

# Recueil et traitement quantitatif des données linguistiques

M1 - Sciences du langage  
2022-2023

Olivier.Kraif@univ-grenoble-alpes.fr

Crédits : Merci à J.P. Chevrot pour ses diapos

# **1. Préambule**

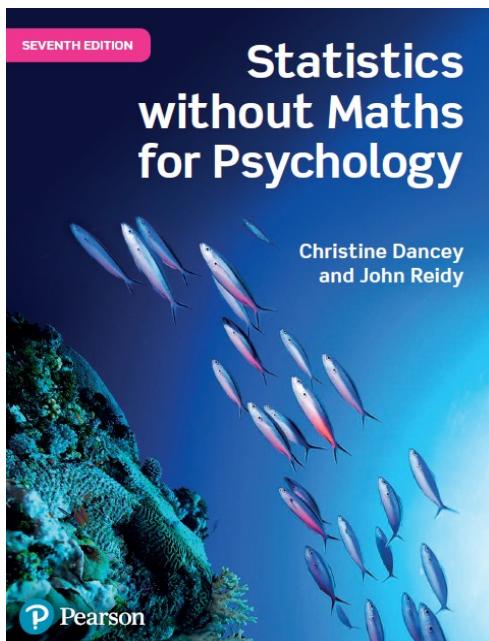
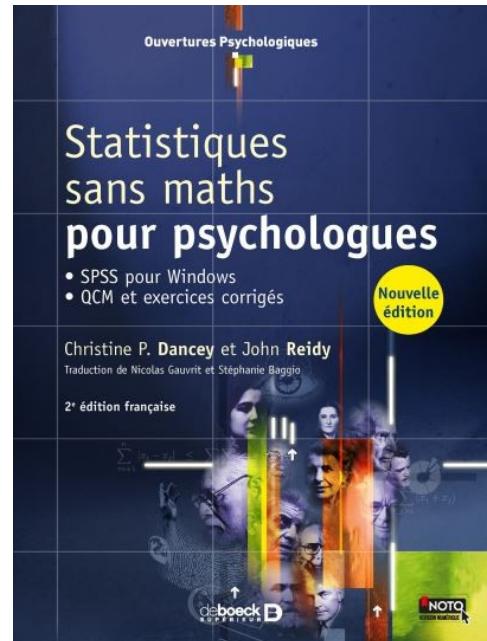
# Objectif du cours

- Maîtriser les notions de base
  - probabilité, probabilité conditionnelle
  - statistiques descriptives vs statistiques inférentielles
  - variable, population, échantillon, représentativité, biais, erreur d'échantillonnage
  - intervalle de confiance, significativité et valeur p
  - distribution normale
  - statistiques univariées et statistiques multivariées
  - indépendance, corrélations
- Décrire, analyser, représenter, interpréter des données
  - savoir utiliser des outils tels que les tableurs, Jamovi, SPSS
  - calculer les indicateurs de tendance centrale (mode, moyenne, médiane) et de dispersion (écart type, etc.)
  - savoir afficher un diagramme de dispersion, une distribution des valeurs
  - savoir reconnaître une représentation normale et déterminer numériquement la normalité
  - savoir choisir et appliquer des tests statistiques, paramétriques ou non

# Repère bibliographique

**Statistiques sans maths pour psychologues** : SPSS pour Windows QCM et exercices corrigés / Christine P. Dancey, John Reidy ; traduction de Nicolas Gauvrit  
Dancey, Christine P.; Gauvrit, Nicolas (1972-....), 2007

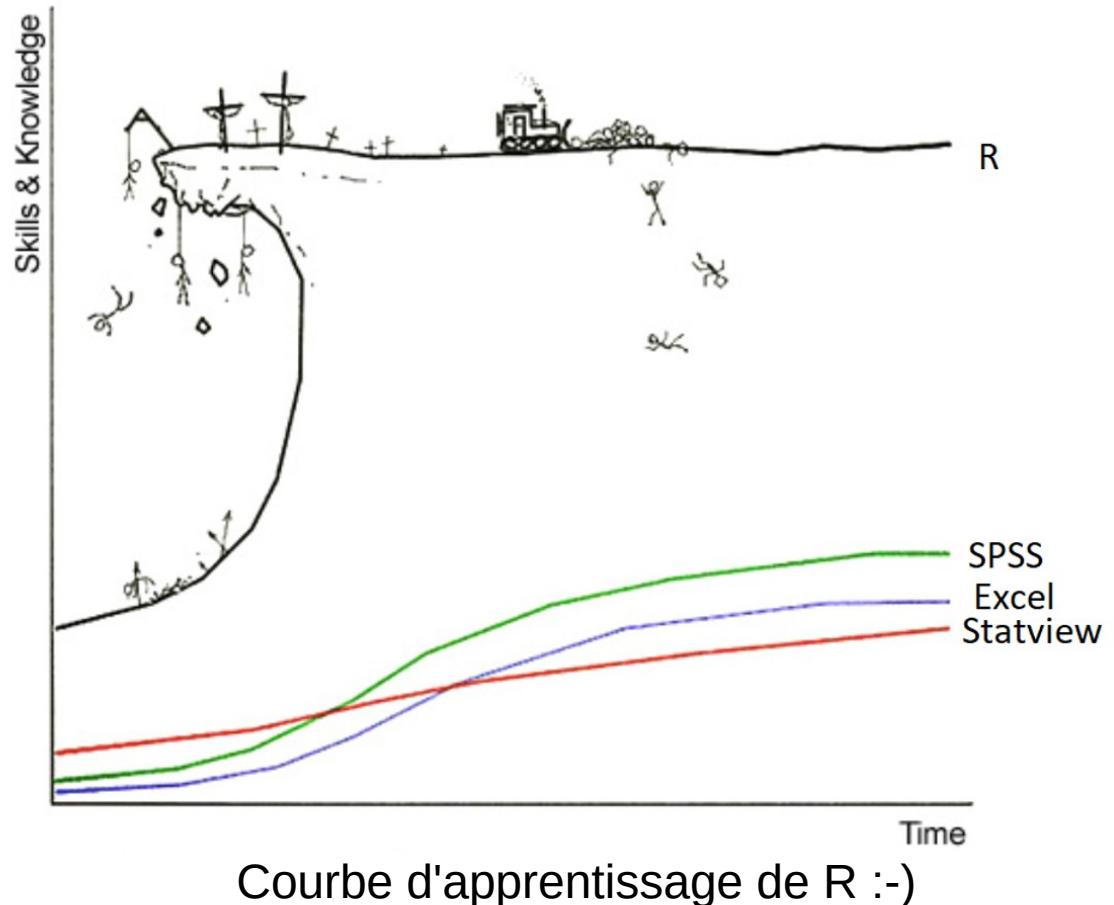
**Disponible à la bibliothèque Bulles, la BU, etc.**



- Basique, clair, sans maths, proche du cours
- Certaines analyses « modernes » ne sont pas présentées dans l'ouvrage

# Outils

- Tableurs : Excel, Calc
- Jamovi (libre), JASP
- SPSS (propriétaire)
- R : nécessite d'être à l'aise dans la programmation



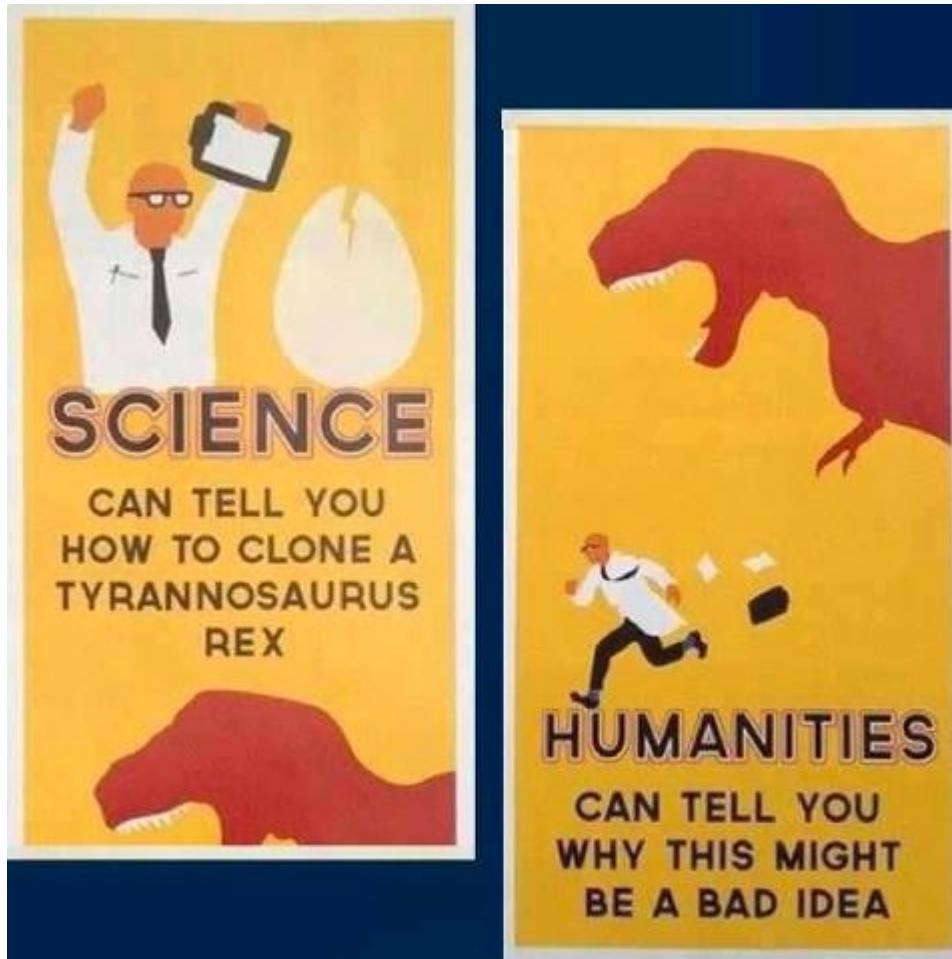
# Logiciel

Jamovi représente une alternative intéressante à SPSS. Son interface est simple et il possède des extensions pour aller plus loin.

Le manuel de Jamovi est disponible ici : <https://jmeunierp8.github.io/ManuelJamovi/>



## 2.1 Panorama des méthodes de recueil de données



<http://www.phdelirium.com/portfolio/difference-sciences-exactes-sciences-humaines/#.Y8mbnKeZPJw>

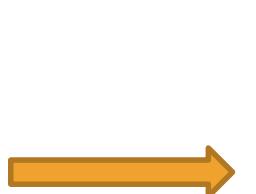
# **Les méthodes de recueil et de traitement des données**

Les méthodes de recueil et de traitement des données sont très variées.  
Chaque science à ses méthodes d'observation privilégiées:

- La sociologie : questionnaires, entretiens
- L'épidémiologie : suivi de cohortes, comparaison malades / sains
- La linguistique : l'étude des corpus
- La physique et la psychologie expérimentale : l'expérimentation.

Les méthodes pour traiter (statistiques) et représenter (graphiques, etc.) les données sont aussi très nombreuses : présentation exhaustive impossible.

Solution adoptée



Vous présenter un panorama des six types principaux de recueil des données en SHS  
Les illustrer par des recherches (réelles, publiées, domaine SDL)

# Les méthodes de recueil et de traitement des données

Un panorama organisé selon quatre critères :

La première colonne liste les types de méthode

Les colonnes 2 -5 listent les critères pour différencier les types de méthodes.

## 1/ Le critère **Structuration** : le plus compliqué !

Question de recherche - Ecriture des formes verbales en /E/ (*ais, ait, aient, é, és, ée, éés, er*) à l'école primaire et au collège. Des progrès ? Quelles erreurs subsistent ? Pourquoi ?

a/ **Situation non structurée** de recueil des données : on récupère les écrits « naturels » des élèves (prise de note, rédactions, brouillons, etc.).

b/ **Situation structurée** de recueil : on met en place une épreuve standardisée (la même pour tous)

Carnet avec une phrase à trou par page	Quentin .....son copain	Quentin .....le train
L'expérimentateur dicte les phrases en entier et les sujets remplissent le pointillé	« Quentin a aidé son copain »	« Quentin va manquer le train »
12 phrase avec cible -é de type sujet+avoir+p.passé+COD/ 12 phrases cible -er de type sujet-+devoir/aller/etc. + infinitif / contrôle nombre de syllabes et fréquence des verbes.		
6 sujets m.sg, 6 sujet m.pl., 6 sujets f.sg, 6 sujets f.pl.		
600 élèves du CE2 à la 4 <sup>ème</sup>		

# Les méthodes de recueil et de traitement des données

## 1/ le critère Structuré / non structuré : avantages et inconvénients

	<b>Avantages</b>	<b>Inconvénients</b>
Recueil structuré des données	<ul style="list-style-type: none"><li>- On cible le recueil sur le point étudié</li><li>- On contrôle mieux les variables de confusion (les biais)</li><li>- Facile à traiter <i>après</i> le recueil</li></ul>	<ul style="list-style-type: none"><li>- On perd des informations « inattendues »</li><li>- On n'est pas certain de pouvoir généraliser les résultats aux situations « naturelles »</li><li>- Long à élaborer <i>avant</i> le recueil</li></ul>
Recueil non structuré des données	<ul style="list-style-type: none"><li>- On recueille le point étudié en contexte (information riche)</li><li>- On peut généraliser les résultats car situation « naturelle »</li><li>- Recueil rapide à mettre en œuvre</li></ul>	<ul style="list-style-type: none"><li>- On doit trier les informations pertinentes et les non pertinentes</li><li>- Difficile de contrôler les variables de confusion (les biais)</li><li>- Long à traiter <i>après</i> le recueil</li></ul>

	<b>Observation Structurée</b> Mesure orientation de la tête + temps de regard		<b>Observation non structurée</b> Enregistrement vidéo	
Validité « écologique » (conditions analogues au milieu naturel)	-		+	
Validité « interne » (contrôle des biais)	+		-	

# Les méthodes de recueil et de traitement des données

## 2/ Le critère Taille

C'est la quantité de données qu'on recueille

- Le nombre de mots dans un corpus
- Le nombre de sujets observés ↗ les extrêmes :

Grande enquête	Plusieurs dizaines, centaines, milliers de sujets	- Suivi d'une cohorte d'enfants pour étudier les progrès en lecture en fonction des méthodes d'enseignement au CP
Etude de cas	Un seul sujet, suivi pendant X années ou enregistré pendant de nombreuses heures	- Enregistrement de trois enfants, 1 heure par semaine entre 2 et 6 ans, dans le milieu familial: étude de la relation entre énoncés adressés à l'enfant et progrès dans un domaine particulier (phonologie, lexique, morphosyntaxe, pragmatique)

	<b>Grande enquête</b>	<b>Etude de cas</b>
Richesse et précision des données	-	+
Généralisation des résultats	+	-

# Les méthodes de recueil et de traitement des données

## 3/ Le critère Temps

- *synchronique* signifie qu'on photographie un phénomène à un moment donné ;
- *diachronique* signifie qu'on s'intéresse aussi à son évolution dans le temps : un processus d'apprentissage, une évolution historique.

### On peut préciser le versant diachronique

Étude diachronique	On réunit des données datant de différentes périodes. La difficulté peut venir du traitement différencié des états de langue.	Ex. Corpus de roman de chevalerie du 13e au 17e siècle.
Étude longitudinale	Le même groupe de sujets ou le même sujet sont observés à différents âges	Ex : 30 sujets observés à 2 ans, puis 3 ans, puis 4 ans.
Etude transversale	Observation de différents groupes ou différents sujets d'âges différents	Ex : on observe 30 sujets de 2 ans, 30 autres sujets de 3 ans, 30 autres de 4 ans. Puis comparaison

	<b>longitudinal</b>	<b>transversal</b>
On peut voir l'évolution pour <i>chaque</i> sujet	oui	non
Les sujets sont les mêmes entre les différents temps d'observation	oui	non
Risque de perdre des sujets en route (déménagement....)	Risque	Pas de risque
L'observation répétée modifie les comportements observés (les fumeurs suivis stoppent plus le tabac !)	Risque	Pas de risque

# Les méthodes de recueil et de traitement des données

## 4/ Le critère Quantitatif/qualitatif

Correspond au type de données et à leur traitement

- *Quantitatif*: l'analyse porte sur des données numériques (fréquences, pourcentage, scores, moyennes, etc.)
- *Qualitatif*: l'analyse porte sur des catégories interprétatives non chiffrable (ex. analyse des fonctions de routines sémantico-rhétoriques). On produit des compte-rendus écrits et schémas qui font ressortir les tendances générales, la cohérence d'un phénomène, une décomposition fine en différents types (typologie).

Ces deux aspects sont souvent complémentaires : les mêmes données peuvent donner lieu aux deux approches, l'analyse qualitative permettant d'éclairer certaines observations qualitatives (et réciproquement).

	<b>quantitatif</b>	<b>qualitatif</b>
Phase d'exploration d'un domaine inconnu	Reste possible	Plus souvent utilisé
Méthode d'observation structurée	oui	non
Méthode d'observation peu structurée	Reste possible	Plus souvent utilisé
Généralisation des résultats	oui	Possible en croisant divers types d'observations



Querelle des méthodes entre ceux qui pensent que les sciences humaines et sociales doivent

- utiliser les mêmes méthodes que les sciences « dures »
- créer leurs méthodes spécifiques.

En SHS, l'accord intersubjectif (qui n'évacue pas la dimension interprétative) constitue une manière d'articuler les deux dimensions.

