Chapitre

jcb

5 mai 2015

Contents

Quand utiliser une démarche quantitative ?

La recherche qualitative s'intéresse aux données mesurables sur lesquelles on peut appliquer des méthodes statistiques pour les décrire, les comparer, les inférer (Tirer une conclusion d'une proposition ou d'un fait, etc. wikitionnaire) ou les modéliser.

Habituellement les travaux de recherche quantitative se répartissent en deux grandes catégories [N.Radziwill. Statistics with R. 2ème ed. 2015]:

- les études observationnelles décrivent les caractéristiques de groupes constitués de façon déterministe (non aléatoire) et ne peuvent que découvrir de possibles associations entre les variables.
- les études expérimentales. L'observateur cherche une relation causale et va pour cela manipuler l'environnement pour comparer deux ou plusieurs traitements. Il ne peut y avoir d'expérimentation que si la constitution des groupes est aléatoire.

Ne pas confondre: il y a la recherche qualitative et la recherche quantitative (plus complémentaires qu'opposées). La recherche quantitative s'appuie sur des données quantitative et qualitatives.

Organiser son travail

Stafford Noble W. A quick Guide to Organizing Computational Biology Projects. PLOS Comp. Biol. (2009) **5**(7). Accédé le 10/2/2015.

Un travail de recherche (TR) peut se résumer à l'équation suivante:

```
TR = Question de recherche + données + analyse (statistique) + publication
```

Le présent chapitre essaie de répondre aux questions suivantes:

- pourquoi la collecte des données est une étape clé du travail de recherche
- comment la qualité du recueil doit s'inscrire dans la philosophie de la recherche reproductible (*reproductible resarch*).
- quelles sont les étapes du traitement des données pour qu'elles soient utilisables par le statisticien
- comment choisir et organiser le support de recueil des données.

Les données (data) sont le corollaire indissociable de tout travail de recherche. Elle peuvent exister avant de poser la question de recherche (travail rétrospectif) ou apparaître après que la question soit posée (travail prospectif). La nature et la pertinence des données à recueillir relève de la QR. On s'intéresse ici à la gestion de ces données en vue de leur exploitation statistique. C'est un étape cruciale pour deux raisons:

- 80% du temps nécessaire au traitement des données consiste à les rendre "présentables", c'est à dire prêtes à être "moulinées" par un logiciel de statistique. Plus on est pressé plus cette organisation des données est importante [ref. Wickham].
- une bonne organisation des données permet à un observateur extérieur de comprendre comment le travail s'est construit et au chercheur de reprendre son travail plusieurs mois ou année plus tard sans se demander comment il a pu obtenir ces résultats. Cette attitude "écologique" connait un développement important actuellement sous le nom de "reproductible research". C'est aussi un moyen de contrôler la fraude scientifique A summary of the evidence that most published research is false (http://simplystatistics.org/2013/12/16/a-summary-of-the-evidence-that-most-published-research-is-false/).

En pratique il faut disposer de trois outils:

- un tableur pour saisir les données et faire un premier nettoyage.
- un logiciel de statistiques pour exploiter les données.
- un traitement de texte qui à la manière d'un cahier de laboratoire permettra de conserver une trace de toutes les opérations effectuées sur les données (appelé aussi *note book* ou *code book*).

Pour organiser son travail on commence par créer un dossier portant le nom de l'étude et qui contient les sous-dossiers suivants:

- **data** pour stocker les données
- **notebook** pour noter tout ce que l'on fait, ce qu'il reste à faire, le dictionnaire des donées, etc.
- **stat** pour la partie statistique
- **production** pour la partie publication
- **doc** documentation générale, références bibliographiques.

Organiser la collecte des données

Primer on Data Management: What you always wanted to know (https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf)

Quels outils ?

