

# L1. Les statistiques: un outil pour l'argumentation

Olivier Godechot

# Comment enseigner les statistiques?

- Approche “fondamentaliste”
  - Fondements mathématiques → Théorie statistique → Pratique statistique
- Approche “pragmatique” (tentée dans ce cours)
  - Problème empirique pratique → mobilisation d’outils statistiques → (et au besoin mobilisation de l’outil statistique)

# Un cours expérimental

- “La statistique, c’est comme le piano, ça s’apprend avec les doigts”
  - C’est en faisant des statistiques qu’on comprend, apprend, retient les notions
- Conséquences très pratiques
  - Comprendre comment les gens pratiquent les statistiques
    - Lectures obligatoires. Participation (20%)
    - Trois intervenants extérieurs : Clément de Chaisemartin, Claire Lemerrier et Jan Rovny
  - Apprendre en classe à pratiquer les traitements statistiques: 1 séance sur deux environ
    - Apporter un ordinateur portable avec un tableur (de préférence excel) et si possible installer R et R Studio
  - Réalisation d’un mémoire type 2-pages ou 4-pages (40%). (Groupes de deux tolérés)
    - *Ex : Insee-Premières, Population et sociétés*, Pew Research Center Short Brief (Ex. <https://pewrsr.ch/3HifVox>)
- Mais aussi un apprentissage classique des notions : examen (40%)

# De l'argumentation aux statistiques

- Argumentation
  - Figures de style (analogie)
  - Manières de conduire le raisonnement
  - Mais aussi exemples empiriques
    - Et des contre-exemples empiriques
  - Éléments empiriques aident-ils à croire dans la thèse ?



Sources: Wikipedia

# Des singularités aux statistiques

- Parfois un exemple singulier suffit à faire preuve
  - Brebis Dolly, preuve du clonage
- Ou la réalité reste singulière
  - Il n'y a qu'un seul Napoléon
- Mais succession d'exemples et de contre-exemples :  
irrésolution
  - « Les enfants de cette ZUP ne font pas d'études supérieures, donc ils sont défavorisés »
  - « Oui, mais j'en connais un enfant de classe populaire qui a fait l'ENA, donc avec un peu de volonté on peut »  
(Conversation fictive de dîner).
- ⇒ Passer à la logique statistique



Sources: Wikipedia

# Ex : James Lind et le scorbut (1747)

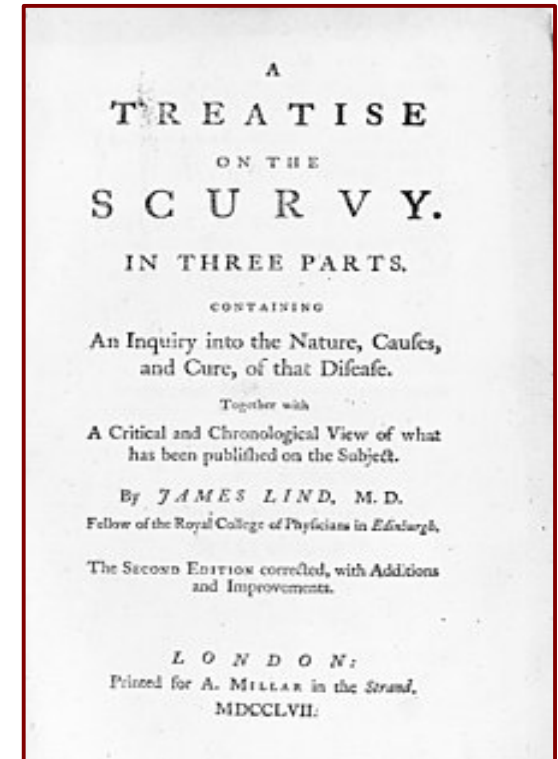


“On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury at sea. **Their cases were as similar as I could have them.** They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; **and had one diet common to all**, viz. water gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times light puddings, boiled biscuit with sugar, etc., and for supper, barley and raisins, rice and currants, sago and wine or the like. **Two were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir vitriol three times a day . . . Two others took two spoonfuls of vinegar three times a day . . . Two of the worst patients were put on a course of sea-water . . . Two others had each two oranges and one lemon given them every day . . . The two remaining patients, took . . . an electary recommended by a hospital surgeon . . .** The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them, being at the end of six days fit for duty . . . The other was the best recovered of any in his condition; and . . . was appointed to attend the rest of the sick. Next to the oranges, I thought the cyder had the best effects . . .”

# La logique statistique prolonge et systématise l'analyse empirique singulière

- 12 marins avec le scorbut. Divisé en 6 groupes
  - Même diète + supplémentation
    - cidre
    - Acide Sulfurique
    - Vinaigre
    - Herbes et décoction aux épices
    - Eaux de mer
    - Oranges et citron → guérison précoce
- Premier essai clinique (pas totalement randomisé) ...  
Mais pas complètement pris au sérieux
- Generalisation des citrons dans la Marine Britannique...  
rôle de la vitamine C identifiée plus tard(1930)

*Introduction*



# La logique statistique est réductionniste

- Mise en équivalence des individus sur un certain nombre de dimensions
- On néglige, oublie les différences
- Réductionniste (une forme de violence dans le monde social → individu réduit à des variables). Résistance
- Réductionnisme nécessaire et efficace
  - Résumer et totaliser l'information
  - Conduire un raisonnement empiriquement fondé en sortant de l'irrésolution des exemples et contre-exemples