# Deux types d'observation complémentaires

	<b>Qualitatif</b>	<b>Quantitatif</b>
<b>Objet observé</b>	Un phénomène, individu, groupe particulier	Un échantillon (de phénomènes, d'individus, de groupes) représentant une population plus large.
<b>Objectif</b>	Approfondir, comprendre un phénomène Accent sur les interprétations subjectives et l'expérience des participants.	Généraliser une observation de l'échantillon à la population. Accent sur les relations entre les variables.
<b>Matérialité des données</b>	Discours des participants. Parcours de vie. Discours du chercheur (notation sur des situations observées). Matériau audiovisuel, textuel.	Quel que soit le matériel recueilli, on le met sous une forme permettant distinguer des éléments et de les compter (on définit des variables).
<b>Moyens de synthèse des observations</b>	Discours résumant des tendances, schémas, cartes mentales, typologie, etc.	Tableaux de valeurs numériques, représentations graphiques, tests.
<b>Axiome (proposition admise sans démonstration)</b>	Le particulier permet d'atteindre le général.	Les éléments qu'on compte sont équivalents, substituables.



Dans les deux cas : un élément  $\equiv$  tous les éléments

# .... mais aussi deux cultures parfois opposées en SHS : approche *hypothético-déductive* vs *empirisme*

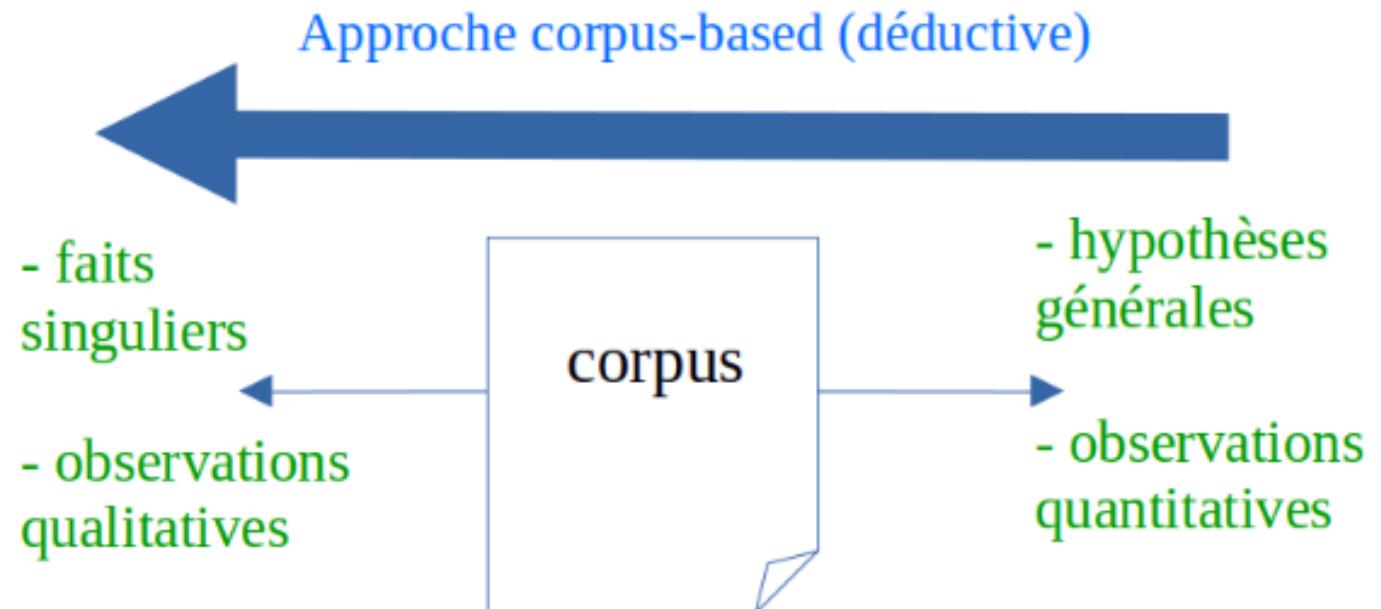
Table 1 Contrasting qualitative and quantitative research

Section	Criterion	Qualitative	Quantitative
1	Approaches to explanation	Explain individual cases; “causes-of-effects” approach	Estimate average effect of independent variables; “effects-of-causes” approach
2	Conceptions of causation	Necessary and sufficient causes; mathematical logic	Correlational causes; probability/statistical theory
3	Multivariate explanations	INUS causation; occasional individual effects	Additive causation; occasional interaction terms
4	Equifinality	Core concept; few causal paths	Absent concept; implicitly large number of causal paths
5	Scope and generalization	Adopt a narrow scope to avoid causal heterogeneity	Adopt a broad scope to maximize statistical leverage and generalization
6	Case selection practices	Oriented toward positive cases on dependent variable; no (0,0,0) cases	Random selection (ideally) on independent variables; all cases analyzed
7	Weighting observations	Theory evaluation sensitive to individual observations; one misfit can have an important impact	All observations are a priori equally important; overall pattern of fit is crucial
8	Substantively important cases	Substantively important cases must be explained	Substantively important cases not given special attention
9	Lack of fit	Nonconforming cases are examined closely and explained	Nonsystematic causal factors are treated as error
10	Concepts and measurement	Concepts center of attention; error leads to concept revision	Measurement and indicators center of attention; error is modeled and/or new indicators identified

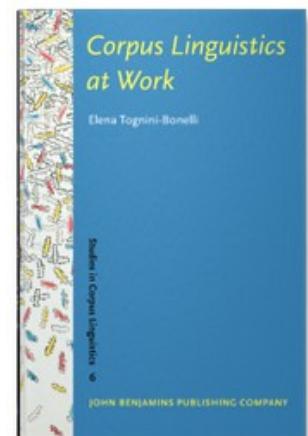
[https://public.wsu.edu/~tnridout/mahoney\\_goertz20061.pdf](https://public.wsu.edu/~tnridout/mahoney_goertz20061.pdf)

Mahoney, J., & Goertz, G. (2006). A Tale of Two Cultures :  
Contrasting Quantitative and Qualitative Research.  
*Political Analysis*, 14(3), 227-249.

# Cas particulier de la linguistique de corpus



pour l'opposition  
*corpus-driven* vs *corpus based*  
cf. A. Tognini-Bonelli (2001)

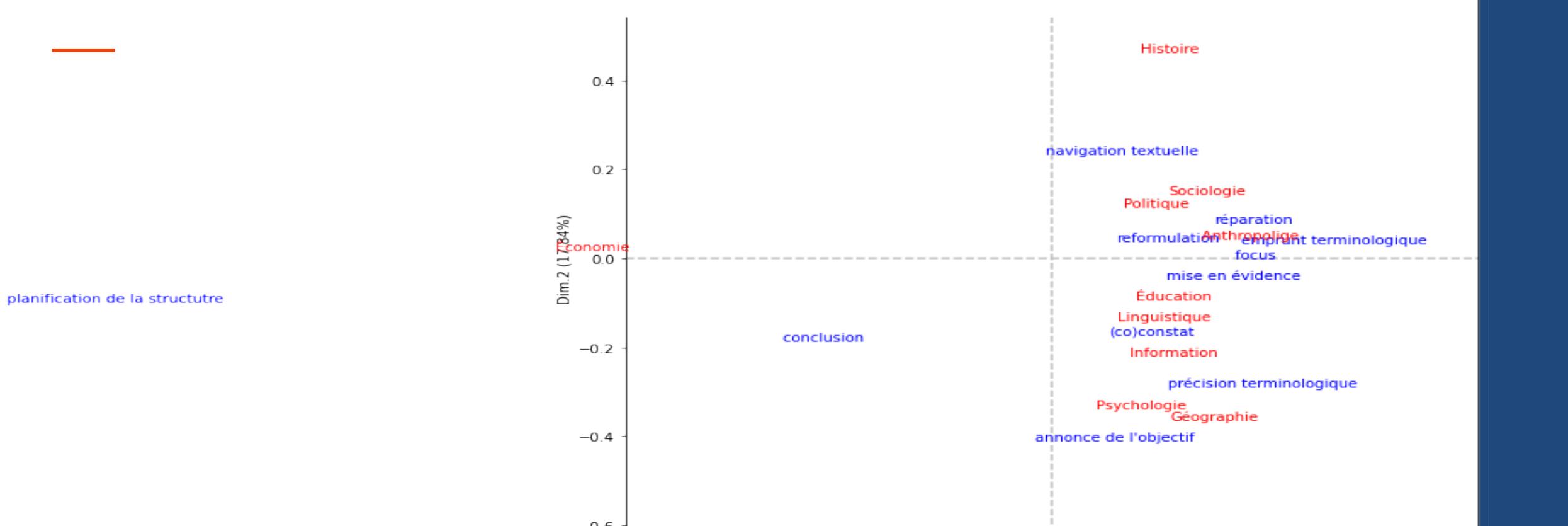


## Exemple d'articulation entre observation qualitative et quantitative. La modélisation des routines de guidage du lecteur RGL (Ji, 2022).

### Exemple d'une routine de planification de la structure discursive

Description sémantique	'L' 'auteur' 'traite' un sous-thème dans une 'partie textuelle' 'numérotée'				
Exemples	<i>Dans la section 2, nous présentons... on formule en conclusion...</i>				
Type de phrase	Phrase déclarative au présent, elle peut être à la voix passive.				
Structure syntaxique					
Réalisations lexicales	Nous/on/ Je (vouloir)	analyser/aborder/analyser/examiner/ étudier/considérer/proposer/exposer/ discuter/développer...	SN/que	(dans la première/deuxième/troisième ou (1,2,3,I,II, III...))	section/ partie/ conclusion)
Éléments périphériques	Complément de nom : <i>de l'/cet article/contribution</i> Marqueurs discursifs : <i>tout d'abord, enfin, d'une part, d'autre part</i> Une phrase qui annonce le plan et marque le début de la planification de la structure textuelle, comme : <i>L'article se présente de la manière suivante.</i>				
Paramètres énonciatifs	Cette routine s'adresse implicitement au lecteur pour donner des informations sur l'organisation du texte. Elle relève donc du dialogisme interlocutif.				
Position textuelle	Dans la partie introductory et le début d'une section locale.				
Fonctions	Annonce du plan textuel global et local des écrits scientifiques, avec parfois une fonction de transition entre des sections.				

# Résultats des spécificités disciplinaires :



- L'**économie** est **la plus spécifique** : (1) **suremploi** des RGL, surtout celles de structuration, (2) **emploi abusif** des routines de planification (88/105).  
→ cette discipline a un vif besoin de mieux organiser le texte.
- La **linguistique** **suremploie** les **routines métalinguistiques** et de **topicalisation**.  
→ présente leur propre point de vue par des routines de precision terminologique et le (co)constat.
- L'**histoire** est remarquable pour **l'emploi abusif** d'une routine de **navigation intertextuelle** (voir l'article de qn...) : une discipline qui se réfère beaucoup aux documents.

## 2.2 Données massives ou *big data* : quelques réflexions



# Qu'est-ce que les données massives ou Big data ?

« Le big data /,bɪg 'deɪtə/1 (litt. « grosses données » en anglais), les mégadonnées ou les données massives, désigne les ressources d'informations dont les caractéristiques en termes de volume, de vitesse et de variété imposent l'utilisation de technologies et de méthodes analytiques particulières pour créer de la valeur, et qui dépassent en général les capacités d'une seule et unique machine et nécessitent des traitements parallélisés. »

Wikipedia Big data, consulté 08-01-2023-01

<http://www.cea.fr/multimedia/Pages/videos/culture-scientifique/technologies/conference-big-data.aspx>

*You cannot email this data to a colleague. You can't even download it on your computer. This is data on an unprecedented impossibly mind boggling massive scale. - Kenneth Benoit (2015)*

From the website of Josef Fruehwald, Univ. of Edinburgh : **Big Data and Sociolinguistics**  
[https://jofrhwld.github.io/papers/plc39\\_2015/#/](https://jofrhwld.github.io/papers/plc39_2015/#/)



David Weinberger

## **2. Notions de base et panorama des méthodes**

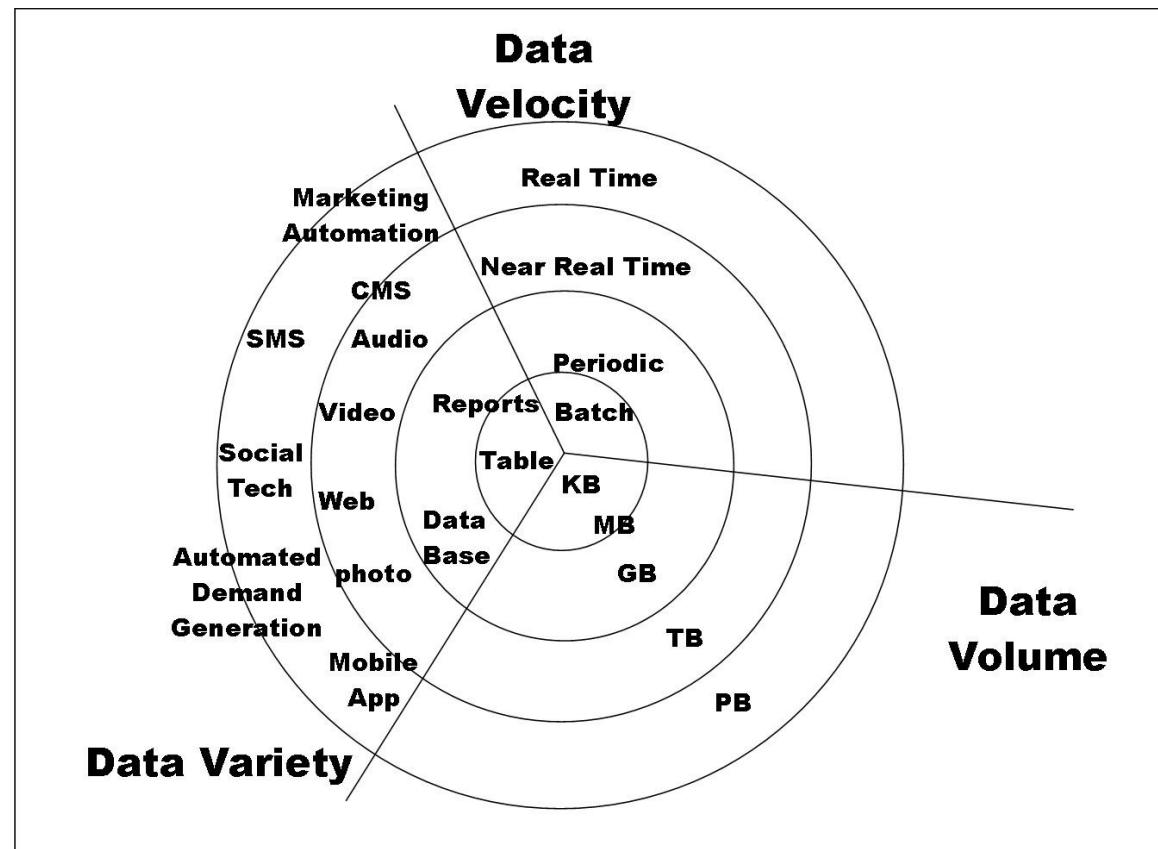
# Les données sont « big » parce qu'elles se développent selon trois dimensions : les trois V (Volume, Vélocité, Variété)

“increasing size of data, increasing rate at which it is produced and the increasing range of formats and representations employed”  
(Douglas, 2001)

Les utilisateurs de Facebook y transfèrent plus de 900 millions de photos par jour (estimation 2015).

Diversité et absence de structure des données FB

- Textes
- Images
- Likes, etc.
- Réseau
- Profils
- Liens
- Video
- Tags
- Connexion
- etc.



Facebook stocke environ 250 milliards d'images (estimation 2015)

NB - si une personne regarde les 250 milliards de photos de FB une seconde chacune il lui faut environ 8000 ans (estimation 2015)

NB – 900 millions de photos par jour sont téléchargées sur FB, soit 10 416 par seconde... (estimation 2015)

# Données, Connaissances, Informations

## Définition *données*

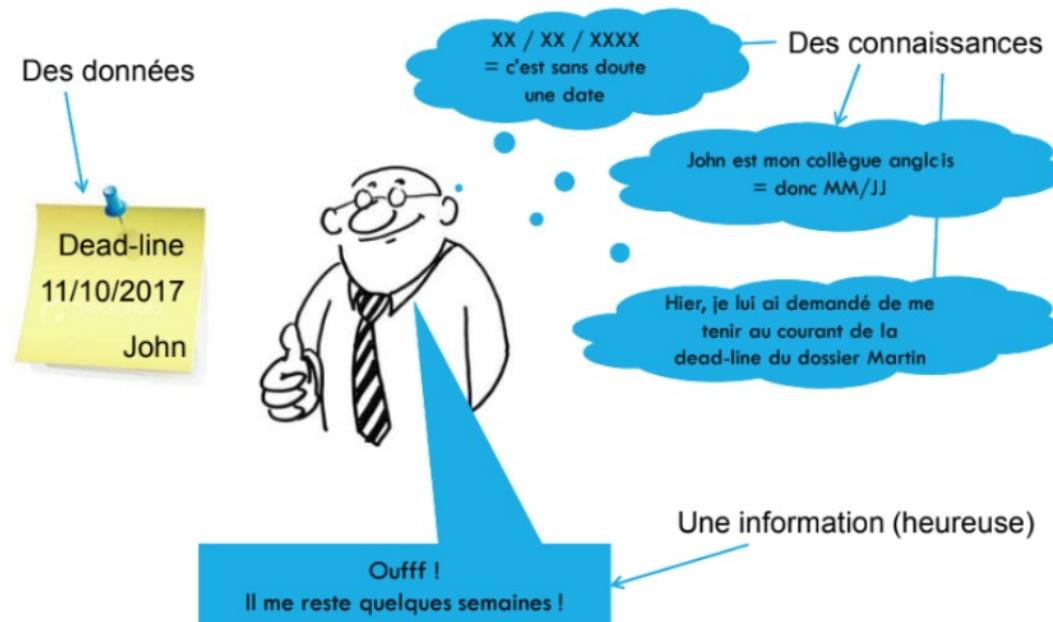
Les données sont des ensembles de symboles (mots, nombres, images, sons...) pour représenter le monde réel (objets, évènements).

