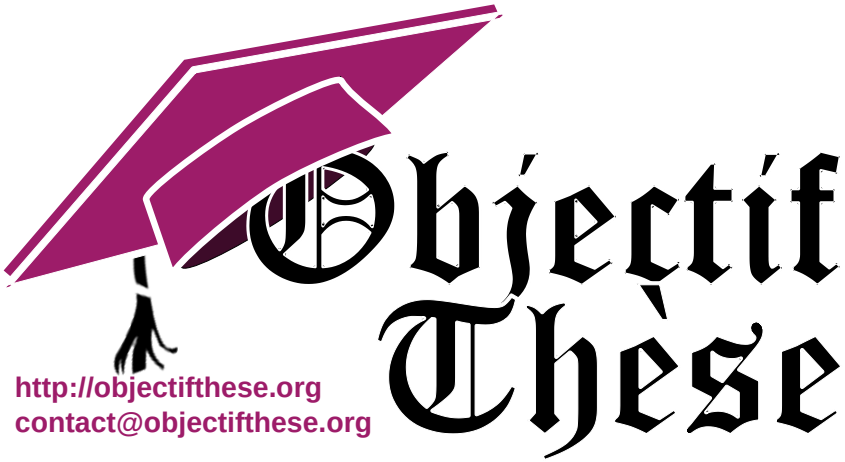
Saisir les données d'une questionnaire dans un tableur pour une analyse statistique



http://objectifthese.org
contact@objectifthese.org

Voir l'ensemble des ressources disponibles et ce fichier mis à jour sur http://objectifthese.org

Positionnement

Conception d'un questionnaire

Des personnes
OU des dossiers

Administration du questionnaire

Saisie de données

	Table de données de type « résultat de questionnaire »

1 ligne par individu statistique
1 colonne par variable

Analyse statistique :
- Description des variables (analyse univariée)
- Description des relations entre variables (analyse bivariée, multivariée)
- Prédiction, modélisation

Objectif

- Savoir saisir des données issues d'un formulaire dans un tableur
- Logiciels : Excel, Calc, SPSS…
- Pour permettre une analyse statistique :
  - Faire l'analyse soi-même dans un de ces logiciels
  - Transmettre les données à un collègue ou une plateforme capable d'analyser les données
  - Utiliser un site web tel http://biostatgv.sentiweb.fr/

Présentation générale du fichier de données

Fichier de données : généralités

- 1 table unique (sauf exceptions suivantes)
- 1 ligne par individu statistique
- 1 colonne par variable
- Noms de variables : sur la première ligne uniquement
- INTERDITS : cellules fusionnées ou fractionnées

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

Exemple 1

Patients vivants

id_patient	taille	hba1c	karnofski
Marcel	165	13.5	90
Justin	188	12	70

Patients décédés

id_patient	taille	hba1c	cause_deces
Christine	155	11.5	IDM
Julienne	168	8.7	AVP

id_patient	taille	hba1c	deces	karnofski	cause_deces
Marcel	165	13.5	0	90	NA
Justin	188	12	0	70	NA
Christine	155	11.5	1	NA	IDM
Julienne	168	8.7	1	NA	AVP

Exemple 2

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

Option 1, multi-table

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

id_patient	sexe
Patient_1	1
Patient_2	0

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

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

Option 2, mono-table

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

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

Exemple 3

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

Option 1, multi-table

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

id_patient	sexe
Patient_1	1
Patient_2	0

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

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

Option 2, mono-table

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

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

Anonymisation des enregistrements

- Données dé-identifiées et entièrement anonymes, selon votre autorisation CNIL. En général :
  - Pas de nom, pas de prénom
  - Pas de numéro de téléphone, d'adresse, de numéro de sécurité sociale, etc.
  - Pas de date de naissance exacte
  - Pas de date exacte de séjour ou consultation
  - Pas de nom de professionnel
  - Résidence du patient : localisation grossière (pas de commune, pas de code postal)
- Mais conserver un identifiant :
  - Nombre arbitraire
  - Table d'équivalence avec les dossiers tenue secrète

