

Initiation à l'analyse de données

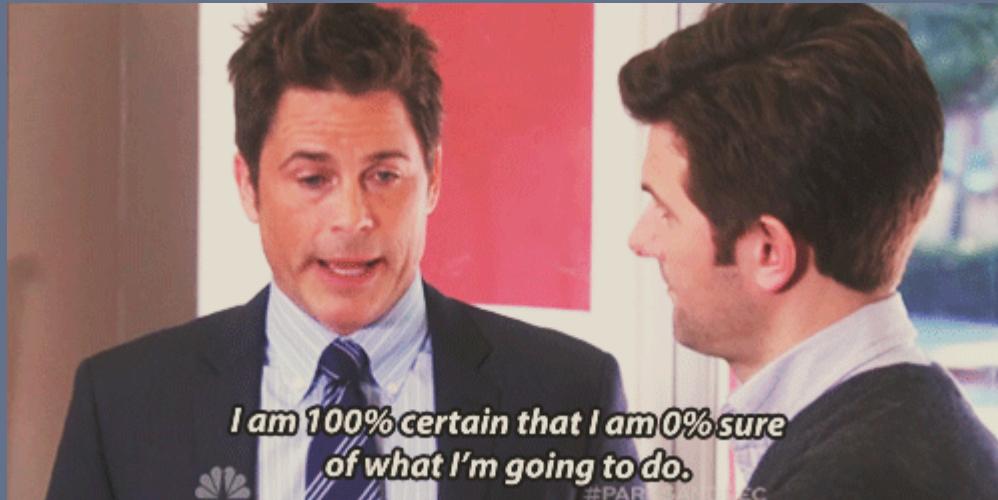


Premier traitement de données de questionnaires

Grégoire Le Campion - UMR Passages CNRS

2022/10/15 (updated: 2022-11-04)

Nos objectifs



Objectifs

Entrevoir les possibilités de l'analyse de données issuent d'une enquête par questionnaire

À part ça :

1. Faire un très bref retour sur les questionnaires
2. Vous parler du format base de données
3. Vous enseignez les premiers indicateurs statistiques utiles pour l'exploration des données
4. Manipuler des outils simples d'analyse de données
5. (Représenter graphiquement vos données)

En bref, à l'issue du cours vous ne serez pas des experts, mais vous aurez quelque armes pour démarrer l'exploration de vos données !

Le questionnaire



Le Questionnaire

Pourquoi fait-on un questionnaire ?

- 1- Souvent d'abord parce que l'on a une problématique l'on souhaite objectiver...
- 2- mais aussi quantifier...
- 3- ou encore lorsque que l'on cherche à valider et à généraliser les résultats

En bref un questionnaire est toujours conçus pour nous fournir des réponses qui illustrent une variable (la variables étant une des choses que l'on souhaite étudier).

Tout l'enjeu de l'analyse de questionnaire (donc de la stat !) va être d'étudier ces variables, de les croiser, de voir les relations qui les relient.

Attention aux données récoltées par questionnaire



Les biais du questionnaire

Le questionnaire est un outil de recueil de données très puissant mais comme toutes les méthodes de recueil de données il peut être soumis à des biais. Et ils sont nombreux...

La définition de biais d'après le Larousse est une position oblique, ou peut aussi renvoyer à une déformation, un travers.

Ils sont particulièrement dangereux dans le recueil de données car ils peuvent nous amener à recueillir de l'information mais qui n'est en réalité pas celle souhaitée.

Quand on construit un questionnaire il faut donc avoir un sorte de hantise permanente du biais tout en acceptant le fait qu'ils sont en réalité quasi inévitable.

Les biais du questionnaire II

Dans un questionnaire (mais comme avec les autres méthodes) les biais peuvent survenir à toutes les étapes :

- 1- Au niveau du choix de "l'univers" de l'enquête, cad la population visée.
- 2- Au niveau de la construction d'un éventuel échantillon.
- 3- Au niveau de la construction du questionnaire.
- 4- Au niveau de l'administration du questionnaire.
- 5- Au niveau de la réalisation elle même de l'enquête, si l'échantillon interrogé réellement ne correspond pas à ce qui a été définis.
- 6- Au niveau de la codification des réponses.
- 7- Au niveau de l'analyse des réponses.

Les biais du questionnaire III

Parmi les biais les plus connus on peut citer :

-La désirabilité sociale : vote « Le Pen », vote Obama (« effet Bradley »)

-Effet de succession de questions (« effet de halo ») : Poser une question sur Hiroshima, suivie d'une question sur la nécessité de développer l'énergie nucléaire

-Effet de sous entendu. Par ex : "Quel Etat présente le plus une menace pour la paix dans le monde ?" Cela oblige le répondant à accepter l'idée que l'existence d'un état présente en soit une menace.

-Lien entre opinion et action. Un sondage sur les intentions de vote n'a aucun sens si l'on n'a pas une idée du taux de mobilisation, c'est-à-dire du pourcentage de gens qui sont sûrs d'aller voter, et du taux de certitude, c'est-à-dire du pourcentage de ceux qui sont sûrs d'avoir arrêté leur choix

Heureusement quand on a conscience des biais on peut tenter des contre mesures.

