délai pendant lequel l'individu a été suivi (pour les exclus vivants et les perdus de vue).

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

2.4.3 Description avec la courbe de Kaplan-Meier

Méthode descriptive de référence :

- **Estimateur de Kaplan-Meier**
- Sa représentation graphique : la « courbe de survie »
- Estime la proportion de survivants au fil du temps
- X=temps ; Y=**probabilité estimée de survivre au moins** ce temps-là

Exemple de mise en œuvre dans un tableur (voir illustration ci-après) :

- Disposer un tableau de données, avec une ligne par individu (onglet séparé)
- Tracer un tableau croisé dynamique (en haut au milieu) permettant, pour chaque délai et pour chaque statut d'événement, d'avoir l'effectif concerné. Modifier ainsi :
  - 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 (on examine les événements, puis les censures, et non l'inverse)
- Recopier statiquement ce tableau (plage A17:H28 dans l'exemple suivant)
- Insérer manuellement une ligne correspondant au délai zéro :
  - Effectif = la taille de l'échantillon
  - Probabilités instantanées et cumulées de survie = 1
- Compléter la deuxième ligne comme en ligne 30 ci-dessous et étendre ces formules. Voici leur signification :
  - Le nombre d'individus correspondant à un temps donné est séparé en nombre d'événements, puis en nombre de censures (colonnes D et E)
  - Effectif_avant correspond à l'effectif résiduel observé à l'issue de la ligne immédiatement au-dessus, après éviction des événements et des censures.
  - 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

Puis tracer le graphique :

- Tracer un « nuage de points »
- Abscisse = délai, ordonnée = probabilité cumulée de survie
- Terminer le graphique en reliant les points à la main. Partir vers la droite puis vers le bas, 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

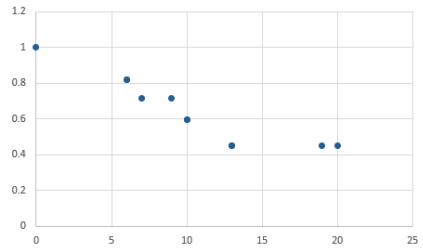
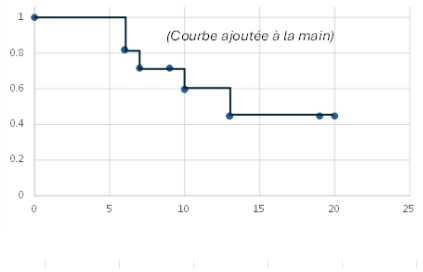


Tableau de survie

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



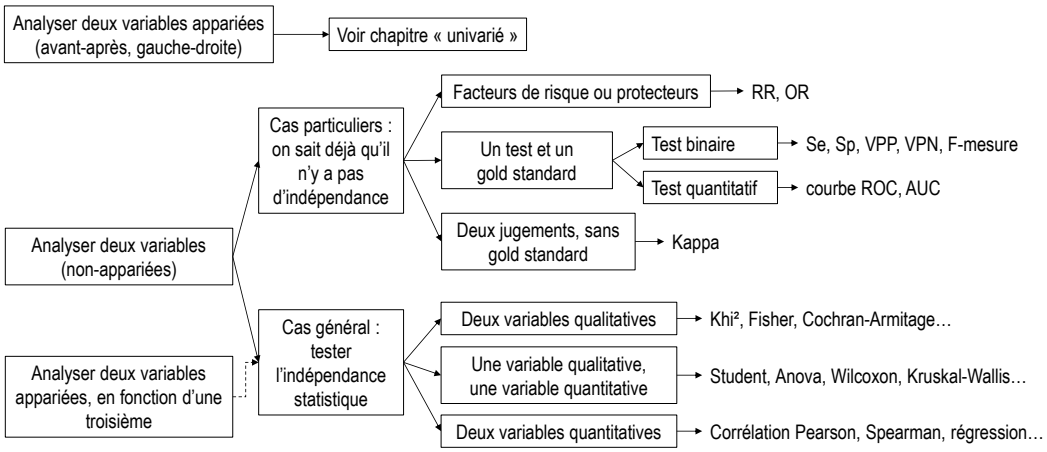
2.4.4 Inférence statistique

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 est complexe, et ne peut pas être réalisé simplement avec un tableur. En pratique, aucun test statistique univarié n'est réalisé sur une variable de survie seule.

3 Analyses statistiques bivariées

3.1 Préambule

Les **analyse bivariées** s'intéressent à la relation entre deux variables :



3.2 Cas général : liaison statistique entre deux colonnes

3.2.1 Préambule

Trois types de variables combinées => **cinq cas de figure**

Quelle affirmation souhaite-t-on pouvoir réfuter ? Voir tableau ci-dessous :

	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]

3.2.2 Deux variables qualitatives

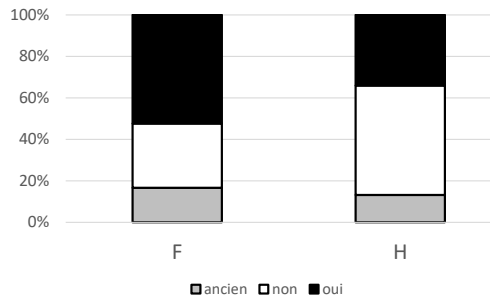
3.2.2.1 Généralités et description

Analyser la relation entre deux variables qualitatives A et B :

- Revient à se demander si la distribution de A est influencée par la valeur de B
- *Exemple : la proportion de blonds/bruns/roux/blancs est-elle la même chez les femmes et les hommes ?*
- Sens de la relation A→B ou B→A : n'a aucune importance

Représentation graphique :

- Objectif : voir la distribution de B dans les sous-groupes de A
- Tracer d'abord un tableau de contingence croisé des effectifs selon A et B
- Réaliser un graphique, par exemple en barres empilées
- *Exemple ci-après : statut tabagique en fonction du sexe*

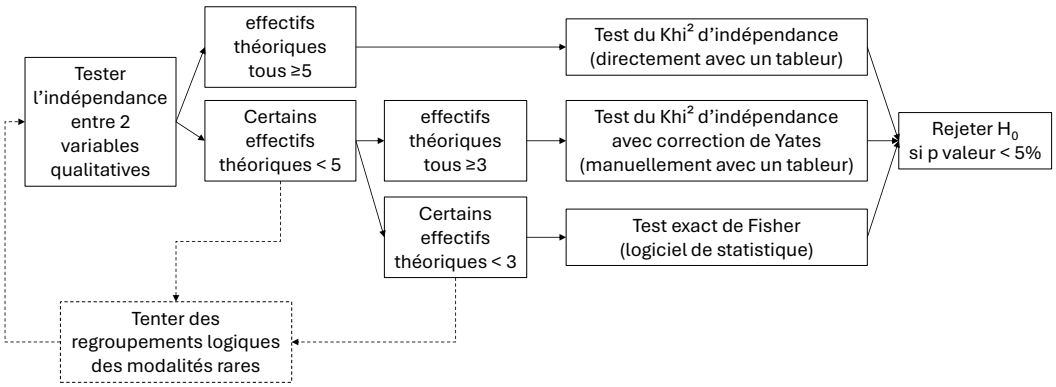


Description par des paramètres :

- Décrire les proportions d'une variable, conditionnellement à l'autre variable
- Calculer les pourcentages de manière intuitive

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. »

3.2.2.2 Tests statistiques : arbre décisionnel



3.2.2.3 Test du Khi² d'indépendance

Test du **Khi² d'indépendance** :

- principal test d'indépendance entre deux variables qualitatives ou binaires
- test **non-paramétrique** et **asymptotique**
- requiert des effectifs minimaux
- on pourra regrouper des modalités de manière cohérente si besoin

Exemple : nous analysons un échantillon de 315 individus, et souhaitons savoir si le statut tabagique est indépendant du sexe.

On procède comme suit dans un tableur :

- Poser **l'hypothèse nulle H_0** : les deux variables sont indépendantes
- Evincer les individus présentant des valeurs manquantes
- Tracer un premier tableau de contingence croisé, présentant les **effectifs observés**
- Tracer un deuxième tableau des **effectifs théoriques** :
  - Reproduire la structure et les sous-totaux
  - Calculer chaque cellule ainsi : *total_ligne x total_colonne / total_général*
  - Garder 1 ou 2 chiffres après la virgule
  - Vérifier la condition de validité : chaque effectif est supérieur ou égal à 5
- Dans une dernière cellule (ici D41, formule en E41), utiliser la formule **chisq.test()** avec Excel ou Calc pour comparer les deux plages de données (observée, théorique)
- Interprétation statistique : p valeur < 5% donc le statut tabagique n'est pas indépendant du sexe

L37 : =$037*L$39/$0$39

	C	D	E	F	G	H	I	J	K	L	M	N	O
34	**Effectifs observés :**							**Effectifs théoriques :**					
35			Tabagisme							Tabagisme			
36			ancien	non	oui	Total				ancien	non	oui	Total
37	sexe	F	7	13	22	42		sexe	F	5.73	20.93	15.33	42
38		H	36	144	93	273			H	37.27	136.07	99.67	273
39		Total	43	157	115	315			Total	43	157	115	315
40													
41	p valeur :	0.0282	=CHISQ.TEST(E37:G38;L37:N38)										

Le calcul de la p valeur procède d'un **raisonnement par l'absurde** : nous supposons H_0 vraie, et calculons la plausibilité de l'observation : c'est la p valeur.

- Si p valeur < 5% : l'observation est trop peu plausible sous H_0, donc H_0 est fausse
- Si p valeur > 5% : nous faisons face à une indétermination

3.2.2.4 Autres tests

D'autres tests sont classiquement enseignés :

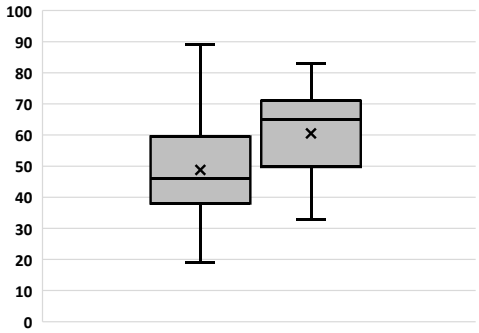
- **Comparaison de deux proportions avec la loi normale** : pas utilisé en pratique.
- **Test exact de Fisher** : idéal mais pas réalisable avec un tableur.
- Le **test de tendance de Cochran–Armitage** : devrait être utilisé lorsqu'une des variables est ordonnée. En pratique, le test du Khi² d'indépendance est plutôt utilisé.

3.2.3 Une variable quantitative et une variable qualitative

3.2.3.1 Généralités et description

Examiner la relation entre une variable quantitative Y et une variable qualitative X :

- revient généralement à **comparer les moyennes** de Y entre les différents sous-groupes déterminés par X
- graphique par excellence : **un ensemble de boxplots**, représentant la distribution de Y, pour chaque modalité de X

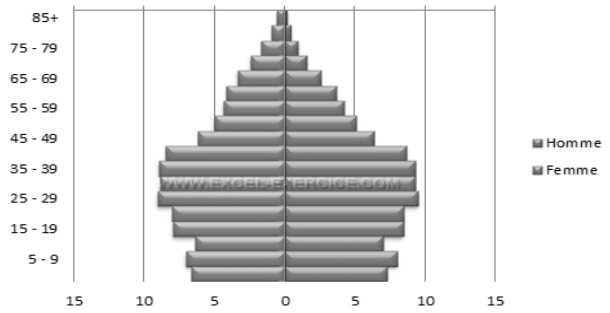


Pour réaliser ce graphique avec un tableur :

- Séparer la série Y en plusieurs séries présentées indépendamment : Y pour $X=x_1$, Y pour $X=x_2$, etc.
- Créer un graphique en boxplot pour une seule des séries
- Puis ajouter 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	

Dans le cas particulier du sexe et de l'âge, la pyramide des âges est appréciée. De nombreux tutoriels peuvent être trouvés sur le web (exemple : https://excel-exercice.com/pyramide-ages/) :



Ensuite, on décrira la variable Y en fonction des modalités de X, comme précédemment.

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) ».

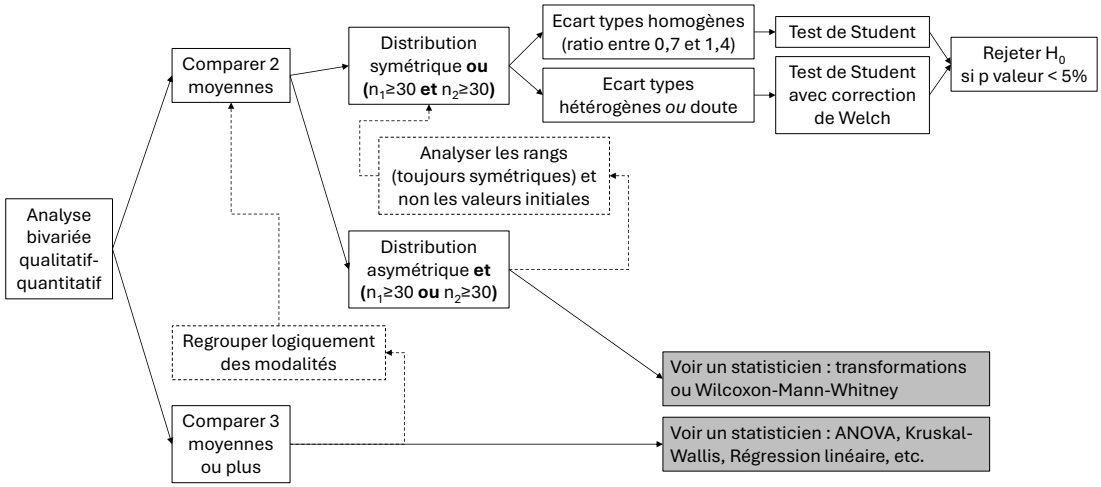
Exemple : « Dans notre échantillon, les hommes ont en médiane 2 enfants (Q1-Q3 : [0 ; 3]), et les femmes ont en médiane 1 enfant (Q1-Q3 : [0 ; 2]) ».

Si vous réalisez un test statistique, ne rapportez plus les dispersions :

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). ».

3.2.3.2 Tests statistiques : arbre décisionnel

Nous vous proposons l'arbre décisionnel suivant, qui est plus simple que dans les cours :



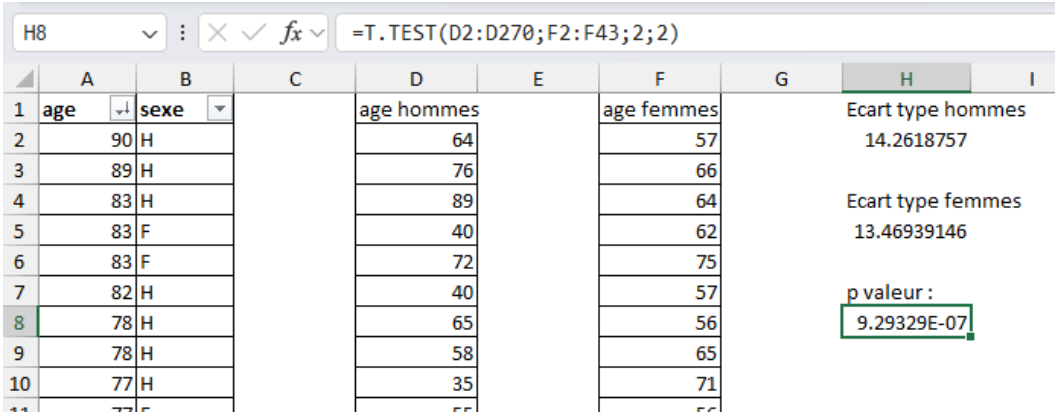
Pour comparer plus de 2 moyennes avec un tableur, vous devrez faire des **regroupements pertinents**, et comparer les sous-groupes **deux-à-deux** en appliquant la **correction de Bonferroni** : au lieu d'interpréter les p valeurs au seuil de 5%, vous les interprèterez au seuil de 5/k%, k étant le nombre de tests réalisés.