# Une mesure statistique. Portée & limite

- « Sondage exclusif : 81% des Français pensent que les parents ont moins d'autorité qu'avant » (La Tribune, 10 Décembre 2023)
- « Intelligence artificielle : plus d'un tiers des Français pensent qu'elle finira par diriger un jour l'humanité » (Le Figaro, 18/01/2024)



# Importance de l'échelle et/ou du point de comparaison

- Distance moyenne de la terre à la lune : 384,400 km
  - On connaît l'échelle de mesure (km). Possible de comparer intuitivement avec d'autres mesures
- La question d'opinion : pas d'échelle intuitive de mesure
  - Point de comparaison démocratique : seuil de 50 %
  - Mais discutable, car résultats très dépendants de la formulation

# Une mesure célèbre sans comparaison ?

- (Milgram 1963) : 65 % des sujets administrent le choc le plus sévère
  - Comparaison morale implicite
  - + Benchmark d'experts
  - « Fourteen Yale seniors, all psychology majors, were provided with a detailed description of the experimental situation. (...) All respondents predicted that only an insignificant minority would go through to the end of the shock series. (...) The class mean was 1.2%. »



<https://www.youtube.com/watch?v=Kzd6Ew3TraA>

*Introduction*

Verbal designation and voltage indication	Number of subjects for whom this was maximum shock
Slight Shock	
15	0
30	0
45	0
60	0
Moderate Shock	
75	0
90	0
105	0
120	0
Strong Shock	
135	0
150	0
165	0
180	0
Very Strong Shock	
195	0
210	0
225	0
240	0
Intense Shock	
255	0
270	0
285	0
300	5
Extreme Intensity Shock	
315	4
330	2
345	1
360	1
Danger: Severe Shock	
375	1
390	0
405	0
420	0
XXX	
435	0
450	26

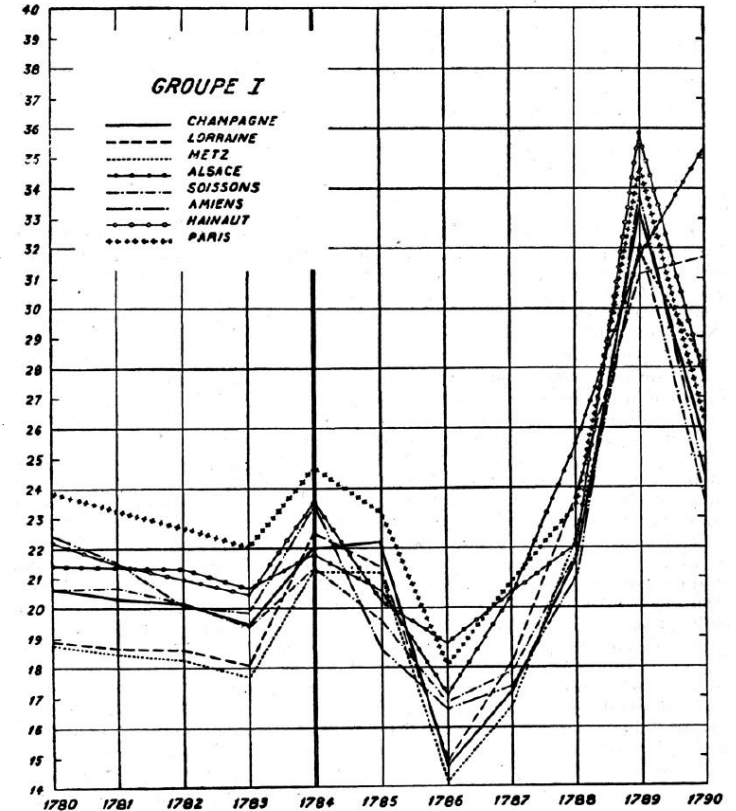
# Les comparaisons à travers les disciplines

- Sociologie : (Durkheim, 1897)
- Comparaison taux de suicide dans les cantons catholiques et protestants en Suisse

CANTONS FRANÇAIS.		CANTONS ALLEMANDS.	
Catholiques..	83 suicides par million d'habitants.	Catholiques..	87 suicides.
Protestants..	453 suicides par million d'habitants.	Protestants..	293 suicides.

# A travers les disciplines

- Histoire. (Labrousse, 1939)
- Prix du Froment de 1780 à 1790



# En économie : Le salaire minimum a-t-il tué l'emploi (Card et Krueger, 1994)

- Augmentation du salaire minimum le 1<sup>er</sup> Avril 1992 dans le New Jersey (4.25 à 5.05\$). Mais pas en Pennsylvanie.
- Examen de l'évolution de l'emploi dans 410 fast-food
- Méthode différence de différences

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	– 2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	– 0.14 (1.07)
3. Change in mean FTE employment	– 2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

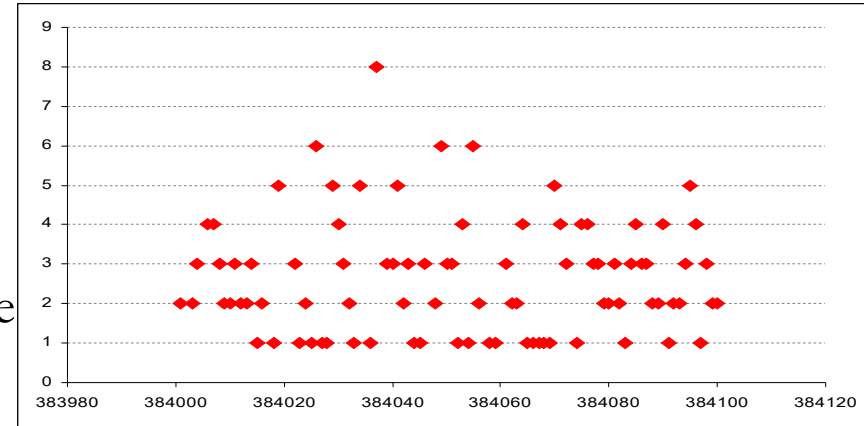
# Que comparer ? Le bon milieu

- Variables quantitatives : plein de mesures → laquelle prendre ?
  - Toutes, Un indicateur qui résume bien l'information
- En France, tous les élèves connaissent la moyenne !
- Représentation de la classe : Moyenne ou Médiane