Les biais du questionnaire IV

Voici une liste non exhaustive de biais, de leur effet et de la méthode pour les contourner :

Biais de confirmation (d'hypothèses)	<ul style="list-style-type: none"> Tendance naturelle qu'ont les individus à privilégier les informations qui confirment leurs idées préconçues, leurs hypothèses et à accorder moins de poids aux points de vue jouant en défaveur de leurs conceptions. 	<ul style="list-style-type: none"> Prendre en considération les informations, les signes qui vont à l'encontre de nos présupposés, de nos hypothèses initiales. Capacité d'interrogation et de remise en question. Poser des questions autres que celles qui confortent nos points de vue initiaux. Interroger le collectif pour la phase de préparation du recueil (trame de questionnaire, d'entretien).
Biais d'auto-complaisance	<ul style="list-style-type: none"> Tendance des individus à attribuer la causalité de leur réussite à leurs qualités propres (causes internes) et leurs échecs à des facteurs ne dépendant pas d'eux (causes externes). 	<ul style="list-style-type: none"> Savoir que les gens endosseront bien souvent la responsabilité de leurs réussites, mais rejettent la responsabilité de leurs échecs. Interroger également les individus sur les causes externes d'une réussite et les causes internes d'un échec.
Biais d'auto-handicap (ou de handicap intentionnel)	<ul style="list-style-type: none"> Cette stratégie consiste à mettre en avant des obstacles à sa propre réussite dans l'optique d'un échec futur, pour éviter des interprétations causales. 	<ul style="list-style-type: none"> Savoir qu'il existe des cas où les individus savent, même avant d'agir, que ce sera un échec. Penser également à interroger les individus sur les facteurs de réussite potentiels.
Effet de primauté	<ul style="list-style-type: none"> Il donne une importance exagérée à ce qui se passe au début du recueil d'information. 	<ul style="list-style-type: none"> Ne pas sous-estimer ce qui se passe entre le début et la fin du recueil d'information.
Effet de récence	<ul style="list-style-type: none"> Il donne une importance exagérée à ce qui se passe à la fin du recueil d'information. 	
Effet de halo	<ul style="list-style-type: none"> Une des caractéristiques de la personne qui détient l'information influe favorablement ou défavorablement la perception totale et globale de celui qui recueille l'information. 	<ul style="list-style-type: none"> Avoir des appréciations plus nuancées des différentes facettes de la personne.

Fixation sur l'objectif	<ul style="list-style-type: none"> Restez focalisé sur la motivation du moment (prendre des notes, poser une question, observer quelque chose en particulier) et ne pas voir tout le reste. 	<ul style="list-style-type: none"> Recueillir l'information à deux afin de limiter les biais de fixation sur l'objectif.
Stéréotypes	<ul style="list-style-type: none"> Croyances concernant les caractéristiques des membres d'un groupe, croyances qui sont généralisées à tous les membres de ce groupe. 	<ul style="list-style-type: none"> Il faut être prudent, car rien ne dit que l'individu partage les croyances et le jugement que l'on se fait de lui.
Préjugés	<ul style="list-style-type: none"> Jugement porté a priori sur autrui, pouvant être le fruit de stéréotypes. 	
Biais de similarité Biais de différence	<ul style="list-style-type: none"> Considérer plus favorablement les personnes qui nous ressemblent. Considérer plus défavorablement les personnes qui ne nous ressemblent pas. 	<ul style="list-style-type: none"> Ne pas considérer une personne seulement parce qu'elle nous ressemble ou qu'elle est différente.
Effet de soumission au groupe	<ul style="list-style-type: none"> Un individu interrogé dans un groupe peut être influencé par les réponses préalables des autres membres, même si celles-ci ne lui conviennent pas. 	<ul style="list-style-type: none"> Indiquer que les avis peuvent être divergents et qu'il n'y a pas de jugement dans les réponses formulées et que chaque individu peut avoir sa vision de la problématique.

Mais que récupère un questionnaire au juste ?

6	0	1	2	3	4	5	7	8	9
4	3	2	7	6	5	2	6	0	4
7	0	1	8	6	7	3	4	0	5
3	5	2	8	4	4	0	7	3	4
0	1	7	4	6	3	6	0	5	6
8	2	1	5	1	4	3	1	6	4
2	7	1	3	3	4	8	0	8	2
0	1	2	7	0	4	1	0	6	0
0	6	0	1	0	4	1	8	3	7
2	5	4	7	3	4	5	6	4	8
1	2	5	4	2	2	4	6	2	3
5	7	1	3	0	7	4	4	3	7
0	2	4	0	8	1	5	4	7	4
5	0	5	2	7	3	2	5	1	1
6	7	8	6	2	8	2	5	4	6
2	7	8	4	5	1	3	7	6	3
3	1	1	6	4	2	3	0	4	7
3	7	7	5	8	5	4	1	0	5
3	3	8	1	4	8	6	3	4	0
5	2	3	7	8	7	0	5	7	1
0	7	5	6	0	6	7	2	0	8
1	3	7	6	4	3	8	5	6	2
1	2	2	1	7	1	7	8	3	4
0	3	2	6	0	1	1	5	7	0
5	1	6	8	0	2	0	2	1	2
7	5	1	7	1	5	7	1	2	1
7	7	1	5	7	1	5	4	7	0
7	1	5	1	5	7	1	5	4	7

Ce que peut récolter un questionnaire.

