Correlation suite, Data Cleaning/Plotting, NHST

26/11/2021

Questions sur notions precedentes

1) Qu’est ce qu’un script et comment en creer un sur RStudio?
2) Donne trois exemples de types de variables dans R: integer (1,2,3), boolean (true, false), double (2.3455, pi)...
3) Comment peut on verifier le type des variables d’une dataframe?
4) Que faire quand on rencontre un message d’erreur sur R?

Set up pour aujourd’hui

Ouvrir R

Ouvrir le R Project du TD ou bien simplement creer un nouveau script

Importer chronolex (soit avec read.csv ou Import Dataset)

Rappel de la semaine dernière

Variance = moyenne des carrés des écarts à la moyenne (Pourquoi doit on élever les écarts au carré?)

La variance est un cas particulier de la covariance. La covariance est une mesure de la manière dont deux variables “bougent” ensemble.

La corrélation est une standardization de la covariance qui nous permet d'interpréter la relation entre deux variables de manière plus clair.

Correlation coefficient (r coefficient) compris entre [-1, 1] et n’a pas d’unite

Fonction r:

- cor() donne le coefficient de correlation, ou une matrice si vous lui passez un tableau/dataframe
- cov() donne la covariance, ou une matrice aussi

Correlation

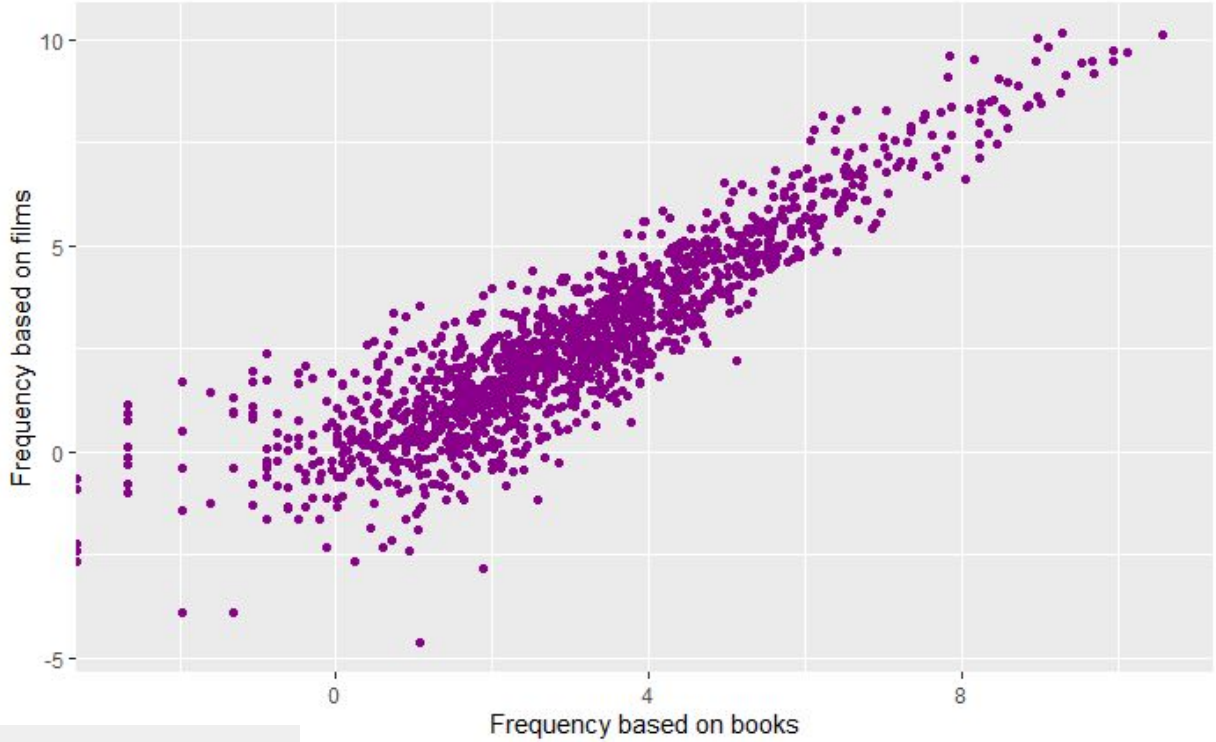
Dans R cree un scatterplot avec les variables:

- X = log(chronolex$freqbooks)
- Y = log(chronolex$freqfilms)

1) Que fait la fonction log?
2) Decrivez le plot

Correlation

Comment peut on caracteriser la relation entre freqfilms et freqbooks d'apres ce graph?




```
ggplot(chronolex, aes(log(freqbooks), log(freqfilms)))+
  geom_point(colour="darkmagenta")+
  xlab("Frequency based on books")+
  ylab("Frequency based on films")
```

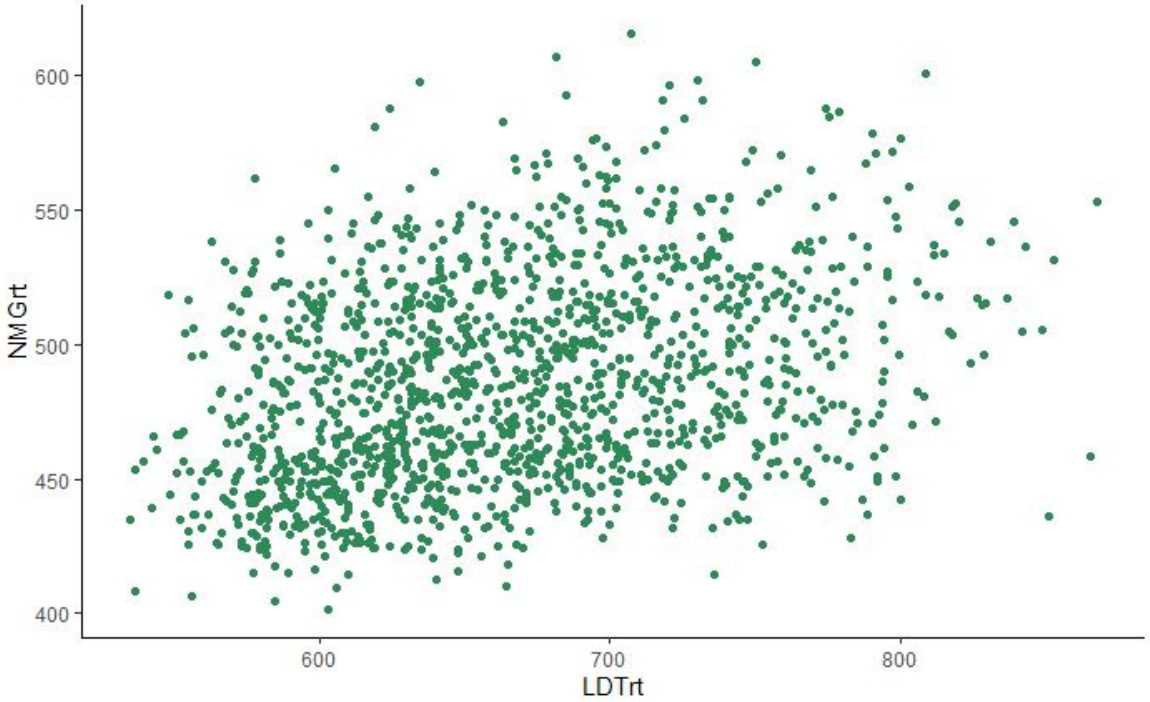
Correlation

Pareil pour

- X = NMGrt
- Y = LDTrt

Correlation

Et entre les temps de reaction en Lexical Decision task vs. ceux en Naming task?




```
ggplot(chronolex, aes(LDTrt, NMGrt))+
  geom_point(color="seagreen")+
  theme_classic()
```