3.2.3.3 Test de Student pour échantillons indépendants, avec ou sans correction de Welch

Exemple : en partant d'un échantillon de 315 individus, nous souhaitons savoir si l'âge est indépendant du sexe, en termes de tendance centrale.

Mise en œuvre du test de Student avec un tableur :

- Poser **l'hypothèse nulle H_0** : la moyenne de l'âge ne dépend pas du sexe
- Séparer l'âge selon les sous-groupes, le présenter dans deux colonnes séparées
- Calculer l'écart type de l'âge dans chaque sous-groupe avec la fonction **stdeva()** de Microsoft Excel, ou **ecartype()** de LibreOffice Calc, ou **ecartype.standard()**. Evaluer si ces valeurs sont proches ou non (dans notre exemple : similaires)
- Calculer la p valeur du test à l'aide de la formule **t.test()** de Microsoft Excel, ou **test.student()** de LibreOffice Calc (voir cellule H8). 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 ou le nombre « 3 » pour appliquer la correction de Welch
- Conclusion statistique : p valeur < 5%, nous concluons que l'âge moyen n'est pas indépendant du sexe
- Conclusion métier : les femmes sont significativement plus jeunes/âgées que les hommes (vérifier les moyennes)



	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			

Si les écart types avaient été très différents, nous aurions utilisé la valeur « 3 » en 4ème paramètre, pour demander la **correction de Welch**. L'interprétation aurait été identique.

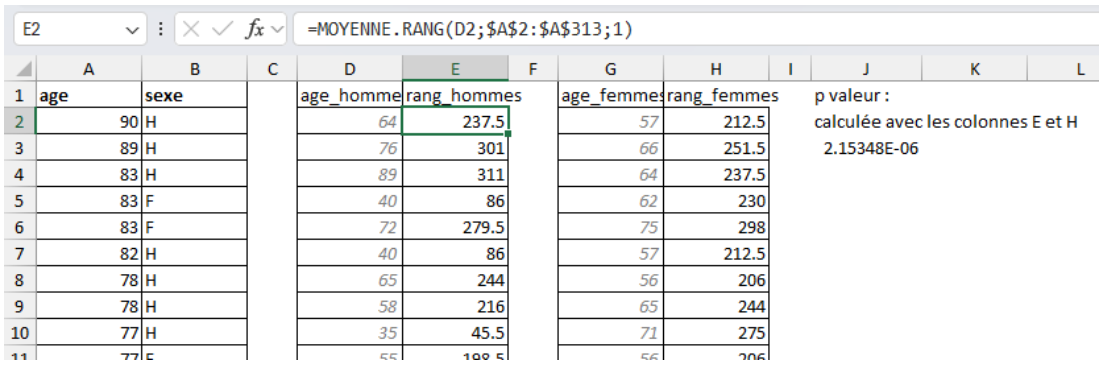
Le calcul de la p valeur procède d'un **raisonnement par l'absurde** : nous supposons H_0 vraie, et calculons la plausibilité de l'observation : c'est la p valeur.

- Si p valeur < 5% : nous observons un fait peu plausible sous H_0, donc H_0 est fausse
- Si p valeur > 5% : nous faisons face à une indétermination

Ce test est **paramétrique** et **asymptotique**.

Pour réaliser un **test de Student sur les rangs** :

- Ne pas considérer la variable quantitative, mais les rangs de cette variable
- Utiliser la fonction **moyenne.rang()** d'Excel ou Calc, avec 3 paramètres :
  - La cellule dont on veut calculer le rang
  - La plage de toutes les valeurs, les groupes étant mélangés (à figer avec les symboles « $ »).
  - Un indicateur d'ordre croissant ou décroissant : peu importe
- Le test est alors toujours valide, quel que soit l'effectif. Pas de correction de Welch
- Le reste du test est identique, y compris l'interprétation



	A	B	C	D	E	F	G	H	I	J	K	L
1	age	sexe		age_homme	rang_hommes		age_femmes	rang_femmes				
2	90	H		64	237.5		57	212.5				
3	89	H		76	301		66	251.5				
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				

3.2.3.4 Autres tests statistiques

Comparaison de deux moyennes avec la loi normale : non réalisée en pratique.

Analyse de la variance ou ANOVA : généralisation du test de Student à plus de deux sous-groupes. Elle ne peut pas être réalisée avec un tableur.

Test de Wilcoxon-Mann-Whitney : test non-paramétrique et exact. Vous pourrez plutôt réaliser un test de Student sur les rangs.

Test de Kruskal-Wallis : test non-paramétrique et exact. Il généralise test de Wilcoxon-Mann-Whitney à plus de deux catégories.

Régression logistique simple : personne ne l'utilise dans ce cas de figure.

Tests post-hoc : si on compare plus de deux moyennes, et si le test revient significatif, cela signifie qu'au moins deux moyennes sont significativement différentes, sans préciser lesquelles. Certains voudront réaliser des tests statistiques deux-à-deux pour mieux décrire cette différence : on les appelle « tests post-hoc ».

3.2.4 Deux variables quantitatives

3.2.4.1 Préambule

Analyse statistique de la **liaison entre deux variables quantitatives** :

- On pourra **mettre en évidence une relation linéaire ou monotone**
- Dans certains cas, modélise une relation linéaire $Y = aX + b$
- Sens de la relation X→Y ou Y→X n'a pas d'importance mathématique, mais doit rester cohérent. Exemple : X explique Y, X précède Y, on souhaite prédire Y, etc.
  Exemple : X sera la taille et Y le poids, car le poids est la conséquence de la taille.

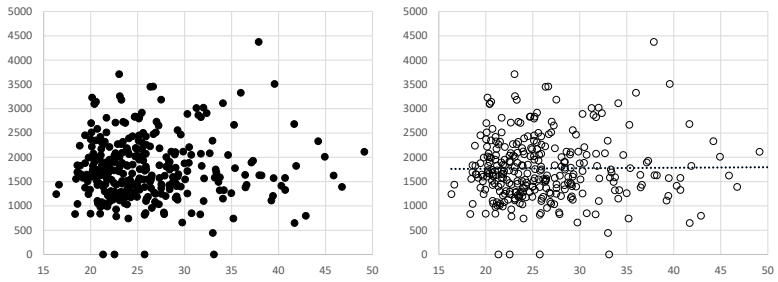
L'absence de relation linéaire n'est pas synonyme d'indépendance statistique :

- 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

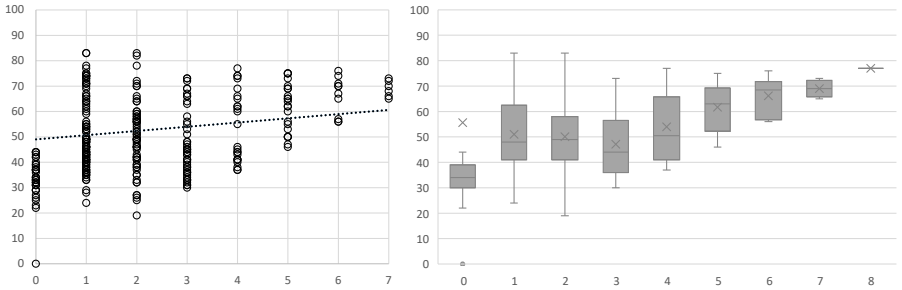
3.2.4.2 Représentation graphique

La représentation graphique

- Préalable indispensable pour comprendre la relation.
- Pour deux variables continues : nuage de points, de préférence avec des cercles (ci-dessous à droite, on voit mieux la densité de points)

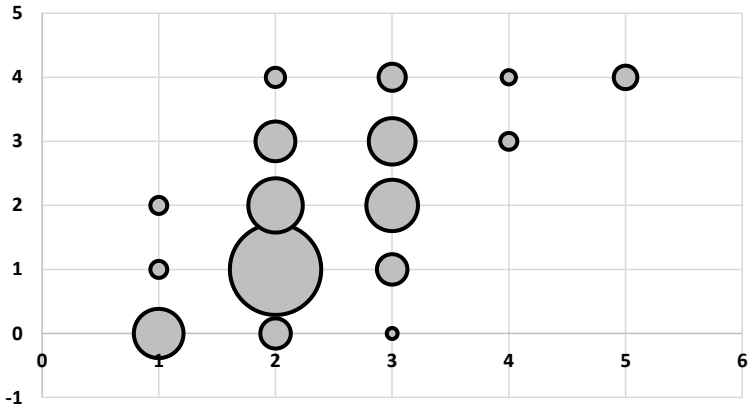


Si une des variables est discrète, envisager d'autres représentations selon le cas :



Si les deux variables sont discrètes :

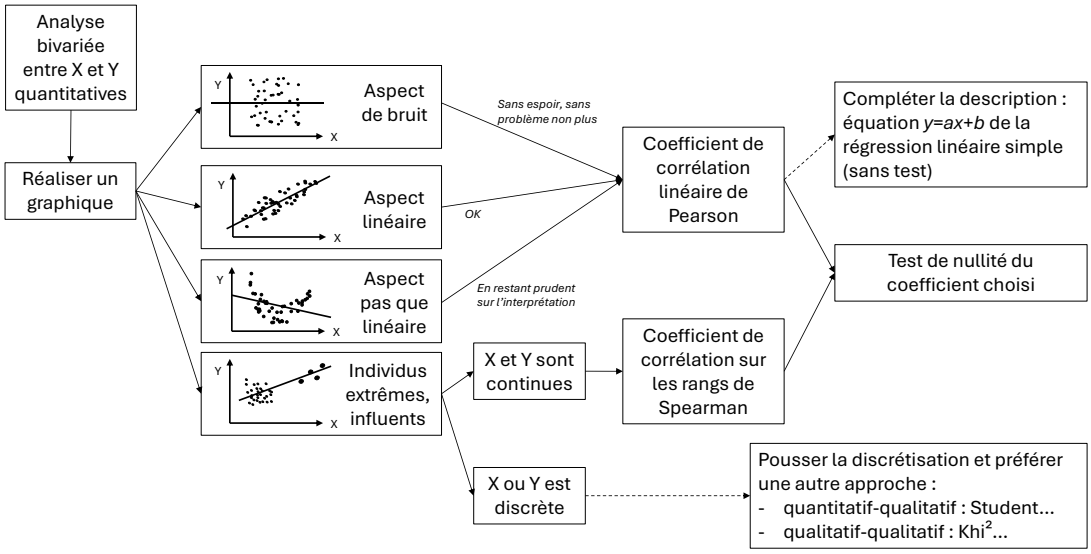
- Réaliser un tableau croisé d'effectifs (par un tableau croisé dynamique)
- Modifier ce tableau :
  - Propriétés du tableau : « disposition classique », pas de sous-total ni de total
  - Propriétés du champ qui apparait le plus à gauche : « répéter les étiquettes d'éléments », pas de sous-total
- Puis demander un **graphique en bulles** (l'effectif étant proportionnel à la surface mais surtout pas au diamètre).



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

3.2.4.3 Description par des paramètres

Le choix des paramètres descriptifs suit l'arbre décisionnel suivant :



Coefficient de corrélation linéaire de Pearson, et **coefficient de corrélation des rangs de Spearman** :

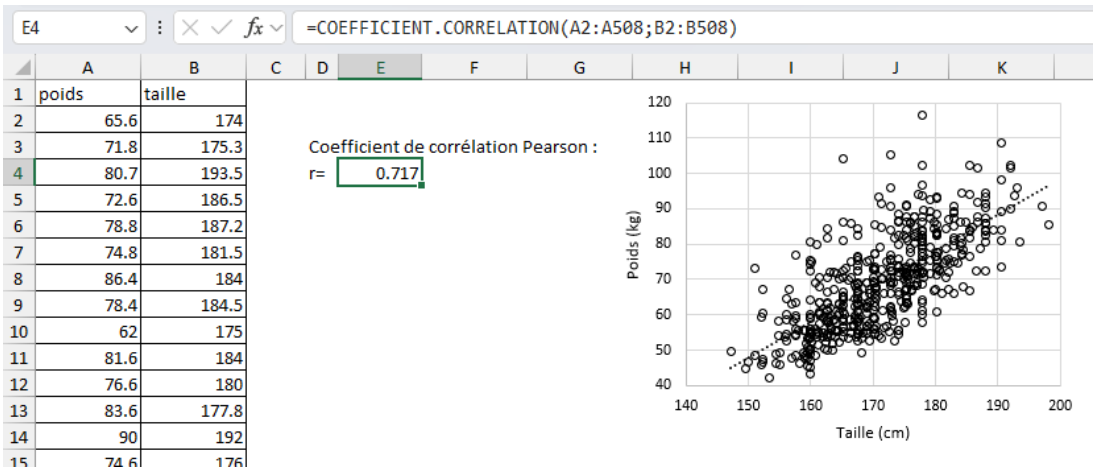
- nombres sans unité, toujours compris entre -1 et +1
- dits « **empiriques** » car calculés dans un échantillon
- interprétation (sous réserve de fluctuations d'échantillonnage) :
  - -1 association parfaite et décroissante
  - $]-1;-0{,}5]$ association décroissante, forte
  - $]-0{,}5;0[$ association décroissante, faible
  - 0 absence d'association linéaire (Pearson) ou monotone (Spearman), mais pas forcément indépendance statistique
  - $]0;0{,}5[$ association croissante, faible
  - $[0{,}5;1[$ association croissante, forte
  - 1 association parfaite et croissante

Exemple : « Dans notre échantillon, le poids et la taille sont corrélés positivement et fortement (r=0,77, p=0,012). »

3.2.4.4 Coefficient de corrélation linéaire de Pearson

Calcul du coefficient de corrélation linéaire de Pearson :

- Il quantifie une relation linéaire
- Disposer le tableau des individus, avec les deux colonnes à analyser
- Dans Microsoft Excel ou LibreOffice Calc, utiliser la fonction **coefficient.correlation()**, ou la fonction **pearson()** :
  - Premier argument : la plage de Y ou X
  - Deuxième argument : la plage de X ou Y



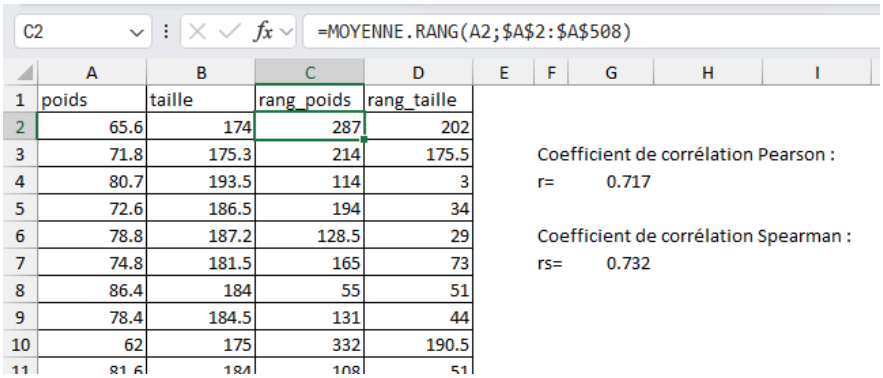
	A	B
1	poids	taille
2	65.6	174
3	71.8	175.3
4	80.7	193.5
5	72.6	186.5
6	78.8	187.2
7	74.8	181.5
8	86.4	184
9	78.4	184.5
10	62	175
11	81.6	184
12	76.6	180
13	83.6	177.8
14	90	192
15	74.6	176

3.2.4.5 Coefficient de corrélation des rangs de Spearman

Le **coefficient de corrélation des rangs de Spearman** est un coefficient de **Pearson** calculé sur les **rangs de X** et les **rangs de Y**.