- items ouverts

Ce sont souvent les réponses les plus riches, les plus nuancées. Permettent plutôt d'identifier les comment ? et pourquoi ? Mais c'est aussi les plus difficile à coder et donc à traiter

- item semi-ouverts (question avec modalités fixées sauf une)

Quand on ne peut pas prévoir toutes les réponses, mais seulement les plus fréquentes et que l'on a l'option autre, précisez...

- items fermés

Très adaptées aux traitements statistiques, ne permettent pas de recueillir des réponses très nuancées. Permettent plutôt d'identifier les qui ? quoi ? où ? quand ? et combien ?

Ainsi le questionnaire offre un grand nombre de possibilités mais à **questions différentes, données différentes et analyses différentes**.

Les variables quantitatives

Une variable quantitative permet de mesurer une **grandeur, une quantité**.

Elle va avoir une forme numérique, et on va pouvoir calculer une somme, une différence, une moyenne...

Une variable quantitative peut elle même appartenir à deux catégories différentes:

- **discrète** : elle correspond alors à un nombre fini de valeurs possibles. Exemple : un nombre de logements, d'enfants, un âge...
- **continue** : correspond à priori, toutes les valeurs possibles. Exemple : une taille, une surface, un salaire, ...

Les variables qualitatives

Une variable qualitative indique des **caractéristiques qui ne sont pas des quantités**. Les différentes valeurs que peut prendre cette variable sont appelées les **catégories** ou **modalités**. Elle peut être :

- **ordonnée** : elle va exprimer un ordre. Exemple : “un peu - beaucoup - passionément”
- **non ordonnée** : elle va exprimer l'appartenance à différents groupes. Exemple : une couleur, un groupe sanguin...

Une variable qualitative ne permet pas de faire des calculs.

| La moyenne d'un groupe sanguin n'a aucun sens.

En revanche, on pourra faire des tableaux de fréquence.

| Combien de personnes sont A+, B-...

Les variables textuelles

Les variables textuelles sont intéressante car permettent de développer plus fortement les réponses des individus... mais plus compliqué à traiter.

Peut importe leur importance il faudra partir sur de l'analyse lexicale ou lexicométrique.

Les analyses lexicométrique peuvent s'avérer très riches et très complexes

Au commencement de la statistique...La donnée



Statistique et données

Le but de la statistique est donc d'analyser des phénomènes observés. La statistique c'est l'étude quantifiée de la variation.

Le premier enjeu de la statistique c'est l'observation du ou des phénomènes que l'on souhaite analyser et comprendre. **Car tout est quantifiable !**

C'est l'étape de **l'investigation**, de la recherche qui passe par **les enquêtes, les questionnaires, l'observation, les recensements**. Et oui un des objectif du questionnaire est bien de quantifier une information.

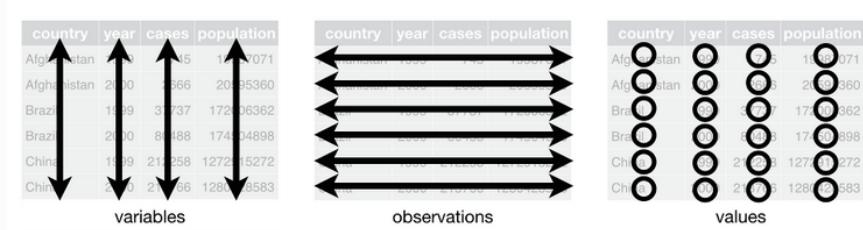
Mais pour fonctionner les observations doivent avoir une forme qui les rends interprétable, analysable et rend possible l'usage de la statistique.

C'est la création de la **base de données**.

Le produit final du questionnaire est la base de données qui met en forme tous les types de données récoltés et donc prêtes pour les différents types d'analyse.

Quelle forme pour nos données ?

A quoi ressemble une jolie base de donnée ?



Chaque colonne est une variable c'est à dire une caractéristique de nos données.

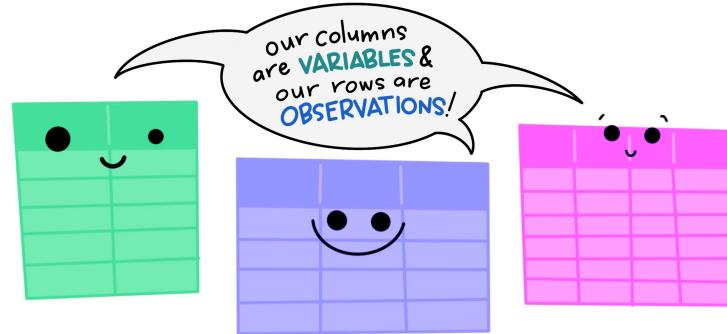
Chaque ligne est une observation, un individu.