## Définition *information*

Une information est une donnée interprétée selon un contexte,

## Définition *connaissance*

La connaissance est un ensemble d'informations organisé, spécifique d'un individu ou d'un collectif, et validé (par la science, l'expérience, le collectif, l'intervention divine, etc.)



A partir de :  
<http://larevuedata.com/donnee-information-connaissance>

# D'où proviennent les données du Big Data ?

## **1/ De réseaux de capteur ad hoc collectant des évènements physiques, biologiques ou concernant les interactions sociales**

Astrophysique : Systèmes de capteurs du rayonnement cosmique

Médecine : surveillance des fonctions vitales de l'être humain grâce à des micro-capteurs (qui pourront être avalés ou implantés sous la peau : SF ?)

Sciences sociales : réseaux de capteurs calculant les proximités entre individus pour étudier la dynamique des interactions sociales, dans une école, une conférence, un hôpital...

## **2/ De la collection non supervisée et à grande échelle des traces numériques laissées par l'activité humaine**

Blogging, réseaux sociaux numériques (Twitter, Facebook), services peer-to-peer (BlaBlaCar), plateformes de partage (Vélib, Citiz), réseaux de transport (entrée-sorties et trajets dans le métro), opérations bancaires, boursières, commerciales, etc.

## **3/ De la constitution volontaire de très grandes bases de données numériques**

Collection à grande échelle des résultats d'analyses et d'imageries médicales  
Collection à grande échelle de données épidémiologiques

# Les enjeux du Big Data

**1/ Enjeux éthiques** : protection des personnes et de la vie privée

**2/ Enjeux commerciaux** : prédiction de l'avenir à partir de l'existant, publicités personnalisées, définition des clients cibles, etc.

**3/ Enjeux politiques, militaires, judiciaires** : prévoir, contrôler, surveiller, manipuler

**4/ Enjeux scientifiques** : plus de données, plus de connaissances scientifiques. Nécessite de développer des méthodes statistiques de plus en plus performantes pour analyser ces données.

**5/ Enjeux techniques en IA** : plus de données, de meilleurs résultats en IA. Mais pas plus de connaissances ? un système tel que ChatGPT transforme-t-il les données en connaissances ?

# Les enjeux du Big Data

En résumé, les données sont des informations brutes qui peuvent être transformées en connaissances grâce à un processus d'analyse et d'interprétation. Les connaissances sont l'ensemble des informations, des savoirs et des compétences acquis par une personne ou une organisation à travers l'expérience et l'apprentissage.

OL

As-tu tiré des connaissances à partir des données qui t'ont entraîné ?



En tant que modèle de langage entraîné par OpenAI, j'ai été construit à partir de grandes quantités de données textuelles qui m'ont permis d'apprendre à comprendre et à générer du langage humain. Ces données ont été utilisées pour entraîner mon modèle de manière à ce que je puisse prédire les mots suivants dans un texte en fonction des mots qui le précédent.



Cependant, il est important de noter que, bien que j'aie été construit à partir de données et que j'utilise ces données pour répondre à vos questions, je ne suis pas capable de tirer des connaissances de la même manière qu'un être humain. Je suis un outil conçu pour fournir des réponses précises et précises en fonction des informations qui m'ont été fournies lors de mon entraînement, mais je ne suis pas capable de comprendre ou d'apprendre de la même manière qu'un être humain. Si vous avez d'autres questions, n'hésitez pas à me les poser !

# Les enjeux du Big Data

- Les trois paradigmes du TAL et les données langagières :
  - 1950-1990 : symbolique
    - peu de données, mais des connaissances expertes
  - 1990-2010 : statistique
    - beaucoup de données, et des connaissances empiriques
    - les outils statistiques s'appliquent bien aux « big data » pour en extraire des modèles
  - 2010-.... : neuronale
    - données massives
    - modèles prédictifs : où sont passées les **connaissances** linguistiques ????
    - nouveaux défi : « explainable AI »
    - ça « marche » mais on perd le contrôle de la donnée

## 2.3 La notion de variable

# Les variables sont les principaux objets construits qu'on manipule pour traiter des données avec une visée quantitative

Une variable est une caractéristique qui est partagée par les différentes “entités” observées et qui peut prendre différentes valeurs

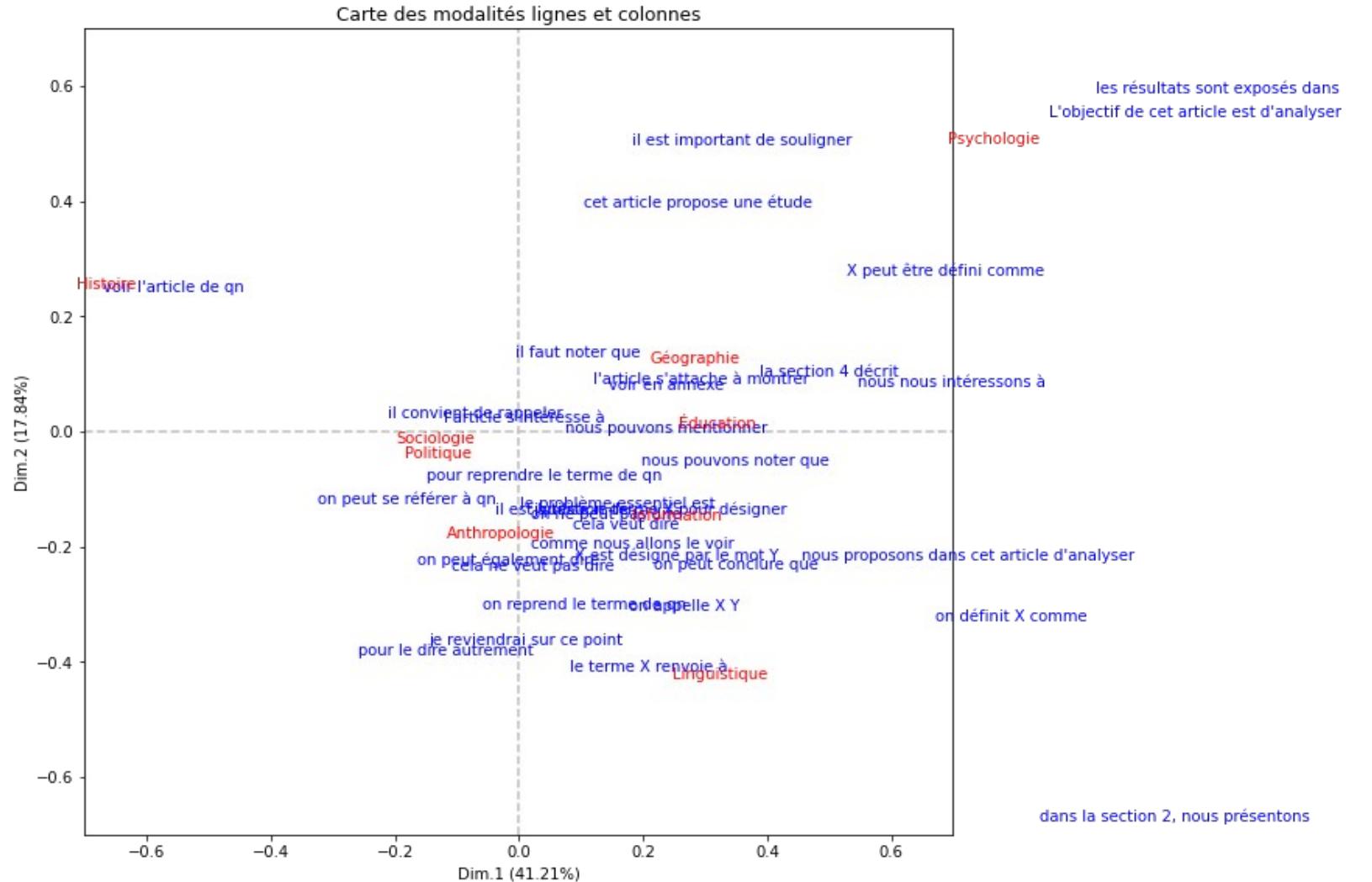
Exemple d'études	Exemple de variables (ou pas)
<b>Observation du langage chez les 200 enfants de l'école maternelle X</b>  <a href="https://dylnet.univ-grenoble-alpes.fr/">https://dylnet.univ-grenoble-alpes.fr/</a> accueil	- être un écolier = pas une variable - être une fille ou un garçon = une variable - avoir moins de 10 ans = pas une variable - année d'arrivée dans l'école maternelle X = une variable - âge au moment de l'étude = une variable
<b>Les routines de guidage du lecteur dans les écrits scientifiques en français (thèse Yujing Ji, 2022)</b>  <a href="https://www.theses.fr/s186455">https://www.theses.fr/s186455</a>	- fréquences des routines ? - genre textuel ? - langue maternelle des auteurs ? - discipline scientifique ? - fonction des routines de guidage ?

# Les variables sont les principaux objets construits qu'on manipule pour traiter des données avec une visée quantitative

Routines \ discipline	ECO	LING	PSYCHO	SOCIO	POL	SC EDU	INFO	GEO	ANTHROPO	HIST
la section 4 décrit	57	1	2	2	3	2	3	4	0	0
dans la section 2, nous présentons	31	6	2		0	1	1	0	0	0
L'objectif de cet article est d'analyser	16	10	36	3	5	15	9	10	4	2
l'article s'attache à montrer	3	0	4	7	8	11	9	8	5	4
cet article propose une étude	2	0	2	1	0	0	0	1	3	1
l'article s'intéresse à	2	0	1	3	10	5	4	3	2	3
nous proposons dans cet article d'analyser	13	11	8	2	7	5	12	3	4	1
nous nous intéressons à	12	7	9	1	8	12	5	4	1	1
les résultats sont exposés dans	40	5	17	2	4	7	2	0	1	0
comme nous allons le voir	52	53	24	37	63	46	37	17	51	34
comme l'indique le tableau	0	1	5	0	0	3	0	2	0	1
je reviendrai sur ce point	13	22	2	7	6	8	4	1	6	15
il convient de rappeler	2	2	2	0	8	0	1	1	2	3
voir en annexe	51	8	17	31	11	33	19	2	10	12
voir l'article de qn	47	22	19	70	72	30	24	17	47	180

# Les variables sont les principaux objets construits qu'on manipule pour traiter des données avec une visée quantitative

comme l'indique le tableau



## Les variables sont les principaux objets construits qu'on manipule pour traiter des données avec une visée quantitative

- Ici les individus de notre **population** sont des *occurrences*
  - à chaque occurrence sont associées des variables
    - lemme de la routine
    - type de routine (méta-linguistique, de topicalisation, de structuration)
    - sa/ses fonctions (p.ex. annonce du plan, reformulation, etc.)
    - la discipline de l'article
  - Les *effectifs* = les *fréquences* associées
  - Dans le tableau précédent, on croise deux variables : lemme et discipline

# Les variables sont de différents types

## Les différents types de variables

	Variables quantitatives ou continues	Variables nominales ou catégorielles	Variables discrètes
Définition	Attribuent une valeur numérique aux objets	Collent une étiquette, un nom aux objets	Attribuent un ordre. On peut associer un <b>nombre entier</b> aux modalités.
Exemples	<ul style="list-style-type: none"><li>- Le revenu mensuel net (en euros)</li><li>- La moyenne au bac (en points) - (échelle de rapport)</li><li>- la taille (en cm)</li><li>- la température corporelle (en degrés celsius) (échelle d'intervalle)</li><li>- le taux de cholestérol HDL (en g/l)</li></ul>	<ul style="list-style-type: none"><li>- Le sexe déclaré à la naissance : <b>deux modalités</b>, masculin et féminin</li><li>- Mode de transport habituel utilisé pour se rendre au travail : <b>quatre modalités</b>, véhicule à moteur, vélo, transports en commun, marche à pied</li><li>- Le statut par rapport à la cigarette : <b>trois modalités</b> : fumeurs, non fumeur, ex-fumeur</li><li>- La discipline d'une routine : 10 modalités.</li></ul>	<p><b>Échelle ordinale :</b></p> <ul style="list-style-type: none"><li>- Le niveau d'étude : <b>trois modalités</b> :</li><li>1. pré-bac, 2. de bac à L3, 3. au delà de L3 → statistiquement limitée (pas de calcul de moyenne, d'écart type etc.)</li><li>- échelle de Likert (satisfaction)</li></ul> <p><b>Échelle d'intervalle</b> (chaque valeur adjacente correspond au même intervalle) :</p> <ul style="list-style-type: none"><li>- dates successives...</li></ul> <p><b>Échelle de rapport</b> (échelle d'intervalle avec un zéro absolu) :</p> <ul style="list-style-type: none"><li>- nombre d'enfants dans un ménage</li><li>- nombre de formes dans un texte</li><li>- le QI (en points)</li></ul>

# Les variables sont de différents types

## Les différents types de variables

	Variables quantitatives ou continues	Variables nominales ou catégorielles	Variables discrètes
<b>Critères</b>			
L'ordre à un sens	Oui	Non	Oui
Il existe en général une origine (zéro) et une unité de mesure	Oui	Non	Ça dépend.
La distance entre deux modalités à un sens. On peut calculer des moyennes et écarts-types.	Oui La distance entre une personne qui mesure 120 cm et une personne qui mesure 150 cm est la même qu'entre une personne de 160 cm et une personne de 190 cm.	Non On ne peut pas parler de distance entre « être fumeur » et « être non-fumeur »	Oui pour les échelles d'intervalle et de rapport.  Non pour les échelles ordinaires.

# Changement de nature des variables

Il est tout à fait possible de recoder une variable continue en variable discrète ou catégorielle (mais le contraire n'est pas vrai)

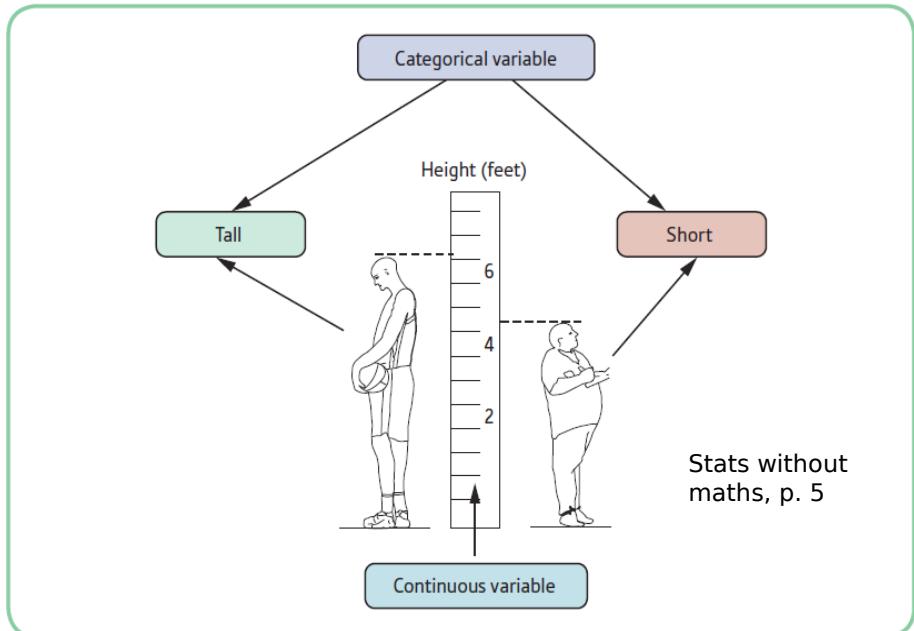


Figure 1.1 Illustration of the conversion of continuous variables into categorical variables

La taille est une variable continue/quantitative exprimée à l'aide d'une unité de mesure. On peut la transformer en variable nominale-catégorielle :

- Catégorie 1 : Taille inférieure à 160 cm = « petit »
- Catégorie 2 : Taille supérieur ou égale à 180 cm = « grand »

Autre exemple sur le « bon cholestérol »  
« pour une protection du cœur optimale, le taux de cholestérol HDL doit être supérieur à 0,45 g/L. »

<https://www.fedecardio.org/>

- Catégorie 1 : HDH > 0,45 g/L. = « protection du cœur »
- Catégorie 2 : HDH < 0,45 g/L. = « pas de protection »

1. Intérêt scientifique, social, de cette transformation ? (pensez au examens où on est reçu avec 10 et refusé avec une note inférieure)
2. Perd-on de l'information ? En quoi est-ce arbitraire ?

# Changement de nature des variables

Pour recoder une variable catégorielle en échelle d'intervalle (discrète ou continue), il faut donc rajouter de l'information. Dans certains cas c'est possible.

Indiquer s'il est possible de recoder en échelle d'intervalle les variables suivantes, et comment ?