Correlation

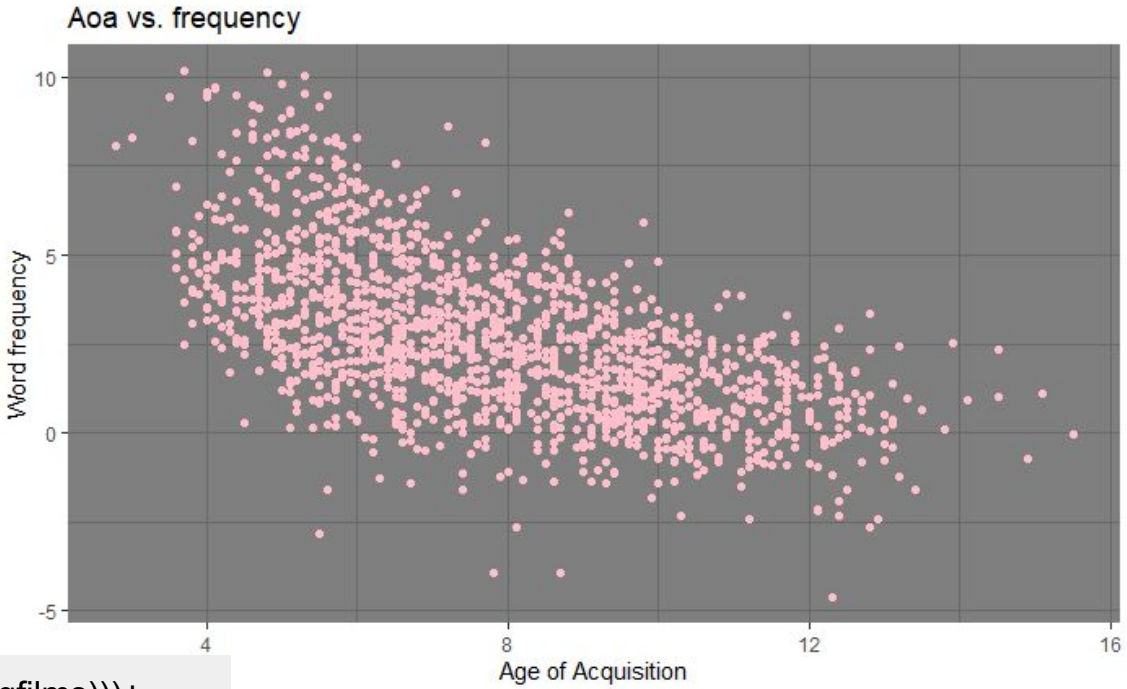
Et entre la frequence des mots (basee sur leur apparition dans les films) et l’age d’acquisition de ces mots?

X = AoA

Y = log(freqfilms)

Correlation

Et entre la frequence des mots (basee sur leur apparition dans les films) et l'age d'acquisition de ces mots?

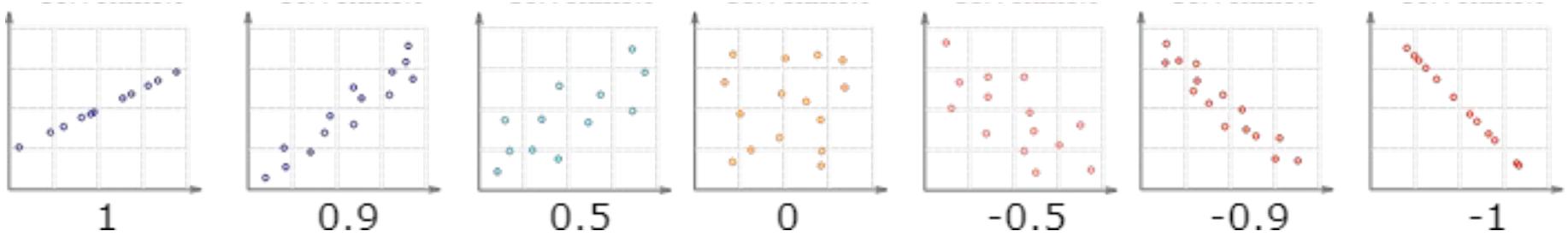



```
ggplot(chronolex, aes(AoA, log(chronolex$freqfilms)))+
  geom_point(color="pink")+
  theme_dark()+
  labs(x = "Age of Acquisition", y = "Word frequency",
title="Aoa vs. frequency")
```

Correlation - resume

Cohen a etabli les criteres suivants souvent utiliser dans la recherche:

- r = .1 → faible correlation
- r = .3 → moyenne correlation
- r = .5 → forte correlation



Plot taken from https://www.mathsisfun.com/data/correlation.html

CORRELATION NOT CAUSATION

Correlation n’est pas causalite

1. A est B sont correles
2. A cause B

⇒ Ce raisonnement est FAUX

Il existe au moins deux explications alternatives a la correlation entre A et B:

1) B cause A (i.e., le probleme de la direction de la cause a effet: une correlation ne nous permet pas de savoir dans quel sens elle serait)
2) A et B ont une cause commune (l’exemple des glaces et de la noyade)

Correlation: Influences

Certaines facteurs peuvent influencer les correlations:

- La presence d’outliers (i.e., de valeurs extremes)
- Probleme de range restriction (i.e., les valeurs que peuvent prendre nos variables)
- Echantillons heterogenes (i.e., il y a des caracteristiques de notre echantillon qui impactent la correlation de facon non negligable).

Correlation: Influences

Certaines facteurs peuvent influencer les correlations:

- La presence d'outliers (i.e., de valeurs extremes)
- **Probleme de range restriction (i.e., les valeurs que peuvent prendre nos variables)**
- Echantillons heterogenes (i.e., il y a des caracteristiques de notre echantillon qui impactent la correlation de facon non negligable).

Range restriction

C’est a dire que tu ne regardes qu’un subset des valeurs que peuvent prendre tes variables. Le coefficient *r* aura tendance a baisser sur des valeurs restreintes.