Chaque cellule (case) une mesure unique.

Quel intérêt ? |

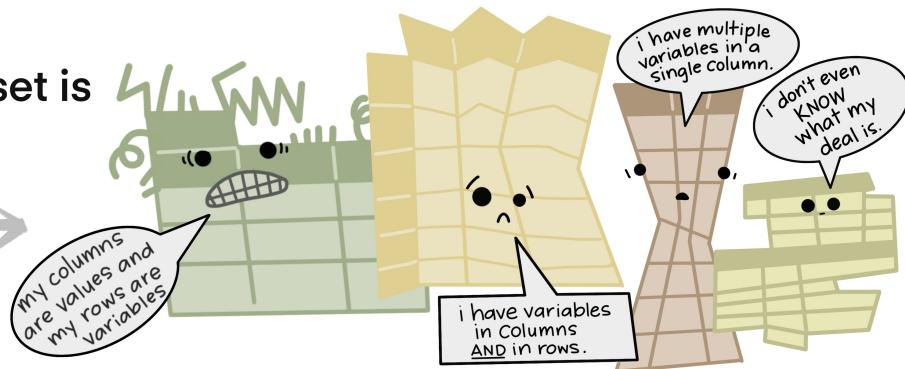
1- Mieux se représenter ces données.

The standard structure of
tidy data means that
“tidy datasets are all alike...”



“...but every messy dataset is
messy in its own way.”

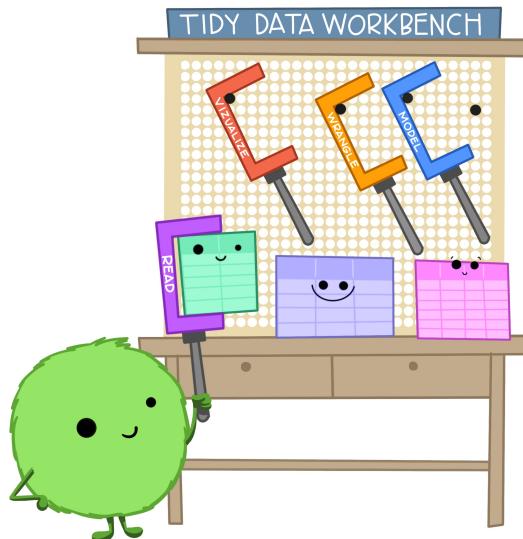
—HADLEY WICKHAM



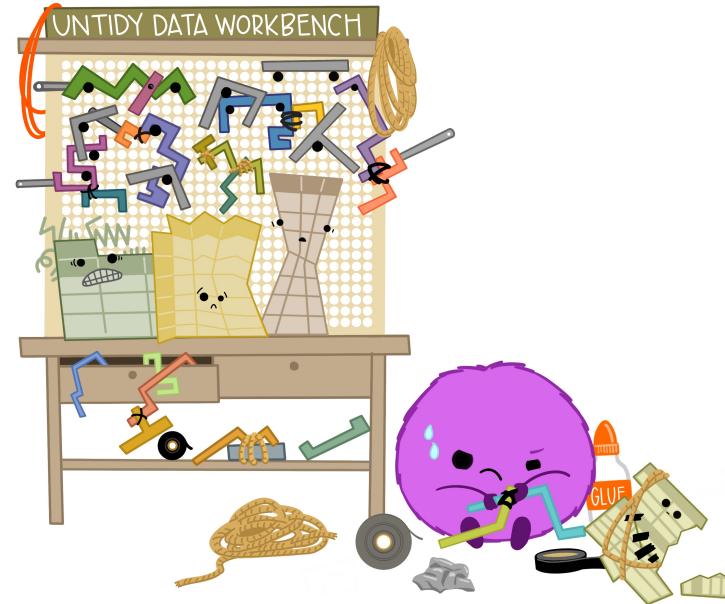
Quel intérêt ? ||

2- Pouvoir utiliser des outils et méthodes adaptés, favoriser la comparaison.

When working with tidy data,
we can use the same tools in
similar ways for different datasets...

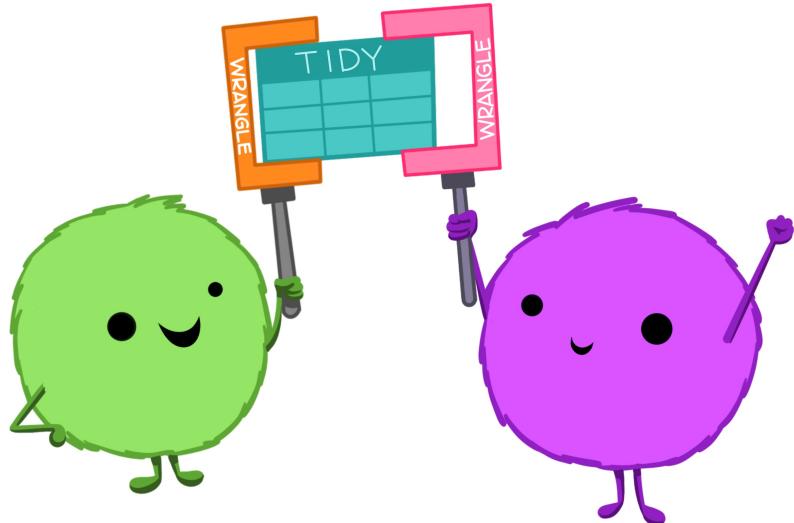


...but working with untidy data often means
reinventing the wheel with one-time
approaches that are hard to iterate or reuse.