Les données collectées ont vocation à être analysées par un logiciel de statistiques. Tous les outils du marché permettent à la fois de saisir les données et de leur appliquer un traitement statistique. Mais aucun outil ne fait ces deux actions correctement:

- **le tableur** (libre Office ®, Excel®, Numbers ®) est parfait pour saisir les données, les organiser, les corriger rapidement ou les compléter. Par contre il est incapable d'assurer la traçabilité des opérations. C'est souvent un outil statistique rudimentaire. Au final on aboutit à une grande feuille de calcul où se mélangent données désorganisées (pour une exploitation statistique), tableaux intermédiaires, graphiques, macros. Cet ensemble inextricable est incompréhensible pour le statisticien et génère de grands retard dans la production de résultats. La reprise d'un tel tableau plus tard est souvent impossible même pour son auteur...
- **le logiciel de statistique** permet d'organiser les données pour les rendre exploitables statistiquement. Par contre ce sont des outils lourds où la saisie des données brutes est nettement moins intuitive et rapide comparée à un tableur.
- le **bon choix** est d'utiliser les qualités des deux: un tableur simple pour saisir les données, un logiciel statistique pour les exploiter Luis A. Apiolaza.

Comment faire communiquer le tableur et le logiciel de statistiques ?

En utilisant un format de stockage des données qui soit commun aux deux. Le format **CSV** = comma separated variable (données séparées par une virgule). [http://fr.wikipedia.org/wiki/Comma-separated_values]

Le format *csv* est le moyen le plus sûr de conserver ses données. C'est un format libre et universel, indépendant des éditeurs et stable dans le temps, que tous les systèmes connaissent aussi bien en lecture qu'en écriture. En pratique tout tableur sait exporter ses données dans ce format et sait également importer des données au format *csv*. De la même manière, tous les logiciels de traitement statistique savent lire le format *csv*.

Comment faire ?: après avoir sauvegardé vos données au format (propriétaire) de votre tableur, aller dans l'onglet **fichier/enregistrer sous** et choisissez dans le menu déroulant **CSV**. Le nom du fichier créé se termine par le suffixe *.csv*

Remarque 1: les tableurs francisés remplacent la virgule du format *csv* par un point-virgule (*semicolon*) pour éviter la confusion avec la virgule décimale française (remplacée par le point décimal anglo-saxon). C'est la variante SSV (semicolon separated variable).

Remarque 2: il existe d'autres séparateurs couramment utilisés comme l'espace, le caractère tabulation (.tsv ou .txt) du format TEXT, le point, etc. Tous ces symboles doivent impérativement ne pas être utilisés dans le nom des colonnes du tableur sous peine de tromper les autres logiciels qui les comprennent comme des séparateurs de colonnes (cf. infra).

Remarque 3: Le format *csv* ne connait pas les formats multifeuilles (classeurs) des tableurs. Si votre document comporte plusieurs feuilles, chaque feuillet du classeur devra être enregistré séparément.

Utiliser un tableur [2]

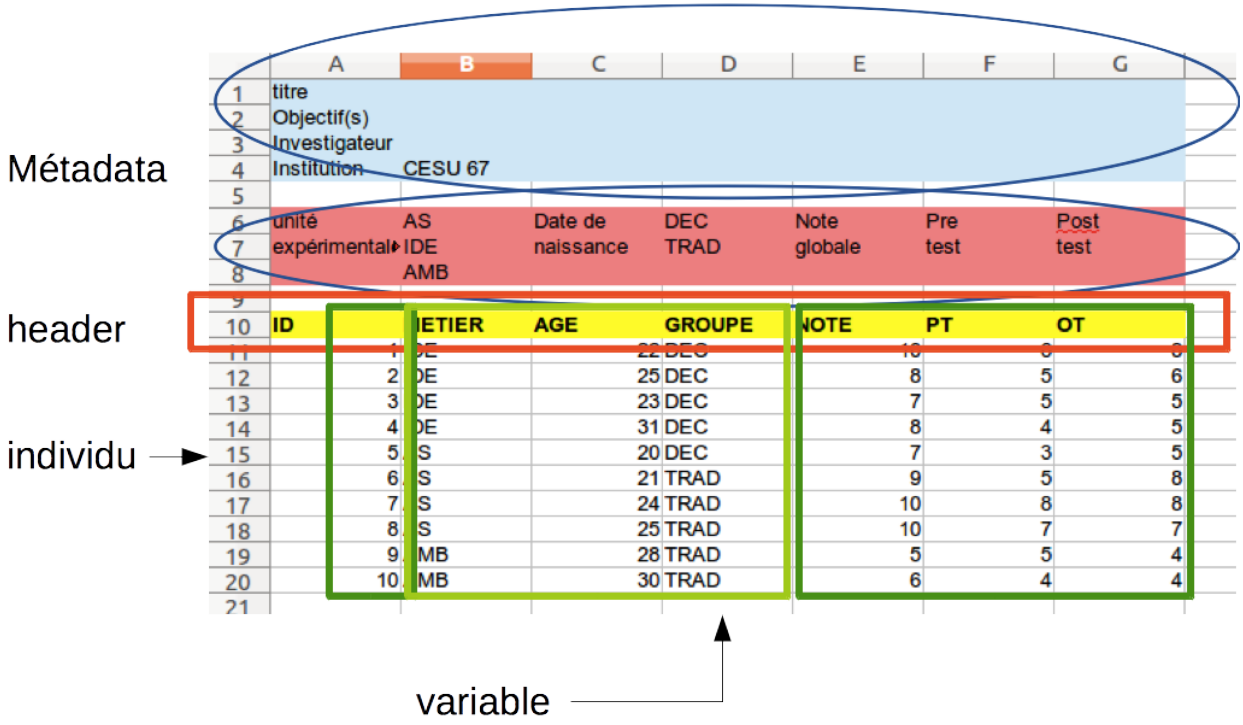
Les données sont généralement conservées sur 2 types de support, les tableurs et les bases de données. Il est aisé de passer de l'un à l'autre. Le tableur est la solution universellement adoptée du fait de sa disponibilité immédiate et de son utilisation intuitive. Cependant un certain nombre de règles doivent être respectées.