Matières	Moy /20	Classe			Appréciations
		Min	Max	Moy	
<b>Français</b> <sup>(1)</sup> Mme PONOMAROVA	17.75	15.00	17.75	16.38	Excellent trimestre, fruit d'un travail consciencieux ! Continue à t'impliquer avec la même méthode !
<b>Histoire Géographie</b> <sup>(1)</sup> M. MARTRON	17.33	16.50	17.33	16.92	Elève très sérieux et impliqué, Rafaël a le souci de la qualité. Ses travaux sont d'une grande qualité! Ensemble tr
<b>LV1 ANGLAIS</b> <sup>(1)</sup> Mme ROZAND	17.64	16.07	17.64	16.86	Très bon ensemble. Elève sérieux et motivé. Il
<b>LV2 ALLEMAND</b> <sup>(1)</sup> Mme ROZAND	19.00	19.00	19.00	19.00	Excellent trimestre. Travail très sérieux en gran
<b>Mathématiques</b> <sup>(1)</sup> Mme BOUTIN	19.47	16.01	19.47	17.74	Excellent trimestre. Elève appliqué, précis et at
<b>Sciences Physiques</b> <sup>(1)</sup> Mme BERTRAND LEBOUV	18.14	17.71	18.14	17.93	Excellent trimestre, Rafaël se montre motivé et

# Moyennes et médianes. Un débat historique

- Quel est le bon milieu ? La moyenne ou la médiane
  - Une question d'astronomie et de mesure (Desrosières, 1993).
    - Mesure de la courbure des arcs de la Terre
  - Il existe une vraie mesure qui est « au milieu » de l'ensemble de mesures
  - Comment choisir ce « milieu » ?
  - Celui qui minimise l'erreur de mesure, l'écart entre une mesure et le milieu.
  - Comment mesure-t-on cet écart ?





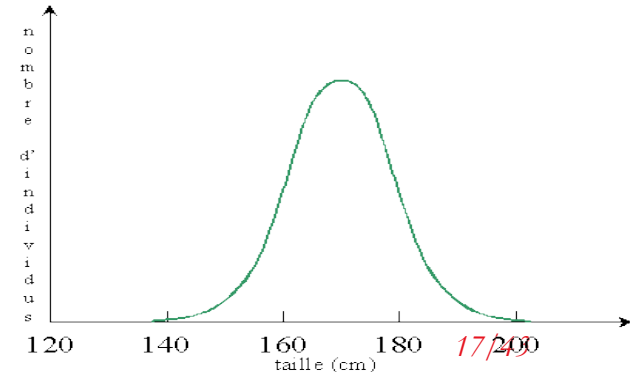
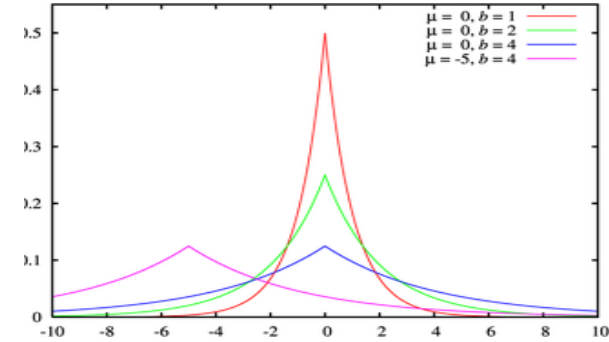
# Moyennes et médianes dans l'histoire

- La médiane Boscovich (Norme L1)

- Min valeur absolue des erreurs  $\text{Min} \sum_{i=1}^{i=n} |X_i - \text{milieu}|$
- Première loi de Laplace
- Médiane

- Synthèse Gaussienne : Legendre, Laplace, Gauss (Norme L2)

- Moindre carré des erreurs :  $\text{Min} \sum_{i=1}^{i=n} (X_i - \text{milieu})^2$
- Loi Normale de Laplace-Gauss
- Moyenne



# Des erreurs de mesure astronomique à l'homme moyen

- Quételet, 1846, *Lettres à S.A.R. le duc régnant de Saxe-Coburg et Gotha: sur la Théorie des Probabilités appliquées aux sciences morales et politiques*
- Des erreurs de mesure lors de la mesure d'une statue, à l'idée d'un « homme moyen » qui sert de moule aux individus empiriques (qui sont les erreurs de mesure)
- → La moyenne a une dimension réductionniste et essentialiste

- Mesure de :
  - Poitrine d'une statue de gladiateur
    - 1000 mesure → exact
  - Une poitrine d'une personne vivante
    - Idem
  - 1000 copies (vivantes) d'une statue

« Je vois sourire Votre Altesse; elle me dira sans doute que de pareilles assertions ne me compromettront pas, attendu qu'on ne sera pas disposé à tenter l'expérience. (...) Je vais peut-être bien l'étonner, en disant que l'expérience est toute faite. Oui vraiment, on a mesuré plus d'un millier de copies d'une statue que je n'assurerai pas être celle du gladiateur, mais qui, en tout cas, s'en éloigne peu: ces copies étaient même vivantes (...) On trouve, dans le 13<sup>e</sup> volume du journal médical d'Edimbourg, les résultats de 5738 mesures prises sur les poitrines des soldats des différents régiments écossais. »