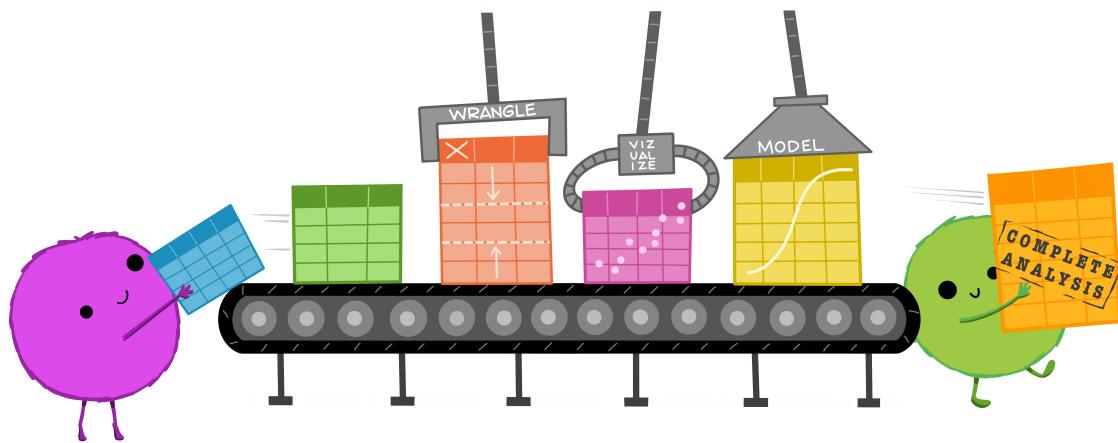
Quel intérêt ? III

3- Faciliter la collaboration avec d'autres personnes.



Quel intérêt ? IV

4- Pouvoir réaliser efficacement toutes les étapes de l'analyse de la donnée, c'est à dire de faire une chaîne de traitement.



L'analyse statistique de données



Que permet l'analyse de données ?

Petit rappel : La statistique c'est un ensemble d'outils dont le but est de nous permettre de **décrire et d'analyser des phénomènes observés** qui sont de même nature.

Ces phénomènes peuvent être sociaux, biologiques, physiques, écologiques, politiques...

Elle permet entre autre :

- De créer une information systématique qui permet donc la comparaison.
- De traiter et analyser l'information : étudier le lien entre les observations, la cause...
- De résumer et représenter l'information : faire des tableaux croisés, des graphiques...
- De vérifier la fiabilité de l'information : notamment en cas de sondage.
- De permettre le premier pas de la réutilisation de l'information dans le but d'une application.

L'intérêt de l'enquête par questionnaire c'est qu'elle va nous fournir des infos pour étudier ces différents points et donc répondre à nos hypothèses.

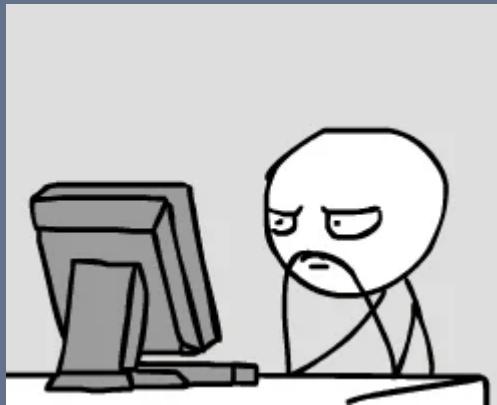
Que permet l'analyse de données ? ||

Pour simplifier on peut distinguer deux grandes branches en statistiques :

- 1- **La statistique descriptive** : qui décrit l'information de manière compréhensible et utilisable.
- 2- **La statistique inférentielle** : qui regroupe l'ensemble des méthodes qui ont pour objectif d'induire des caractéristiques d'un groupe général à partir d'un échantillon de ce groupe.

Au vue de la richesse et de l'hétérogénéité de ce que peut rapporter en terme de reponses un questionnaire on va donc avoir la chance de pouvoir se balader le long de ses 2 branches en fonction de la forme des données récoltées.

Come get some statistic !



1 er étape de l'analyse de données

La 1er étape de l'analyse va toujours être d'abord de mieux connaître nos données.

Avant de se lancer à cœur perdu dans le calcul d'indicateurs, la première étape devrait toujours être de simplement regarder et décrire ses données: c'est la statistique descriptive

Il faut à minima se poser ces questions :

-De quoi parle exactement ma base de donnée?

-Combien de variables constituent ma base de données ?

-Quel type de variables? Sont-elles quantitatives ou qualitatives ? Discrètes ou continues ? Ordonnées ou non ordonnées ?

-Combien d'observations ?