La première (et très importante): un tableur doit servir à saisir des données et ne servir qu'à cela. La plupart des ennuis commencent lorsqu'on pense qu'un tableur peut tout faire, en particulier des statistiques. Il faut abandonner cette idée et se placer uniquement sous l'angle de la saisie des données. Pour faire des statistiques, il y a des logiciels pour cela.

Anatomie et physiologie d'un tableur

L'organisation des données dans une feuille de calcul pour qu'elle puissent être traitées immédiatement par logiciel de statistiques est une opération semée d'embûches, si on ne respecte pas quelques règles de base.

Les données portent généralement sur des mesures appelées **variables** effectuées sur des **individus**. Les données sont stockées dans un tableau rectangulaire appelé **table** pour une base de données, **feuille** pour un tableur, **matrice** en mathématiques ou **dataframe** en statistiques. What's In A Tableaccédé le 1/3/2015.



Par convention[5], une feuille est constituées de **cellules** délimitées par des **lignes** et des **colonnes** formant un tableau rectangulaire où:

- chaque ligne du tableau représente un individu unique
- chaque colonne contient les différentes valeurs d'une **même** variable.
- la première ligne appelée **header** est réservée pour le nom des colonnes.
- une table ne doit contenir qu'un type de données

- si plusieurs tables sont nécessaires, elles doivent inclure une colonne, identique dans chaque table, qui permette de les lier (merging)
- le total des lignes doit être égal au total des observation + 1 (header)

Question: combien de variable dans le tableau suivant:

lésions	hommes	femmes
oui	4	1
non	2	5

La façon dont le tableau est présenté, il semble que il y ait seulement deux variables. La bonne réponse est trois: lésions, sexe, nombre.

sujet	lésion	sexe	nombre
1	oui	H	4
2	oui	F	1
3	non	H	2
4	non	F	5

Ce format est également appelé large (**wide**) et s'oppose au format long (**long**). Le format *wide* est plus adapté au traitement statistique. La plupart des logiciels ont des instructions permettant de passer d'un format à l'autre (ex. format pivot dans excel).

La feuille de saisie est divisée en 3 zones horizontales:

- la première, **facultative**, sert à stocker les métadonnées [http://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e]. Sa présence est à éviter (utiliser un document annexe) et sera éliminée au moment de l'analyse.
- la seconde stocke le nom des colonnes: c'est l'**en-tête** ou **header**.
- la troisième contient les données proprement dites.

La troisième zone est divisée verticalement en 2:

- les colonnes les plus à gauche contiennent les **variables qualitatives** appelées aussi **facteurs** (sexe, profession, type de pédagogie) qui permettront de faire des analyses en sous-groupes.
- les colonnes les plus à droite contiennent les variables quantitatives (age, note) qui serviront aux calculs (moyenne, variance...)