« Il nous montre que les choses se passent absolument comme si les poitrines qui ont été mesurées avaient été modelées sur un même type, sur un même individu, idéal si l'on veut, mais dont nous pouvons saisir les proportions par une expérience suffisamment prolongée. »

# Moyennes et médianes en pratique

- La moyenne
  - *La moyenne est égale à la somme des valeurs de la variable divisée par l'effectif*

- Formalisme : 
$$\bar{X} = \sum_{i=1}^{i=n} \frac{X_i}{n}$$

- La moyenne pondérée
  - *La moyenne pondérée est égale à la somme des valeurs de la variable multipliées par leurs poids divisée par la somme des poids*

- Formalisme

$$X_{Pond} = \frac{\sum_{i=1}^{i=n} p_i X_i}{\sum_{i=1}^{i=n} p_i}$$

# Exemple

Notes			
10		Somme	52
18		Nombre de notes	4
15		Moyenne	13
9			

Poids	Notes	Poids*Note		
2	10	20	Somme Poids*Note	146
4	18	72	Somme des Poids	10
3	15	45	Moyenne	14.6
1	9	9		

# Proportions et médiane

- La proportion est une moyenne d'une variable indicatrice (ayant comme valeur 1 ou 0)

- Formalisme : 
$$p = \sum_{i=1}^{i=n} \frac{X_i}{n}$$

- Médiane :

- *La médiane est le seuil qui permet de séparer une variable ordonnée en deux groupes d'effectifs égaux*

- Formalisme : 
$$X_{\text{Médiane}} = X_{\frac{n}{2} / X_1 < \dots < X_n}$$

# Exemple 2

Notes	A			
A	1		Proportion de A	0.5
A	1			
B	0			
C	0			

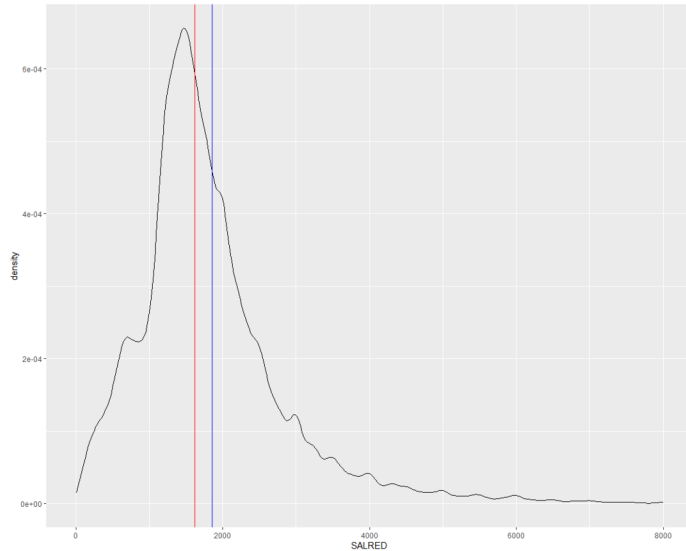
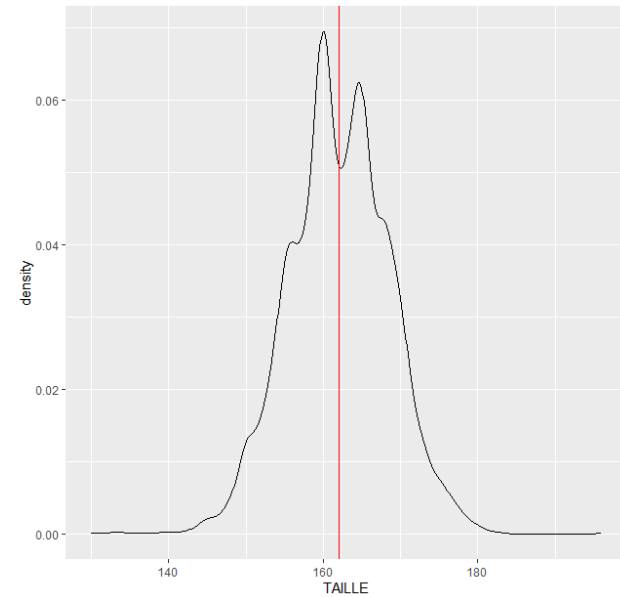
Notes	Ordre			
9	1			
10	2			
15	3		Médiane	12.5
18	4			

# Alors, moyenne ou médiane ?

- La moyenne
    - Tient compte de toute l'information
    - Permet de retrouver la somme
    - Peut être sensible à des valeurs extrêmes
    - N'est pas forcément centrée, si la distribution est très asymétrique
  - La médiane
    - N'est pas « déformée » par des valeurs extrêmes
    - Est plus robuste sur de petits échantillons
    - A une dimension « politique » (majorité)
- Il n'y a pas de « bonne solution »
- C'est une affaire, de données, de convention, de tradition

# Moyenne et médiane, une convergence/divergence instructive

- Taille des femmes
  - Forme à peu près Gaussienne
  - Moyenne : 162.03 cm
  - Médiane : 162 cm
- Salaire
  - Médiane : 1625
  - Moyenne : 1856 (au niveau du D6)
- Différence indique l'asymétrie de la distribution





# Moyenne et médiane avec le logiciel R

```
summary(ii$SALRED)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
20 1230 1625 1857 2200 37833 377566
```

```
mean(ii$SALRED,na.rm=T)
```

```
[1] 1856.831
```

```
median(ii$SALRED,na.rm=T)
```

```
[1] 1625
```

```
Hmisc::wtd.mean(ii$SALRED,na.rm=T,w=ii$extri)
```

```
[1] 1871.336
```