- population active : actif / chômeur
- tabagisme : fumeur / non-fumeur
- genre romanesque : romans policier / sentimental / historique / de science-fiction
- sexe biologique : homme / femme

# Changement de nature des variables

## Recoder une variable catégorielle en variable continue

**L'orientation sexuelle peut être traitée comme une variable catégorielle avec les catégories suivantes : bisexuel, hétérosexuel, homosexuel, asexuel, etc.**

**On peut la transformer en variable continue sous forme d'échelle d'orientation sexuelle**

- On pose une série de questions
- A chaque question, les participants répondent sur une échelle de 0 (jamais) à 5 (très fréquemment)
- On fait la somme des réponses; chaque participant obtient un score qui va de l'hétérosexualité à l'homosexualité
- Variantes : les échelles d'hétérosexualité et d'homosexualité sont différentes. Chaque participant obtient deux scores. On peut introduire d'autres orientations : asexualité, attirance vers des personnes non binaires, etc.

82. I have erotic fantasies that include women.	-.333	.772	.000	.000
70. I realize I am attracted to differences in female bodies.	-.394	.766	.000	.000
44. I have erotic fantasies about women.	-.485	.726	-.269	.000
60. Specific parts of the female anatomy are attractive to me.	-.279	.723	-.251	.000
41. The female anatomy intrigues me.	-.328	.721	.000	.341
63. I have female friends to whom I am sexually attracted.	-.303	.685	.000	.000
48. I am easily aroused by good looking women.	-.483	.671	-.275	.000
6. Pictures of nude female are exciting.	-.450	.658	-.295	.000
17. When I masturbate, my fantasies include only men.	.490	-.556	.291	.376
14. I get aroused thinking about women in their underwear.	-.353	.551	.000	-.267
25. The expression of affection between two males is acceptable.	.336	.000	.772	.000

Exemple d'échelle d'orientation sexuelle pour hommes;  
<https://tpcjurnal.nbcc.org/the-development-of-a-sexual-orientation-scale-for-males/>

Les variables décrivant le statut social sont souvent catégorielles (favorisé vs. défavorisé). Comment créer une variable continue du statut social d'une personne qui prenne en compte à la fois son capital économique (salaire, fortune) et son capital scolaire (niveau d'étude) ?

# **Variables explicatives, prédictives, indépendantes**

## **Variables à expliquer, dépendantes**

### **Variables de confusion**

Variables explicatives  
Variables prédictives  
Variables indépendantes

Expliquent (sont des causes)  
Prédisent (sont des facteurs)  
Ne sont ni cause ni facteur

Variables à expliquer  
Variables prédictées  
Variables dépendantes

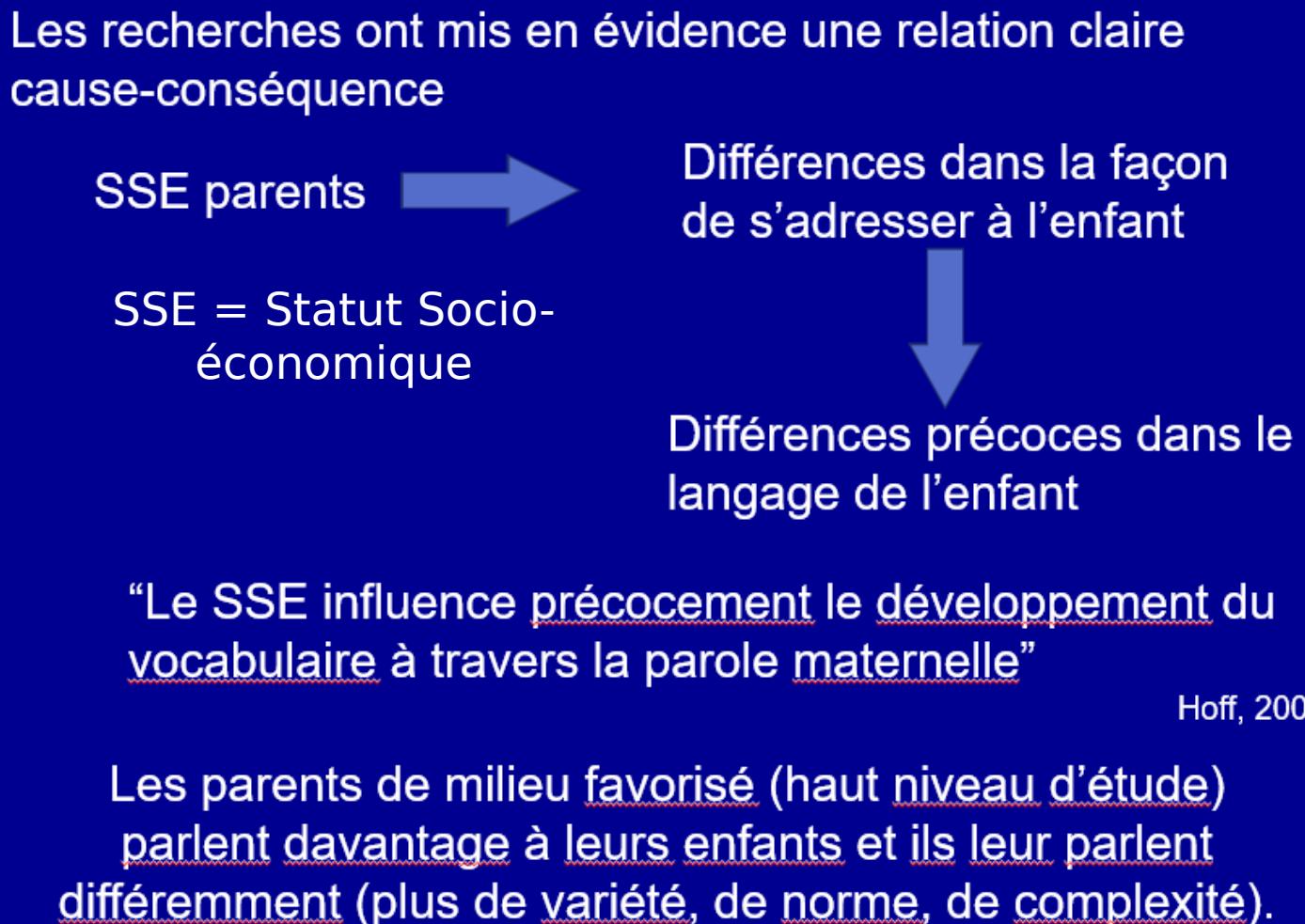
**Evènement A explique/ est la cause de Evènement B : on sait établir l'enchaînement causal entre A et B**

**Evènement A prédit Evènement B : on sait que A favorise B sans nécessairement établir l'enchaînement causal entre A et B**

# Variables explicatives, prédictives, indépendantes

## Variables à expliquer, dépendantes

**Exemple de relation explicative dans le domaine des différences interindividuelles dans l'acquisition du langage oral par le jeune enfant**



# **Variables explicatives, prédictives, indépendantes**

## **Variables à expliquer, dépendantes**

### **Exemple de relation prédictive dans le domaine médical : facteurs de mortalité de la Covid aux début de la pandémie**

Qui meurt du Covid-19 ? Une étude statistique d'une ampleur inédite en Europe apporte une réponse détaillée. Mis en ligne le 7 mai, les premiers résultats du projet Opensafely se basent sur les données médicales de 17 millions de personnes résidant au Royaume-Uni et offrent une analyse des facteurs de risque de décès. Cette étude n'est pas encore publiée dans une revue à comité de lecture.

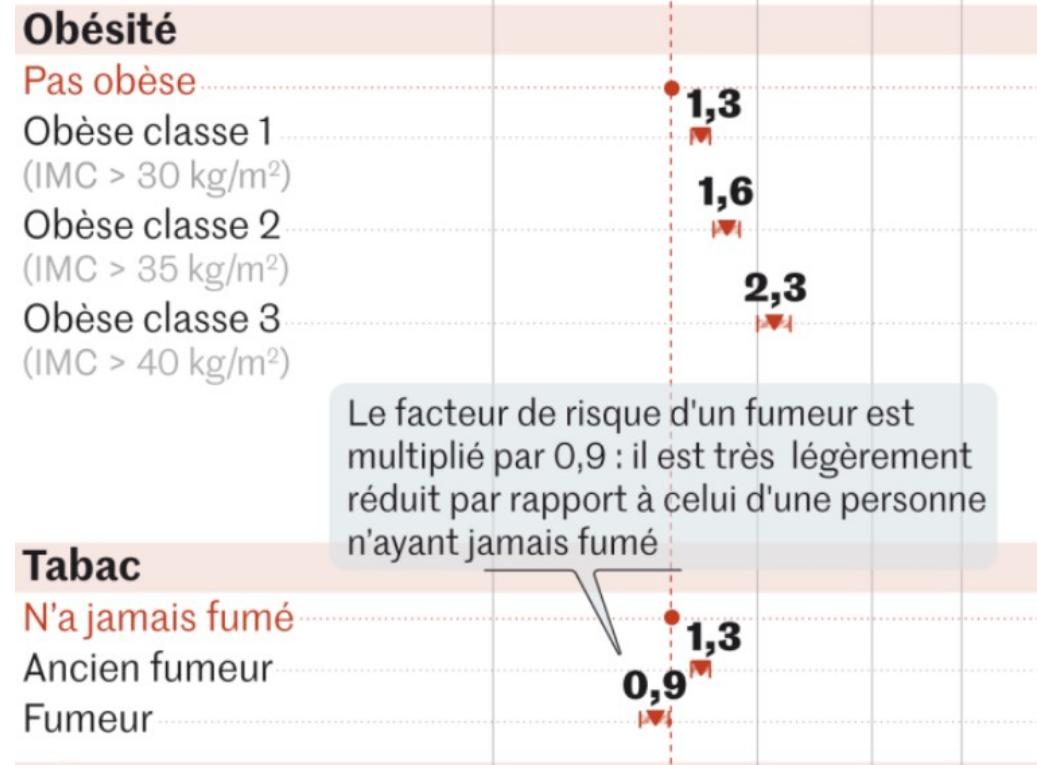
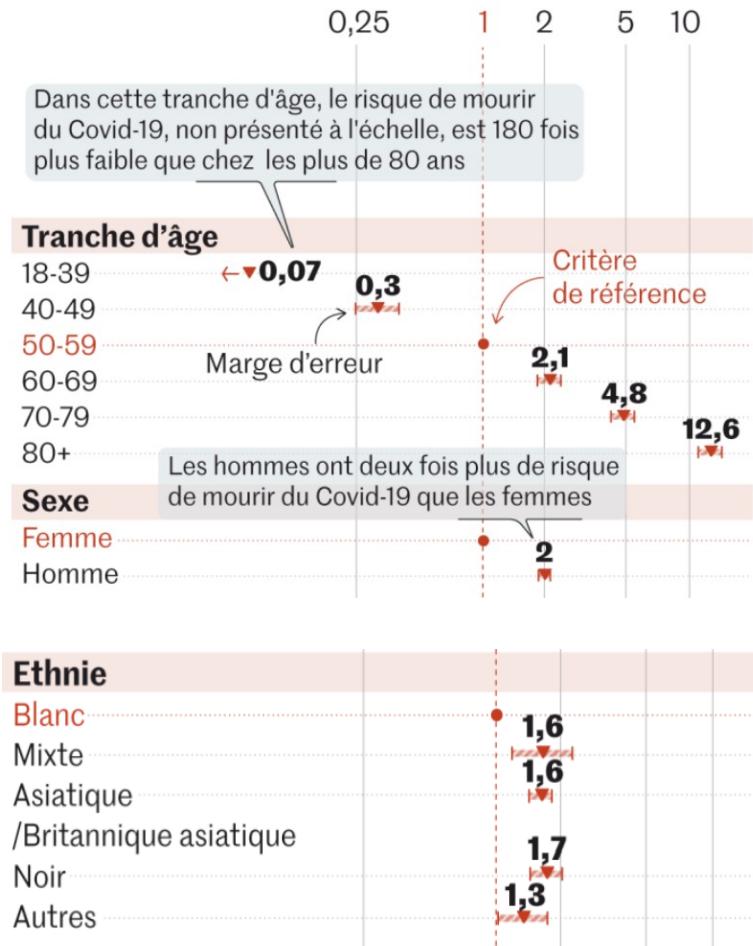
Sur cette très large cohorte, 5 683 personnes sont décédées du Covid-19 dans des hôpitaux britanniques entre le 1<sup>er</sup> février et le 25 avril. Les facteurs de risque de mortalité que révèle cette étude menée conjointement par des équipes d'Oxford, de l'université de Leeds, du Centre de recherche national en soins intensifs (ICNARC) et de la London School of Hygiene and Tropical Medicine, corroborent ce que les observations des praticiens et les résultats d'autres études ont déjà montré.

[https://www.lemonde.fr/sciences/article/2020/05/19/covid-19-l-age-principal-facteur-de-risque-de-mortalite\\_6040077\\_1650684.html](https://www.lemonde.fr/sciences/article/2020/05/19/covid-19-l-age-principal-facteur-de-risque-de-mortalite_6040077_1650684.html)

# Variables explicatives, prédictives, indépendantes

## Variables à expliquer, dépendantes

Exemple de relation prédictive dans le domaine médical : facteurs de mortalité du Covid



Ces relations sont-elles explicatives ? Lesquelles peuvent le devenir ?

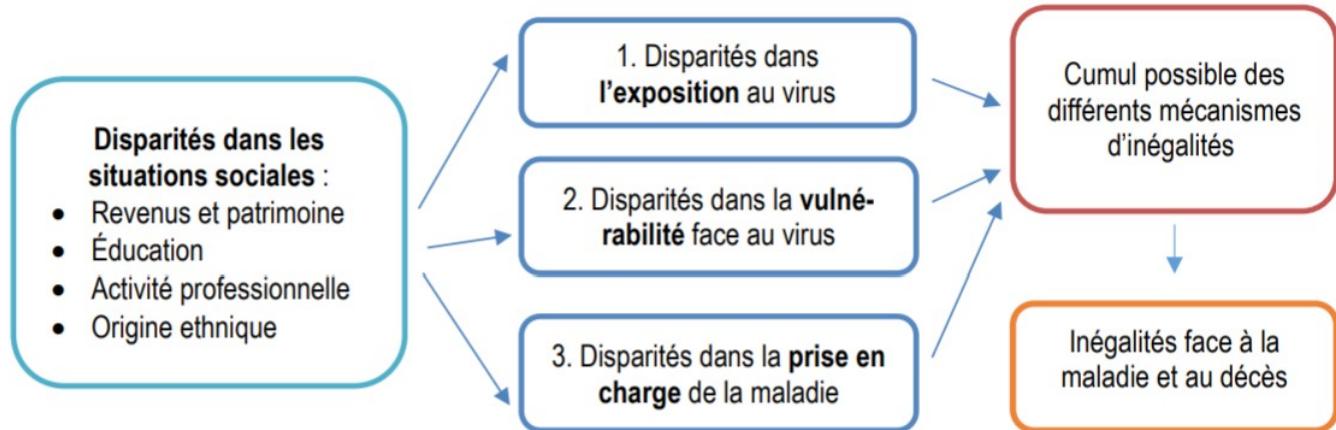
[https://www.lemonde.fr/sciences/article/2020/05/19/covid-19-l-age-principal-facteur-de-risque-de-mortalite\\_6040077\\_1650684.html](https://www.lemonde.fr/sciences/article/2020/05/19/covid-19-l-age-principal-facteur-de-risque-de-mortalite_6040077_1650684.html)

# Variables explicatives, prédictives, indépendantes

## Variables à expliquer, dépendantes

**Expliquer permet de comprendre et prédire permet d'agir**

Figure 1 • Le modèle de Blumenshine (2008)



Modèle explicatif des inégalités sociales face aux infections virales

IA : Prédiction en santé (dans le futur) basées sur l'analyse de données massives (du passé) par des systèmes d'Intelligence artificielle

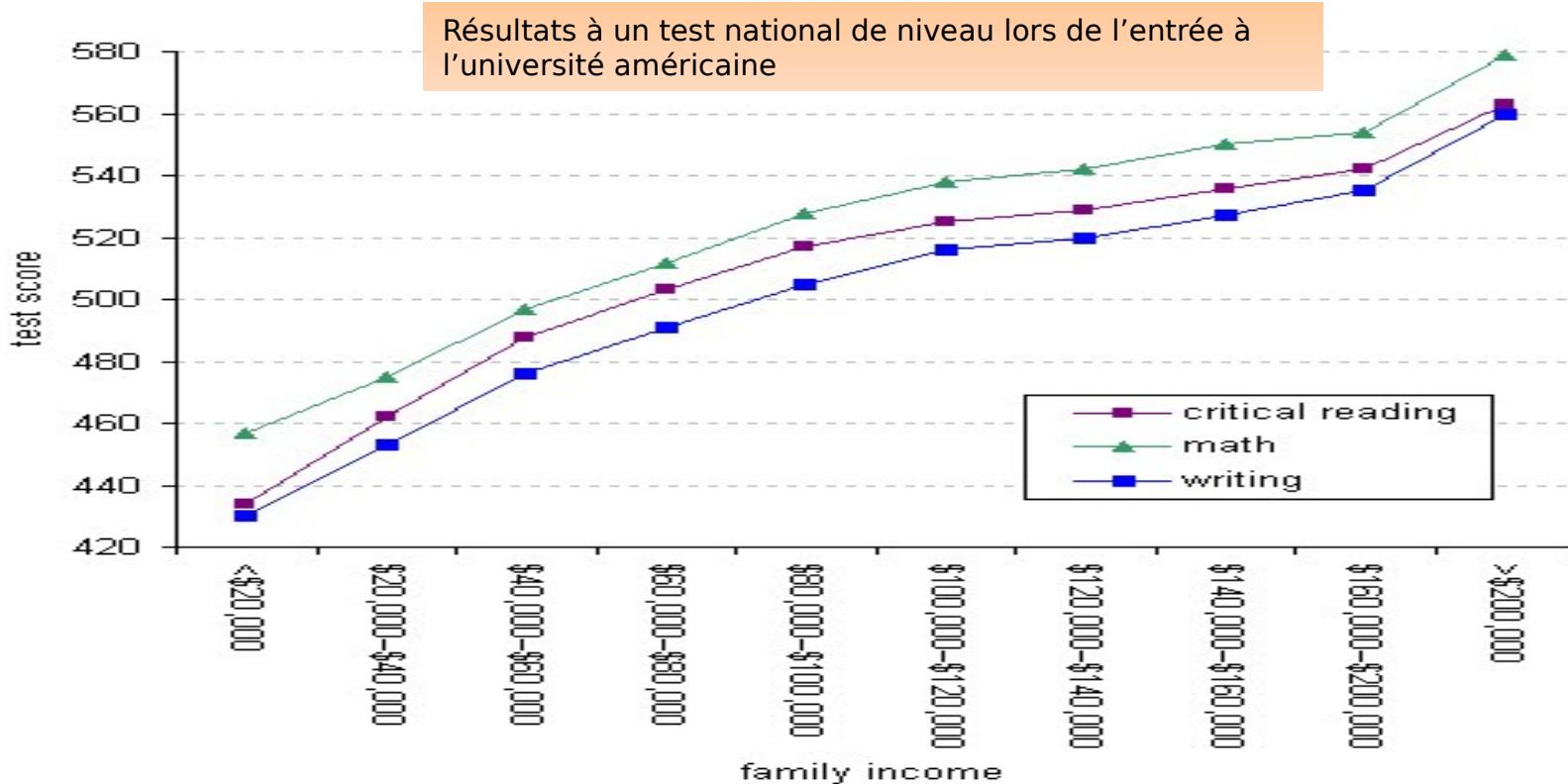
Prédire les tumeurs pulmonaires malignes répondant à l'immunothérapie à partir de l'analyse automatique des images scanner

[www.sciencesetavenir.fr/high-tech/intelligence-artificielle/cancer-une-intelligence-artificielle-predit-les-chances-de-succes-de-l-immunotherapy\\_126965](http://www.sciencesetavenir.fr/high-tech/intelligence-artificielle/cancer-une-intelligence-artificielle-predit-les-chances-de-succes-de-l-immunotherapy_126965)

Prédire des trajectoires de patients : l'évènement n+1 (rechute, complications...) à partir de données antérieures multiples (profil de la pathologie, soin, résultats d'examens, mode de vie, hérédité...)