Ce gabarit doit être bien pensé dès le départ pour éviter toute réorganisation importante par la suite, source d'erreur par confusion des versions.

Le nom des colonnes (header)

- il doit tenir sur une seule ligne, la première appelée **header**.
- Sert à stocker le nom opérationnel d'une variable (celle qui apparaîtra par défaut sur les graphiques).

- le nom doit être à la fois succinct (5-10 caractères) et explicite: AgeAuDiagnostic plutôt que ADx.
- il ne doit pas contenir d'espace ou de caractères qui pourrait être mal interprété (, ; / @ &, espace)
- remplacer les espace par des *underscore* (caractère 8) ou des points: **Ceci_est_un_exemple** ou comme cela **nom.marital**.
- **pas de caractères accentués**. Pour éviter toute tentation, il est recommandé d'utiliser des lettres majuscules ou le terme anglais correspondant: **PRE_TEST**, **firstname** à la place de prénom.
- Ne pas fusionner les cellules pour éviter une erreur pour incohérence du tableau. Le nombre de cellules du *header* doit être strictement égal au nombre de colonnes de données.

les mesures

- toujours utiliser le point décimal (jamais la virgule). Pour corriger des données déjà saisies, utiliser la fonction *rechercher/remplacer* du tableur.
- une colonne ne peut contenir que des chiffres ou du texte, jamais des deux. La seule exception est le code **NA** (not avalaible) qui peut remplacer une valeur manquante dans une colonne de chiffres.
- Le dates de préférence au format ISO:
  - AAAA-MM-JJ ex. 2014-02-06
  - AAAA-MM-JJ HH:MM:SS ex. 2014-02-06 10:25:36
- Un logiciel ne sait pas distinguer les codes de couleur
- Un logiciel gère très bien les lettres:
  - Non: 1, 1, 2, 2, 2, 0, 1, 2
  - Oui: H, H, F, F, F, NA, H, F
- Ne pas mélanger les majuscules et les minuscules:
  - Homme <> homme <> HOMME sont compris comme trois variables différentes
- Valeur manquante:
  - La déclarer explicitement (ne pas laisser de blanc)
  - Ne pas utiliser le zero ou une valeur négative (-99)
  - Valeur par défaut: NA (not avalaible).

en résumé sont interdits

- les accents
- les blancs (espace)
- la virgule: confondue avec les séparateurs de colonne. A remplacer par le point décimal.
- les caractères réservés
- les macros
- la surbrillance ou les couleurs
- les lignes ou les colonnes vides (pour créer des séparations)

les astuces utiles du tableur

- fonction **rechercher/remplacer**
- Encodage: UTF-8

Note Book

C'est le carnet de laboratoire de l'analyste. Le format habituel pour ce document est un fichier type taitement de texte (Libre Office, Word) . Il devrait comporter deux sections intitulées:

- « Conception de l'étude» qui décrit de façon détaillée la façon dont vous avez recueilli les données et ce vous voulez en faire.
- « codage» (dictionnaire, lexique) qui décrit chaque variable et ses unités.

Conception de l'étude

Cette partie contient les **métadonnées**. A Gentle Introduction to MetadataJeff Good University of California, Berkeley. Accédé le 28/4/2015 Littéralement, ce sont des données à propos des données. Elles servent à comprendre le contexte dans lequel se fait l'étude:

- nom de l'auteur, des investigateurs
- version du questionnaire
- méthodes de description des variables: 'pour la rubrique sexe préciser H ou F'.

En pratique il est recommandé de ne pas faire figurer cette zone dans le tableur mais de les faire figurer dans un document annexe comme le **dictionnaire** pour la description des données et le **notebook** pour expliquer comment les données ont été recueillies,par qui, dans quelles conditions, etc.

Dans le contexte d'un travail de recherche, les métadonnées correspondent au contenu du paragraphe matériel et méthode. Nicole Radziwill propose d'utiliser la méthode des 5W/1H:

- who
- what
- where
- when
- why
- how

Les métadonnées sont une composante essentielle de la recherche reproductible. C'est une information indispensablepour le statisticien.