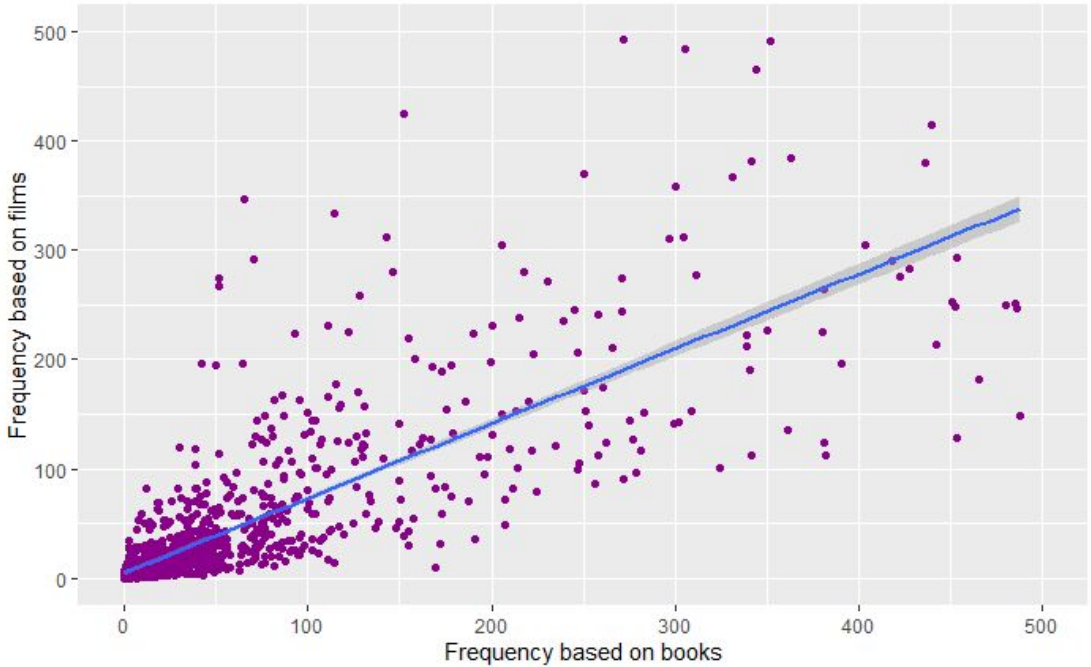
Exemple: Cree un scatterplot de freqbooks et freqfilms + quel est leur coefficient de correlation?

Range restriction

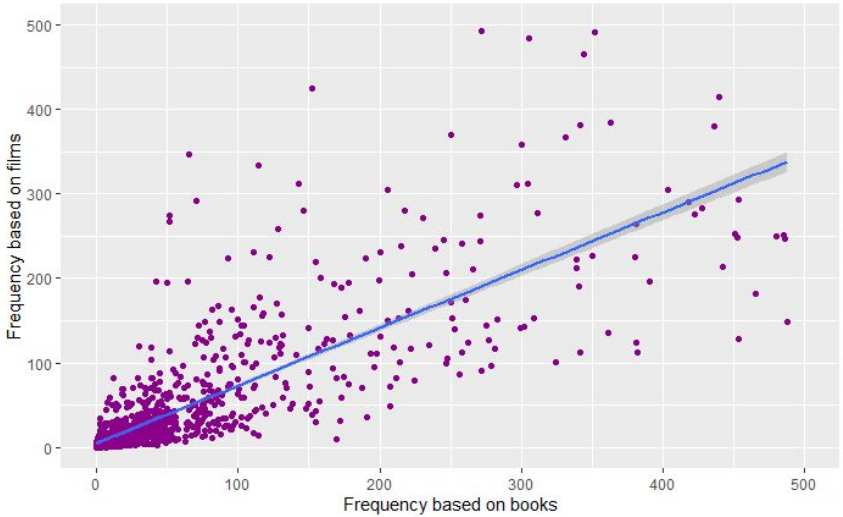
R = ~0.85

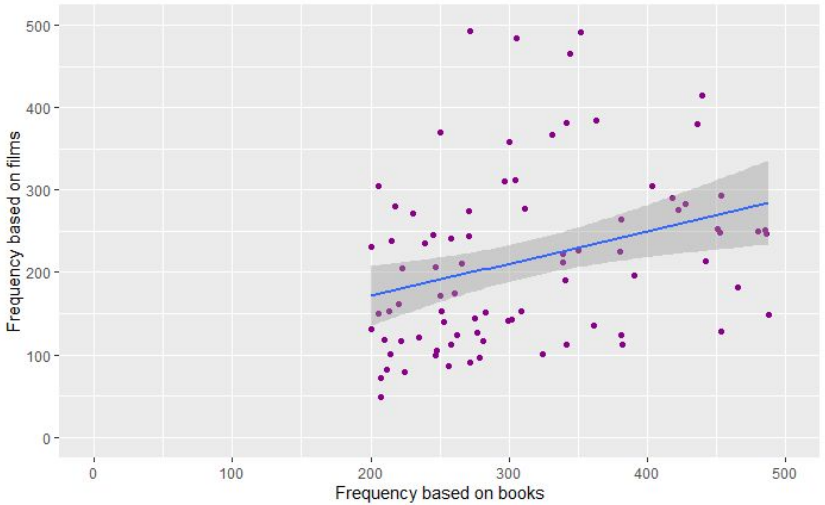
Pretty high.

Now let's say we want to look at a subset of words with frequency higher than 200



Range restriction





Quel est r maintenant?

r = 0.81 (vs. 0.85 avant). Faire attention quand on restreint l'echelle on peut changer les statistiques.

Correlation: Influences

Certaines facteurs peuvent influencer les correlations:

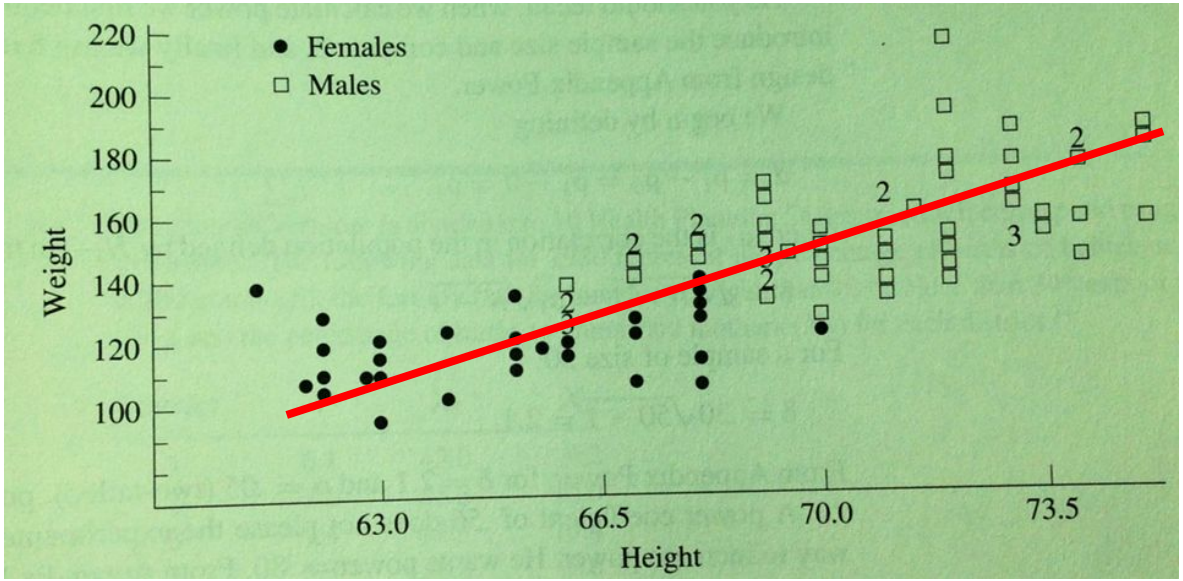
- La presence d'outliers (i.e., de valeurs extremes)
- Probleme de range restriction (i.e., les valeurs que peuvent prendre nos variables)
- **Echantillons heterogenes (i.e., il y a des caracteristiques de notre echantillon qui impactent la correlation de facon non negligable).**

Echantillon heterogene (i.e., extra variable)

Imagine un échantillon de 92 etudiants ou tu mesures le poids et la taille des gens. On obtient ca:

Ca donne un r = .78

Mais...