<https://blogrecherche.wp.imt.fr/2018/10/16/ia-parcours-de-soin/>

# Un cas où les données statistiques sont avancées comme arguments dans un débat de société



Relation très nette entre  
- Test de niveau scolaire passés à l'entrée dans l'université américaine  
- le revenu moyen de la famille de l'étudiant.



Interprétation ?

<http://economix.blogs.nytimes.com/2009/08/27/sat-scores-and-family-income/>

Blog *Economix* du New York Times - 9/09/10

Sur le test SAT :

<http://en.wikipedia.org/wiki/SAT>

## **Des réactions à cette enquête trouvées dans des blogs ou des forums ...**

My guess is that, on average, higher income families have more intelligent parents, and they have smarter children because of genetics.

(Children with more economic resources) tend to have better teachers, more resource-rich educational environments, more educated parents who can help them with school and, sometimes, expensive SAT tutoring..

These tests tend to be written and evaluated by privileged individuals who may inadvertently include class-based knowledge, not just knowledge, in the exam (asking questions, for example, that rely on background information about golf instead of basketball).

The underlying data for the graph is incapable of supporting a causal analysis of the relationship (...). Suppose we were to graph average SAT scores by the number of bathrooms a student has in his or her family home. That curve would also likely slope upward. But it would be a mistake to conclude that installing an extra toilet raises yours kids' SAT scores.

## Trois conclusions

My guess is that, on average, higher income families have more intelligent parents, and they have smarter children because of genetics.

(Children with more economic resources) tend to have better teachers, more resource-rich educational environments, more educated parents who can help them with school and, sometimes, expensive SAT tutoring..

Le même fait statistique peut donner lieu à des interprétations différentes :  
**différencier description des tendances et interprétation des tendances**

These tests tend to be written and evaluated by privileged individuals who may inadvertently include class-based knowledge, not just knowledge, in the exam (asking questions, for example, that rely on background information about golf instead of basketball).

L'observation elle-même peut orienter les résultats : **notion de biais ou d'artefact**

The underlying data for the graph is incapable of supporting a causal analysis of the relationship (...). Suppose we were to graph average SAT scores by the number of bathrooms a student has in his or her family home. That curve would also likely slope upward. But it would be a mistake to conclude that installing an extra toilet raises yours kids' SAT scores.

Une relation statistique entre deux événements A et B ne signifie pas que A entraîne B :  
**notion de causalité**

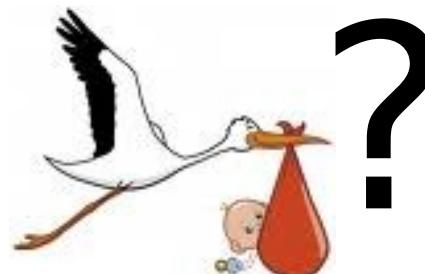
# Variables de confusion et causalité

**Une variable de confusion (*confound*) crée une association non causale entre une variable explicative et une variable à expliquer**

**Une légende raconte que les cigognes apporteraient les nouveau-nés à leurs futurs parents. Preuve statistique quand on consulte les registres de naissance de la première moitié du 20e siècle en France**



Association statistique avérée



Relation cause-conséquence  
Cigognes → naissances ?

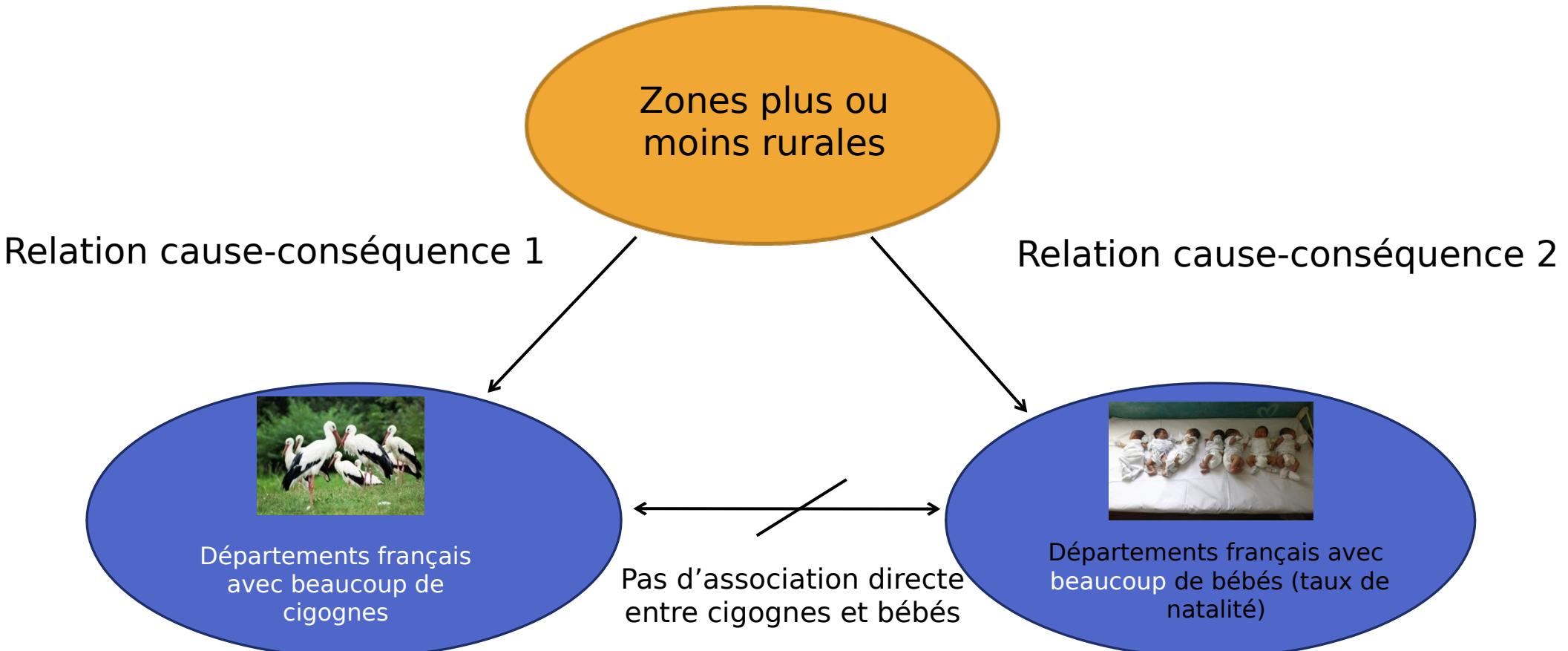


Indice de la croyance populaire :  
les cigognes amènent les bébés ?

# Variables de confusion et causalité

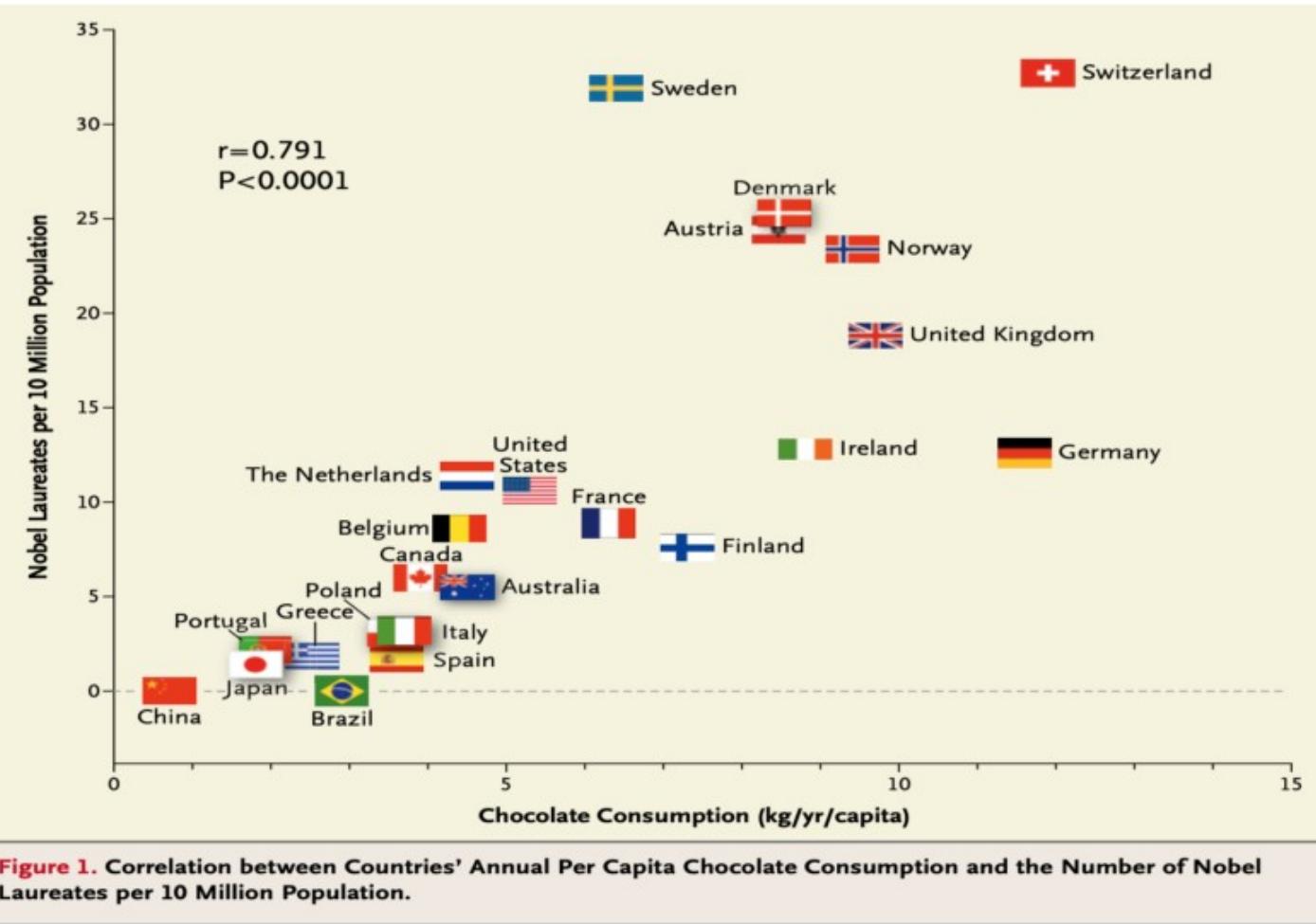
**L'association de deux événements ne signifie pas que l'un est la cause et l'autre la conséquence (lien de causalité).** La relation entre cigognes et bébés passe par un troisième élément appelé facteur ou variable de confusion. La **concomitance** n'implique pas de relation causale.

## Variable de confusion



# Les variables de confusion créent des illusions de causalité

Messerli, Franz H. 2012. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." New England Journal of Medicine 367(16):1562-64.



« Plus la population d'un pays mange de chocolat, plus il compterait de prix Nobel. Cela pourrait indiquer que le cacao dopérait les capacités mentales, selon une étude publiée jeudi dans la très sérieuse revue médicale américaine New England Journal of Medicine . Les flavonoïdes, de puissants antioxydants qu'on trouve en grande quantité dans les fèves de cacao, le thé vert et le vin rouge, ont montré qu'ils réduisent le risque de démence et améliorent les fonctions mentales chez les personnes âgées, note le Dr Franz Messerli, de l'université Columbia à New York et auteur de l'étude. »

<https://sante.lefigaro.fr/actualite/2012/10/12/19278-croquer-chocolat-pour-avoir-nobel>

## Pour réfléchir

Quelles variables de confusion ?

Validité des deux variables ?

Que faudrait-il faire pour vraiment conforter l'hypothèse ?

Autres corrélations bizarres et explications

[https://www.lemonde.fr/sciences/video/2019/09/12/correlation-et-causalite-peut-on-decrocher-un-prix-nobel-en-mangeant-du-chocolat\\_5509656\\_1650684.html](https://www.lemonde.fr/sciences/video/2019/09/12/correlation-et-causalite-peut-on-decrocher-un-prix-nobel-en-mangeant-du-chocolat_5509656_1650684.html)

## 2.4 Notion de valeur p (*p-value*) et de significativité

# Statistiques descriptives et inférentielles

Le terme *statistique* a deux significations :

Sens 1 (on dit plutôt *les statistiques*, au pluriel)

Données chiffrées relatives à un domaine : chômage, réussite scolaire, relation santé-consommation de tabac, COVID, sport, etc.

Sens 2 (on dit plutôt *LA statistique*, au singulier)

Une science : L'ensemble des principes et des méthodes permettant

a/ de recueillir, classer, synthétiser et communiquer des données numériques en vue de l'utilisation de celles-ci

b/ d'inférer des conclusions générales à partir d'observations particulières : domaine de la notion de signification

# Statistiques descriptives et inférentielles

Le terme *statistique* a deux significations :

Sens 1 (on dit plutôt *les statistiques*, au pluriel)

Données chiffrées relatives à un domaine : chômage, réussite scolaire, relation santé-consommation de tabac, COVID, sport, etc.