Calcul du coefficient de Spearman avec un tableur :

- Il quantifie une relation monotone (croissante ou décroissante)
- Disposer un tableau des individus avec les variables X et Y
- Ajouter une colonne représentant les rangs de X à l'aide de la fonction **moyenne.rang()** (ou **rang()** sur les versions plus anciennes)
- Ajouter une colonne représentant les rangs de Y de même
- Calculer le coefficient de corrélation de Pearson sur ces rangs et non les variables initiales, avec **coefficient.correlation()**, ou **pearson()**



	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

3.2.4.6 Test de nullité du coefficient de corrélation et synthèse

Nous souhaitons savoir si le coefficient calculé (Pearson ou Spearman) est significativement différent de zéro :

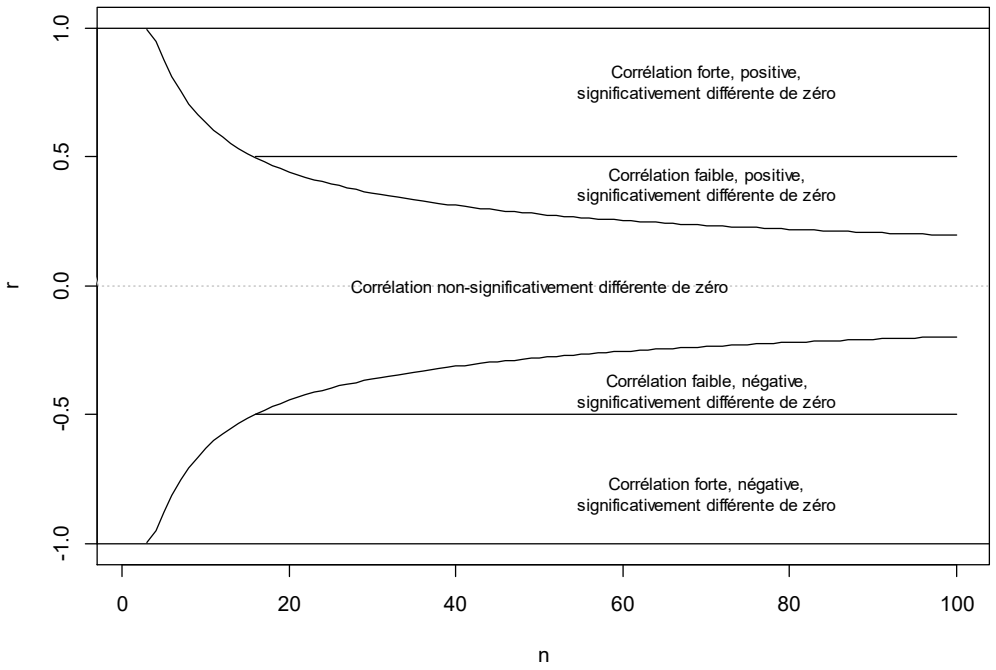
- Hypothèse nulle H_0 : en population, la corrélation entre X et Y est nulle
- Calculer le coefficient de corrélation souhaité, Pearson ou Spearman (G10 ci-après)
- Saisir le nombre d'individus
- Calculer la statistique de test (cellule G12 ci-après)
- Calculer la p valeur (cellule G13 ci-dessous)
- Ici, p valeur<5%, nous rejetons l'hypothèse nulle, la corrélation est significativement différente de zéro

	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							

De manière plus générale :

- Si **p valeur < 5% : on rejette H_0**, donc X et Y ne sont pas indépendantes
- Si **p valeur > 5%** : on fait face à une **indétermination**

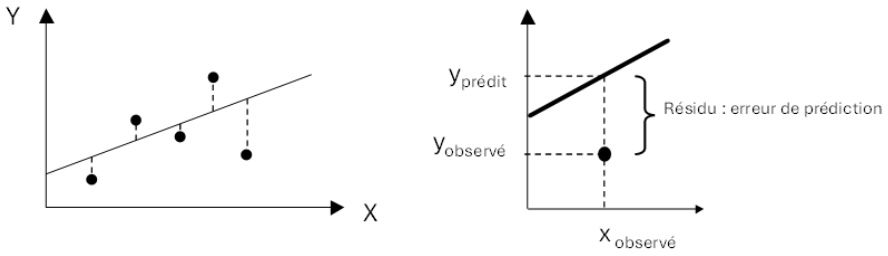
Ainsi, pour chaque réalisation du test statistique, nous avons seulement deux paramètres d'entrée : le coefficient de corrélation (r pour Pearson ou r_s pour Spearman) et la taille de l'échantillon (n). Il vous suffira de vous reporter à la figure ci-après : positionnez un point unique correspondant à votre expérience, et vous obtiendrez en une seule opération le résultat du test de nullité et l'interprétation de votre coefficient de corrélation.



3.2.4.7 Régression linéaire simple

La **régression linéaire simple** :

- Uniquement si le coefficient de Pearson est valide (cf. arbre décisionnel précédent)
- Calcule l'équation de la droite $Y = aX + b$ qui résume le mieux le nuage de points
- Pour ce faire, elle réduit le carré de l'erreur verticale, ou résidu
- Permet, si x_0 est connue, de prédire une prédire une nouvelle valeur y_0 inconnue



L'équation de la droite de régression s'obtient avec Microsoft Excel et LibreOffice Calc à l'aide des fonctions **pente()** et **ordonnee.origine()** ou plus récemment **droite.reg()** :

	E	F	G	H	I	J	K
14							
15		Equation de la droite de régression y=ax+b					
16		a	pente	1.02	=PENTE(A2:A508;B2:B508)		
17		b	ordonnée à l'origine	-105.01	=ORDONNEE.ORIGINE(A2:A508;B2:B508)		

Dans l'exemple précédent, on obtient l'équation de droite $y = 1{,}02.x - 105{,}01$. Pour un individu dont la taille est connue et le poids inconnu, on peut ainsi prédire son poids.

3.2.5 Une variable de survie et une variable qualitative

Comparer la survie de plusieurs groupes d'individus revient à tester l'indépendance entre une variable de survie et une variable qualitative.

Exemple : 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

Pour répondre à ce type de question :

- Tracer conjointement des courbes de Kaplan Meier, une pour chaque sous-groupe (voir 2.4 Variables de survie en page 45)
- Tests statistique réalisables : le **test du Log Rank** et **modèle de Cox**, avec l'aide d'un biostatisticien

3.2.6 Une variable de survie et une variable quantitative

Cette question revient à savoir si une variable quantitative influe de manière log-linéaire la survie, en termes de *hazard ratio*, ou rapports des risque instantanés. Le **modèle de Cox** répond à cette question. Nous vous conseillons plutôt de discrétiser la variable quantitative (par exemple en quartiles) pour retourner au cas précédent.

3.3 Deux variables appariées, dans plusieurs groupes

3.3.1 Préambule

Précédemment, nous avions déconseillé de procéder à des comparaisons avant-après dans un seul groupe. En revanche, les analyses de variables appariées **dans plusieurs groupes** restent tout à fait **valides**.

3.3.2 Deux variables binaires appariées, et une variable de groupe

Exemple : on s'intéresse à l'infection à Toxoplasma Gondii (le plus souvent asymptomatique, et qui guérit spontanément). On inclut des patients, qu'on randomise dans deux groupes. Après leur randomisation, on réalise une sérologie. Le premier groupe est soumis à un traitement antiinfectieux préventif et curatif. Six mois plus tard, on réalise une deuxième sérologie chez tous les patients.

On procède ainsi :

- Représenter les individus avec notamment :
  - Une colonne définissant le groupe : placebo ou traitement
  - Une colonne définissant le statut à l'inclusion : 0/1
  - Une colonne définissant le statut après 6 mois : 0/1
- Créer une nouvelle colonne de différence après – avant :
  - Elle vaut -1 pour les patients positifs qui deviennent négatifs
  - Elle vaut +1 pour les patients négatifs qui deviennent positifs
  - Elle vaut 0 pour les patients qui restent stables, positifs ou négatifs
- Hypothèse nulle H_0 : l'évolution est indépendante du groupe de traitement
- Réaliser un test du Khi² d'indépendance entre la colonne de groupe, et la nouvelle colonne, après éviction des effectifs des sujets stables, comme défini dans le chapitre 3.2.2.3 en page 48.

3.3.3 Deux variables quantitatives appariées, et une variable de groupe

Exemple : on s'intéresse à des patients diabétiques de type 2, asymptomatiques, et à leur mesure d'hémoglobine glyquée, HbA1c. 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 traitement qu'on souhaite évaluer. 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.

On procède ainsi :

- Représenter les individus avec notamment :
  - Une colonne définissant le groupe : placebo ou traitement
  - Une colonne définissant le dosage à l'inclusion
  - Une colonne définissant le dosage après 6 mois
- Créer une nouvelle colonne d'évolution après – avant
- Hypothèse nulle H_0 : la moyenne de l'évolution est indépendante du groupe de traitement
- Réaliser un test de Student, comme défini dans le chapitre 3.2.3.3 en page 51. S'il y avait plus de deux groupes, on réaliserait une ANOVA.

3.4 Cas particuliers d'analyses bivariées

Dans des cas particuliers d'analyses bivariées :

- Les tests statistiques définis précédemment sont valides mais peu intéressants
- **Il est déjà évident que les variables étudiées ne sont pas indépendantes**
- On cherche plutôt à **quantifier la force de leur association**

Méthodes spécifiques à chaque situation :

- **Facteurs de risque** ou protecteurs en épidémiologie (voir 3.4.1 page 58)
- **Tests diagnostiques** à réponse binaire (voir 3.4.2 page 60)
- **Outils de détection** d'un nombre indéterminé d'événements (voir 3.4.3 page 61)
- **Tests diagnostiques** à réponse quantitative (voir 3.4.4 page 61)
- **Accord entre deux juges** avec le coefficient **Kappa** (voir 3.4.5 page 63)

3.4.1 Facteurs de risque ou protecteurs en épidémiologie

Nous nous intéressons au lien entre :

- Une exposition, notée « E+ » (exposés) ou « E- » (non-exposés)
- Une maladie, notée « M » (malades) ou « NM » (non-malades)
- Question : le facteur est-il un **facteur de risque**, ou un **facteur protecteur** ?
- Traçons un tableau de contingence croisé :

	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)

Nous calculons :

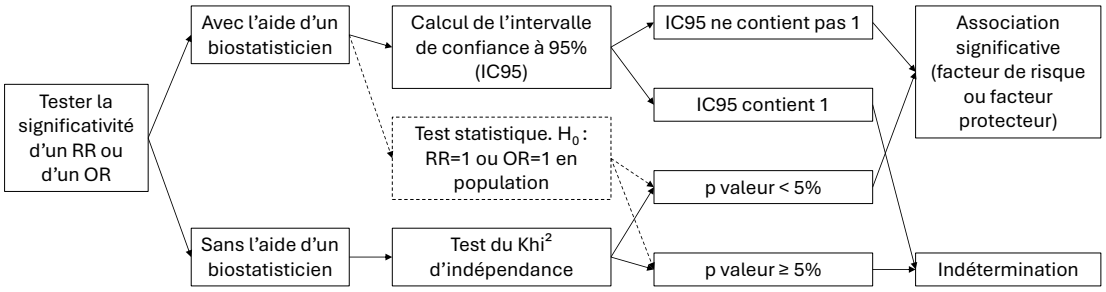
- Le **risque relatif (RR)** :
  - proportion de malades chez les exposés, divisée par la proportion de malades chez les non-exposés (cellule H2 ci-dessus)
  - calcul interdit dans les études de cas-témoin
- L'**odds ratio (OR)** :
  - produit des effectifs favorables à une relation de risque divisé par le produit des effectifs favorables à une relation protectrice (cellule H3 ci-dessus)
  - calcul toujours autorisé

L'odds ratio et le risque relatif sont du même ordre de grandeur dans les situations les plus fréquentes. Ils s'interprètent ainsi :

- $[0;1[$ facteur protecteur
- 1 facteur indépendant
- $]1;+\infty[$ facteur de risque

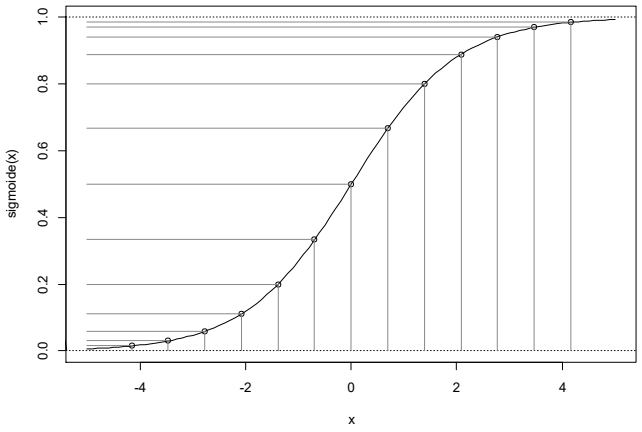
Attention, une valeur de 0,5 (le risque est divisé par 2) est aussi importante en termes de taille d'effet qu'une valeur de 2 (le risque est multiplié par 2).

L'arbre décisionnel suivant indique comment tester la significativité de cette association :



L'OR se comprend différemment en fonction de la prévalence :

- Lorsqu'une maladie est rare en population (0%-10%) :
  - **l'odds ratio s'interprète comme un risque relatif**
  - l'OR multiplie par autant la fréquence de la maladie, lorsqu'on passe du groupe non-exposé au groupe exposé
- Lorsque la maladie est fréquente (30%-70%) :
  - L'OR a un effet linéaire : il « ajoute » un certain nombre à la fréquence
- Lorsque la maladie est ubiquitaire (90%-100%) :
  - l'OR divise par autant la fréquence de l'état non-malade, lorsqu'on passe du groupe non-exposé au groupe exposé
- Le graphique ci-dessous illustre l'effet sur la prévalence d'une maladie (en ordonnée) qu'ont des odds ratios de 2 (décalages successifs vers la droite en abscisse).