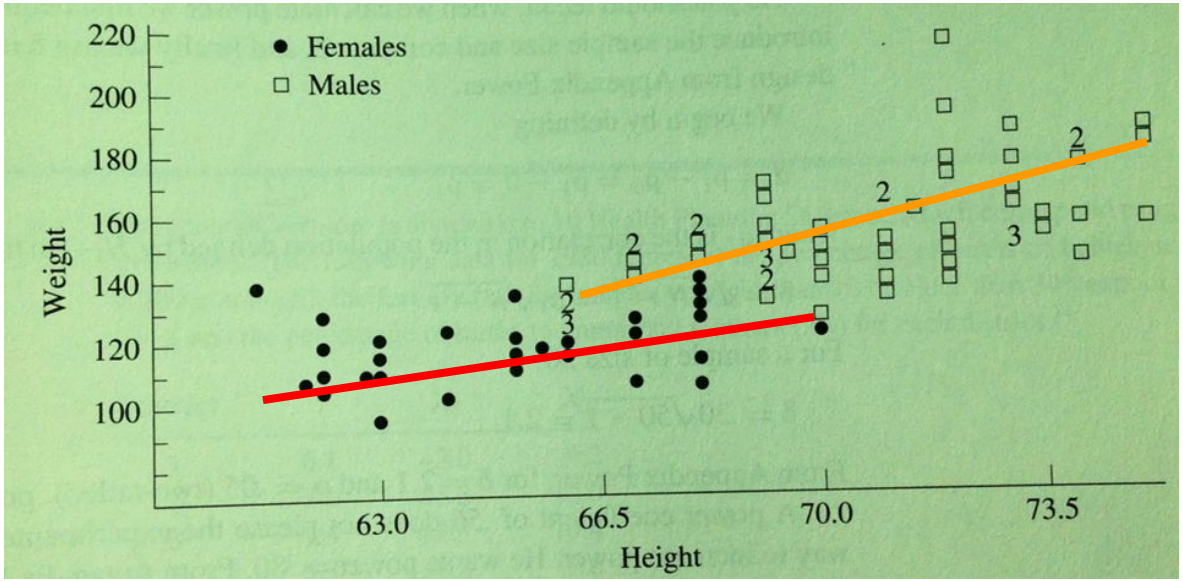


Echantillon heterogene (i.e., extra variable)

Si on considere Males & Females a part on obtient:

rMales = .60

rFemales = . 49



Correlation: Influences

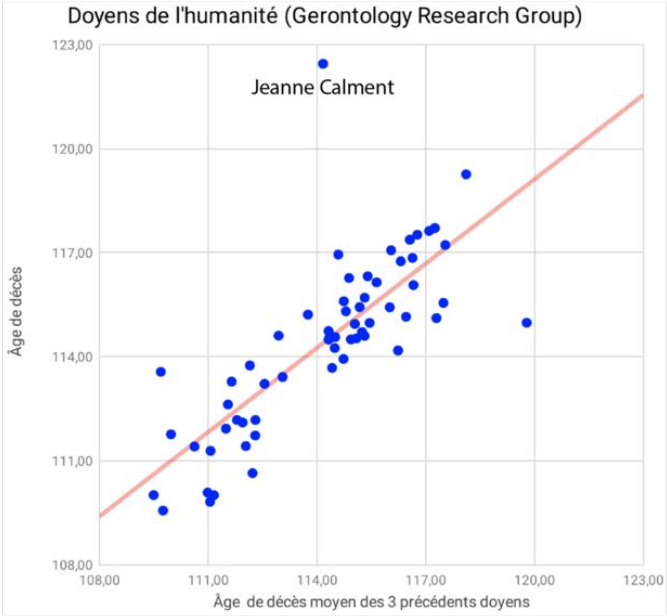
Certaines facteurs peuvent influencer les correlations:

- **La presence d'outliers (i.e., de valeurs extremes)**
- Probleme de range restriction (i.e., les valeurs que peuvent prendre nos variables)
- Echantillons heterogenes (i.e., il y a des caracteristiques de notre echantillon qui impactent la correlation de facon non negligable).

Outlier - Valeurs aberrantes/distantes

Definition: une valeur/observation qui est “loin” des autres observations pour le meme phenomene/meme experience/meme question de recherche.

Cause: probleme de mesure/d'expérimentation, erreur, fraude, au melange de population dans l’echantillon, ou simplement etre du a la variabilite du phenomene observe

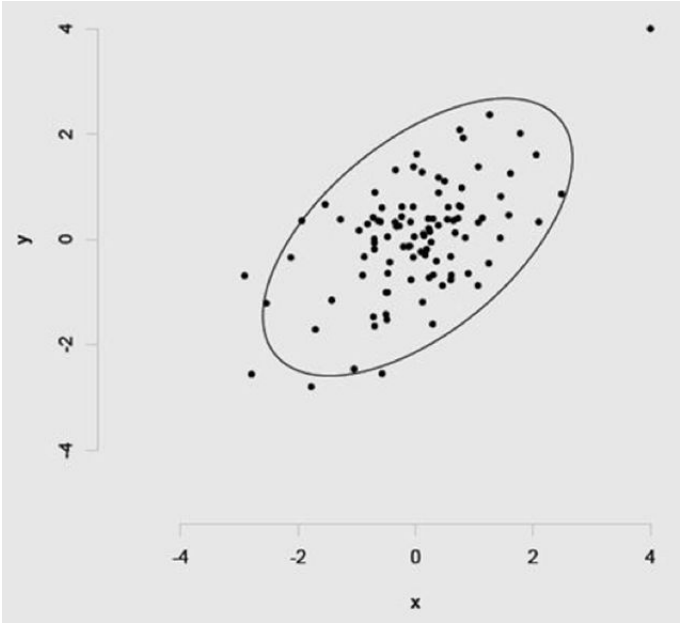


Correlation et outliers

Les outliers peuvent poser probleme parce qu'ils peuvent generer des correlations qui n'existent en fait pas. Ou les faire paraitre plus forte qu'elles ne sont.