Le dictionnaire (Note Book)

C'est la partie du _notebook___ consacrée à la description des variables utilisées. Par exemple la variable sexe est codée "H" ou "F", ou bien homme = 1, femme = 2. La définition d'une variable doit comporter les éléments suivants:

- le nom complet
- l'abréviation utilisée dans le *header*
- pour les *variables quantitatives*:l'unité de mesure utilisée et la fourchette de validité
- pour les *variables quantitatives*: l'ensemble des niveaux attendus pour une variable qualitative (nom et abréviation)

exemple:

- poids du nouveau-né, PNN, en grammes (500 g - 5000 g)
- catégorie socio-professionnelle (CSP):
  - infirmier(ere) diplomé(e) d'état (IDE)
  - aide-soignant(e) (AS)
  - infirmier-anesthésiste (IADE)
  - sage-femme (SF)

Sauvegardez vos données [1]

Il est généralement conseillé de faire trois copies de vos données. Par exemple, vous pourriez avoir le jeu original, une copie sur un disque dur externe local, et une autre copie sur un disque externe situé ailleurs. Ces copies doivent avoir des localisations géographiques différentes. Assurez-vous que le système de back-up est fiable et que les données sont réellement récupérables. Le "Cloud" est une solution envisageable (Google drive, Bit Bucked, GitHub).

Sécurisez vos données

En général le cryptage des données n'est pas conseillé. Cependant il peut être indispensable pour certaines données sensibles (données nominatives, personnelles, recommandation de la CNIL) de les crypter. Gardez des mots de passe et les clés sur le papier (2 exemplaires), et dans un fichier numérique crypté comme PGP (Pretty Good Privacy). Ne comptez pas sur le cryptage 3ème partie seul.

Il est également recommandé de stocker des données dans un format non compressé. Si vous avez besoin d'économiser de l'espace, nous recommandons de limiter la compression à votre copie de sauvegarde.

Formats de fichier [1]

Certains formats de fichiers sont plus adaptés à la conservation que d'autres. Par exemple, certains formats sont propriétaires, ce qui signifie que leur structure ne est pas accessible au public pour la documentation ou la réplication. Si vous enregistrez vos données dans un tel format et que la société qui le détient cesse d'exister, il peut être impossible de les récupérer. Idéalement les données:

- doivent rester accessibles dans le futur (20 ans)
- dans un format non propriétaire
- respecter des standards ouverts, documentés
- être commun, utilisé par la communauté de recherche pertinent ou des communautés;
- utiliser des polices de caractère ouvertes et standard (ASCII, Unicode);
- ne pas dépendre d'un support unique pour ˆêtre affichées ou manipulées.
- si possible ni cryptées, ni compressées.

Les standards ouverts sont par exemple:

- texte: ASCII ou Unicode (UTF8)
- Pour l'audio: MPEG-4, pas Quicktime;
- Pour les images: TIFF ou JPEG2000 ou PNG, (pas GIF ou JPG)
- Pour données structurées: XML, JSON ou RDF (pas SGBDR)
- traitement de texte: libre office (pas Word)
- tableur: libre office (pas excel)

Nommage des fichiers [1]

Le nom donné aux fichiers est très important car c'est par cet identifiant que les données sont lues par un logiciel statistique et que l'on peut savoir si la version utilisée est la bonne. Certaines disciplines ont défini des standard très élaborés (DOE's Atmospheric Radiation Measurement (ARM) program).

Voici quelques recommandations:

- Utilisez un système de nommage des fichiers qui reste cohérent.
- la lecture du nom doit être suffisament explicite pour rester comréhensible dutant plusieurs années. Par exemple: projet_questionaire_lieu_date_version.ext
- Évitez ces symboles dans les noms de fichiers: “/ : *? “<> [] & $. Ces caractères ont une signification particulière dans certains systèmes d'exploitation d'ordinateur qui pourrait entraîner une mauvaise lecture ou la suppression de ces fichiers.
- évitez les caractères accentués qui sont mal supportés par les logiciels statistiques
- Utilisez le caractère soulignement (_), pas des espaces pour séparer les termes.
- Essayez de garder les noms de dossiers courts (15-20 caractères ou moins) et descriptif de ce qui est à l'intérieur.
- Essayez de garder les noms de fichiers de moins de 25 caractères.
- Inclure les dates dans les noms de fichiers ou de dossiers; utiliser le format AAAA-MM-JJ.
- Si vous ne utilisez pas le versionnage automatique, inclure un numéro de version à la fin du nom de fichier tel que v01. Changez ce numéro de version chaque fois que le fichier est enregistré. Ne utilisez pas d'étiquettes confusion: la révision, finale, final2, etc.
- Pour la version finale, remplacer le dernier mot pour le numéro de version. (Ceci est particulièrement important si les fichiers sont partagés.)
- Les noms de fichiers doivent contenir qu'un seul point, avant l'extension de fichier (par exemple name_paper.doc PAS name.paper.doc)

Organiser ses données

Votre travail de recherche est important et des données bien organisées faciliteront le travail du statisticien et de vos pairs qui doivent pouvoir reproduire et évaluer vos résultats (ref[1]).

Comment coder les variables

Niveau de mesure	Groupes mutuellement exclusifs	Rangs ordonnés	Valeurs équidistantes	Référence nulle non arbitraire	exemples
Nominal	X				Status marital
Ordinal	X	X			Niveau de stress (1-7)
Intervalle	X	X	X		Echelle de dépression (1-100)
Ratio	X	X	X	X	Poids (kg

La division des types de données a été proposée en 1946 par le psychologue américain Stanley Smith Stevens Stevens 1946. Accédé le 1/5/2015. Voir aussi Level of measurement.

Dans une feuille de calcul, on peut distinguer quelques grandes catégories de données en fonction de leur nature data type:

- quantitatives:
  - continue. Ce sont toute variable mesurée sur une échelle quantitative continue comme par exemple le poids mesuré en kg.
  - discrètes (nombre de frères et soeurs)
- qualitatives ou factorielles (sexe, parité) ventilées en catégories (levels) mutuellent exclusives) divisées en:
  - Nominales (catégorielles: l'ordre n'a pas de sens): sexe, CSP, binaires (fumeur ou non). Categorical data sont des variables où il existe plusieurs catégories, mais qui ne sont pas ordonnées. L'exemple classique est le sexe: homme ou femme.
  - Ordinales (catégorielles ordonnées: l'ordre à un sens): échelle de Likert (mauvais, passable, bon), score de Glasgow.variables ordinales. Le nombre de niveaux ne dépasse pas 100.
  - Intervalles
  - Ratio
- Données manquantes Missing data sont des données manquantes et irrécupérables. Elles sont codées avec le symbole NA (not avalaible).
- Données censurées [Censored data](http://en.wikipedia.org/wiki/Censoring_(statistics)) sont des données manquantes mais ont sait pourquoi. exemple classique: un patient perdu de vue. Elles doivent aussi être codés `NA` quand vous n'avez pas les données. Mais vous devriez également ajouter une nouvelle colonne à vos données appelé, "VariableNameCensored" qui devrait avoir des valeurs de `true` si censuré et `false` si pas. Dans le lexique, il faut expliquer pourquoi ces valeurs sont manquantes. Il est absolument essentiel de mentionner à l'analyste, si il y a une raison connue pour que ces certaines données soient manquantes.

Remarque: dans la mesure du possible il vaut mieux éviter de représenter les variables qualitatives par des chiffres: H ou F est plus explicite que 1 ou 2, sans compter le risque de confusion ou d'erreur d'interprétation (parfois évident comme le calcul de la moyenne sur le sexe, parfois plus subtile comme calculer un score de Glasgow moyen). Les logiciels destatistiques savent très bien manipuler les symboles.

Cette classification des variables n'a pas seulement un intérêt sémantique. La nature d'une variable détermine le choix d'un indicateur, d'une représentation graphique ou d'un test. La moyenne n'est utilisable que pour les variables quantitatives alors que la médiane est le paramètre de référence pour une variable catégorielles. Il n'y a de corrélation qu'entre variables continues,alors que le test du Khi2 analyse la relation entre 2 caractères quantitatifs.