Première ligne : variables

- Noms des variables sur la première ligne
  - Interdiction de fusionner ou fractionner des cellules
  - Si première ligne de « chapeau », doit pouvoir être supprimée sans incidence (elle le sera !)

visite_anesth			visite_chirurgien		
vis_anesth_date	vis_anesth_nom	vis_anesth_duree	vis_chir_date	vis_chir_nom	vis_chir_duree

- Chaque nom de variable doit être unique
- Noms des variables :
  - Minuscules comprenant des lettres non accentuées de a à z, et des chiffres de 0 à 9, et éventuellement l'underscore « _ », à l'exclusion de tout autre caractère
  - Commencent par une lettre
  - Sont en particulier proscrits : caractères accentués, « e et o collés », « a et e collés », majuscules, espaces, tirets, points, ponctuations, caractères spéciaux
- Habitudes :
  - Identifiants en premier, commençant par « id_ »
  - Décomptes commençant par « nb_ »
  - Autres variables plutôt courtes, concaténation avec l'underscore « _ », intelligibles.
    Ex : date_prem_consult, date_der_consult, nb_visites…

Mise en forme par type de variable

Les identifiants

- Nature : chaînes de caractères (le plus souvent nombre entiers)
- Habituellement en première colonne, de la forme « id_ »
- valeurs manquantes interdites

id_patient	age
1	63
2	59
3	69
4	45
5	76
6	NA
7	NA
8	47
9	46
10	89
11	97
12	64

Les variables quantitatives (discrètes ou continues)

- Dans un tableur, si correctement saisies, s'alignent automatiquement à droite
- Formatage :
  - Ne pas utiliser de séparateur de milliers
  - Notation scientifique acceptée (1.23E05 = $1.23 * 10^5$)
  - Ne jamais écrire les unités
  - Durées : utiliser des nombres décimaux (un mois dure 30.44 jours, et une année 365.25 jours)
  - Pas de symbole pourcentage « % ». Pas « 50% » mais « 0.5 »
  - Indiquer toute la précision disponible, sans arrondi
- Les données manquantes, quel qu'en soit le motif, doivent être signalées par « NA » (pour un traitement en tableur, laissez vide)

id_patient	age
1	63
2	59
3	69
4	45
5	76
6	NA
7	NA
8	47
9	46
10	89
11	97
12	64

Les variables binaires

- Dans un tableur, si correctement saisies, s'alignent automatiquement à droite
- Formatage :
  - Utiliser « 0 » pour non/absent/faux/pas réalisé
  - Utiliser « 1 » pour oui/présent/vrai/réalisé
- Les données manquantes, quel qu'en soit le motif, doivent être signalées par « NA » (pour un traitement en tableur, laissez vide)

id_patient	chirurgie
1	0
2	0
3	1
4	0
5	1
6	NA
7	NA
8	0
9	1
10	1
11	1
12	0

Les dates

- En avez-vous vraiment besoin ?
- Dans un tableur :
  - Si correctement saisies, s'alignent automatiquement à droite
  - Elles deviennent alors des nombres « affichés comme des dates »
- Formatage :
  - Utilisez le format pour lequel votre tableur ou votre OS est configuré
  - Vérifiez avant export que le format est le même dans toute la colonne
  - Sinon, préférez le format « aaaa-mm-jj »
- Les données manquantes, quel qu'en soit le motif, doivent être signalées par « NA » (pour un traitement en tableur, laissez vide)

id_patient	date_chir
1	2020-06-27
2	2018-06-21
3	2018-02-04
4	2019-12-04
5	2019-12-03
6	NA
7	NA
8	2020-06-27
9	2018-06-21
10	2018-02-04
11	2019-12-04
12	2019-12-03