Sans l'outlier: r = .50

Avecl'outlier: r = .57

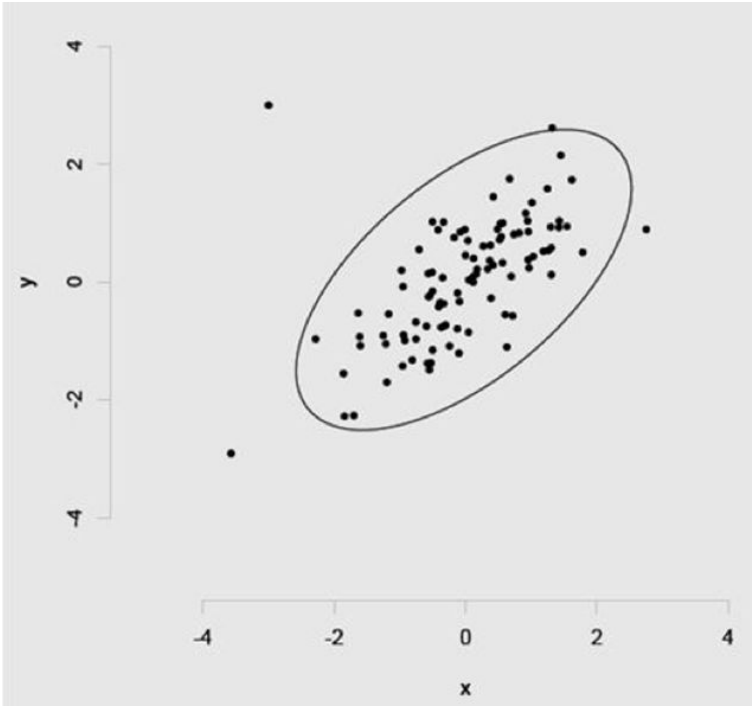


Outliers

A l’inverse ils peuvent atténuer une correlation qui est en fait sans doute assez forte

Avec l’outlier: r = .61

Sans l’outlier: r = .75

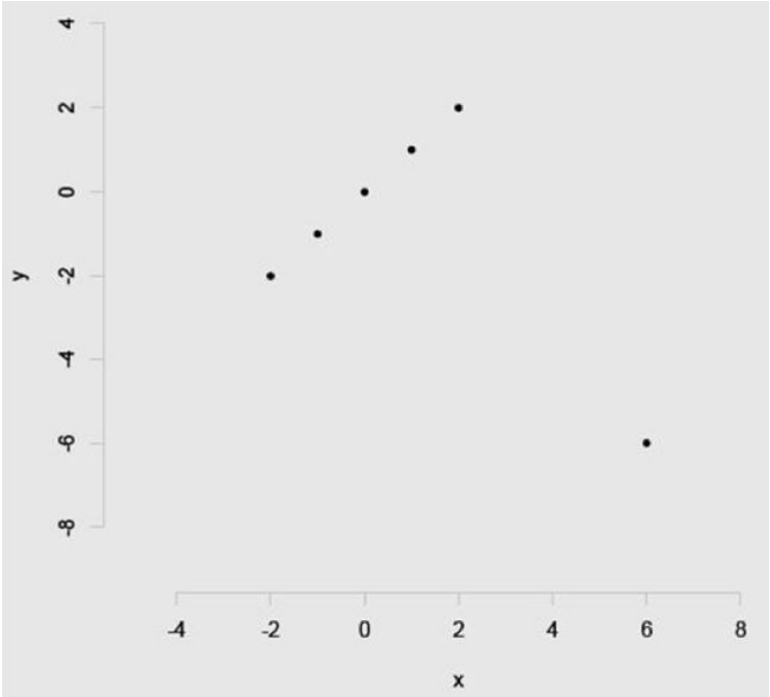


Outliers

Peuvent vraiment poser de gros soucis, allant jusqu'a change le signe du coefficient de correlation

Avec l'outlier: r = - 0.5

Sans l'outlier: r = 1



Outliers et autres problemes

Les outliers peuvent produire des resultats qui reflètent fortement un petits nombres de cas extreme plutot que la relation general observée dans les données.

Comment s’assurer que l’on travaille avec un set de donnees qui est “propre” et pas influence par des outliers ou autre soucis d’echantillonage

First rule of fight club

Toujours avoir une bonne connaissance de vos donnees, c'est a dire:

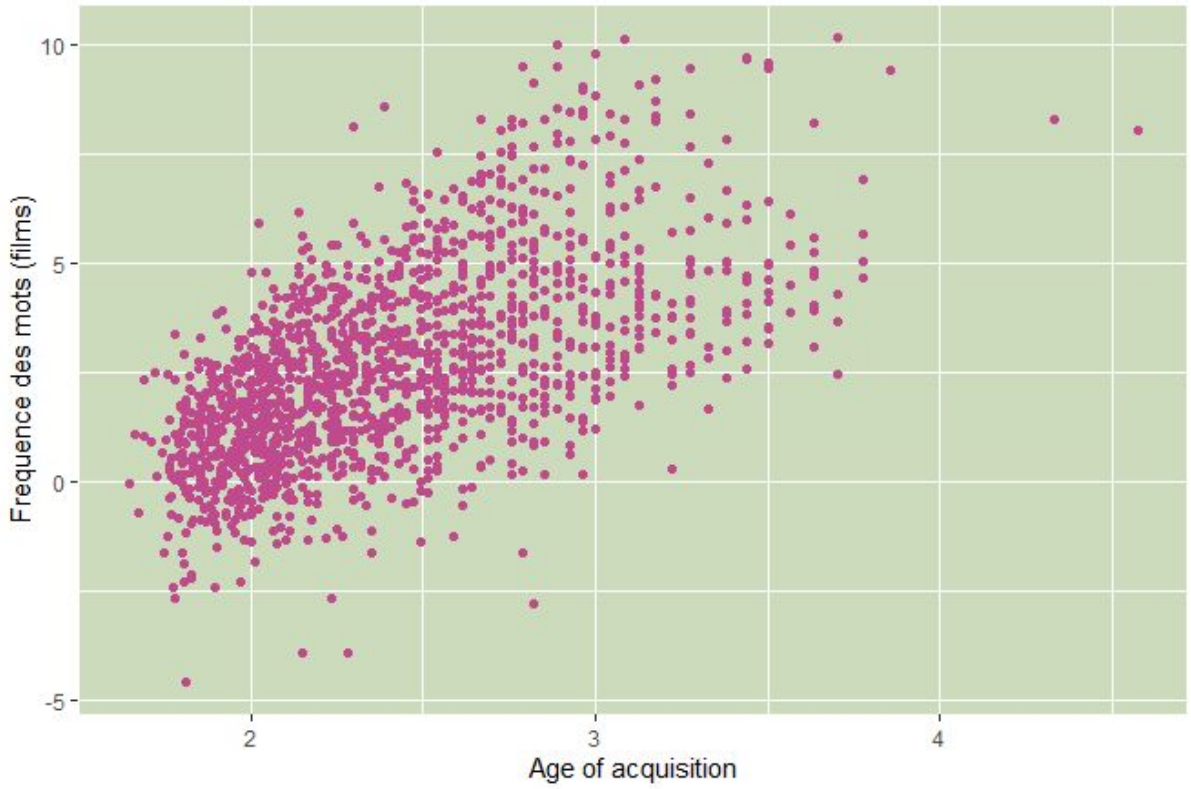
- Comprendre vos variables et en avoir une connaissance theorique (e.g., un temps de reaction de moins de 200ms pour appuyer sur un bouton est improbable et reflete probablement une erreur)
- Comment ont-elles ete mesure? (e.g., quelles sont les valeurs possibles etc)
- Regarder la distribution des variables, a-t-elle un sens?