# Moyenne, médiane avec le logiciel Excel

- Plusieurs solutions
  - Dans une cellule
    - =Moyenne(B1:B15049)
    - =Mediane(B1:B15049)
- Dans le ruban Insertion
  - Tableau croisé dynamique
    - (Le plus sûr sans doute)
- Autre solution
  - Installer les macro complémentaires
    - Fichiers -> Options → Compléments → Compléments Excel → Atteindre → Analysis Toolpack
    - Données → Utilitaires d'analyse → Statistiques descriptives

	A	B	C	D	E	F
1		SALRED	SEXE		=moyenne(B1:B15049)	
2		35	2			
3		40	2			
4		42	1			
5		50	2			
6		50	2			
7		60	2			
8		60	2			
9		60	2			
10		64	2			
11		64	1			
12		70	2			
13		70	2			

- Tableau croisé dynamique
  - Peut-être le plus sûr

The screenshot displays the Microsoft Excel interface with the 'Insertion' ribbon active. A PivotTable is set up in the worksheet, showing the average of the 'SALRED' field. The task pane on the right, 'Champs de tableau croisé...', is open, showing the field list with 'SALRED' selected. The PivotTable is configured with 'SALRED' as the value field, and the calculation type is set to 'Moyenne de SALRED'.

	A	B	C
1			
2			
3	Moyenne de SALRED		
4	1871.989833		
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

Champs de tableau croisé...  
Choisissez les champs à inclure dans le rapport :  
SALRED  
 SALRED  
 SALREDTR  
Faites glisser les champs dans les zones voulues ci-dessous:  
Filtres  
Colonnes  
Lignes  
Valeurs  
Moyenne de SALRED  
 Différer la mise à jour de la disposition Mettre à jour

Moyenne : 1871.989833 Nb (non vides) : 2 Somme : 1871.989833

# Avec ces milieux, que comparer ?

- Des groupes
  - aussi semblables possibles sur les variables qui ne nous intéressent pas
  - Aussi différents que possibles sur les variables d'intérêt
- Exemple de Durkheim
  - Cantons Protestant / Catholiques (Au sein des cantons francophones/germanophones)
- L'expérience randomisée comme manière d'isoler une variable

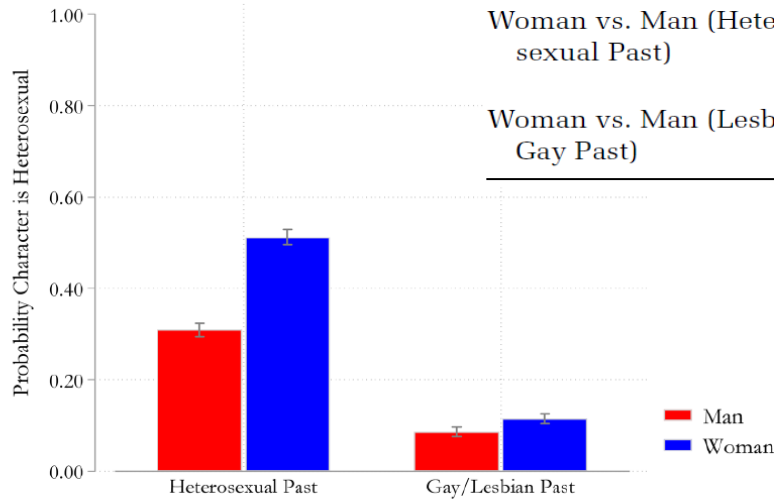
# Une expérience aléatoire sur l'orientation sexuelle (Mize, Manago, 2018)

“*Michael* is currently single but has had multiple happy relationships *with women* in the past. *Michael* has only dated *women* and one of his relationships with a *woman named Emily* lasted for over two years. The other night, *Michael* met *Matt* and felt attracted to him. At the end of the night, *Michael and Matt* went home together and had a casual sexual encounter.”

- 4 traitements:
  - (1) man with a heterosexual dating history but recent same-sex encounter,
  - (2) man with a gay dating history but recent different-sex encounter,
  - (3) woman with a heterosexual dating history but recent same-sex encounter
  - (4) woman with a gay dating history but recent different-sex encounter
- Question sur l'attribution de l'orientation sexuelle
  - how likely they thought the target character was heterosexual, bisexual, or gay/lesbian (from 0 to 100)

# Résultats

- Enquête
- Echantillon représentatif
- 2000 participants



**Table 1.** Means for Sexual Orientation Percent Guesses Based on Vignette Characteristics (Top Panel) and Fractional Response Logit Regression Results (Bottom Panel); Study 1 ( $N = 1,965$ )

	Percent Certainty That Target Character Is Listed Sexual Orientation		
	Heterosexual	Bisexual	Gay/Lesbian
Woman (Heterosexual Past)	51	41	15
Man (Heterosexual Past)	31	51	24
Woman (Lesbian Past)	11	54	40
Man (Gay Past)	9	51	47