Événements temps-dépendants, données censurées à droite

- Patients suivis un certain temps (délai connu), certains subissent un événement (décès par exemple), d'autres non
- DEUX variables sont nécessaires (préfixe commun xxx_) :
  - Une variable d'événement « xxx_evt », binaire
  - Une variable de délai « xxx_delai », quantitative continue, exprimée dans la même unité pour tous (ex : mois, jours, etc.)
- Utilisation conjointe :
  - Si l'événement est observé au bout d'un délai *d* : xxx_evt=1 et xxx_delai=d
  - Si l'événement n'est toujours pas observé malgré un suivi sur une durée *d'* : xxx_evt=0 et xxx_delai=d' (que ces patient soient perdus de vue, ou exclus vivants, c'est-à-dire sans événement à la fin de l'étude)
- Données manquantes interdites :
  - si statut inconnu à une date donnée, prendre la date de point antérieure et le statut xxx_evt=0
  - en l'absence totale d'information, xxx_evt=0 et xxx_delai=0

Ex : on observe la survie de patients à l'aide des variables dc_evt et dc_delai. Délai en mois. Les patients 2, 3 et 4 sont décédés :

id_patient	dc_evt	dc_delai
1	0	36
2	1	12
3	1	4
4	1	20
5	0	8
6	0	3
7	0	35
8	0	21
9	0	9
10	0	18
11	0	16
12	0	5

Les variables qualitatives non-ordonnées (nominales)

- Dans un tableur, restent alignées à gauche
- Données s'exprimant en texte court. Ex : « bleu », « sage-femme ». Absence de notion d'ordre.
- Formatage :
  - Pas de norme particulière
  - Ne pas remplacer par des nombres !
  - Analysables seulement si nombre limité de modalités (max 20) => faire des regroupements
  - Bien vérifier que typographie homogène : « Souvent », « souvent », « souvent[espace] » et « souvent. » sont différents
  - La chaîne de caractères « NA » ne peut être utilisée, réservée aux données manquantes
- Les données manquantes, quel qu'en soit le motif, doivent être signalées par « NA » (pour un traitement en tableur, laissez vide)

id_patient	cheveux
1	Blonds
2	Bruns
3	Bruns
4	Bruns
5	Gris
6	Bruns
7	Absents
8	Absents
9	NA
10	Bruns
11	Bruns
12	NA

Les variables qualitatives ordonnées

- Même caractéristiques que précédemment
- Formatage :
  - Préférer des noms de modalités dont le tri alphabétique est pertinent
  - Exemple niveau d'études :
    - 0_aucun_diplome
    - 1_brevet
    - 2_bac
    - 3_licence
    - 4_master
    - 5_these
  - Exemple échelle de Likert :
    - 1_pasdutout_OK
    - 2_plutotpas_OK
    - 3_ni_OK_ni_pasOK
    - 4_plutot_OK
    - 5_toutafait_OK
    - (ou préférer en quantitatif -2, -1, 0, 1, 2 sans libellé, dans ce cas très particulier)

id	etudes	qbac
1	4_master	5_toutafait_OK
2	4_master	4_plutot_OK
3	1_brevet	2_plutotpas_OK
4	2_bac	1_pasdutout_OK
5	2_bac	4_plutot_OK
6	3_licence	4_plutot_OK
7	2_bac	NA
8	4_master	5_toutafait_OK
9	NA	3_ni_OK_ni_pas OK
10	NA	NA
11	1_brevet	2_plutotpas_OK
12	NA	3_ni_OK_ni_pas OK

Les variables qualitatives multivaluées

- Questions à plusieurs réponses possibles
- K cases à cocher => k variables binaires (sauf cases spéciales)
- Garder un préfixe commun pour les noms de variables
- Données manquantes :
  - si aucune saisie ou aberrent => toutes les variables en NA
  - Sinon, absence de coche = valeur 0
- Traiter intelligemment les cases « aucun », « tous », « pas concerné », etc.

id	atcd
1	hta ; tabac
2	NA
3	tabac ; ID
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