⇒ quand vous collectez des donnees ou recuperez un dataset, il est toujours possible que des erreurs existent (e.g., quelqu'un a tape 100 au lieu de 10 etc…)

Data Cleaning

Imaginez que vous obteniez ce plot.

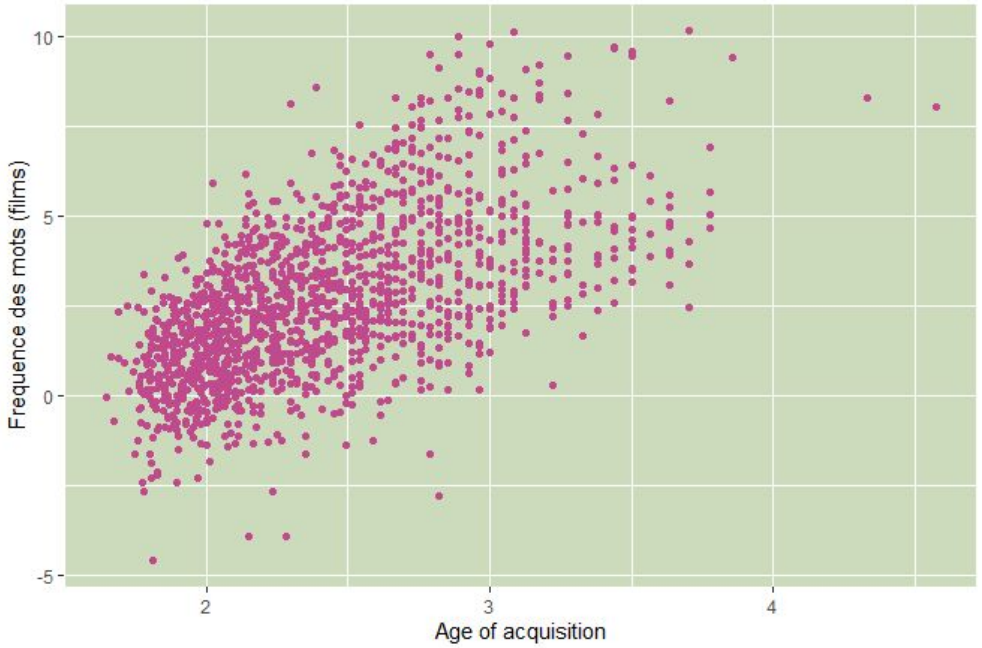
Est ce probable?



Data Cleaning

Est ce probable? Non

```
ggplot(chronolex, aes(1/AoA*10+1, log(freqfilms)))+
  geom_point(fill="#F5D5ED", colour="#C1498D" )+
  ylab("Frequence des mots (films)")+
  xlab("Age of acquisition")+
  theme(panel.background = element_rect(fill = "#CDDBBD"))
```



J'ai fais une erreur avec AoA, que j'ai transforme de facon bizarre ("accidentellement").

⇒ ne faites pas une confiance aveugle aux plots/nombres etc.. et basez vous aussi sur vos connaissances theoriques des phenomenes.

Faites des graphiques encore et toujours

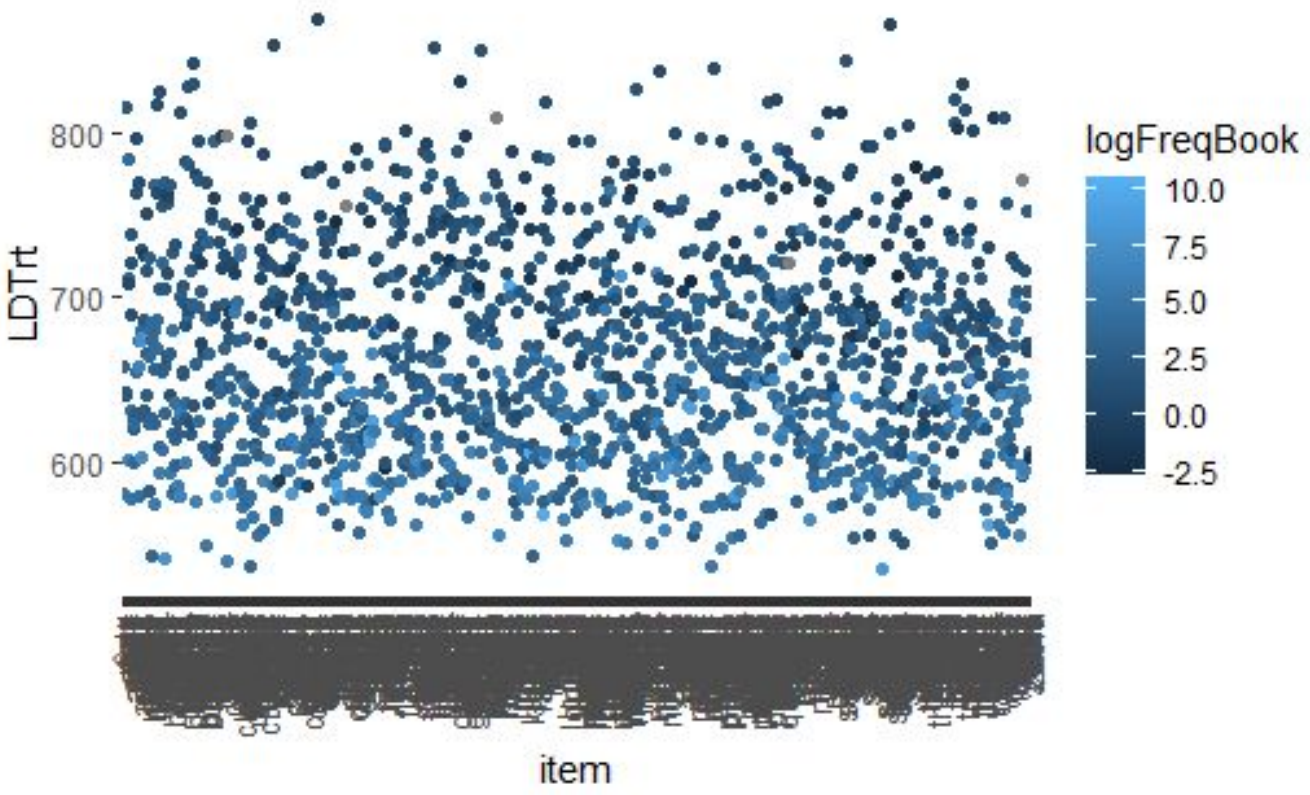
C’est le moyen le plus simple de se familiariser avec ses donnees.

Faites un scatterplot de LDTrt (temps de reaction en LD) par mot (item)

Puis faites un scatterplot de NMGrt (temps de reaction en Naming) par mot (item).