-Y a t'il des données manquantes ?

A ce stade on utilisera tous les indicateurs statistiques tel que la moyenne, la médiane, l'écart type et bien sur aussi la représentation graphique qui sera d'une aide très précieuse. Mais aussi le tableau croisé !

Analyse de données et tests statistique

Une fois que nous avons décrit nos variables et que nous connaissons mieux nos données nous pouvons passer aux tests statistiques. On parle aussi de test d'hypothèse.

Le principe de base est en réalité assez simple, il faut :

- 1- Poser un hypothèse, nommée H_0 ou hypothèse nulle,
- 2- Réaliser le test,
- 3- Voir si le résultat du test est significatif,
- 4- En fonction de la significativité du test accepter ou refuter notre hypothèse.

Des tests il en existe un très grand nombre pour répondre à peu près toutes les questions et hypothèses que l'on peut se poser.

De notre côté nous nous intéresserons seulement au test de comparaison de moyennes ou de proportions de deux échantillons indépendants ainsi qu'à l'étude de relation entre variables.

La comparaison de deux échantillons indépendants

La comparaison de moyenne ou de pourcentage entre deux échantillons est simple il s'agit simplement de savoir si deux groupes ont obtenu un score similaire ou différent.

Sur le fond cela l'est beaucoup moins car tout est toujours différents : 2 est différent de 3, 0.001 de 0.0012, 1000 de 250000

Le problème de la comparaison est de savoir si deux scores au delà de leurs différences observées ont une différence que l'on peut considérer comme significative.

Est ce que finalement l'écart que l'on note intuitivement entre deux scores reflète une différence réelle entre deux populations.

Le t de student

Le test de comparaison de moyenne le plus connu est le **t de Student**. Il en existe plusieurs versions : le test t à 1 échantillon (on compare à un chiffre défini, un standard), le test t apparié (comparé deux groupes liés, permet d'étudier une évolution avant/après) et le test t de deux échantillons indépendants qui nous intéressera ici.

Pour l'anecdote, le t de Student a été développé par William Gosset en 1908 qui travaillait chez Guinness à Dublin. Guinness lui a interdit de publier sous son vrai nom, il a alors choisi le pseudo Student.

Le test t de deux échantillons indépendants permet de comparer la moyenne de deux populations différentes mais quelques conditions indispensables :

- 1- D'abord pouvoir calculer une moyenne ! La variable que l'on étudie est continue
- 2- Les données sont normalement distribuées
- 3- La variance des deux groupes est homogène c'est à dire égale.

Le t de student II

Que faire si les conditions ne sont pas respectés ?

Pas de panique on pourra utiliser les test non paramétriques qui sont fait pour ce genre de situation (le Wilcoxon ou Mann et Whitney) !

S'il existe plus de deux groupes, inutiles de faire plusieurs comparaison deux à deux on utilisera l'Anova qui a en simplifiant les conditions similaires au t de Student.

Si les conditions ne sont pas remplis la encore on peut compter sur son équivalent non paramétrique : le Kruskall-Wallis.

L'analyse bivariée

Faire une analyse bivariée c'est étudier la relation entre deux variables

Sont-elles liées ? les valeurs de l'une influencent-elles les valeurs de l'autre ? ou sont-elles au contraire indépendantes ?

À noter qu'on va parler ici d'influence ou de lien, mais pas de relation de cause à effet. Les outils présentés permettent de visualiser ou de déterminer une relation, mais la mise en évidence de liens de causalité proprement dit est nettement plus complexe.

Là encore en fonction de vos variables on partira sur une méthode ou une autre :

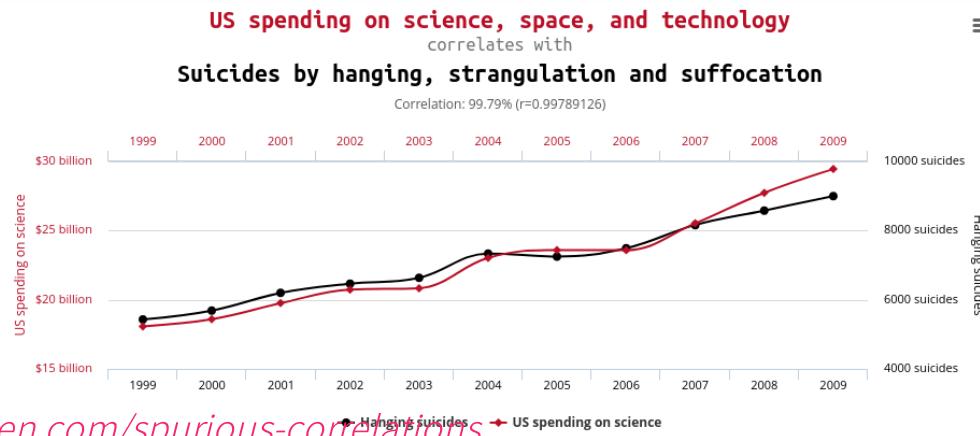
- Si on croise deux variables quantitatives : la corrélation
- Si on croise deux variables qualitatives : le chi²

La corrélation

La corrélation est une quantification de la relation entre des variables.