Cycle de vie des données [3]

Les données vont traverser plusieurs étapes avant de pouvoir être analysées:

- données brutes (raw data)
- données désorganisées (messy data)
- donnée nettoyées et exploitables (tidy data)

Carly Strasser et coll.[3] décrivent 8 étapes dans l'acquisition et le traitement des données:

1. planifier

2. collecter: les données sont recueillies à la main ou récupérées via des capteurs et stockées sur un support numérique
3. Vérifier la qualité et l'intégrité des données
4. Décrire: les données sont soigneusement et précisément décrites sous forme de métadonnées standardisées
5. Préserver: les données placées sur un support permettant de les stocker sur le long terme
6. Découvrir: les données potentiellemet utiles sont identifiées et obtenues avec les métadonnées qui les accompagnent.
7. intégrer: les données provenant de sources disparates sont combinées pour former un ensemble de données homogène qui puisse être analysé
8. Analyser: analyse des données

Nettoyer les données [5]

Il n'y a que dans les ouvrages de statistique où l'on trouve des données prêtes à être analysées. Dans un travail de recherche les données passent par plusieurs états avant de pouvoir être analysées:

1. les données brutes (raw data): c'est la saisie initiale de l'expérimentateur. Il manque souvent le header, certains type de données sont erronés, erreur de catégorisation des données, caractère mal encodé ou inconnu (caractères accentués, réservés), etc. La lecture directe de ces données par un logiciel de statistique échoue.

2. les données techniquement correctes: ce sont des données brutes corrigées de sorte qu'elle sony techiquement acceptables par un programme de traitement. Les "dirty data" sont transformées en "messy data" (données propres):

- données manquantes: il manque toujours des données... Chaque fois que c'est possible essayer de retrouver la donnée. On ne doit pas supprimer les valeurs manquantes. Si la donnée est irrécupérable, la remplacer par un symbole compréhensible par le logiciel statistique comme NA (Not Avalaible). Ne jamais laisser une case vide. Eviter les symboles qui prêtent à confusion: 0, 99999, -1, etc. Ne pas confondre NA et Inconnu ou NSP (ne sais pas) ou NSPR (ne souhaite pas répondre). Ne pas supprimer des observations parceque des données manquantes.
- données spéciales: concernent surtout les valeurs numériques pour les nombres n'appartenant pas à l'ensemble des réels: NA, NaN (not a number), Inf.
- données extrèmes (outliers)

3. Données cohérentes: dans cette dernière étape on vérifie la cohérence des données techniquement correctes. Par exemple qu'il n'existe pas d'age négatif ou supérieur à une certaine limite, une pathologie incompatible ave un sexe donné. Cette étapepeut se révéler particulièrement fastidieuse et complexe.

Ce n'est qu'à la fin de cette troisième étape que les données peuvent être soumises à une analyse statistique. Ce processus de nettoyage ralentit considérablement le processus d'analyse statistique des données (Hadley Wickham Tidy Data). Inversement le respect des règles édictées dans les paragraphes précédents permet d'en gagner beaucoup.

2 remarques: - La correction des données natives peut être un exercice périlleux puisque les données brutes originales risquent d'être écrasées et il n'y aurait aucun moyen de vérifier ce processus ou de récupérer des erreurs commises pendant cette phase. Les corrections doivent toujours se faire sur une copie des données originales afin de pouvoir revenir en arrière en cas de problème. - toutes les étapes doivent être documentées dans le *notebook* pour comprendre quelles manipulations ou transformations ont été néccessaires pour passer des données brutes aux données exploitables.

Organiser le contrôle de versions

La gestion de versions est une technique qui permet de sauvegarder les différentes étapes de son travail, tout en conservant une trace chronologique, permettant de revenir à une version antérieure. Un des outils les plus utilisé est **git** (Git en français). Disponible sur toutes les plateformes, c'est un logiciel libre et gratuit, qui permet d'enregister un ensemble de fichiers (tableur, doc, pdf, etc.) avec un numéro de version horodaté. Git permet de retrouver toutes les versions d'un fichier et d'indiquer quelles sont les modifications apportées entre deux versions. Il est ainsi possible de retourner à une version antérieure en cas d'erreur, de développer une branche parralèle puis de de fusinner avec la branche principale. De ce fait il est très bien adapté au travail d'équipe, chacun pouvant développer une branche du projet avant de fusionner. Git peut être hébergé sur une machine locale et/ou être synchronisé avec un dépot dans le cloud (Github, Bitbucket)

Git-les bases pour bien gérer les versions de votre projet. Linux Pratique (2013) n°83 59-63.