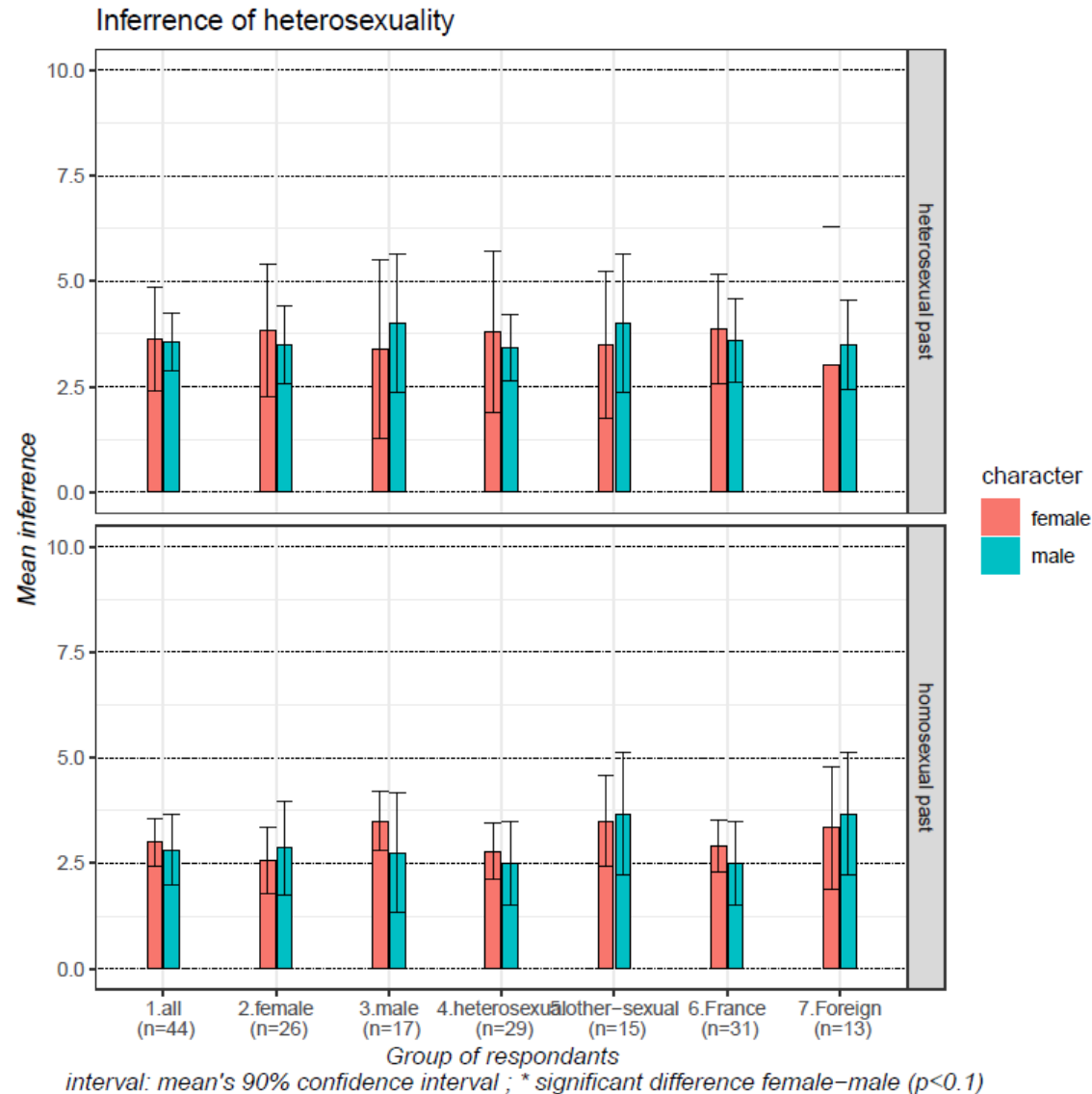
  

	Fractional Response Logit Results		
	Heterosexual	Bisexual	Gay/Lesbian
Woman vs. Man (Heterosexual Past)	.205*** (.022)	-.101*** (.023)	-.078*** (.017)
Woman vs. Man (Lesbian/Gay Past)	.029* (.014)	.027 (.023)	-.068** (.023)

**Figure 2.** Probability Rating That Vignette Character Is Heterosexual, Study 1

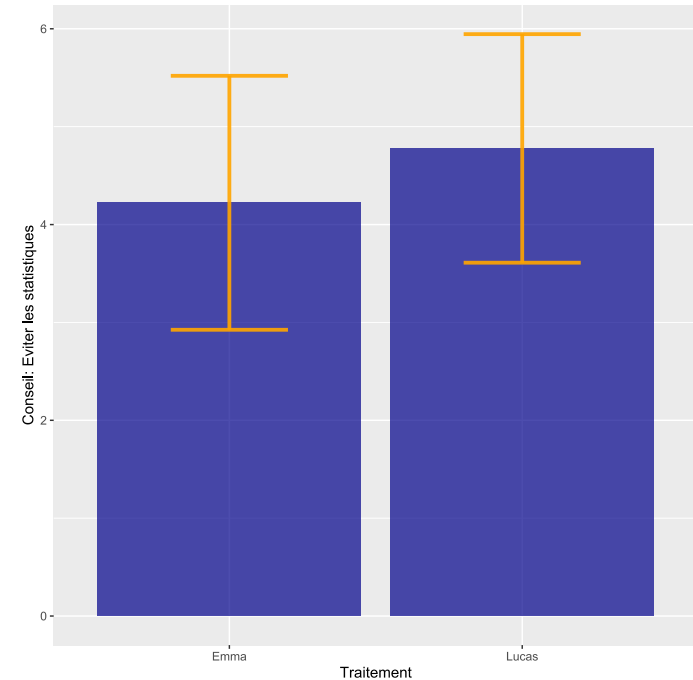
# Réplication à Sciences Po Dijon

- Pas de différence d'inférence d'hétérosexualité pour une vignette homme et femme
  - Voir inversion pour l'inférence d'homosexualité (plus pour les vignettes femmes)
- Spécificité de Sciences Po
- + Effectif faible (44)



# Une petite expérience cachée

- **Emma [Lucas]** est passionnée par la sociologie. **Elle [II]** veut devenir sociologue. Mais elle **[II]** n'est pas à l'aise avec les chiffres, tout ce qui est logiciel et mathématiques.
- Si vous deviez lui donner un conseil, où se situerait-il [sur une échelle de 1 à 10 ]?
  - 1. Pour réaliser ce projet, il faut absolument se former aux statistiques et en maîtriser parfaitement les fondements mathématiques  
2 ... 9
  - 10. Nul besoin d'apprendre les statistiques et d'autres méthodes sont disponibles dans cette discipline
- N=18 (!)