Le calcul du coefficient de corrélation repose sur le calcul de la covariance entre les variables. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation.

Lorsque l'on mesure une corrélation on étudie la variation commune des variables. On peut théoriquement mesurer une corrélation entre n'importe quelle variable.



La corrélation II

Il existe plusieurs méthodes de calcul pour analyser des corrélations.

Le choix d'un coefficient est déterminé par les spécificités des variables étudiées. Un mauvais choix risque de fausser vos résultats et vos interprétations.

Pour faire le bon choix, vous devez prendre en compte :

- Les types de variables (quantitative, qualitative)
- Les types de données (numérique, chaîne de caractère et facteur)
- La forme des distributions des séries statistiques analysées

La corrélation III

Les coefficients de corrélation les plus connus et utilisés en SHS sont : Le R de Bravais-Pearson / Le Rho de Spearman / Le Tau de Kendall

Si vous cherchez à étudier une relation linéaire entre deux variables quantitatives et continues et qu'au moins l'une des deux suit une distribution normale, on peut réaliser la corrélation de Pearson. Il s'agit du coefficient de référence lorsque l'on parle de corrélation.

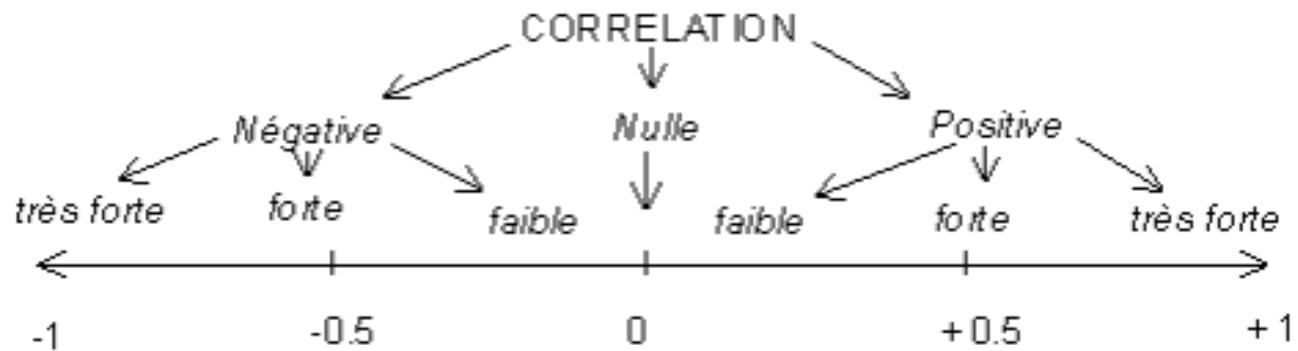
En revanche, si vos données ne suivent pas une loi normale on utilisera plutôt le Rho de Spearman ou le Tau de kendall

Dans certaines disciplines (comme par exemple en psychologie) on considère que Spearman s'utilise dans le cas de variables ordinaires où la distance “ressentie” entre les classes de nos données qualitatives est la même entre tous les intervalles (par exemple les réponses aux échelles de Likert)

Interprétation de la corrélation

Imaginons que nous étudions la corrélation entre le nombre de suicides par strangulation aux Etats-Unis et le niveau d'investissement du gouvernement américain dans la recherche.

Voici une image pour synthétiser :



Interprétation d'une corrélation

Interprétation de la corrélation II

En bref la corrélation peut :

- Etre supérieur à 0** : les variables sont associées positivement, plus les investissements dans la recherche augmentent, plus le nombre de suicides par strangulation augmente et inversement
- Etre inférieur à 0** : les variables sont associées négativement, plus les investissements dans la recherche augmentent, plus le nombre de suicides par strangulation diminue, et inversement
- Etre égale à 0** : il n'y a absolument aucun lien entre les variables, le niveau d'investissement dans la recherche n'a strictement aucune influence sur le nombre de suicides par strangulation, et inversement

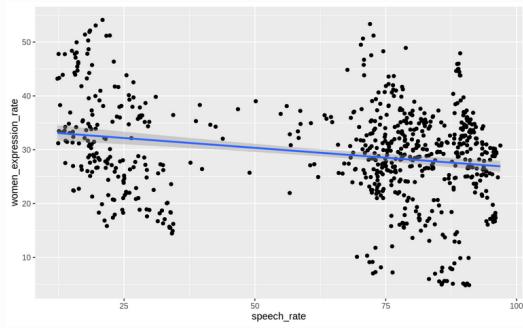
Interprétation de la corrélation III

Deux informations importantes sont donc à analyser :

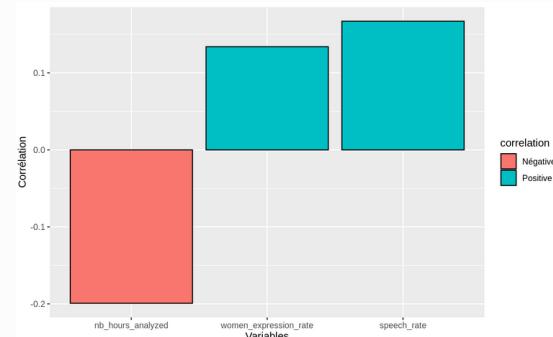
- Le sens de la relation** : la corrélation est-elle positive ou négative (coefficient supérieur ou inférieur à 0) ?
- La force de la relation** : plus la valeur du coefficient est proche de + 1 ou de - 1, plus les deux variables sont associées fortement. Au contraire, plus le coefficient est près de 0, moins les variables partagent de covariance et, donc, moins l'association est forte.