3.4.2 Tests diagnostiques à réponse binaire

Evaluation classique d'un test diagnostique :

- Test diagnostique binaire (examen biologique, examen radiologique, test clinique, question posée au patient, etc.)
- Pathologie binaire, authentifiée par un **gold standard**, test réputé exact

On définit quatre sous-groupes d'individus (voir tableau suivant) :

- Vrais positifs (VP) : individus malades qui ont un test positif (à juste titre)
- Vrais négatifs (VN) : individus non-malades qui ont un test négatif (à juste titre)
- Faux positifs (FP) : individus non-malades qui, à tort, ont un test positif
- Faux négatifs (FN) : individus malades qui, à tort, ont un test négatif

	M	**NM**
T+	VP	FP
T-	FN	VN

Quatre quantités rapportent la proportion de bien-classés au sein d'un sous-groupe :

- La **sensibilité (Se)** : proportion de bien classés parmi les malades $Se = \frac{VP}{VP+FN}$
- La **spécificité (Sp)** : proportion de bien classés parmi les non-malades $Sp = \frac{VN}{VN+FP}$
- La **valeur prédictive positive (VPP)** : proportion de bien classés parmi les tests positifs $VPP = \frac{VP}{VP+FP}$
- La **valeur prédictive négative (VPN)** : proportion de bien classés parmi les tests négatifs $VPN = \frac{VN}{VN+FN}$

La **VPP et la VPN :**

- sont **les seules mesures utiles pour la pratique clinique**
- sont **influencées par la prévalence** de la maladie
- plus une maladie est rare, plus sa VPP diminue

Exemple : nous étudions la relation entre un test et une maladie. Nous traçons un tableau de contingence, et ajoutons les totaux des lignes et des colonnes. Nous calculons les quatre quantités d'intérêt (colonnes G, H et I ci-après).

			malades	non-malades				
			M	NM		Se :	0.950	=C3/C5
	tests positifs	T+	950	250	1200	Sp :	0.444	=D4/D5
	tests négatifs	T-	50	200	250	VPP :	0.792	=C3/E3
			1000	450	1450	VPN :	0.800	=D4/E4

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.

Exemple : On analyse un millier de courriers de sortie rédigés par des médecins. Chaque courrier contient des concepts médicaux en nombre variable (ex : « appendicite », « anesthésie locale », etc.). On demande à des humains d'identifier et coder ces concepts. On souhaite évaluer un logiciel d'IA qui réalise la même tâche.

			Gold standard					
			GS+	Gs-		Se = rappel =	0.950	=C15/C17
	Outil de	T+	950	250	1200	VPP = précision =	0.792	=C15/E15
	détection	T-	50	???				
			1000			F-score =	0.864	=MOYENNE.HARMONIQUE(H14;H15)

Dans le tableau de contingence :

- Le nombre de concepts détectés par les humains est GS+
- Le nombre de concepts détectés par le logiciel est T+
- Leur intersection donne les vrais positifs (VP) : on peut donc calculer Se et VPP
- Il est impossible d'estimer le nombre de vrais négatifs VN :
  - Faut-il prendre chaque mot du courrier ? Faut-il prendre chaque groupe nominal ? Faut-il plafonner ce nombre ?
  - **Soit le nombre de VN est inconnu, soit il est extrêmement élevé**
  - Donc, soit la Sp et la VPN sont incalculables, soit elles sont très élevées, ce qui ne permet pas d'évaluer l'outil de détection

On calcule alors plutôt :

- La **Se** :
  - Si un item existe, quelle est la probabilité qu'il soit détecté ?
  - Dans ce contexte, on l'appelle également **rappel** (*recall* en Anglais)
- La **VPP** :
  - Si l'outil détecte un item, quelle est la probabilité que cet item soit réel ?
  - Dans ce contexte, on l'appelle également **précision** (*precision* en Anglais)
- La **F-mesure**, ou **F-Score**, ou **mesure F1** :
  - Est une **moyenne harmonique de Se (rappel) et VPP (précision)** : c'est l'inverse de la moyenne des inverses
  - Se calcule avec la fonction **moyenne.harmonique()** d'Excel ou Calc

3.4.4 Tests diagnostiques à réponse quantitative

Situation typique :

- On souhaite prédire une variable binaire (par exemple, le statut malade/non-malade)
- Le test fournit une réponse quantitative
- On peut **binariser** cette réponse à l'aide d'un **seuil** : le choix du seuil résulte d'un compromis entre sensibilité et spécificité

La **courbe ROC** (*Receiver Operating Characteristic*) :

- Permet de visualiser l'impact du choix du seuil sur la performance du test

- Quantifie, globalement, la performance du test grâce à l'aire sous la courbe AUC (*Area under the Curve*)
- Peut aider à choisir le meilleur seuil

Exemple : On mesure le taux de Bêta-HCG chez 500 femmes enceintes, puis on observe la présence d'une trisomie 21. Ce taux est un nombre réel positif, qu'on peut discrétiser en utilisant notamment deux seuils :

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**

Dans cet exemple, avec deux seuils, on obtient deux tableaux de contingence :

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

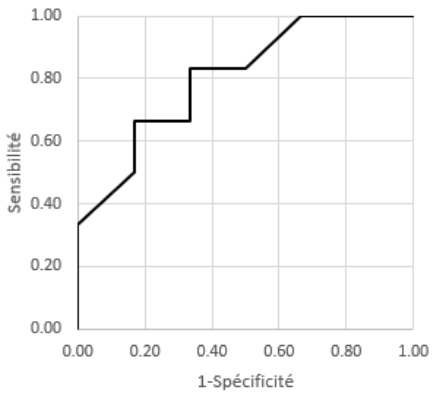
Voici comment tracer la courbe ROC avec un tableur :

- Réaliser plusieurs tableaux de contingence à différents seuils
- Reporter les couples sensibilité-spécificité obtenus dans un nouveau tableau
- Ajouter une colonne **1-Sp** (colonne AG ci-après)
- Tracer un **nuage de points** avec lignes droites et points masqués :
  Ordonnée = *Se* Abscisse = *1-Sp*
- Redimensionner le graphique pour qu'il ait une forme carrée

	AD	AE	AF	AG	AH	AI
	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

…

	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AI
20												
21		Seuil	Se	Sp	1-Sp	$(1-Se)^2+(1-Sp)^2$	Surface					
22		0	1.00	0.00	1.00	1.000	0					
23		8	1.00	0.00	1.00	1.000	0					
24		10	1.00	0.17	0.83	0.694	0					
25		14	1.00	0.33	0.67	0.444	0.069					
26		25	0.83	0.50	0.50	0.278	0					
27		**38**	0.83	0.67	0.33	**0.139**	0.111					
28		51	0.67	0.67	0.33	0.222	0					
29		**53**	0.67	0.83	0.17	**0.139**	0.139					
30		72	0.50	0.83	0.17	0.278	0.153					
31		81	0.33	1.00	0.00	0.444	0.167					
32		83	0.17	1.00	0.00	0.694	0.167					
33		90	0.00	1.00	0.00	1.000	XXXXX					
34							0.806					

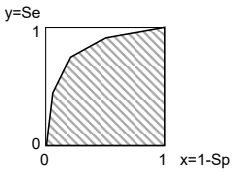


Pour identifier **le point le plus proche du point parfait** :

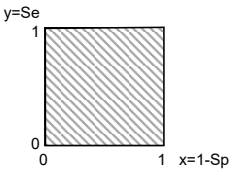
- Calculer pour chaque seuil la quantité $(1 - Se)^2 + (1 - Sp)^2$ (colonne AH)
- Sélectionner le seuil qui minimise cette quantité (seuils 38 et 53 dans l'exemple)
- Ce seuil est un bon compromis Se-Sp, pas forcément le meilleur

Pour calculer l'**aire sous la courbe ROC, notée AUC** (*area under the curve*) :

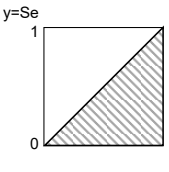
- Ajouter sur la première ligne la formule présentée en cellule AI22
- La prolonger vers le bas, hormis la dernière ligne
- L'aire sous la courbe est la somme de la colonne (ici 0,806, ou 80,6%)
- L'AUC s'interprète comme suit :



Courbe ROC habituelle, pour un bon test : $0,5 \leq AUC \leq 1$



Courbe ROC d'un test parfait : AUC=1



Courbe ROC d'un test aléatoire : AUC=0,5

3.4.5 Accord entre deux juges, sans gold standard

Situation typique :

- Deux variables binaires (ou qualitatives) représentent chacune un jugement visant à classer les individus dans des catégories (ex : malade et non-malade)
- **On ne sait pas quel est le véritable statut** des individus : pas de gold standard
- On souhaite seulement **quantifier la concordance** entre les deux jugements

Exemples classiques de cadres d'utilisation de la concordance :

- Concordance **entre deux experts**
- Concordance **entre deux méthodes**, deux appareils, deux tests, etc.
- Concordance entre deux utilisateurs du même appareil ou de la même méthode : **reproductibilité inter-opérateurs**
- Concordance entre deux itérations de la même méthode de mesure : **reproductibilité** intra-opérateur, **répétabilité** de la mesure

La **proportion d'accord** (ou proportion de concordance observée) :

- Proportion de cas où les deux juges sont d'accord (P_0, définie plus bas)
- Interprétable seulement lorsque les deux statuts sont bien équilibrés
- **Ne doit jamais être utilisée** : elle n'est pas un critère acceptable
- Solution : le **coefficient Kappa de Cohen**.

Voici comment calculer le **Coefficient Kappa** avec un tableur :

- Tracer un tableau de contingence croisant les deux jugements (colonnes ABCD)
- Créer un deuxième tableau de même structure (colonnes FGHI), présentant des effectifs théoriques qu'on obtiendrait si les deux jugements étaient indépendants (reporter les totaux, calculer deux effectifs sans les arrondir)
- Calculer la proportion de concordance observée P_0 (cellule L3)
- Calculer la proportion de concordance théorique P_c (cellule L4)
- Calculer le coefficient Kappa $K = (P_o - P_c)/(1 - P_c)$ (cellule L5)

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	Effecti					Effect						
3		A+	A-	total			A+	A-	total		Proportion de concordance observée P0=	=(B4+C5)/D6
4	B+	45	15	60		B+	=$I4*G$6/$I$6	...	=D4		Proportion de concordance aléatoire Pc=	=(G4+H5)/I6
5	B-	5	35	40		B-	...	=$I5*H$6/$I$6	=D5		Kappa=(Po-Pc)/(1-Pc)=	=(L3-L4)/(1-L4)
6	total	50	50	100		total	=B6	=C6	=D6			

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	Effectifs observés :					Effectifs aléatoires :						
3		A+	A-	total			A+	A-	total		Proportion de concordance observée P0=	0.8
4	B+	45	15	60		B+	30.00	...	60		Proportion de concordance aléatoire Pc=	0.5
5	B-	5	35	40		B-	...	20.00	40		Kappa=(Po-Pc)/(1-Pc)=	0.6
6	total	50	50	100		total	50	50	100			

Interprétation du coefficient Kappa :

- C'est une sorte de **proportion d'accord entre les deux juges**, après retrait du hasard
- En cas de valeur négative, l'interpréter comme $K = 0$
- L'intervalle d'interprétation est [0 ; 1]. 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 $0{,}0 \leq K \leq 0{,}2$

Le coefficient Kappa est **doublement symétrique** : on peut intervertir les juges, mais on peut également inverser les réponses pour chaque jugement.

4 Analyses statistiques multivariées, en bref

Analyses multivariées non-supervisées : étudient les associations entre variables, les groupes d'individus, sont souvent des méthodes exploratoires.

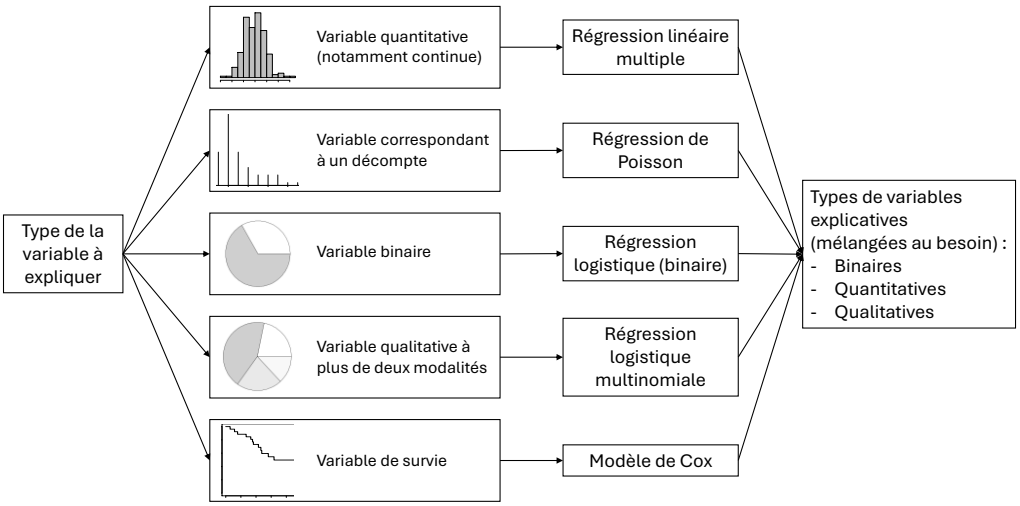
Analyses multivariées supervisées :

- Expliquent **une variable** (dite "**à expliquer**"), par l'ensemble des autres variables
- Permettent ensuite de **prédire cette variable**, en utilisant les autres variables
- Souvent utilisées pour expliquer une variable d'intérêt par une exposition d'intérêt, ajustée sur d'autres facteurs (l'âge, le sexe, les pathologies chroniques, etc.) : l'exposition d'intérêt et les facteurs d'ajustement sont alors mélangés
- Notamment les **régressions**, très utilisées en santé en **design observationnel**[2]

Parmi les analyses multivariées supervisées, les **méthodes de régressions** :

- De loin les plus utilisées en recherche en santé
- Choix de la méthode : selon le **type de la variable à expliquer** (cf. post)
- Complexes à réaliser, valider et interpréter : faire appel à un biostatisticien

[2] Dans les essais randomisés contrôlés, l'affectation aléatoire des individus est censée gommer les différences entre individus. L'analyse répondant à l'objectif principal est généralement une simple analyse bivariée (Khi² ou Student).



5 Réflexions sur certains tests statistiques ou leur paramétrage

5.1 Tests de comparaison à une norme

Tests de comparaison à une norme :

- Comparaison d'une proportion observée à une proportion attendue : test binomial et test du Khi² d'adéquation
- Comparaison d'une moyenne observée à une moyenne attendue : test de Student observé-attendu
- Théoriquement valides, mais…
- Généralement utilisés dans des **recherches de piètre qualité méthodologique**

Exemple : Un médicament homéopathique miraculeux, le Phallus Elongator 22CH (PE22CH), est vendu pour allonger le pénis. Nous mesurons le pénis de 50 patients qui ont pris du PE22CH. En juin, vers 14h, les participants sont installés dans une cabine et mesurent eux-mêmes leur membre. La taille moyenne obtenue est significativement supérieure à la taille moyenne publiée dans une autre étude (p<0,001). Dans cette autre étude, le pénis des participants avait été mesuré en février à 8h00 par un infirmier du Service de Santé des Armées.