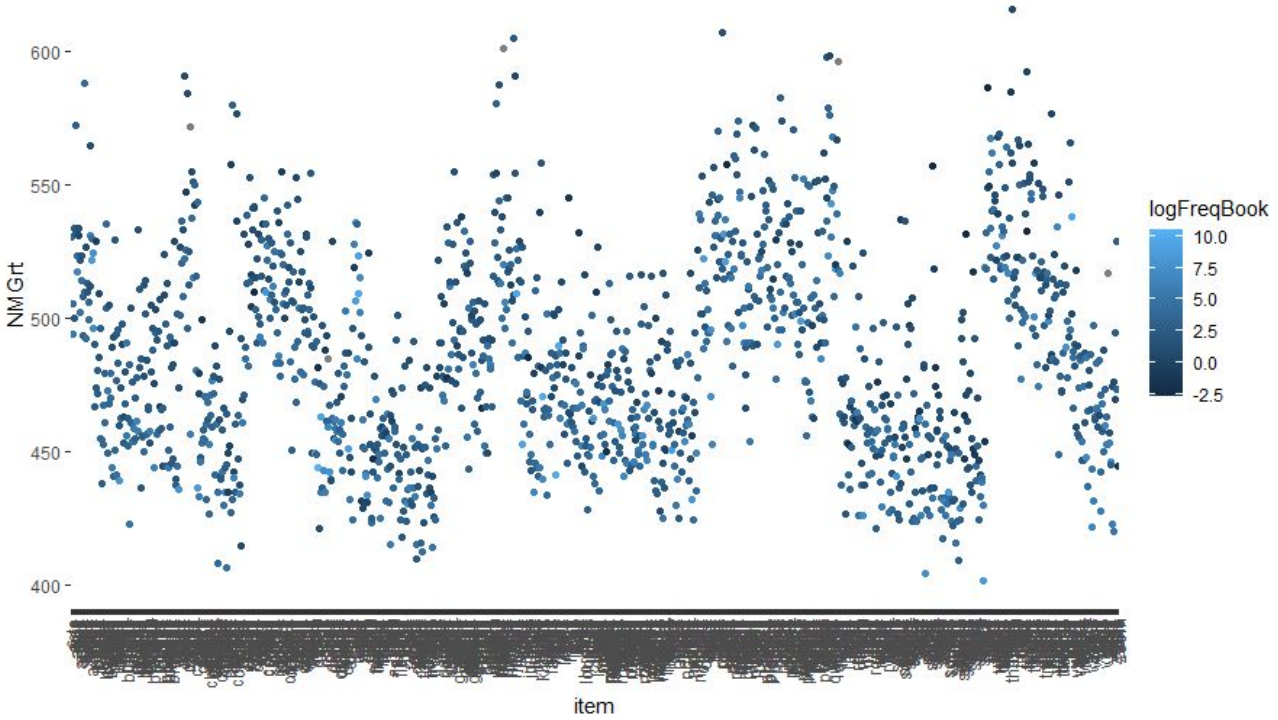
LDTrt

Plot classique pour des moyennes de RT



Faites des graphiques encore et toujours

Que peut on dire de ce plot?



Importance des graphs: Quartet d’Anscombe

Anscombe a cree 4 dataset de 11 observations chacuns.

Importez le csv anscombe.csv dans R.

Quelles sont la moyenne et l’ecart type de x pour chaque dataset?

De y?

Qu’en deduit-on?

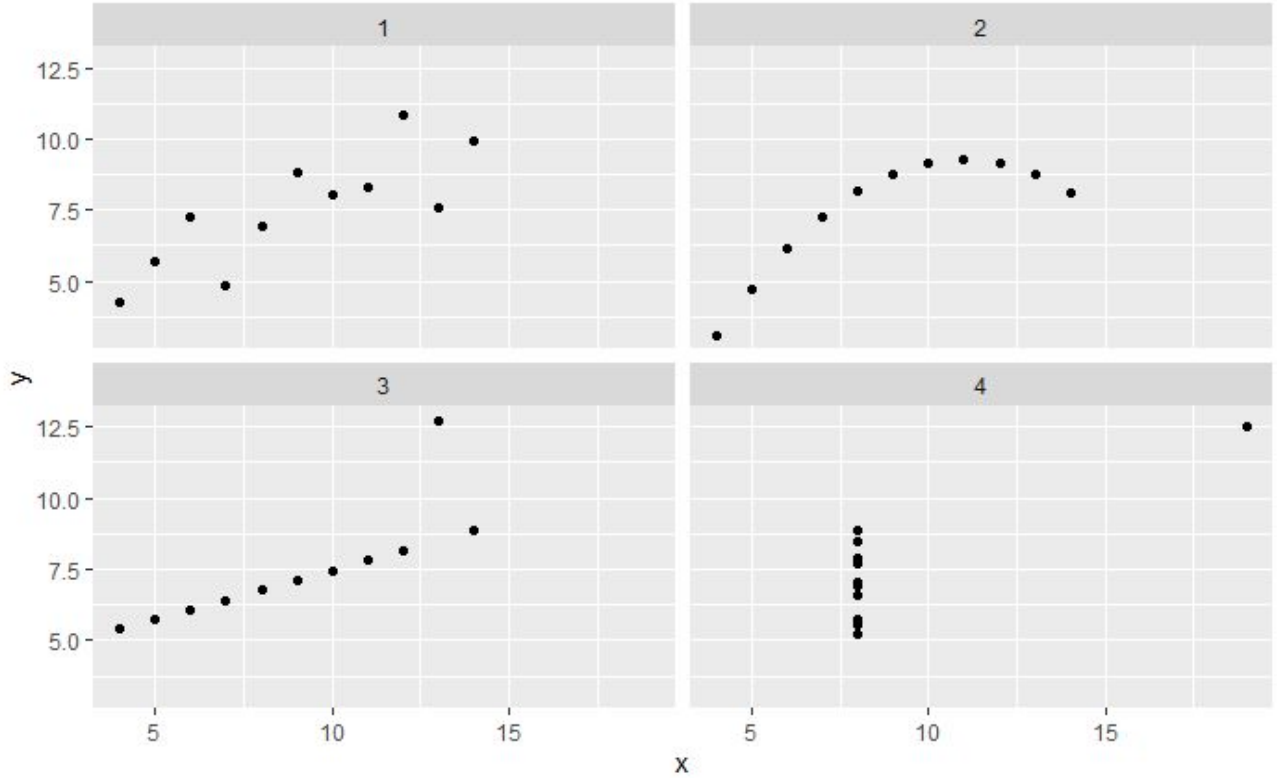
Representation graphique des donnees Anscombe

Maintenant faisons les plots de X et Y pour chaque data set.

```
ggplot(anscombe, aes(x, y))+
  geom_point()+
  facet_wrap(~Dataset)
```

Representation graphique des donnees Anscombe

Que voit on?



Data cleaning

L’inspection graphique est suffisante pour les valeurs clairement aberrantes (l’example de Jeanne Calment) mais que faire quand c’est moins evident?

Il y a une batterie de diagnostiques pour savoir si certaines valeurs sont des outliers. Ils mesurent:

- Leverage: A quel point une observation est elle inattendu en fonction des valeurs generales de cette variable?
- Discrepancy: La difference entre les valeurs observees et les valeurs predites pour la variable observee
- Influence: A quel point les coefficients statistiques (e.g., correlation) changeraient si on enlevait l’outlier.