Sens 2 (on dit plutôt *LA statistique*, au singulier)

Une science : L'ensemble des principes et des méthodes permettant

a/ de recueillir, classer, synthétiser et communiquer des données numériques en vue de l'utilisation de celles-ci

## Statistiques descriptives

b/ d'inférer des conclusions générales à partir d'observations particulières : domaine de la notion de signification

## Statistiques inférentielles



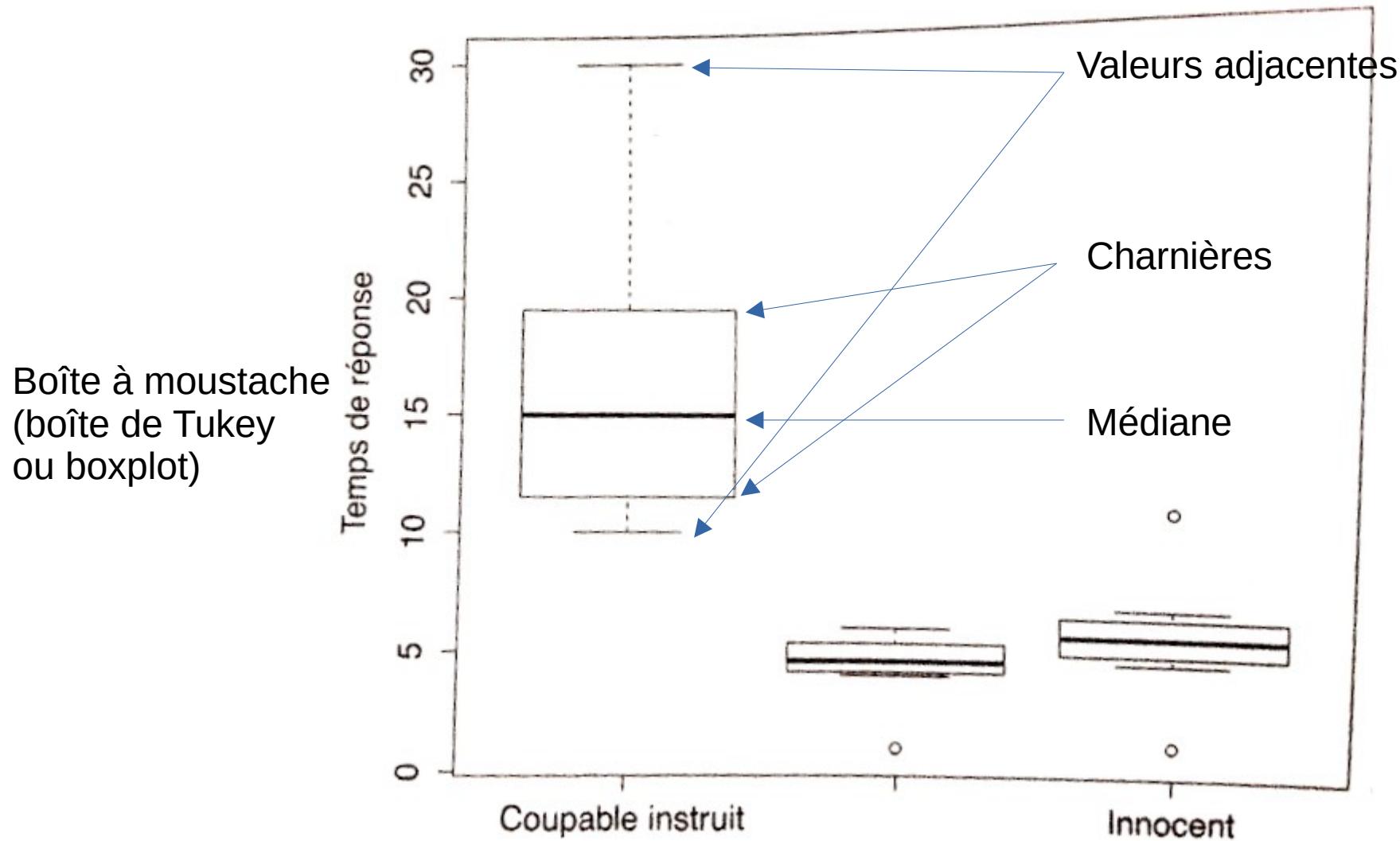
On approfondira ces 2 aspects

# Statistiques descriptives et inférentielles

- Pour des variables suivant une échelle d'intervalle, on peut calculer des **indicateurs de tendance centrale** tels que :
  - **mode** : valeur la plus fréquente (si var. discrètes)
  - **médiane** : valeur qui se situe au milieu ( $\approx 50\%$  au dessous,  $\approx 50\%$  au dessus)
  - **moyenne**
- En complément on a des **indicateurs de dispersion**
  - **écart-type** : racine carré de la variance (moyenne des carrés des écarts à la moyenne)
  - **étendue** : différence entre maximum et minimum
  - **écart inter quartiles**. L'écart interquartile est la différence entre le troisième et le premier quartiles. Il correspond à l'étendue de la série statistique après élimination de 25 % des valeurs les plus faibles et de 25 % des valeurs les plus fortes. Les deux bornes sont appelées **charnière**. Au delà de  $1,5 * \text{écart interquartile}$  de chaque charnière se situent les **valeurs extrêmes** (ou **aberrantes**). Les **valeurs adjacentes** établissent la frontière avec ces valeurs extrêmes (plus petite et plus grande valeurs non extrêmes). Cette mesure est plus robuste que l'étendue, qui est sensible aux valeurs extrêmes.
- Représentation synthétique : **boîte à moustache**

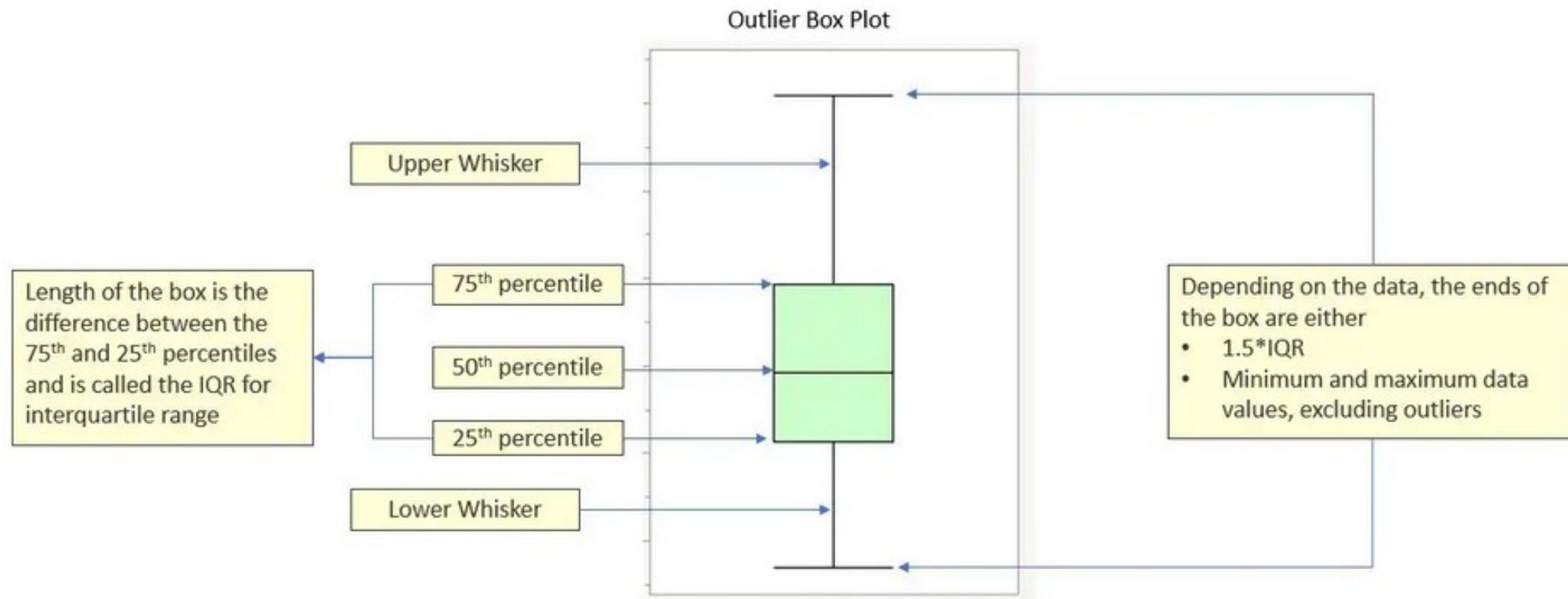
# Statistiques descriptives et inférentielles

- Exemple : étude de Winograd et Rosenfeld (2011), sur les méthodes consistant à tromper les détecteurs de mensonges. Trois groupes : 1/ coupable instruits (mentent mais ont été formés à tromper le détecteur) 2/ coupables non instruits 3/ innocents.



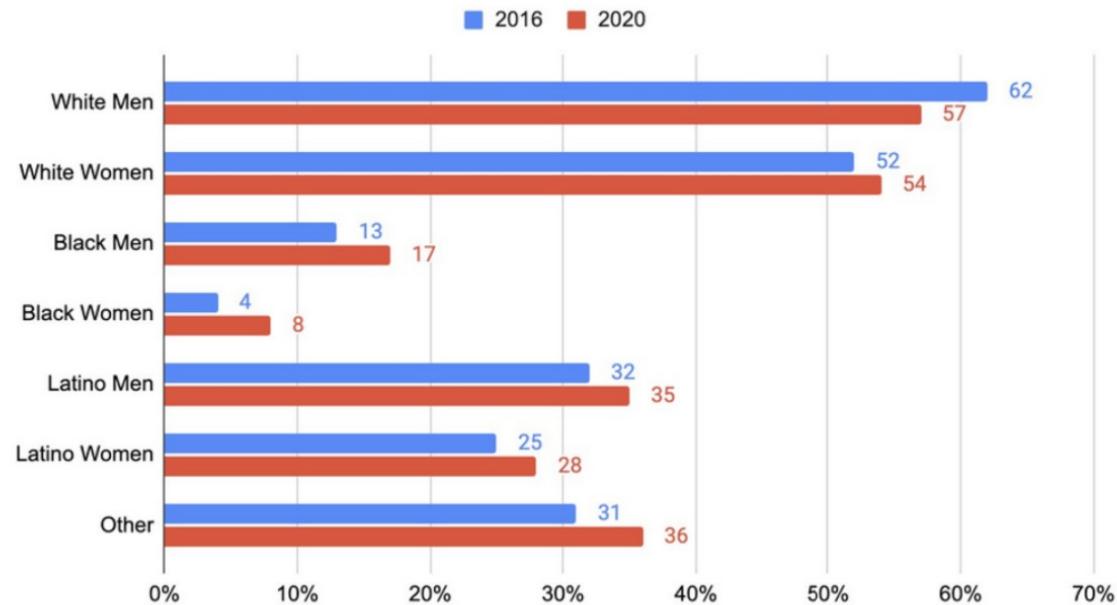
# Statistiques descriptives et inférentielles

- Boîte à moustache ou boxplot. Selon le cas, les moustaches représentent soit  $1,5 \times$  l'écart interquartile, soit les valeurs minimum et maximum, si celles-ci tombent dans cet intervalle. Les valeurs qui dépassent éventuellement les moustaches sont dites « aberrantes » (outliers), et sont représentées par des points.



# Statistiques descriptives pour une var. nominale

Support for Trump by Race and Gender (Edison Exit Poll)



The National Election Pool (NEP) is a consortium of American (...) organizations (ABC News, CBS News, CNN, and NBC News, etc. ) to provide exit polling information for US elections. The Pool contracts with Edison Research to conduct the exit polling and to perform vote tabulations.

[wiki/National\\_Election\\_Pool](#)

Election exit pool = un sondage des électeurs effectué immédiatement après leur sortie des bureaux de vote.

**Recueillir des données** : 100 000 votants interviewés en 2020 à la sortie de 700 bureaux de vote

**Classer des données** : classement des affirmations de votes par origine ethnique et par sexe.

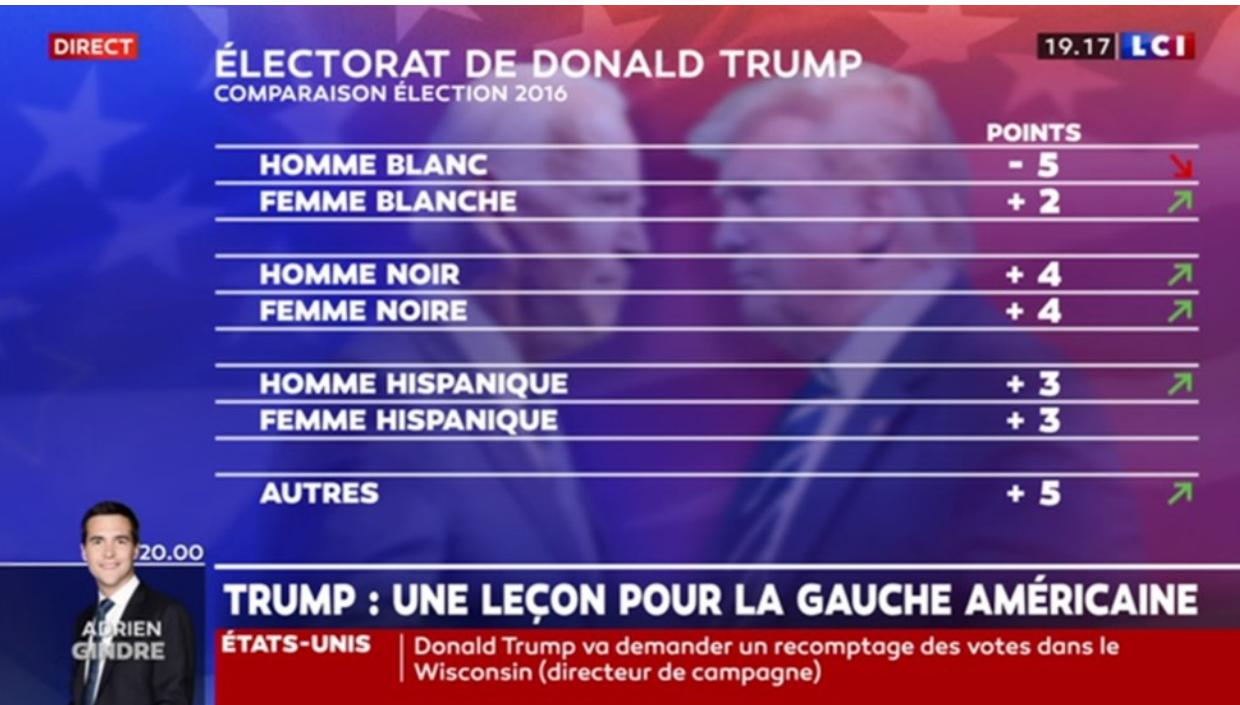
**Synthétiser de données** : ici, exprimer les affirmations de vote par des pourcentages en faveur des candidats. Si la variable suit une échelle d'intervalle, on pourra indiquer le mode, la moyenne, la médiane, l'écart-type, etc. Différentes représentations possibles : **diagramme en bâton, camembert, carte colorée**, etc.

# Statistiques descriptives

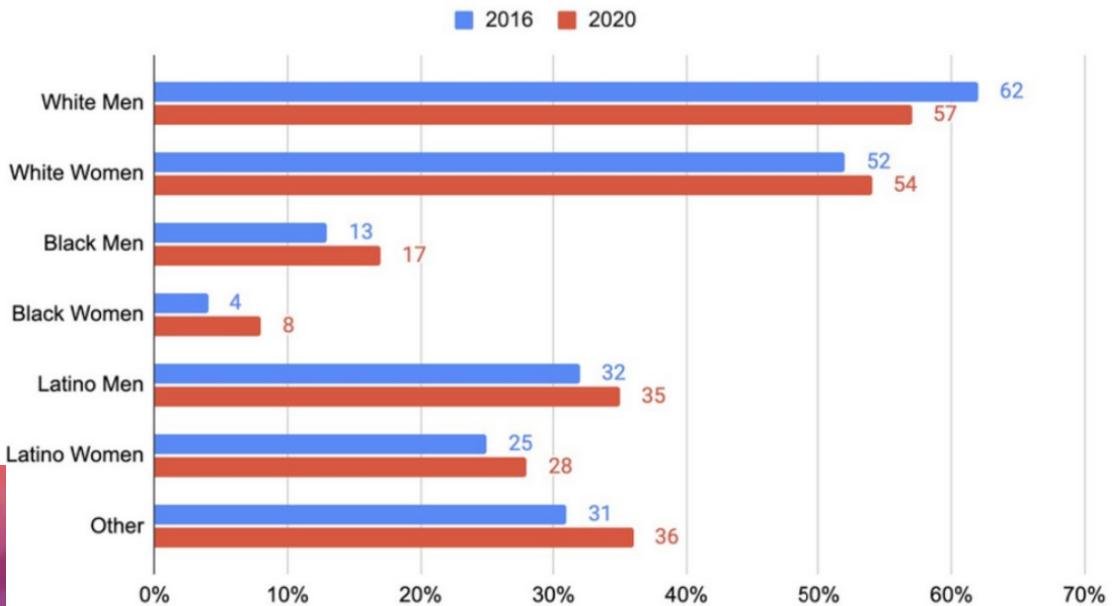
## Communiquer des données

### Graphique initial

Résultats statistique publié par LCI et repris par d'autres média



Support for Trump by Race and Gender (Edison Exit Poll)

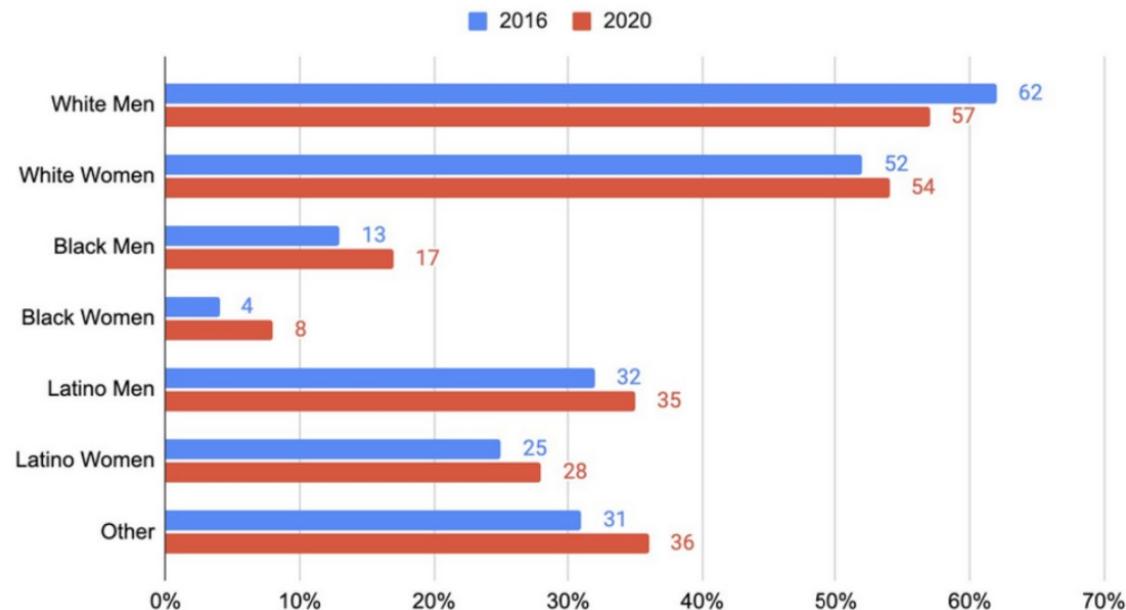


Différences dans ce qui est souligné ou minimisé ?

57

# Question de la statistique inférentielle : peut-on généraliser ces résultats de l'échantillon à la population ?

Support for Trump by Race and Gender (Edison Exit Poll)



En 2016, quelque 138 millions d'Américains avaient voté de manière anticipée (par courrier ou en personne).