La différence est essentiellement imputable aux conditions de mesure. Entre deux études différentes, les conditions de mesure sont toujours différentes. C'est le principal écueil de ce type d'étude.

5.2 Tests appariés dans un seul groupe, avant-après

Tests de comparaison avant-après :

- Comparaison d'une proportion avant-après : test de McNemar
- Comparaison d'une moyenne avant-après : test de Student apparié
- Théoriquement valides
- Souvent utilisés dans des **recherches de piètre qualité méthodologique**

Exemple : Le Phallus Elongator 22CH (PE22CH), médicament homéopathique miraculeux, est vendu pour allonger le pénis. Nous incluons 100 patients. Nous mesurons leur pénis le matin. Puis nous leur administrons le PE22CH. Nous leur servons un repas puis évaluons leur longueur pénienne en début d'après-midi, selon un protocole identique. La taille mesurée est en moyenne plus élevée (p<0,001).

De nombreux facteurs ont changé entre les deux mesures : l'heure, le stress des participants et le contact avec les investigateurs. Ces facteurs suffisent à expliquer l'amélioration de la mesure.

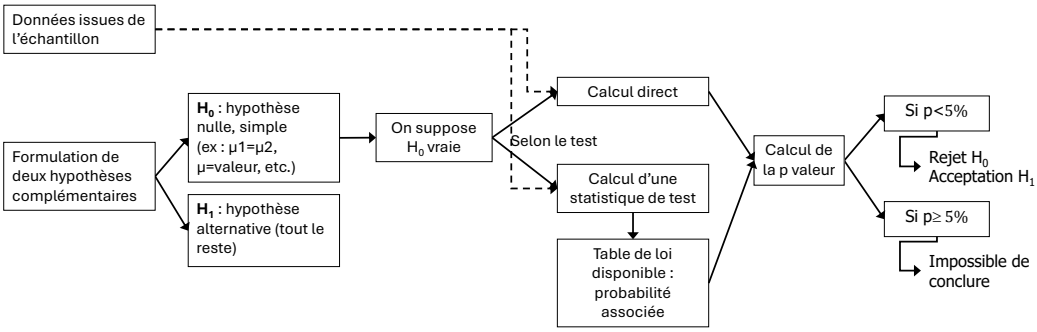
Le problème dans les **études avant-après dans un seul groupe**, est qu'il existe :

- un **effet du temps qui passe** (guérison spontanée)
- un **effet placebo** et un **effet « cocooning »**
- un **effet lié à la répétition** : **banalisation** des mesures, **gain d'expérience**

En revanche, ces mêmes études réalisées dans deux groupes, un exposé et un non-exposé, peuvent permettre d'évaluer l'effet de l'exposition (comparaison de 4 moyennes ou 4 proportions appariées dans deux groupes).

5.3 Tests qu'on réalise en espérant ne pas rejeter H_0

L'ensemble des tests que nous avons vus étaient exécutés en espérant rejeter l'hypothèse nulle H_0. Lorsque la p valeur excède 5%, on est dans une indétermination : **on ne peut jamais prouver que H_0 est vraie**.



Pourtant, certains tests sont souvent réalisés pour prouver que H_0 est vraie :

- **Kolmogorov Smirnov** et **Shapiro-Wilk**, pour montrer qu'une distribution observée dans un échantillon suit une distribution normale (par exemple)
- **Fisher-Snedecor**, pour montrer que deux variances sont similaires
- **Hosmer-Lemeshow**, pour montrer qu'une régression logistique est bien calibrée

Ces utilisations sont **incorrectes**, car on ne peut jamais prouver que H_0 est vraie.

5.4 Test paramétrique, non-paramétrique, asymptotique, exact

Test paramétrique ou non-paramétrique ?

- **Test paramétrique** : s'appuie sur l'estimation d'un paramètre (moyenne, proportion)
- **Test non-paramétrique** : s'intéresse directement aux effectifs ou aux rangs
- Notion peu importante, mais souvent confondue avec la suivante

Test exact ou asymptotique ?

- **Test exact** : la p valeur est calculée de manière exacte, soit durant la procédure, soit en utilisant une table déjà calculée
- **Test asymptotique** : la p valeur peut être approchée lorsque l'échantillon est suffisamment grand (ex : n≥30) ou que la variable étudiée suit une loi spécifique.

Voici comment on peut classer les tests vus 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
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

5.5 Test unilatéral ou bilatéral ? Et pourquoi 5% ?

Calcul de la p valeur :

- On suppose que H_0 est vraie
- On additionne les probabilités de l'observation et des observations moins attendues :
  - « du même côté » si le test est unilatéral
  - « des deux côtés » si le test est bilatéral (voir 2.2.3.2 page 36)
- Donc dans les tests unilatéraux :
  - Les **p valeurs** sont **plus faibles**
  - On **rejette plus facilement** H_0

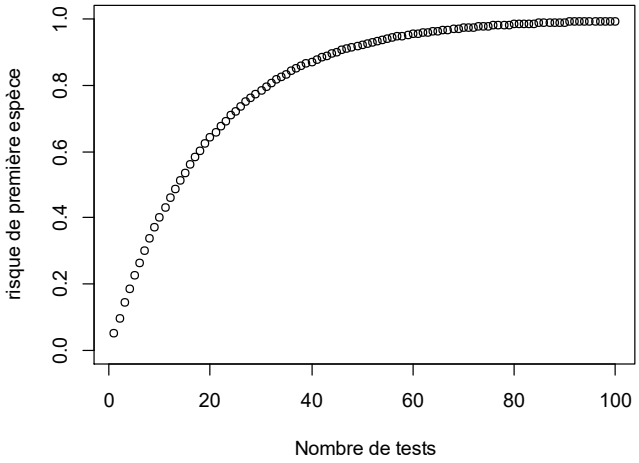
Les bonnes pratiques (ex : exigences de la FDA et l'EMA) :

- Recommandé : réaliser des **tests bilatéraux au seuil de 5%**
- Possible (mais sans intérêt) : réaliser des tests unilatéraux au seuil de 2,5%
- Interdit : réaliser des tests unilatéraux au seuil de 5%
  (hormis les tests de non-infériorité : cadre très particulier)

5.6 Correction de Bonferroni

Position du problème :

- Soit un jeu de données **aléatoires** : il ne comporte aucune association statistique
- On réalise k tests statistiques d'indépendance entre les différentes variables
- Pour chaque test :
  - H_0 est vraie par construction : ces variables sont effectivement indépendantes
  - Donc la probabilité de rejeter H_0 (sachant que H_0 est vraie) est de 5%
- Pour l'ensemble des k tests :
  - La probabilité qu'au moins un des tests soit significatifs est : $1 - 0{,}95^k$
  - Il en résulte une inflation du risque de première espèce :



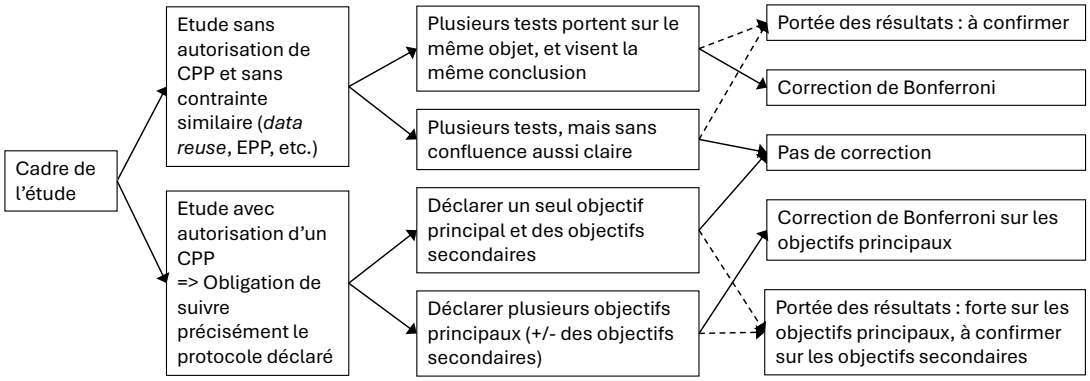
Pour préserver un risque total de 5% :

- On peut appliquer la correction de Šidák :
  réaliser chaque test statistique au seuil $1 - 0{,}95^{1/k}$
- Ce seuil est très proche de $0{,}05/k$ tout simplement ! Bonferroni l'a démontré.

La correction de Bonferroni doit être utilisée chaque fois qu'on réalise plusieurs tests, en espérant qu'au moins un des tests soit significatif : les tests se rapportent **au même objet** et peuvent, en cas de positivité, amener **la même conclusion**.

Exemple (mauvais) : Pour montrer la supériorité d'un somnifère A sur un somnifère B, on les administre à deux groupes de patients. On compare plusieurs caractéristiques : délai d'endormissement, durée de sommeil profond, délai avant le premier réveil, délai jusqu'au lever, nombre de cauchemars, etc. Si A est meilleur que B sur au moins une de ces caractéristiques, on conclura que A est meilleur que B, globalement.

Voici la conduite à tenir :



6 Interpréter une association statistique en général

6.1 Discuter la significativité statistique

Exemple des tests bivariés en général : lorsque la p-valeur est **inférieure à 5%**,

- On rejette l'hypothèse nulle : les deux variables **ne sont pas indépendantes**
- Cela ne confirme ni l'existence, ni l'importance de la taille d'effet
- La p valeur rend compte du **niveau de confiance qu'on a lorsqu'on rejette H_0**
- La **taille d'effet** est évaluée par des indicateurs plus simples (ex : différence de deux moyennes, risque relatif, etc.), +/- assortis d'un IC95

Une **p valeur inférieure à 5%** :

- A une portée décisionnelle très forte dans une étude **déclarée en CPP** et sur l'**objectif principal** uniquement :
  - pour de nombreuses raisons méthodologiques : protocole écrit en aveugle des données, nombre de sujets limité par le calcul du NSN, etc.
  - même si la p valeur vaut à peine 4,9%
- En-dehors de l'objectif principal : la portée décisionnelle est moindre
- Dans les autres études : la portée décisionnelle est bien moindre

Une **p valeur supérieure à 5%** ne permet jamais de conclure : on ne sait pas si l'association n'existe pas, si l'échantillon est trop petit, si le test est mal spécifié, etc.

6.2 De la significativité statistique à la causalité et à l'explication

Pour prouver une causalité :

- Elle n'est jamais prouvable par une méthode statistique seulement
- Elle peut être prouvée selon la **méthodologie de l'étude** :
  - Uniquement dans l'objectif principal d'un **essai randomisé contrôlé contre placebo ou traitement de référence, et en double aveugle** (ou triple)
  - Car l'exposition est modifiée par l'expérimentation et les patients sont affectés aléatoirement aux groupes => cela neutralise les facteurs de confusion

Dans les **études observationnelles** :

- Il existe un biais majeur : le **biais d'indication**
  - Les patients qui ont reçu le traitement à évaluer sont différents des autres
  - La différence observée résulte probablement de cette différence initiale
  - Les **scores de propension** peuvent atténuer le biais d'indication, mais on ne peut jamais savoir s'ils l'annihilent : les termes « **inférence causale** » et « **émulations d'essais cliniques** » sont excessifs
- On peut cependant approcher la causalité par un faisceau de preuves : association statistique + confirmation par la littérature + arguments physiopathologiques + relation dose-effet, etc.

Une fois la causalité acquise, **il n'existe aucune méthode pour affirmer avec certitude une explication**. L'explication nécessite à la fois une causalité, et de nombreux éléments de connaissance. Elle relève donc de la discussion et non du résultat scientifique. Elle peut évoluer dans le temps. Mais au fond, l'explication n'est pas toujours primordiale :

Exemple : Alertés par l'incidence des morts subites du nourrisson, certains ont déduit que, couché sur le dos, en cas de régurgitation, le nouveau-né inhalait ses vomissements. S'en sont suivies des recommandations généralisées, appuyées par des campagnes d'information dans les années 1960 : il fallait coucher les nouveau-nés sur le ventre.
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. Cette simple association statistique, sans explication, a permis de promouvoir exactement l'inverse. En France, entre 1991 et 1997, il a été possible de diviser par quatre la mortalité !

Dans cet exemple :

- Le raisonnement qui cherchait une explication a tué des milliers d'humains
- Le raisonnement qui ignorait l'explication a sauvé des milliers de vies dans le monde
- L'explication n'avait, au fond, aucune importance

6.3 Principaux biais en épidémiologie et en recherche clinique

Dans toute étude, il faut rechercher des biais pour s'assurer que le résultat peut être interprété correctement et généralisé au-delà de l'échantillon observé, en population.

6.3.1 Définitions : erreur, biais, biais différentiel, biais conservateur

L'**erreur** :

- est un processus aléatoire qui altère la mesure
- variables quantitatives : cette erreur est le fait d'un **bruit**, qui suit une **loi normale**
- peut être liée à l'appareil de mesure, à la précision de la mesure par l'opérateur, etc.
- est **équilibrée**, **imprévisible**, **indépendante** des facteurs connus
- est **parfaitement gérée** par les méthodes statistiques
- dans le cas d'un échantillon, une erreur supplémentaire est liée à l'échantillonnage (sélection aléatoire d'un échantillon relativement petit)

Un **biais** :

- est une erreur **déséquilibrée**, ou qu'on peut en partie prédire
- n'est **pas géré** par les méthodes statistiques

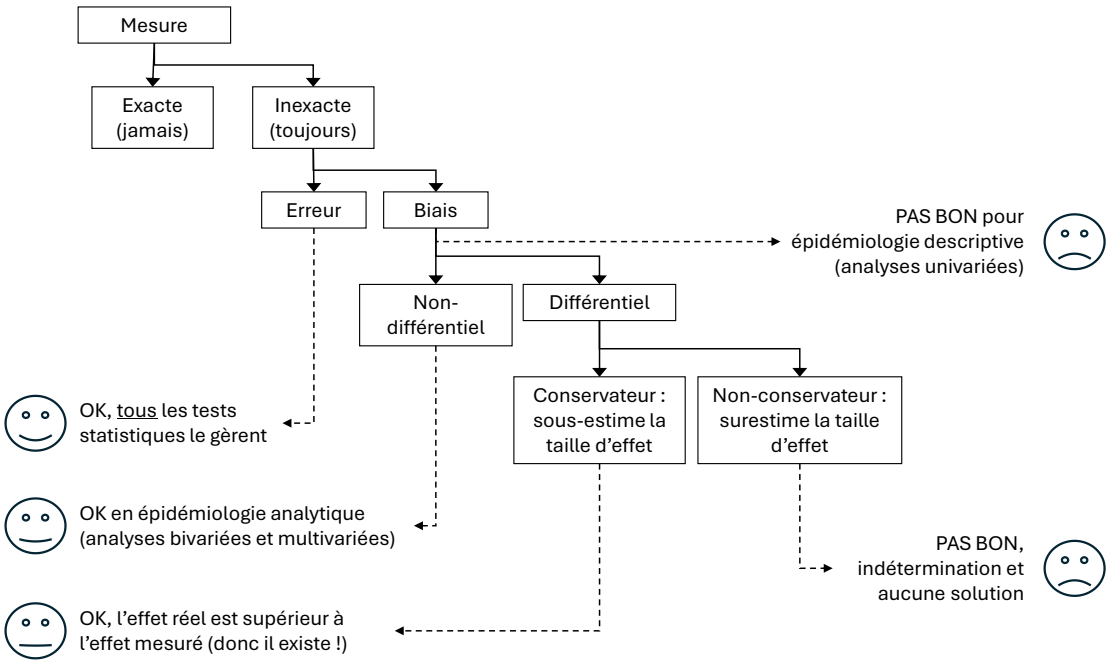
Exemple : on s'intéresse au poids moyen des humains adultes, et on suppose qu'en population il vaille 71kg. Dans des échantillons de 30 individus, on mesure 71,05kg, puis 68,64kg, etc. Le mode de recrutement des participants exclut de fait les personnes obèses, et les personnes qui utilisent tout le temps la voiture. La moyenne des moyennes, sur un très grand nombre, est plus faible que 71kg. Ces moyennes sont biaisées.