Les étapes de l'analyse statistique

C'est la question de recherche qui va dicter la profondeur de l'analyse statistique. Par ordre de complexité croissante, on peut la décrire ainsi:

- statistique descriptive: il s'agit de décrire les données au travers de quelques chiffres qui les résument et sans aller plus loin dans l'interprétation.
  - L'analyse univariée est une étape fastidieuse mais indispensable concernant chaque variable pour en déterminer les principales caractéristiques et detecter des anomalies (valeurs aberrantes, hors limites (outliners)). Les principaux indicateurs sont:

Mesures	Var.Quantitative	Var.Qualitative
centralité	moyenne	médiane
dispersion	variance, écart-type	quantiles
a	minimum, maximum	répartition par facteur
graphisme	histogramme, scatterplot	boxplot

```
- l'analyse bivariée étudie de la liaison possible entre deux variables: covariance, corrélation, régres
- Une représentation graphique est utile pour apprécier laforme d'une distribution: histogramme, diagra
```

- statistique inférentielle: un travail porte toujours sur un échantillon que l'on espère aussi représentatif d'un population que l'on ne connaîtra probablement jamais. L'inférence rassemble les techniques statistiques qui permettent d'extrapoler les conclusions basées sur un échantillon à une population. C'est la forme de statistique la plus souvent observée dans la littérature scientifique [Jeff Leeks. The elements of data analytic style. A guide for peope who want to analyze data.2015. Leanpub ed.]
- statistique inférentielle: un travail porte toujours sur un échantillon que l'on espère aussi représentatif d'un population que l'on ne connaîtra probablement jamais. L'inférence rassemble les techniques statistiques qui permettent d'extrapoler les conclusins basées sur un échantillon à une population.
- statistique prédictive: tandis que l'analyse inférentielle quantifie les relations entre des mesures à l'échelle d'une population, l'analyse préditictive utilise un sous-ensemble des données pour prédire d'autres mesures sur le même ensemble (ex. sondages d'opinion)

- statistique causale essaie de pévoir quelles sont les conséquences dela modification d'une variable sur les autres variables. C'est le cas des essais cliniques randomisés: influence de l'arrêt du tabac sur la survie. A la différence des méthodesinférentielles ou prédictives, lesméthodes causales identifient l'importortance et le sens des relations entre variables.

modélisation: c'est trouver la meilleure combinaison possible de plusieurs variables pour prédire le comportement d'une variable d'intérêt. Par exemple quelle sont les chances qu'une personne se plaignant d'une douleur thoracique soit porteuse d'une pathologie coronarienne sachant que l'on a mesuré chez elle un certain nombre de variables (sexe, age, poids, taux de cholestérol, etc.) dont on a montré préalablement (par les méthodes statistiques précédentes) qu'elles avaient un lien avec la pathologie suspectée. Il existe de nombreuses techniques de modélisation. L'une des plus répandue est le modèle linéaire.

inférence: estimer les paramètres d'une population à partir d'un échantillon. La précision de l'estimation dépend de la taille de l'échantillon. Inversement, si on fixe la précision souhaitée on peut déterminer la taille l'échantillon. D'une manière générale, pour doubler la précision (c'est à dire diviser par 2 l'intervalle dans lequel se trouve le paramètre d'intérêt), il faut quadrupler la taille de l'échantillon.

modélisation: c'est trouver la meilleure combinaison possible de plusieurs variables pour prédire le comportement d'une variable d'intérêt. Par exemple quelle sont les chances qu'une personne se plaignant d'une douleur thoracique soit porteuse d'une pathologie coronarienne sachant que l'on a mesuré chez elle un certain nombre de variables (sexe, age, poids, taux de cholestérol, etc.) dont on a montré préalablement (par les méthodes statistiques précédentes) qu'elles avaient un lien avec la pathologie suspectée. Il existe de nombreuses techniques de modélisation. L'une des plus répandue est le modèle linéaire.