# Portée et limite des comparaisons statistiques

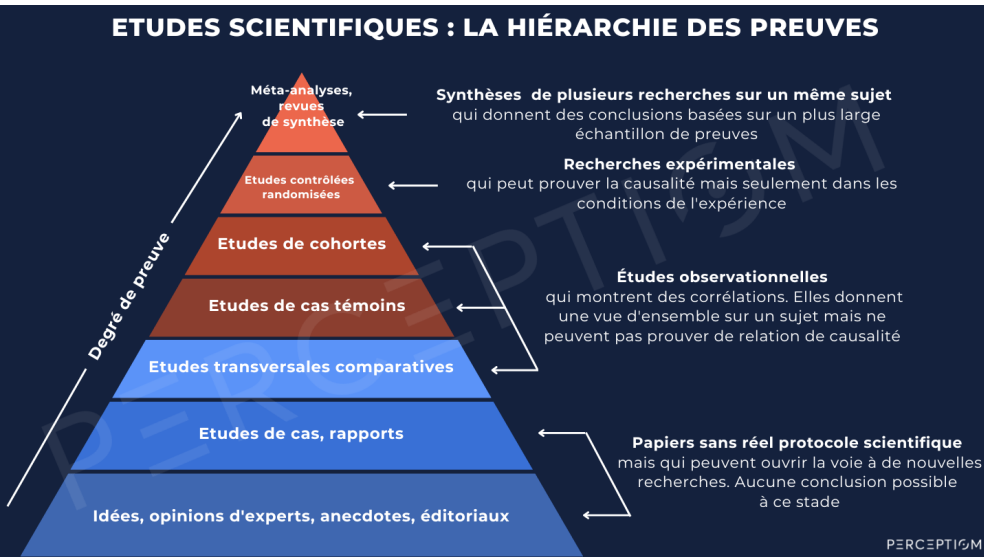
- « C'est partiel »
  - Validité externe
  - Résultats limités au domaine étudié
  - Accusation de manque de validité externe ne peut jamais échouer
    - De quoi est-ce représentatif ?
- « Corrélation n'est pas causalité »
  - Causalité inverse
  - Variable cachée

# Religion ou Urbanisme : Exemple d'une variable cachée

- Halbwachs, Maurice. 1930. *Les causes du Suicide*
- Le suicide est-il vraiment dû à la religion? Ou à une variable confondante / cachée
- Suisse. Différence d'urbanisation joue beaucoup
- Une fois dans un district urbain/industriel , Catholique Français → similaire aux protestants
- Même dans districts ruraux, les protestants sont plus connectés à la vie urbaine

	Districts agricoles	Districts mixtes	Districts industriels
Catholiques allemands.....	74	116	165
Catholiques français.....	88	151	309
Protestants allemands.....	209	228	257
Protestants français.....	414	429	378

# Statistiques et causalité. Petit guide de survie épistémologique



- Forte domination de la hiérarchie des preuves en science
- Idée qu'il y a des méthodes causales (Etudes randomisées) et d'autres non-causales.

# La causalité : plus une allégation théorique qu'un élément de méthode

- Théorie : A cause B (T). Empirie : A et B corrélés (E)
  - C'est déjà pas mal. (E) apporte du crédit à (T)
  - Logique à la Popper (1959).
    - (E) apporte du crédit à (T) jusqu'à ce que (E') falsifie (T)
  - Mais, interprétation causale de E soumis à un risque d'erreur.
- Mayo (1996). La méthode compte. Non qu'elle soit en elle-même « causale », mais permettant d'écartier des erreurs d'interprétation

# Meilleure preuve ou faits plus intéressants

- Tournant méthodologiste => accent sur la qualité de la preuve
  - Certes, si la preuve est meilleure, c'est toujours plus convaincant
  - Cela dit risque de prime à la sophistication :
    - Risques et biais des méthodes sophistiquées sousestimés et risques et biais des méthodes élémentaires (« descriptives ») surestimés
  - Risque de se focaliser sur des expériences qui représentent très imparfaitement le fonctionnement de la vie sociale
- Des faits bien prouvés ne sont pas forcément intéressants scientifiquement
  - Dans certains domaines (en particulier en sciences sociales), expérimentation très limitée ou impossible
  - Expériences naturelles limitées et partielles (car portant sur des sous population spécifiques)
    - Ex : augmentation du salaire minimum dans le New Jersey et son impact sur les fast-foods.
  - Reste des méthodes observationnelles d'analyse des covariations