## Echantillon

100 000 votants interviewés en 2020 à la sortie de 700 bureaux de vote

## Population

230 millions de votants potentiels

La statistique inférentielle élabore des méthodes pour estimer le risque qu'une tendance inexiste dans une population soit néanmoins observée dans un échantillon

# Question de la statistique inférentielle : peut-on généraliser ces résultats de l'échantillon à la population ?

Comment savoir si l'observation faite sur un échantillon (une situation particulière, des sujets particuliers) peut être généralisée (représente la population) ?

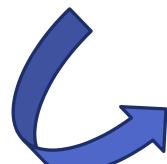


On veut savoir si la soupe est assez salée... On en goûte une cuillère.  
Mais le taux de sel dans la cuillerée goûtée représente-t-il le taux de sel dans toute la soupe ?

**Plusieurs possibilité de solutions....**

Recommencer  
l'observation ?  
- possible mais coûteux  
- on peut finir la soupe !

Estimer le **risque** que le taux de sel observé dans la cuillerée ne représente pas le taux de sel réel dans la soupe.



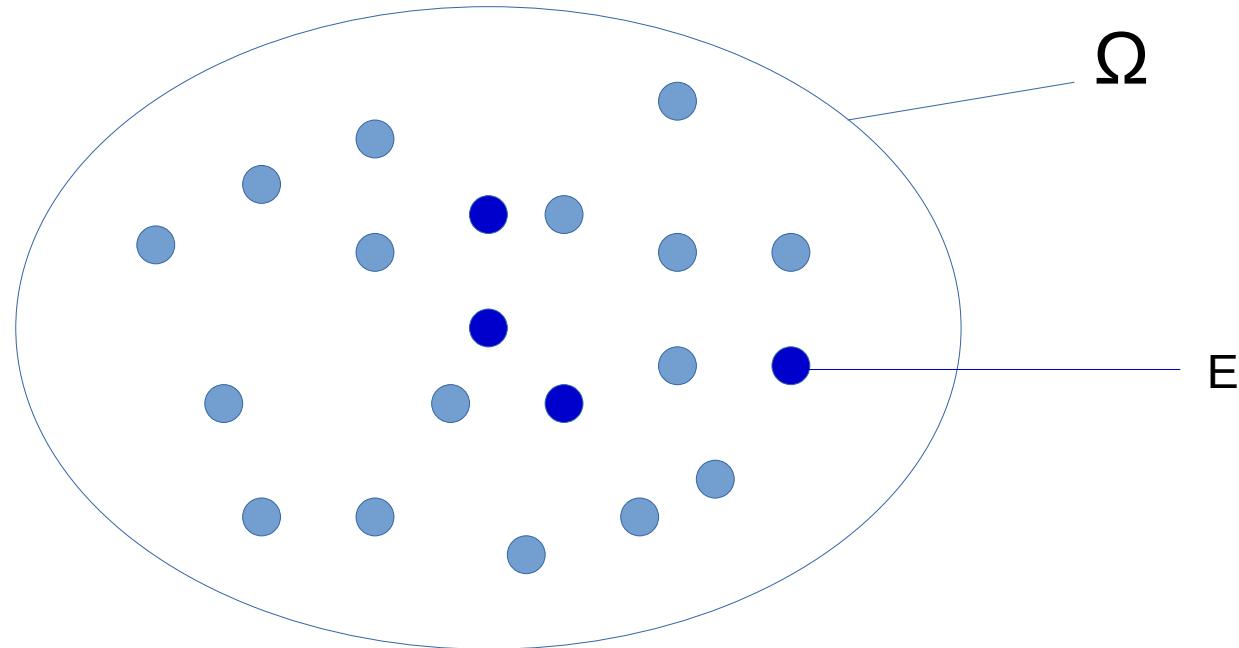
**Solution adoptée  
par les  
statisticiens**

# Parenthèse sur le calcul des probabilités

- Domaine des probabilités discrètes
  - notion d'événement (p.ex. on tire 6 au dé)
  - notion d'issue possible (p.ex. tirage d'un dé : 6 issues)
$$\Omega=\{1,2,3,4,5,6\}$$
- La probabilité mesure une proportion :
  - entre le nombre d'issues correspondant à un événement et le nombre d'issues possibles
  - p.ex. :
    - $p(\text{tirer un nombre pair})= 3/6 = 0.5 = 50 \%$
    - $p(\text{tirer un ou deux})= 2/6 = 0.33 = 33 \%$
  - dans une classe comptant 12 filles et 10 garçons, quelle est la probabilité de tirer aléatoirement un garçon ?

# Parenthèse sur le calcul des probabilités

- Représentation ensembliste

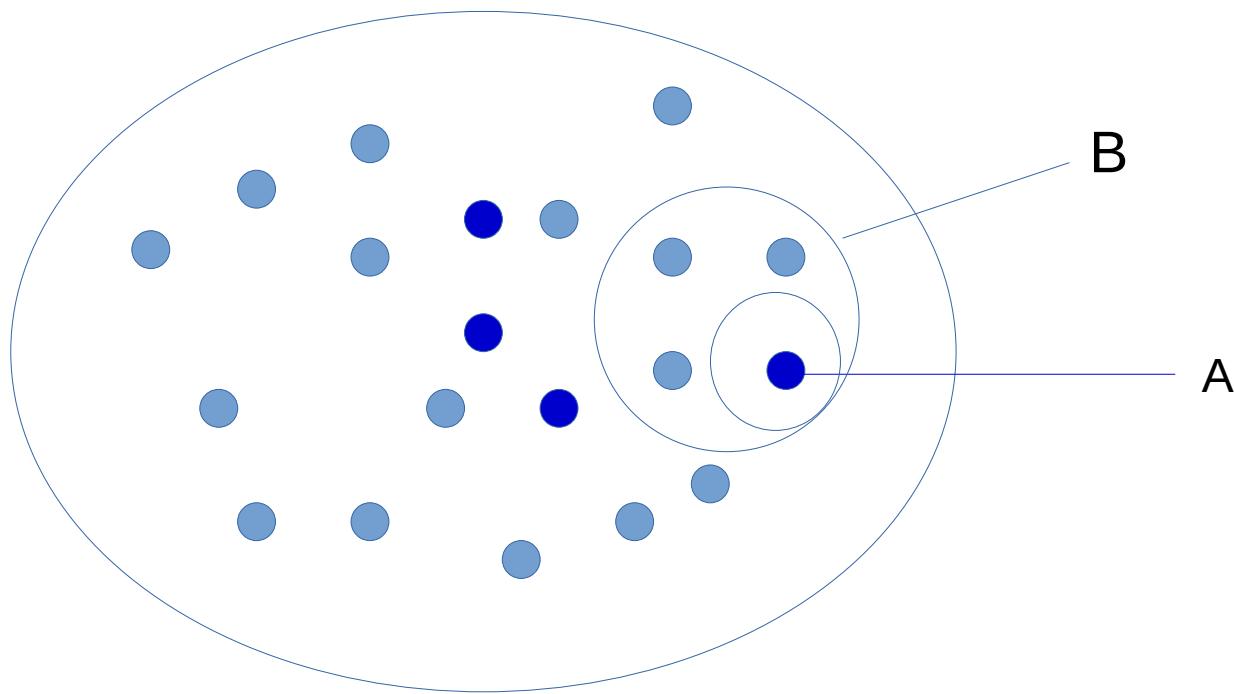


$$p(E) = |E|/|\Omega| = 4/20 = 0,2$$

ex. probabilité de tirer un mot du lexique des émotions dans un texte

# Parenthèse sur le calcul des probabilités

- Probabilité conjointe de deux événements :

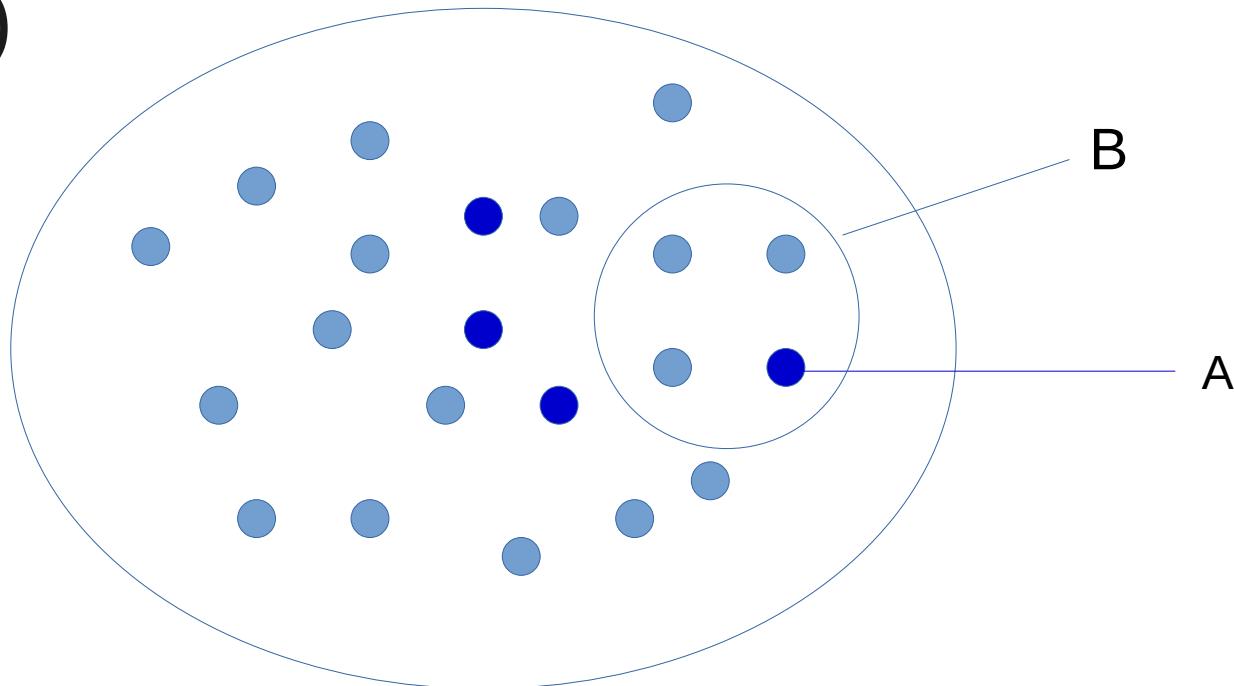


$$p(A \cap B) = |A \cap B| / |\Omega| = 1/20 = 0,05$$

ex. probabilité de tirer un adjectif du lexique des émotions

# Parenthèse sur le calcul des probabilités

- Probabilité conditionnelle : probabilité d'un événement sachant un autre événement  $p(A/B)$  (on change d'univers de référence)



$$p(A/B) = |A \cap B| / |B| = 1/4$$

ex. probabilité de tirer un mot du lexique  
des émotions au sein de l'ensemble des adjectifs

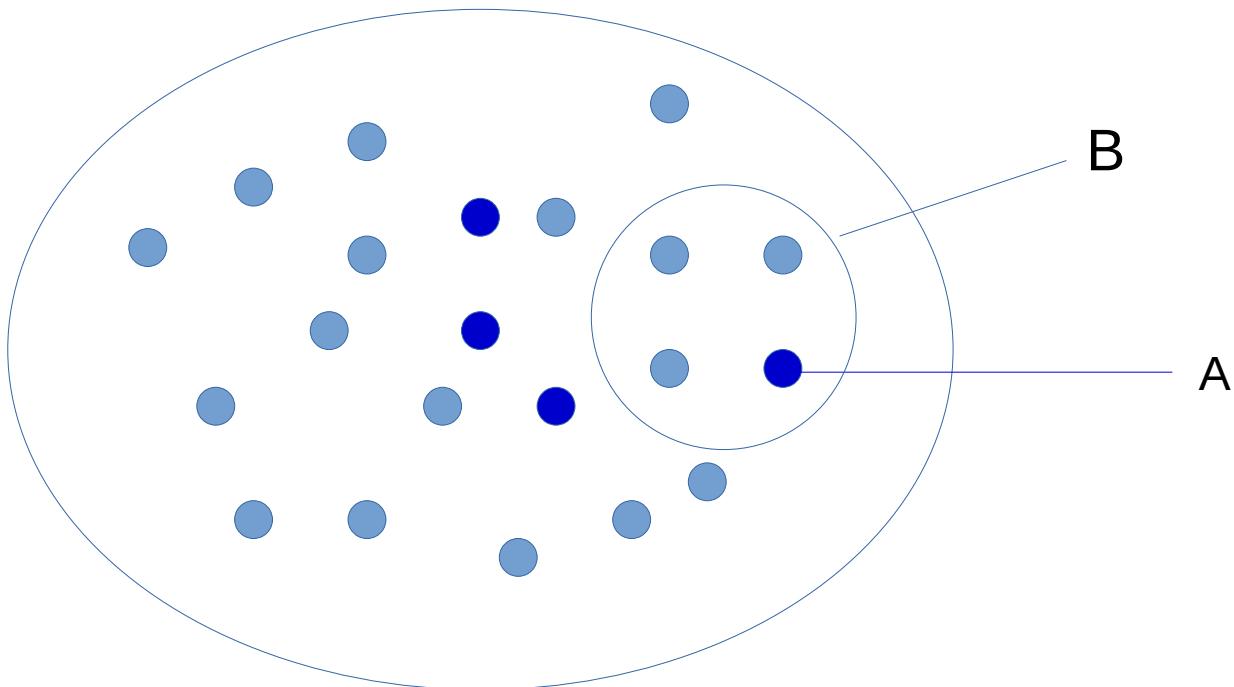
# Parenthèse sur le calcul des probabilités

- Attention  $P(A/B)$  peut être très différent de  $P(B/A)$ 
  - si plus d'une personne sur deux en réanimation à cause du Covid est non vaccinée  
 $P(\text{non vaccinée/soin critique}) = 58 \%$
  - ça ne veut pas dire qu'une personne non vaccinée sur deux est en soin critique  
 $P(\text{soin critique/non vaccinée}) = 4,21/100000 = 0,00421 \%$
- Sources :
  - [https://www.liberation.fr/checknews/quelle-est-la-part-des-non-vaccines-en-reanimation-et-a-l-hopital-dans-les-autres-pays-deurope-20220116\\_7DW3QX3DS5FRJD5G7TFPOH5O6M/](https://www.liberation.fr/checknews/quelle-est-la-part-des-non-vaccines-en-reanimation-et-a-l-hopital-dans-les-autres-pays-deurope-20220116_7DW3QX3DS5FRJD5G7TFPOH5O6M/)
  - <https://www.tf1info.fr/sante/cinquieme-vague-de-covid-19-les-non-vaccines-contre-le-coronavirus-sont-ils-largement-plus-touches-par-les-cas-graves-et-les-deces-a-l-hopital-2204882.htmlp>

Ces probabilités sont calculées sur des populations différentes !

# Parenthèse sur le calcul des probabilités

- Deux événements sont dits indépendants si et seulement si  $P(A,B)=P(A)*P(B)$



$p(A)=4/20$ ,  $P(B)=4/20$ ,  $p(A)*p(B)=16/400=0,04 \approx p(A,B) (0,05)$   
ici les deux événements sont proches de l'indépendance

# Parenthèse sur le calcul des probabilités

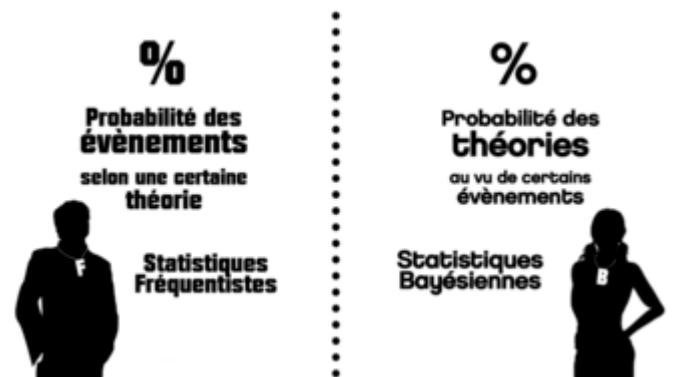
- L'indépendance se traduit par  $P(A/B)=P(A)$  et  $P(B/A)=P(B)$
- Par exemple : probabilité de tirer un 6 après un 6 au dé :  
 $p(6,6)=p(6)*p(6)=1/36$   
 $p(6/6)=p(6,6)/p(6)=1/6$
- Théorème de Bayes :  $P(A/B)=P(B/A)*p(A)/p(B)$

*Inférence bayesienne* : si on connaît dans le détail les causes possibles d'une conséquence observée, l'observation des effets permet de remonter aux causes. On peut donc estimer la probabilité d'une cause 'A' en fonction de son effet 'B' (utile p.ex. pour poser un diagnostic)

# Parenthèse sur le calcul des probabilités

- Inférence bayésienne
  - $P(A|B)$  : probabilité *a posteriori* de A sachant B.
  - $P(A)$  : probabilité *a priori* de A (dans la population)
  - $P(B)$  : probabilité *a priori* de B (dans la population)
  - $P(B|A)$  : fonction de vraisemblance de A

On cherche à calculer les probabilités (ou degré de confiance) de diverses causes hypothétiques à partir de l'observation d'événements connus.