Data cleaning

Avant de pouvoir parler de ces differents diagnostiques il faut se pencher sur la regression.

Mais avant ca on va parler de Null Hypothesis Significance Testing.

Tests d’hypotheses

- Utile lorsque l’etude de la population entiere est impossible et qu’on doit se baser sur un echantillon
- Echantillonage = marge d’erreur
- Test d’hypothese = quelle est la probabilite d’obtenir les statistiques observees?

“De façon générale, un test d’hypothèse statistique vise à déterminer si une variation observée dans un échantillon de données est compatible avec un modèle “par défaut” (l’hypothèse nulle), ou si les observations sont si improbables selon cette hypothèse nulle qu’elle doit être rejetée au profit d’une hypothèse alternative.” (https://pmarchand1.github.io/ECL7102/notes_cours/4-Tests_hypothese.html)

Fisher, Neyman & Pearson

Fisher: le test d’hypothese a pour but de refuter une hypothese donnee, pas d’en prouver une concurrente.

Neyman & Pearson: Ajoute a l’approche de Fisher.

- Deux hypotheses, celle testee (H0) et une contre-hypothese/hypothese alternative (H1). Elles doivent etre mutuellement exclusive.
- Du coup deux types d’erreurs:
  - Erreur de type I ou Faux positif (α) : rejeter H0 a tort
  - Erreur de type II ou Faux negatif (β) : conserver H0 a tort.

Elements d’un test d’hypothese

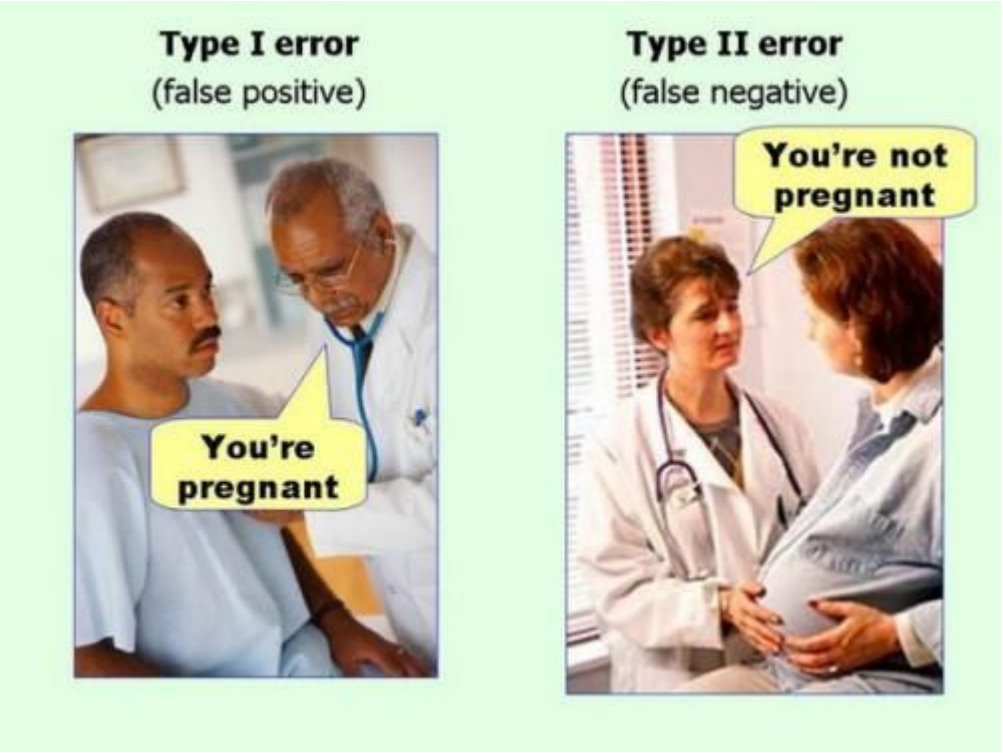
Aujourd’hui on utilise en gros un mix de Fisher, Neyman et Pearson.

Trois elements principaux:

1) Une statistique qui mesure l’ecart des observations par rapport a l’hypothese nulle (H0)
2) La distribution de cette statistique sous l’hypothese nulle (H0 - une distribution normale par exemple)
3) Un seuil de significativite (alpha)

Erreur type I et type II

Exemple classique:



Type 1 et type 2

Decision/Realite	H0 vraie	H0 fausse
Rejet H0	Erreur 1 (alpha)	Bonne decision
Non Rejet H0	Bonne decision	Erreur 2 (beta)

α = probabilite de faire une erreur de type 1 si H0 est vraie. Souvent fixe a 0.05

β = probabilite de faire une erreur de type 2. Mais on s'interesse plus a 1 - β , i.e., la probabilite de rejeter H0 lorsqu'elle est fausse (i.e., de detecter un effet significatif quand il y a en un). C'est la **puissance** du test.

Test d'hypotheses

Donc pour recapituler, en general:

1) Les chercheur.euses ont une hypothese de depart, par exemple "les mots frequents sont reconnus plus rapidement" (i.e., moyenne de temps de reaction plus grande pour les mots non frequent). C'est H1
2) H0 est donc: il n'y a pas de difference de temps de reaction entre les mots du fait de la frequence (i.e., les moyennes seront les memes)
3) Alpha est a 0.05
4) On doit choisir un test statistique en fonction de nos hypotheses sur la distribution des observations (normale, binomiale, etc…): Anova, Chi-square, regression, logistic regression….
5) On obtient la fameuse *p-value.*

P-value

ATTENTION: Tres souvent mal comprise.

La p-value represente la probabilite que *sachant l'hypothese nulle* (i.e., H0 est vraie) on obtienne les observations que l'on a.

P-value = P(Data|H0)

Donc si la valeur est petite, cela veut dire qu'en partant du principe que H0 est vraie, la probabilite d'observe les donnees que l'on a est faible ⇒ donc H0 est surement fausse.

Correlation & Effect size/Taille de l’effet

Attention: important de ne pas confondre la *significativite* et l’*importance* d’un effet. On peut avoir un resultat qui est statistiquement significatif, mais pas forcement important.

Apparaît donc le concept de taille de l’effet: i.e., a quel degre un phenomene est il présent dans la population (Cohen, 1988).

Correlation & Effect size/Taille de l’effet

La taille de l’effet vous permet d’etablir le percentage de la variance dans vos donnees qui est expliquee par vos variables.

Pour la correlation de Pearson c’est pratique: l’effect size est r au carre

Donc si r = 0.3, r^2 = .09 ⇒ une correlation de .3 signifie que moins de 10% de la variance d’une variable est expliquée par l’autre variable (meme si souvenez vous que selon Cohen en recherche 0.3 est considéré comme une corrélation moyenne).

Semaine prochaine

Exercice a rendre (je posterai le TD cette apres-midi ou demain, desolee du delai)

→ TD 3 = note sur 3 donc un peu plus long.

La semaine prochaine on parlera:

- De test statistique tel que t-test, ANOVA et regression.
- D’effet significatif vs. pas (comme vu aujourd’hui)

Je pense aussi changer de dataset pour voir un peu de nouvelles variables.