En **épidémiologie descriptive**, l'existence d'un biais empêche d'estimer les quantités qui nous intéressent (moyennes, prévalences, incidences, etc.).

En **épidémiologie analytique** (recherche de facteurs de risque ou protecteurs), il est possible qu'un biais ne pose pas problème, s'il est **non-différentiel**, c'est-à-dire s'il ne perturbe pas l'estimation des quantités qui nous intéressent (RR ou OR).

Exemple : 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 en 2020. Le tabagisme est largement sous-codé (5% des séjours), mais indépendamment de la maladie d'Alzheimer. L'odds ratio est donc conservé, ce biais est non-différentiel.

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 (ex : il rapproche le RR ou l'OR de la valeur 1). Dans ce cas, il peut être qualifié de **biais différentiel conservateur** : c'est plutôt une bonne nouvelle. Si on observe un effet, alors l'effet réel est plus fort que l'effet observé. Cette distinction, conservateur ou non, relève de l'interprétation fine du cas étudié, en intégrant des connaissances externes.



6.3.2 Biais de sélection

Ils **résultent d'une anomalie de sélection** des participants. L'échantillon n'est plus aléatoirement issu de la population d'intérêt. Exemples :

- **Biais de recrutement** : on recrute des patients différents de la population visée
- **Biais de non-réponse** : les personnes qui répondent au questionnaire sont différentes de celles qui ne répondent pas, et le taux de réponse est faible
- **Biais de censure informative** : les patients perdus de vue sont différents des patients qui sont correctement suivis dans l'étude

6.3.3 Biais d'information

Ils **résultent d'une anomalie dans le recueil d'information** sur les participants. Exemples :

- **Biais de classement** : l'erreur concerne le groupe dans lequel le patient est affecté
- **Biais de désirabilité sociale** : le sondé donne une réponse fausse car la vraie réponse lui fait honte (ex : opinions politiques, mode de vie, sexualité, etc.)
- **Biais de mémorisation** : le participant se souvient mal de la réponse à donner
- **Biais de codage** : en réutilisation de données, certains soins et pathologies sont sous-codées (ex : tabagisme, ECG) ou sur-codées
- **Effet nocebo, effet placebo, effet cocooning** : la simple inclusion du patient dans une étude induit des effets positifs ou négatifs
- **Biais liés à l'enquêteur** : l'enquêteur peut être influencé s'il sait quel traitement reçoit le patient (d'où les méthodologies en double aveugle)
- **Biais d'indication des examens complémentaires** : en vie réelle, les examens complémentaires sont réalisés lorsque leur probabilité d'être anormaux augmente

6.3.4 Biais de confusion

Les facteurs de confusion n'altèrent pas les estimations faites dans l'étude, mais plutôt leur interprétation. Exemples :

- **Biais protopathique** : une maladie silencieuse provoque des symptômes, et donc la prise d'un traitement, finalement la maladie émerge, et est diagnostiquée. Le traitement symptomatique devient un facteur de risque (statistique) de la maladie.
- **Biais d'immortalité** : 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, faisant croire que cet événement les a rendus immortels avant sa survenue. Ce biais existe dans les études rétrospectives.
- **Biais d'indication** : ce biais très important survient dans la quasi-totalité des études observationnelles comparatives. Nous lui dédions la partie suivante.

6.3.5 Un des biais de confusion, le biais d'indication

Le **biais d'indication** :

- Toujours présent dans les **études observationnelles comparatives**
- **Impact majeur**, limitant l'interprétation de la comparaison entre deux interventions
- Patients exposés ou non à l'intervention à évaluer :
  - Choix jamais aléatoire : choix fait par les soignants car c'est la meilleure option
  - Choix qui reflète, en fait, de profondes différences entre les patients (maladie, comorbidités, autres traitements, etc.)
- Conséquence : la différence observée pourrait plus dépendre de ces différences que de la supériorité du traitement à évaluer

Méthodes traditionnelles pour atténuer le biais d'indication :

- Elles tentent de prendre en compte les variables qui pourraient motiver le choix du traitement (âge, sexe, caractéristiques de la maladie)
- **Apparier** les patients : pour chaque patient d'un groupe, imposer 1 (ou 2 ou 3) patient similaire dans l'autre groupe
- **Stratifier** les analyses : comparer au sein de sous-groupes homogènes (ex : une analyse chez les hommes de 20 à 50 ans, etc.)
- **Ajuster** les analyses : inclure ces variables aux côtés de la variable de groupe, dans des analyses multivariées

On peut également utiliser des **scores de propension**, visant une prétendue « **inférence causale** » ou « **émulation d'essai clinique** ». On pourra là aussi **apparier** les patients sur ce score, ou **ajuster** sur ce score.

Ces méthodes ne peuvent s'appliquer **que sur les données disponibles**, qui rendent insuffisamment compte des critères de choix du traitement. Elles atténuent le biais d'indication, sans le faire disparaître.

6.4 Analyses de sensibilité

Analyse de sensibilité :

- analyse supplémentaire réalisée en modifiant le protocole, pour voir quel impact cela a sur les résultats (ex : exclure certains patients, modifier un critère de jugement, etc.)
- aucune bonne pratique n'indique quelles analyses réaliser
- aucune règle n'indique comment interpréter les résultats, à quel seuil
- relève de la pure expertise du chercheur

Rédiger et présenter le document

Les conseils présentés ici sont autant valables pour la rédaction de votre mémoire académique que pour un article scientifique. Nous évoquerons :

- Microsoft Word et LibreOffice Writer, logiciels de traitement de texte (partie 1 en page 73)
- Zotero, logiciel de gestion de la bibliographie (partie 2 en page 75)
- La rédaction scientifique (partie 3 en page 79)
- L'impression du document (partie 4 en page 86)
- La préparation du diaporama et la soutenance orale (partie 5 en page 87)

1 Utiliser un traitement de texte de manière appropriée

Retrouvez sur http://www.objectifthese.org des vidéos explicatives, ainsi qu'un modèle de document déjà prêt à l'emploi, qui vous fera gagner beaucoup de temps.

Nous verrons comment utiliser **Microsoft Word** et **LibreOffice Writer** (ou OpenOffice Writer) de manière appropriée pour réaliser un mémoire académique. Nous supposons que vous savez déjà réaliser les opérations de base, et présenterons les outils indispensables pour monter en gamme et gagner du temps.

1.1 Généralités sur les styles

Vous ne devrez plus faire de formatage local, mais utiliser les styles :

- Application d'un style à un paragraphe => mise en forme automatique et homogène
- Modification de la définition du style => modification de tout le texte concerné
- Gain de temps, amélioration de la qualité de mise en forme, homogénéité
- Activation de certaines fonctionnalités, par exemple pour les styles « Titre X »
- Style par défaut : style « **Normal** »

1.2 Le cas particulier des styles « Titre X »

Les styles nommés « Titre », « Titre 1 », « Titre 2 » :

- Existent déjà dans le modèle de document par défaut
- Améliorés dans le modèle de document proposé sur le site http://objectifthese.org
- Permettent en particulier (voyez la vidéo) :
  - **Numérotation hiérarchique** automatique cohérente dans tout le document
  - Mise en place d'un **sommaire** dynamique
  - Utilisation du **mode plan** pour réorganiser tout le document
  - Utilisation du **volet de navigation**, sur la gauche
  - **Renvois automatiques** et dynamiques vers des titres dans le document

1.3 Afficher les caractères non-imprimables

Vous devez afficher les caractères non-imprimables à l'aide du bouton ci-contre. Ces caractères sont :

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]

En particulier :

- **Espace insécable** (aspect de bulle) : espace qui n'est pas coupé par le passage à la ligne. Exemple : devant les ponctuations doubles, séparateur de milliers, etc.
- **Retour de chariot** : passe à la ligne sans changer de paragraphe. Exemple : très utile dans les légendes, les listes à puce, etc. Faites-le précéder d'une tabulation si vous souhaitez éviter les problèmes liés à l'alignement justifié.
- **Saut de page** : permet de passer à la page suivante.

1.4 Afficher les champs dynamiques sur trame grise

Ceci est indispensable pour distinguer le contenu dynamique : numérotation des pages, des figures et tables, renvois, table des matières, insertions Zotero, etc.

Dans Microsoft Word : *Fichier > options > avancées > Champs avec trame : toujours*

Dans LibreOffice Writer : *Affichage > trame de fonds de champ (ctrl+F8)*

Pour demander une **mise à jour des champs**, sélectionnez l'élément (ou tout le document avec [contrôle]+[a]), puis appuyez sur la touche F9, ou utilisez un clic droit.

1.5 Figures et légendes

Lorsqu'on insère une figure dans votre document, il faut :

- **L'aligner sur le texte**, dans un paragraphe unique, sans encadrement :
  - 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*
- Lorsque cette option est disponible, coller l'image **en tant que métafichier**, ce qui permet d'éviter la pixélisation, et de redimensionner a volo l'objet :
  - Dans Microsoft Word : *Coller > collage spécial > Image (métafichier)*
  - Dans LibreOffice : *Edition > collage spécial > collage spécial > métafichier*
- **Générer une légende** avec numérotation automatique. Pour ce faire,
  - dans Microsoft Word ou dans LibreOffice : *Clic droit sur l'image > Insérer une légende*
  - ensuite choisir le type d'élément (Figure/Tableau) et compléter la légende (utiliser le **retour de chariot** ctrl+Entrée si besoin). NB : en dépit des apparences, cette légende sera indépendante de l'image
- **Citer** chaque figure dans le texte, à l'aide d'un renvoi :
  - Dans Microsoft Word ou dans LibreOffice : *Menu insertion > renvoi*
    Puis choisir le type d'élément (Figure/Tableau)
  - Le plus souvent, on affiche le mot clef (« Figure ») et le numéro :
    - Dans Microsoft Word : *texte et numéro uniquement*
    - Dans LibreOffice : *catégorie et numéro*
- En fin de document, insérer une **table des illustrations dynamique** (déjà présente dans le modèle de document http://objectifthese.org) :

- 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")*

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. Tout comme pour les figures, vous pourrez insérer des légendes, des renvois, et une table des illustrations (voir précédemment).

1.7 Rappels sur la ponctuation et les espacements

En Français, il faut suivre les règles de ponctuation suivantes.

- Un espace après les ponctuations simples (. , …), rien avant. Exemple :
  Voici,·et·voilà.·Encore·voilà.¶
- Un espace insécable avant, et un espace après les ponctuations doubles (; : ! ?). Exemple :
  Je·demande·simplement°:·Venez-vous°?·Non°!¶
- Un espace à l'extérieur des marques d'encadrement ([] () -- "" ‘’). Exemple :
  Je·"devine"·-difficilement-·deux·couleurs·(rouge,·jaune).¶
- Un espace à l'intérieur et à l'extérieur des guillemets français. Exemple :
  Je·«°devine°»·une·certaine·frustration…¶

1.8 Typographie des nombres dans le texte

En Français, le séparateur décimal est la virgule. Le séparateur des milliers est facultatif, vous pouvez utiliser l'espace insécable mais évitez le point.

Ex : Cette maison coûte 1 500 000,01€.

Arrondissez les nombres **à 1 ou 2 chiffres** après la virgule, en choisissant une unité intuitive. Maintenez les zéros qui rendent compte de la précision :

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% »

Pour les **p valeurs**, conservez 2 chiffres significatifs pour les p valeurs supérieures à 1%, et 1 chiffre significatif pour les p valeurs inférieures à 1%. Exemples :

p valeurs : 0,82 0,045 0,002 2E-5

2 Installer et utiliser Zotero, logiciel de bibliographie

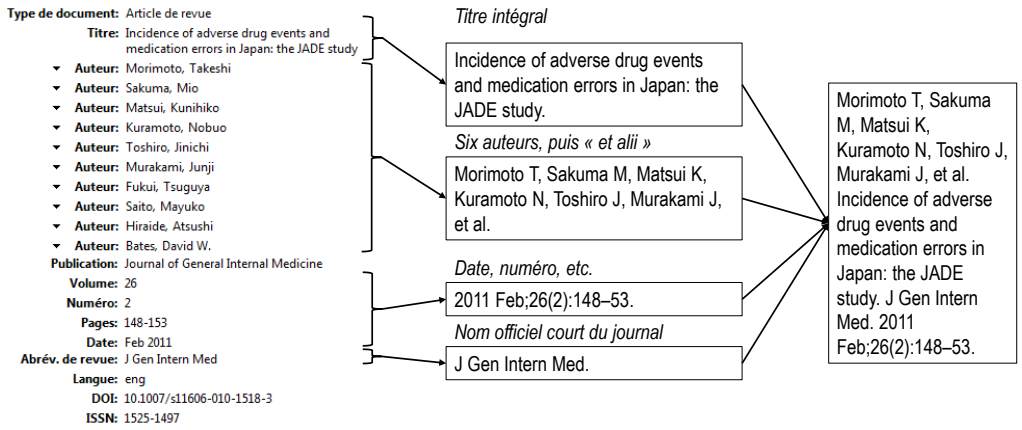
2.1 Difficultés liées à l'affichage de la bibliographie

Un mémoire académique doit s'appuyer sur des **références** bibliographiques, listées en fin de document. Elles sont rappelées par les **citations** dans le texte.

Introduction
Chaque année, les effets indésirables médicamenteux seraient responsables de 98 000 décès aux USA [1]. En milieu hospitalier, il est possible de les prévenir à l'aide de règles d'alerte. Ces règles sont néanmoins écrites par des experts, en s'appuyant sur une connaissance académique elle-même fondée sur des déclarations insuffisantes [2,3]. L'objectif de ce travail est de produire de telles règles par data mining [4].
Matériel, Méthodes, Résultats, Discussion
(...)

Références
1 Kohn LT, Corrigan JM, Donaldson MS. To err is Human. Washington DC: National Academy Press ; 1999
2 Morimoto T, Gandhi TK, Seger AC, Hsieh TC, Bates DW. Adverse drug events and medication errors: detection and classification methods. Qual Saf Health Care 2004.
3 Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. Journal of Biomedical Informatics 2003.
4 Adriaans P, Zantige. Data mining. Edingburgh: Addison Wesley ; 1996.