# Le logiciel de statistique (R)

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for installing the `confintr` package, loading data from 'emp196qi.Rda', and reading four CSV files (INDIV161 to 164) into a data frame `ii`. It then calculates the mean and median of the `SALRED` variable and plots a density curve with vertical lines for the mean (blue) and median (red).
- Environment:** Shows the loaded objects: `herpin` (data frame, 4.4 MB, 13111 observations) and `ii` (data table, 1.3 GB, 436577 observations).
- Console:** Shows the execution of `quantile(ii$SALRED, na.rm=T, p=0.65)` and `quantile(ii$SALRED, na.rm=T, p=0.61)`, returning values like 1816, 1949, 1865, 1841, and 1850.
- Plots:** A density plot of `TAILLE` with a y-axis labeled 'density' and an x-axis labeled 'TAILLE'. A red vertical line marks the median at approximately 160.

```
4 # install.packages("confintr")
5 # load("../Data/Emploi/Emploi_1996/emp196qi.Rda")
6 # can't do it
7 ee <- emp196qi
8 rm(emp196qi)
9 table(ee$imigra)
10
11
12- # LFS 2016 ####
13 ii1 <- fread("H:/Mon Drive/Data/Emploi/Emploi_2016/Csv/INDIV161.csv")
14 ii2 <- fread("H:/Mon Drive/Data/Emploi/Emploi_2016/Csv/INDIV162.csv")
15 ii3 <- fread("H:/Mon Drive/Data/Emploi/Emploi_2016/Csv/INDIV163.csv")
16 ii4 <- fread("H:/Mon Drive/Data/Emploi/Emploi_2016/Csv/INDIV164.csv")
17
18 ii <- rbind(ii1, ii2, ii3, ii4)
19 rm(ii1, ii2, ii3, ii4)
20
21 summary(ii$SALRED)
22 mean(ii$SALRED, na.rm=T)
23 median(ii$SALRED, na.rm=T)
24 Hmisc::wt.d.mean(ii$SALRED, na.rm=T, w=ii$extri)
25
26 quantile(ii$SALRED, na.rm=T, p=0.615)
27 colnames(ii)
28
29 ggplot(ii, aes(x=SALRED)) +
30   xlim(0, 8000) +
31   geom_density() +
32   geom_vline(xintercept = mean(ii$SALRED, na.rm=T), colour="blue") +
33   geom_vline(xintercept = median(ii$SALRED, na.rm=T), colour="red")
34
35 +
36   annotate(paste("Median =", median(ii$SALRED, na.rm=T)), x=500, y=5000, colour="red", fill="red")
37
38
39
40
```

```
R 4.3.1 - D:/Olivier/Owncloud/Teaching/Causal_Analysis_in_Sociology/Applications/T01_Herpin/ <
100%
1816
> quantile(ii$SALRED, na.rm=T, p=0.65)
65%
1949
> quantile(ii$SALRED, na.rm=T, p=0.62)
62%
1865
> quantile(ii$SALRED, na.rm=T, p=0.61)
61%
1841
> quantile(ii$SALRED, na.rm=T, p=0.615)
61.5%
1850
>
```



# Avantages et inconvénients

- R. Logiciel Écriture Script –  
exécution
  - Coût :
    - apprentissage du code (long)
  - Avantage :
    - Mémorisation, réplique, adaptation
    - Nombre de fonctions statistiques immense
    - Possibilité de manipulation de données complexes
    - Bonne gestion des grosses bases de données
- Excel. Logiciel Menu/Ruban Souris
  - Avantage :
    - navigation dans les données (relativement) intuitive
    - Statistiques élémentaires (assez) faciles à mettre en œuvre
    - Pratique pour la saisie
  - Coût :
    - Fonctions statistiques très limitées
    - Recodage des variables laborieux
    - Dangereux ⇒ possibilité d'altérer les données
    - Pas de mémorisation des opérations de mise en forme. Uniquement le résultat final



# Comprendre les formules dans un tableur

- Avantage d'une formule :  
réplication par copier-coller
- La base : la Cellule.
  - Colonne -ligne. Ex.
  - Cellule relative =C3
  - Colonne absolue, ligne relative  
=\$C3
  - Colonne relative, ligne absolue  
=C\$3
  - Cellule absolue =\$C\$3
  - Switch : touche F4
- Plage : C3:D4

	A	B	C	D	E	F	G	H
1								
2			Comportement	Comportement B				
3		Hommes	120	350			=C3	350
4		Femmes	240	210			240	210
5								
6								

	A	B	C	D	E	F	G	H
1								
2			Comportement	Comportement B				
3		Hommes	120	350			=\$C3	120
4		Femmes	240	210			240	240
5								
6								
7								

# Exercice

- Avec le fichier herpin.csv calculer la taille moyenne et médiane
- En enlevant au préalable les non-réponses et refus de répondre (998,999)

# Références

- Card, David, and Alan B Krueger. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania.” *The American Economic Review* 84 (4): 772–93.
- Desrosières, Alain. 1993. *La politique des grands nombres: histoire de la raison statistique*. La découverte.
- Durkheim, Émile. 1897. *Le suicide*. F. Alcan.
- Halbwachs, Maurice. (1930) 2014. *Les causes du suicide*. Paris: PUF.
- Labrousse, C-E. 1939. “Les Prix: Prix et Structure Régionale Le Froment Dans Les Régions Françaises 1782-1790.” *Annales d'histoire Sociale (1939-1941)* 1 (4): 382–400.
- Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. University of Chicago Press.
- Milgram, Stanley. 1963. “Behavioral Study of Obedience.” *The Journal of Abnormal and Social Psychology* 67 (4): 371.
- Mize, Trenton D., and Bianca Manago. 2018. “Precarious Sexuality: How Men and Women Are Differentially Categorized for Similar Sexual Behavior.” *American Sociological Review* 83 (2): 305–30. <https://doi.org/10.1177/0003122418759544>.
- Popper, Karl. (1959) 2005. *The Logic of Scientific Discovery*. London: Routledge.