Représentation graphique de la corrélation

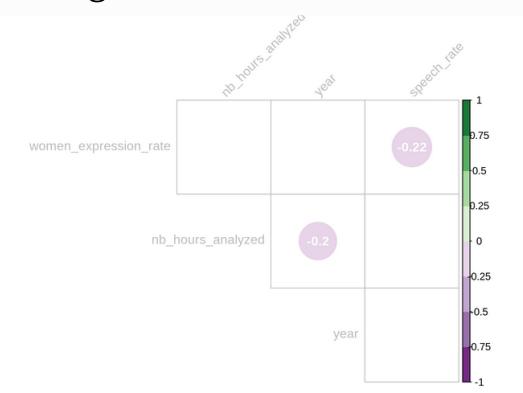
Le nuage de point :



L'histogramme :



le corrélogramme :



Le réseau:



Le Chi²

Le chi² est très intéressant car permet d'étudier des variables qualitatives qui sont souvent négligées.

C'est d'ailleurs un test qui est lui aussi souvent négligé alors qu'il peut s'avérer très puissant.

Le chi² ne sert qu'à une chose : étudier s'il y a indépendance entre les variables que nous analysons.

Nous sommes bien d'accord que par indépendance nous entendons ici que l'appartenance à tel modalité de notre 1ere variable n'a pas d'influence sur l'appartenance à tel modalité de notre seconde variable.

Le Chi² II

Pour fonctionner ce test vas se baser sur des tableaux croisés

Le TC est puissant car permet de comparer une distribution réelle (ce que nous avons observé) à une distribution théorique qui correspond aux réponses que nous aurions s'il n'y avait aucun lien entre les variables que nous croisons.

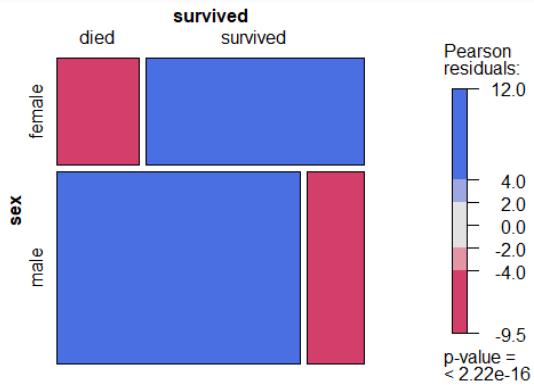
Le chi² permet de nous indiquer si la répartition de nos effectifs dans les cellules de notre tableau est significativement différente de la répartition des effectifs théoriques que nous étudions.

Si l'écart entre réel et théorique est faible alors il y a de fortes chances qu'il n'y ait pas de liens.

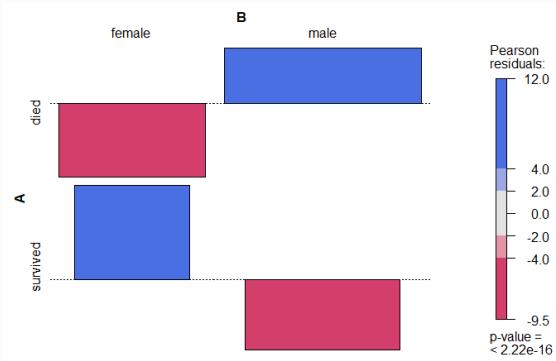
En passant sur les détails le chi² va donc chercher à comparer des effectifs théoriques aux effectifs observés. Ce qui va nous permettre d'identifier des endroits où on a plus ou moins d'observation que l'on devrait théoriquement avoir si nos deux variables n'avait strictement aucun lien : on parle des résidus.

Représentations graphiques du Chi²

Le mosaic plot classique :

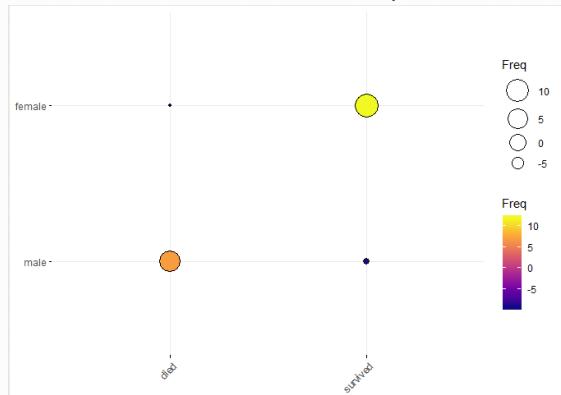


Le mosaic plot upgrade :



Le balloon plot classique :

Autre version du balloon plot:



Let's go with Jamovi!