La présentation des références est propre à chaque journal :



La **norme Vancouver** est préférée en santé :

- les citations sont numérotées par ordre d'apparition
- les références sont ordonnées de même
- les citations doivent être condensées :
  [1][2] devient [1, 2] *[1][2][3] devient [1-3]* *[1][2][3][8] devient [1-3, 8]*
- dès que les citations changent, il faut tout corriger, ce qui peut être fastidieux

Ces difficultés sont ingérables à la main, mais aisément surmontées par les utilisateurs de Zotero, logiciel gratuit et open source de gestion de la bibliographie.

2.2 Installer Zotero

Zotero (pour Windows, Linux et Mac OS) :

- permet de stocker et citer des références bibliographiques
- permet de synchroniser ces références entre différents appareils
- s'interface avec les navigateurs web (pour enregistrer rapidement des références) et 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.

2.3 Créer un compte (facultatif et gratuit)

La création d'un compte sur le site https://www.zotero.org/user/register est facultative et gratuite. Nous vous le conseillons vivement dans l'une des situations suivantes :

- Vous souhaitez une sauvegarde automatique dans le cloud
- Vous travaillez sur plusieurs postes alternativement
- Vous travaillez à plusieurs simultanément sur une bibliographie
- Vous souhaitez créer une bibliothèque publique

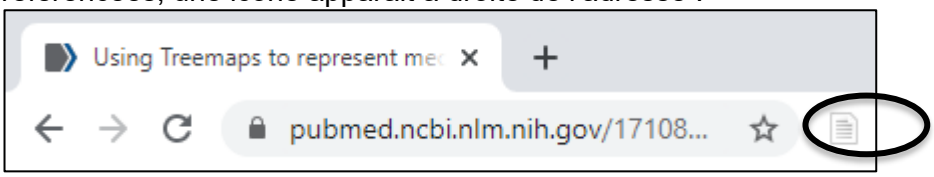
Créez tout d'abord votre compte sur la page https://www.zotero.org/user/register. Définissez un mot de passe spécifique à Zotero. Puis lancez Zotero, et cliquez sur le menu *Edition* > *Paramètres* > *Synchronisation*. Saisissez alors le login et le mot de passe définis lors de l'enregistrement.

Pour **partager votre bibliographie** avec vos collègues ou encadrants, cliquez sur le menu : *Fichier* > *Nouvelle bibliothèque* > *Nouveau groupe…* Vous pourrez alors paramétrer cette bibliothèque partagée et inviter d'autres personnes à y contribuer.

2.4 Utiliser Zotero pour créer et maintenir votre bibliothèque

Pour **enregistrer automatiquement un document** dans Zotero :

- Lancez tout d'abord Zotero
- Positionnez-vous dans le dossier dans lequel vous souhaitez insérer le document
- Naviguez sur internet avec votre navigateur : lorsque des pages visitées sont prêtes à être référencées, une icône apparait à droite de l'adresse :



- Cliquez sur cette icône : le document est alors automatiquement inséré dans Zotero, dans le dossier courant

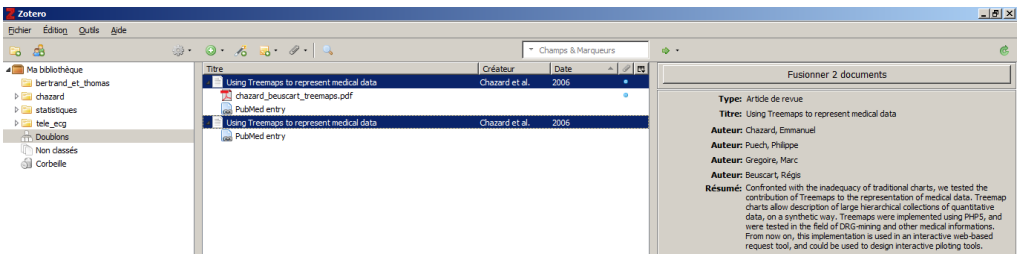
Pour **enrichir un document** déjà enregistré :

- Dans l'interface de Zotero, sélectionnez l'élément
- Vous pourrez éditer et corriger si besoin les métadonnées
- Vous pourrez ajouter une note (par exemple vos commentaires personnels)
- Vous pourrez ajouter un PDF : cliquez sur le bouton ci-contre puis « joindre un fichier »

- Ces modifications se synchronisent automatiquement et sans effort avec votre compte Zotero si vous en avez créé un, ou si vous le créez plus tard

Pour **gérer les doublons** :

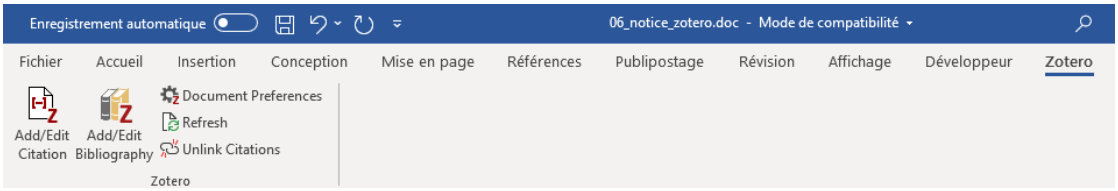
- Détectez les doublons : cliquez sur la collection « doublons », ils s'affichent automatiquement. Ne supprimez pas les doublons !
- Fusionnez les doublons : sélectionnez les articles et demandez leur fusion
- Lors de la prochaine synchronisation, la correction sera faite automatiquement dans les documents qui utilisent déjà ces références



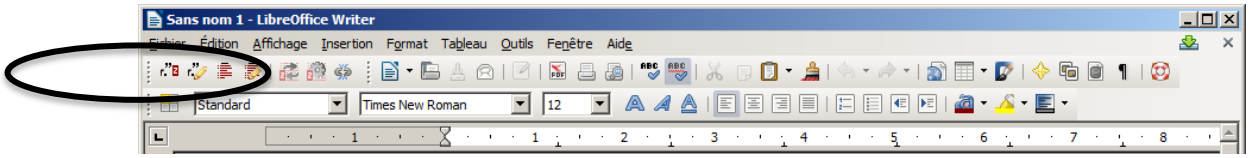
2.5 Citer les références dans un traitement de texte

Intégration de Zotero dans le traitement de texte :

Dans Microsoft Word, un ensemble de boutons devient désormais accessible dans le ruban, sous l'intitulé « Zotero » :



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 ».



Ces boutons correspondent aux actions suivantes :


Insérer ou éditer une citation dans le texte


Préférences de Zotero pour ce document texte


Insérer ou éditer une bibliographie (en fin de document)


Synchroniser avec Zotero (forcer la mise à jour)


Remplacer tous les liens dynamiques par du texte statique (utile sur une copie d'un document)

Insérer ou modifier une citation dans le texte :

- positionnez-vous là où vous souhaitez citer la référence
- cliquez sur le bouton ci-contre. La première fois, Zotero vous demande de choisir un style de citation, choisissez « Vancouver »
- recherchez le document (ex : saisissez des mots du titre), puis validez :

Insérer une bibliographie en fin de document

- rendez-vous à l'endroit où vous souhaitez créer la bibliographie
- cliquez sur le bouton ci-contre dans votre traitement de texte
- la bibliographie se mettra à jour chaque fois que nécessaire

Pour un mémoire académique, nous conseillons le style « *Vancouver (superscript, brackets, only year in date)* ». Le style « *Elsevier Vancouver* » est également apprécié. Pour un article scientifique, utilisez le style portant le nom du journal.

3 Rédiger les différentes parties du mémoire

3.1 Organisation selon le plan IMMRaD

Les articles scientifiques suivent presque tous le plan **IMMRaD**. Il correspond à celui d'une **recette de cuisine** :

- **I**ntroduction : contexte, avantages et défauts connus d'autres recettes
- **M**atériel (au sens « matériaux ») : ingrédients (œufs, farine, beurre, etc.)
- **M**éthodes : actions (couper, hacher, mélanger, cuire, laisser reposer, etc.)
- **R**ésultats : résultat brut (un cake de 500g prêt en 30 minutes)
- **a**nd **D**iscussion : commentaire des résultats et des méthodes, perspectives

Vous devez suivre strictement ce plan pour votre mémoire, pour notamment :

- Permettre de trouver l'information en lecture non-linéaire et non-exhaustive
- Séparer les éléments objectifs des éléments interprétatifs ou subjectifs
- Séparer les données internes à l'étude des données externes (bibliographie)

3.2 Rédaction de l'introduction

L'introduction :

- **pose le contexte** et l'**état de l'art** en exposant les connaissances préexistantes
- nécessite une **recherche bibliographique** approfondie
- comporte de très nombreuses références bibliographiques
- comporte des éléments **subjectifs** qui justifient l'existence du travail qui sera réalisé
- tient sur une demi-page dans un article, et une trentaine de pages dans une thèse

Cette partie devra principalement répondre à quatre critiques majeures :

- **Critique n°1 : Ce travail ne sert à rien.** Vous devrez dresser l'état de l'art pour faire apparaître un manque, et convaincre le lecteur que ce travail était indispensable.
- **Critique n°2 : Ce travail a déjà été fait.** Il est possible de montrer qu'un travail a déjà été fait en le citant, mais il est impossible de montrer avec certitude qu'il n'a pas été réalisé. Montrez que vous avez lu de nombreux travaux dans le même domaine, et qu'aucun d'eux n'est similaire au travail en cours. Il faudra les citer.
- **Critique n°3 : Ce travail n'est pas réalisable ou n'apportera pas la réponse attendue.** Montrez qu'un faisceau d'éléments laisse penser que le travail est réalisable.
- **Critique n°4 : Vous n'êtes pas la bonne personne.** Si vous avez su déminer les critiques précédentes, vous êtes devenu crédible. Par votre maîtrise de la bibliographie, vous pourrez produire des résultats parfaitement intégrables dans le mur des connaissances collectives, telle une brique de petite taille, mais solide et des mêmes dimensions que les autres briques.

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.

3.3 Rédaction de la partie Matériel et méthodes

3.3.1 Généralités

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 les méthodes, il ne s'agit pas de décrire en détail **ce que vous avez réellement fait**, mais plutôt **ce qu'il faudrait faire pour atteindre ce résultat**. Epargnez au lecteur vos tâtonnements, pour exposer le plus simplement possible une méthode qui permet de retrouver exactement vos résultats.

3.3.2 Exemple de structuration

Dans une **recherche sur des patients**, le plan pourra être le suivant :

- *Design* de l'étude : forme générale (voir section 4 en page 16), critères d'inclusion ou exclusion, critères d'exposition (variable définissant les groupes notamment), critère principal de jugement (survie, score, etc.), critères secondaires de jugement, etc.
- Patients et établissements (personnes incluses, services, hôpitaux)
- Données (données recueillies sur les patients)
- Analyse de données (ce qu'on souhaite calculer ou tester)
- Analyse statistique (méthodes employées, sans citer les variables)
- Cadre réglementaire (financement, CPP, CNIL, consentement)

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)
- Analyse de données (ce qu'on souhaite calculer ou tester)
- Analyse statistique (méthodes employées, sans citer les variables)
- Cadre réglementaire (financement, CNIL, information des sujets)

3.3.3 Paragraphe « Analyse statistique » en particulier

Le modèle de documents fourni sur le site http://objectifthese.org contient déjà un texte prêt à l'emploi, que voici. Il faudra le raccourcir en fonction des méthodes mobilisées. Ce paragraphe doit rester abstrait et ne pas citer les noms des variables.

3.3.3.1 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.*

3.3.3.2 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.

3.3.3.3 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 coefficients assortis d'intervalles 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 ratios (OR) assortis d'intervalles 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 ratios (HR) assortis d'intervalles 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*

3.3.3.4 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%.

3.4 Rédaction de la partie résultats

3.4.1 Généralités

La partie résultats, comme dans une recette de cuisine :

- Montre les résultats obtenus, ni plus, ni moins
- Sans rappeler **les noms des méthodes** utilisées
- De la manière la plus **froide et neutre** possible
- En garantissant que, si on applique les méthodes, on obtient ces résultats
- Sans **aucun élément d'interprétation**
- Sans **aucun élément de contexte**, **aucune comparaison** des méthodes à la littérature, **aucune comparaison** des résultats à la littérature
- Avec **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

Cet exemple de plan conviendra pour la plupart des **études cliniques**, ou **réutilisant des données** portant sur des personnes :

- *Flowchart* (diagramme de flux d'inclusion des patients : cf. 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* des participants (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* des articles (cf. chapitre suivant)
- Processus d'annotation
- Caractéristiques des articles retenus
- Critères de qualité des articles retenus
- [puis réponse aux questions posées en objectif]

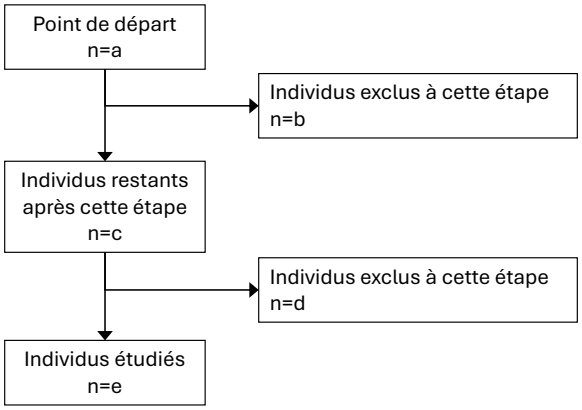
Au sein de chaque partie, l'organisation doit sembler naturelle au lecteur. Par exemple, lorsqu'on souhaite décrire des patients, on peut utiliser un ordre chronologique :

- 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
- prise en charge initiale
- prise en charge principale, dans un ordre naturel
- suites immédiates puis long terme

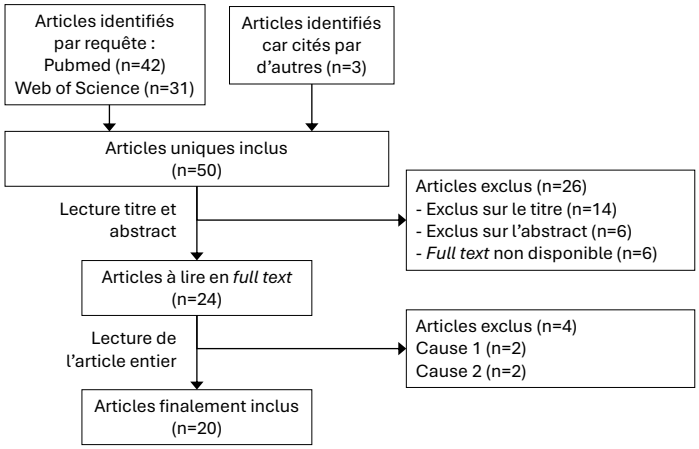
3.4.3 Flowchart

Le ***flowchart***, ou **diagramme de flux** :

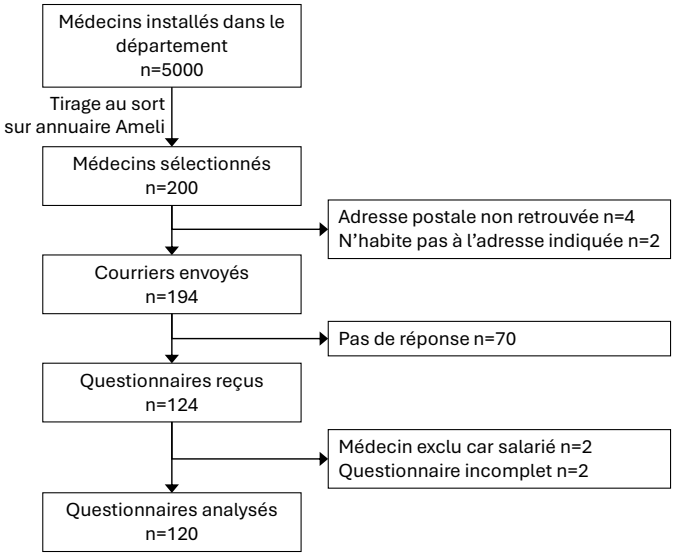
- constitue très souvent la première partie d'une section « résultats »
- fait progresser l'échantillon vers le bas, en excluant les individus vers la droite
- comporte des effectifs explicites, cohérents et vérifiables



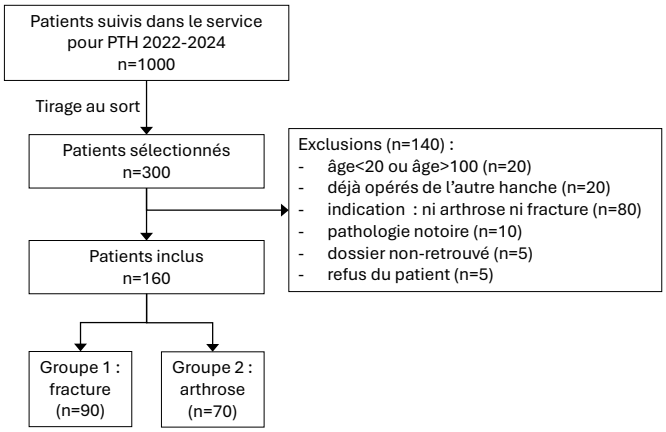
Exemple d'une revue de la littérature :



Exemple d'un questionnaire papier (le flowchart permet de calculer le taux de réponse, voir page 18) :



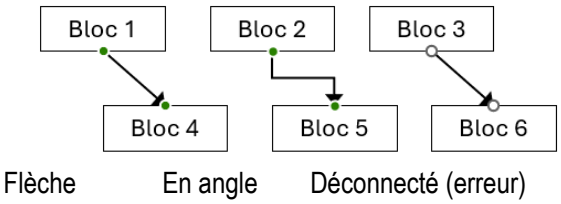
Exemple d'une étude réalisée sur des patients ou des dossiers :



Nous vous conseillons d'utiliser un logiciel de présentation tel **Microsoft Powerpoint** ou **LibreOffice Impress**. Destinés à préparer des diaporamas, ils sont très efficaces pour cet usage.

Le site Objectif Thèse vous propose un document Powerpoint contenant déjà des *flowcharts* réutilisables. La vidéo jointe montre certaines manipulations utiles à la construction de *flowcharts* : http://www.objectifthese.org

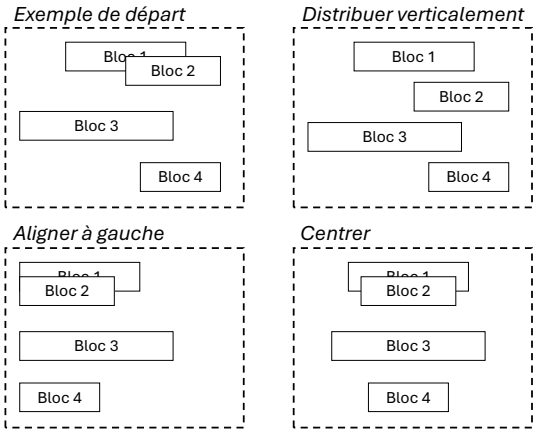
Vous utiliserez notamment des **connecteurs**. Dans Impress ou dans les anciennes versions de Powerpoint, il s'agissait de composants spécifiques. Dans les versions actuelles de Powerpoint, il s'agit désormais de flèches. Si les rectangles sont déplacés ou redimensionnés, ces connecteurs suivent automatiquement, ce qui laisse le diagramme cohérent.



Une fois votre *flowchart* grossièrement réalisé, sélectionnez les rectangles (pas les connecteurs : ils suivront), puis utilisez une des **huit fonctions d'alignement ou répartition** disponibles, dont trois sont présentées ci-dessous :

Dans Microsoft Powerpoint : *menu Forme > Aligner > [choisir la fonction]*

Dans LibreOffice Impress : *menu Format > Aligner [ou] Répartition*



3.4.4 Résultats d'analyses statistiques

Présentation des résultats d'analyse statistique :

- Suivez les recommandations, ne produisez pas trop d'indicateurs (voir le cours dédié à la présentation des résultats statistiques, sur la page http://objectifthese.org)
- **Suivez strictement** les consignes que vous avez-vous-mêmes énoncées dans la section « matériel et méthodes »
- Rédigez des phrases **simples, et prévisibles**, permettant une lecture non-linéaire
- **N'interprétez pas** (ce sera fait en discussion)
- Ne rappelez pas les **méthodes statistiques** (ce sera fait dans les méthodes)

Voici 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 (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 (p=0,32).*
  - **Quantitatif-quantitatif :**
    L'âge et la durée de séjour sont en relation linéaire croissante (r=0,620, r²=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 suivants 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 suivants 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 suivants 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 : (…)

3.5 Rédaction de la discussion

Dans la discussion, réutilisez sans limite les références bibliographiques de l'introduction. Vous pouvez utiliser le plan suivant :

- **Principaux résultats :** énumérez ce que vous vouliez faire, ce que vous avez fait, et ce que vous avez obtenu, pour mettre en avant la cohérence du travail
- **Discussion de la méthode :**
  - Listez de manière honnête et constructive les **biais** (voir page 69)
  - **Comparez** votre méthode à celle des autres publications
- **Discussion des résultats :**
  - **Interprétez** vos résultats (voir page 69)
  - **Comparez** vos résultats à ceux des autres publications
- **Perspectives :**
  - **Orientez** les autres chercheurs vers des **travaux** désormais nécessaires
  - Restez prudent, et évitez de conseiller directement une attitude aux cliniciens

3.6 Rédaction de la conclusion

Lorsqu'elle est présente, elle prend généralement la forme d'une phrase unique, à emporter à la maison : un « *take home message* ».

4 Imprimer et diffuser le document

4.1 Finalisation du document

Dernières opérations avant l'impression :

- Paginez en **recto simple, simple interligne** (comme le modèle http://objectifthese.org). Evitez vraiment le double interligne.
- Insérez des sauts de pages lorsque nécessaire, de telle sorte que :
  - les tableaux ne soient pas coupés
  - les tableaux et figures ne soient pas séparées de leur légende
  - aucun paragraphe trop petit ne figure seul sur une page
- Sélectionnez tout le document ([contrôle]+[a]) puis **mettez à jour tous les champs** (F9 ou clic droit et « mettre à jour les champs »)
- Recherchez le mot **« erreur » ou « introuvable »** qui témoigne de renvois rompus
- Recherchez la présence de « **n.d.** » dans votre **bibliographie Zotero** : le cas échéant, complétez les métadonnées dans l'interface de Zotero puis resynchronisez avec votre document avec le bouton Zotero Refresh
- Contrôlez scrupuleusement l'**orthographe** sur une version papier et non à l'écran
- En même temps, notez **tous les sigles** sur papier, pour contrôler ensuite votre liste en début de document

Une fois prêt à imprimer, enregistrez également en format PDF :

Microsoft Word : *Fichier > Enregistrer sous > [Choisir le format PDF]*

LibreOffice Writer : *Fichier > Exporter vers > Exporter au format PDF*

4.2 Impression non-professionnelle

Si vous imprimez vous-même votre mémoire :

- Vérifiez que votre **jury accepte** les impressions maison
- Vérifiez la **qualité de l'impression** sur toutes les pages (trainées, inclinaison, etc.)
- Pour relier, idéalement, utilisez une **baguette de reliure** plastique ou, si vous disposez d'une machine, une **baguette de reliure à anneaux en plastique** ou, mieux, d'une **baguette de reliure à anneaux en métal**
- Si nécessaire, tirez à part certaines figures en A3 couleur

4.3 Impression professionnelle

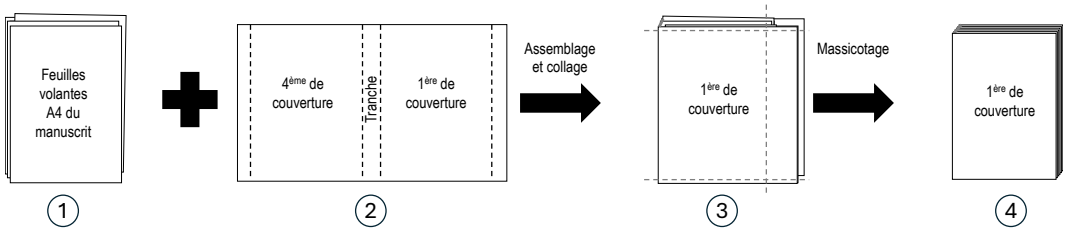
L'impression professionnelle :

- peut être réalisé par de **nombreuses boutiques** ou associations étudiantes
- offre une **meilleure qualité et coûte souvent moins cher** qu'une impression maison
- propose généralement une reliure en **dos carré collé**

Pour réaliser un dos carré collé, l'imprimeur :

- imprime tout le document sur feuilles volantes A4
- met en page lui-même la couverture
- assemble la liasse de feuilles volantes et la couverture, les aligne et les colle
- recoupe le tout au massicot électrique

Ceci nécessite des marges : choisissez 2,5cm partout..



4.4 Envoi par courrier postal, le cas échéant

Rarement, il faut envoyer le document par courrier postal au jury :

- Préférez les **adresses personnelles** (avec leur autorisation) aux universités et hôpitaux (fréquentes pertes de courrier)
- **Evitez** impérativement les **accusés de réception**, inutiles et casse-pied
- Joignez un courrier et demandez l'envoi d'un **SMS** pour confirmer la réception
- Utilisez une **simple enveloppe** en papier kraft, sans rembourrage (les courriers trop gros doivent être retirés en bureau postal)

5 Utiliser un logiciel de présentation pour la soutenance orale

5.1 Concevoir le diaporama, sur le fond

Voici comment concevoir le diaporama :

- Réutilisez le **plan IMMRaD** du document
- Comptez généralement **1 minute par diapositive**, hors titres
- Introduction : raccourcissez-la, et insistez sur le contexte et les éléments visuels
- Méthodes : simplifiez énormément. L'assistance les comprendra via les résultats.
- **Résultats : insistez** sur cette partie, en privilégiant les graphiques
- Discussion : simplifiez énormément
- **Ne rédigez pas** les diapositives : utilisez un style **télégraphique**, de nombreuses **listes à puces** hiérarchiques, et écrivez très gros

5.2 Concevoir le diaporama avec un logiciel de conception

Le site Objectif Thèse vous propose un document Powerpoint prêt à l'emploi. La vidéo jointe montre certaines manipulations utiles : http://www.objectifthese.org

Utilisez **Microsoft Powerpoint** ou **LibreOffice Impress** pour concevoir votre diaporama (évitez les solutions en ligne) :

- Utilisez les **masques de diapositives** (Powerpoint), ou **diapos maîtresses** (Impress) :
  - Equivalents des styles des traitements de texte, mais pour la disposition complète
  - Permettent un **gain de temps**, d'**homogénéité** et de **qualité**
  - Dans Microsoft Powerpoint : *menu Affichage > Masque des dispositives*
  - Dans LibreOffice Impress : *menu Affichage > Diapo maîtresse*
- Dans chaque bloc de texte de votre masque, activez le rétrécissement automatique du texte lorsqu'il est trop long :
  - 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** :
  - Dans Microsoft Powerpoint : *menu Masque des dispositives > Fermer le mode masque*
  - Dans LibreOffice Impress : *menu Affichage > Normal*
- Ajoutez un **numéro de diapositive** :
  - 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
- Conseils de mise en forme :
  - Evitez le rouge foncé (qui s'affiche souvent en marron)
  - Evitez les **contrastes trop faibles** entre le texte et son fond
  - Utilisez des **polices standard et sans sérif** (Arial, Cambria, Helvetica, etc.)
- Conseils quant au contenu :
  - Privilégiez les **illustrations** au texte, en citant les sources
  - Evitez les **animations** et les **démonstrations** en direct
- Faites plusieurs **répétitions à voix haute en temps réel**
- In fine, enregistrez en plus **une version PDF**
  - Dans Microsoft Powerpoint : *Fichier > Enregistrer sous > [Choisir PDF]*
  - Dans LibreOffice Impress : *Fichier > Exporter vers > Exporter au format PDF*

5.3 Présenter le diaporama, avec le logiciel

Vous pouvez activer (c'est le cas par défaut) le **mode présentateur** de Powerpoint, ou la **console de présentation** d'Impress. L'écran principal (ordinateur, à gauche ci-dessous) est destiné au présentateur, et l'écran secondaire (projecteur, à droite ci-dessous) au public.

Si vous souhaitez le lire :

- **imprimez votre texte sur du papier** car le mode présentateur ne sera pas toujours disponible (visioconférence notamment)
- **vérifiez le résultat** : un texte trop long sera tronqué sans message d'alerte

Si vous utilisez un **pointeur laser** :

- limitez les **tremblements** en tenant le pointeur à deux mains, coudes plaqués contre votre bassin
- attention : un pointeur laser est **inutilisable** sur écran LCD, en visio, en enregistrement, ou sur plusieurs écrans simultanément
- autres possibilités : le pointeur intégré au logiciel (avec la souris) ou un pointeur numérique[3]

[3] Exemple : l'excellent pointeur Logitech Spotlight

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 : s'il vous manque des explications détaillées, vous y trouverez notamment l'ouvrage « **Objectif Thèse Niveau 2 : poulet consciencieux** », qui reprend la même organisation mais avec de nombreux détails. Sa version PDF est gratuite. L'ouvrage « **Objectif Thèse Niveau 3 : coq méthodique** » s'adresse à ceux qui souhaitent réaliser les analyses statistiques avec R, et réaliser des analyses multivariées.

Si cet ouvrage vous a plu, n'hésitez pas à en faire la promotion. En regard du travail fourni, 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 ;-)**

Glossaire

Cet ouvrage de **110 pages**, comprenant **125 illustrations** dont **24 arbres décisionnels**, vous accompagnera pour réaliser toutes les étapes de votre **mémoire académique quantitatif en santé** (M1, M2, thèse d'exercice ou d'université). L'approche simple, didactique, percutante mais rigoureuse, vous permettra notamment de réaliser toutes les **analyses statistiques avec un tableur**, sans logiciel de statistique et sans l'aide d'un biostatisticien. Cet ouvrage est la **version condensée et simplifiée** du livre Objectif Thèse niveau 2 Poulet consciencieux.

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 : 978-2-9579934-2-0 110 pages 19.80 €

9 782957 993420