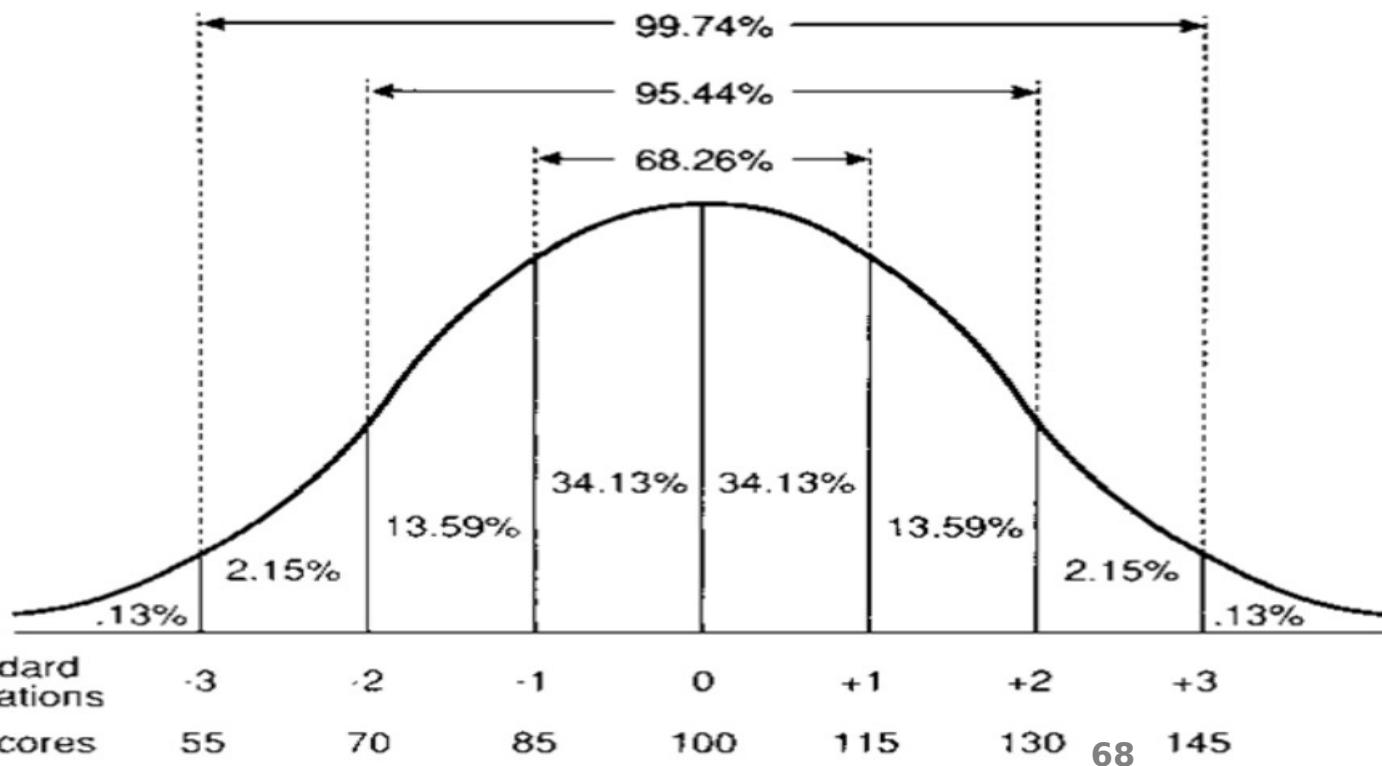


# Retour à la notion de statistique inférentielle : peut-on généraliser les résultats d'un échantillon à la population ?

Ex : expérience d'échantillonnage autour de la notion de Quotient Intellectuel : WAIS (Wechsler Adult Intelligence Scale)

Distribution de l'intelligence au sein d'une population caucasienne.

Moyenne: 100 Ecart-type: 15



Voir TP échantillonnage

# TP - Feuille Excel - Cas 1

- 1/ On a choisi une population de 5 sujets, 2 personnes Blondes 3 personnes Brunes
  - 2/ Dans cette population il n'y a aucune différence entre le QI des Blonds et le QI des Bruns. La moyenne est de 100 dans les deux groupes.
  - 3/ On s'est demandé dans combien d'échantillons de la population on retrouvait une différence entre Blonds et Bruns au profit des Blonds.  
On a fait la liste de tous les échantillons de 2, 3 ou 4 sujets contenant au moins un Blond et au moins un Brun: 20 échantillons  
Parmi ces échantillons, on se demande combien manifestent une différence entre Blonds et Bruns au profit des Blonds: 45% des échantillons
  - 4/ Conclusion
    - Dans 45 % des échantillons de la population on retrouve une différence au profit des Blonds
    - Dans 55 % des échantillons, on retrouve une différence au profit des bruns ou aucune différence.
- Un observateur qui voudrait inférer la tendance réelle de la population à partir d'un échantillon aurait 45 % de risques d'inférer à tort qu'il y a une différences entre Bruns et Blonds au profit des Blonds.

# Cas 1 : en bref

Valeur p = risque de généraliser à tort une tendance d'un échantillon à la population : autrement dit, risque qu'une tendance qui n'existe pas dans la population soit observée dans un échantillon

## Comment calculer ce risque ?

- 1/ On considère une population théorique (même effectif, où la variable est distribuée avec les mêmes caractéristiques : moyenne, variance, normalité, etc.) dans laquelle la tendance en question n'existe pas.
- 2/ On estime le nombre d'échantillons dans lesquels on retrouve la tendance en question au hasard (ou une tendance supérieure).
- 3/ La proportion d'échantillons qui répondent à ce critère représente le risque en question.

## Seuil conventionnel de significativité

Dans les SHS, on admet qu'un résultat est significatif quand la valeur p est inférieure ou égal au seuil conventionnel de 5%

$$p \leq 0.05 / p \leq .05$$

« Risque de rejeter à tort l'hypothèse nulle » = risque de premier type (erreur de première espèce) (p est parfois notée  $\alpha$ )

## Cas 2 : existence d'une corrélation

Dans la deuxième étude, on se situe dans le cas (fantaisiste!) où la couleur des cheveux serait vraiment corrélée au QI.

Cette hypothèse, dite alternative (par rapport à l'hypothèse nulle), est notée  $H_1$ . On effectue un échantillonnage, et :

- on constate que dans 10 % des observations, l'hypothèse  $H_1$  est rejetée.
- on a donc une probabilité de 10 % de rejeter à tort l'hypothèse  $H_1$

On parle dans ce cas de « risque de second type » ou « erreur de deuxième espèce ». La valeur  $p$  ne mesure pas le risque de second type , mais seulement le risque de premier type. On note  $\beta$  le risque de second type.

# En résumé

**Cas 1/** L'hypothèse nulle ( $H_0$ ) postule l'égalité entre des paramètres statistiques (par exemple, la moyenne) de deux échantillons dont on fait l'hypothèse qu'ils proviennent de populations équivalentes.

Il y a malgré tout un risque d'observer une différence dans un ou plusieurs échantillons (rejeter à tort  $H_0$  = **risque de premier type**)

**La valeur p estime ce risque**

**Cas 2/** L'hypothèse alternative ( $H_1$ ) postule une différence entre des paramètres statistiques (par exemple, la moyenne) de deux échantillons dont on fait l'hypothèse qu'ils proviennent de deux populations différentes.

Il existe malgré tout un risque de ne pas observer cette tendance dans un échantillon (rejeter à tort  $H_1$  = **risque de second type**)

Ce **n'est pas** ce risque que la valeur p estime

# En images

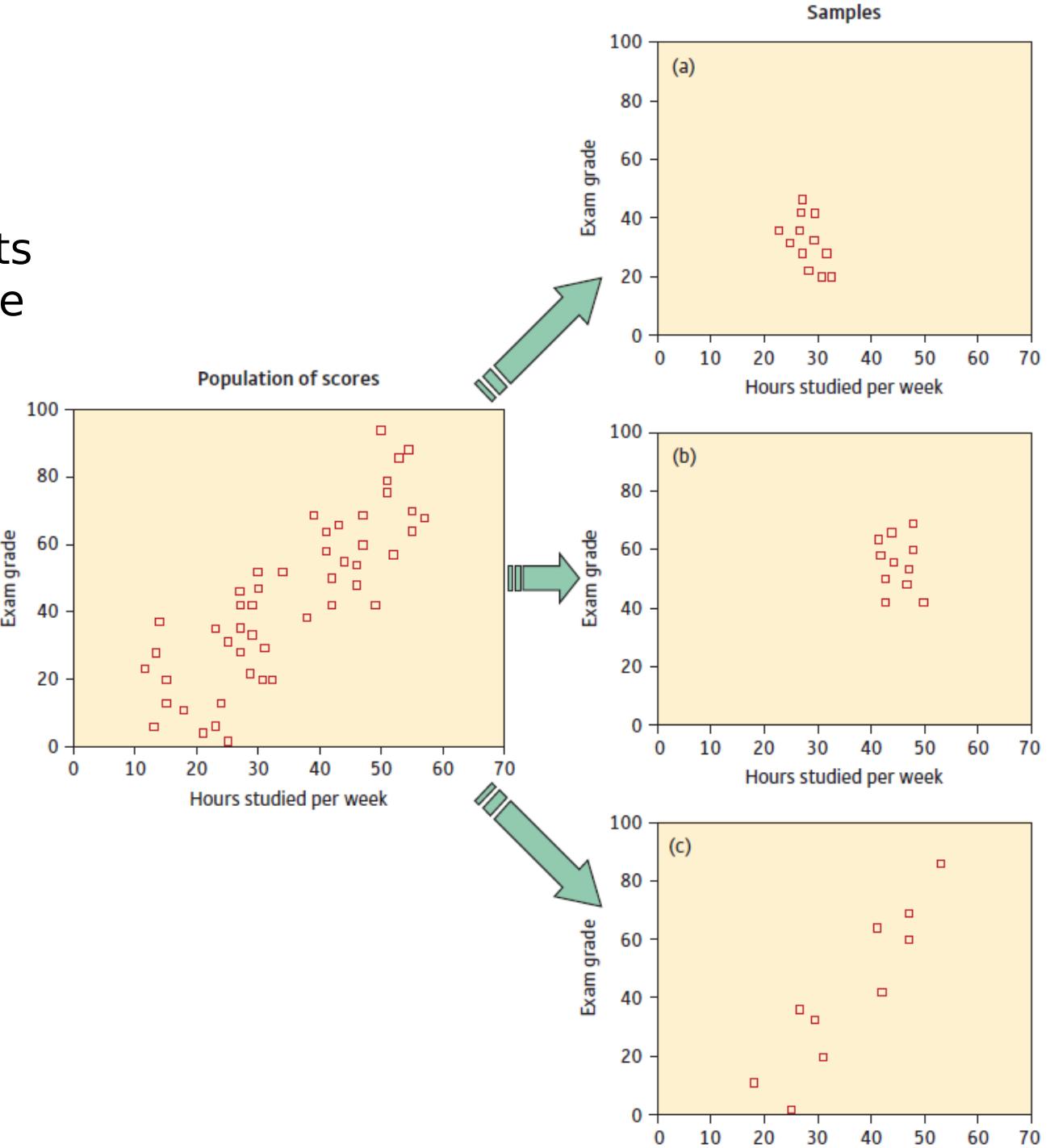
A gauche une population d'étudiants mettant en évidence un lien entre le nombre d'heures par semaine passées à étudier et les notes à l'examen final

A droite, trois échantillons de la population.

Lequel vous semble plus représentatif ?

Quel type de risque les deux autres échantillons font-ils courir ?

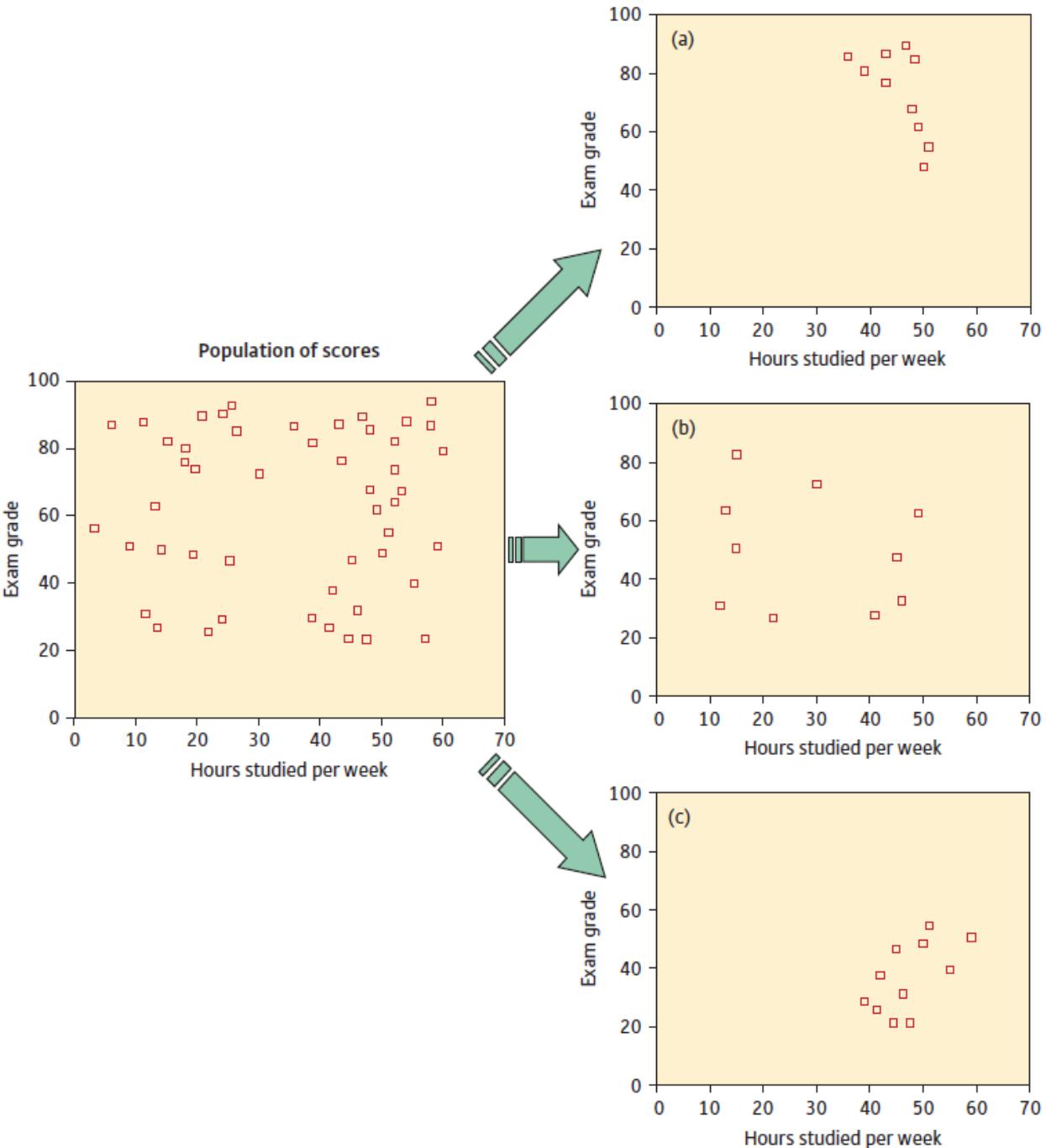
Est-ce ce risque qu'estime la valeur p ?



# En images

A gauche une population d'étudiants **ne mettant en évidence aucun lien** entre le nombre d'heures par semaine passées à étudier et les notes à l'examen final

A droite, trois échantillons de la population.  
Un des échantillons vous fait-il courir un risque et lequel ?  
Est-ce ce risque qu'estime la valeur p ?



# En Quizz - Vrai ou faux ?

**Dire qu'un résultat est significatif c'est :**

- 1- Affirmer qu'une tendance est forte (par exemple grande différence de moyenne)
- 2- Affirmer que le risque de trouver un résultat dans un échantillon sachant que la tendance **n'est pas** vérifiée dans la population est inférieur ou égal au seuil usuel de 0.05 ou 5%
- 3- Affirmer que le risque de ne pas trouver un résultat dans un échantillon sachant que la tendance est vérifiée dans la population est inférieur au seuil usuel de 0.05 ou 5%
- 4- Affirmer que le risque de généraliser (inférer) à tort un résultat à partir d'un échantillon est inférieur au seuil usuel de 0.05 ou 5%
- 5- Affirmer que le risque qu'un résultat soit faux est inférieur au seuil usuel de 0.05 ou 5%
- 6- Affirmer que le risque de rejeter à tort H<sub>0</sub> est inférieur au seuil usuel de 0.05 ou 5%

# En Quizz - Vrai ou faux ?

**Dire qu'un résultat est significatif c'est :**

- 1- Affirmer qu'une tendance est forte (par exemple grande différence de moyenne)
- 2- Affirmer que le risque de trouver un résultat dans un échantillon sachant que la tendance **n'est pas** vérifiée dans la population est inférieur ou égal au seuil usuel de 0.05 ou 5%
- 3- Affirmer que le risque de ne pas trouver un résultat dans un échantillon sachant que la tendance est vérifiée dans la population est inférieur au seuil usuel de 0.05 ou 5%
- 4- Affirmer que le risque de généraliser (inférer) à tort un résultat à partir d'un échantillon est inférieur au seuil usuel de 0.05 ou 5%
- 5- Affirmer que le risque qu'un résultat soit faux est inférieur au seuil usuel de 0.05 ou 5%
- 6- Affirmer que le risque de rejeter à tort H0 est inférieur au seuil usuel de 0.05 ou 5%

# **Conclusion sur l'inférence statistique et les tests de significativité**

**La proba p estimant la significativité d'un résultat, c'est :**

- Le risque de trouver un résultat dans un échantillon sachant que la tendance **n'est pas** vérifiée dans la population (dire que deux soupes sont salées différemment alors que ce n'est pas le cas)
- Le risque de généraliser (inférer) à tort un résultat à partir d'un échantillon.
- Le risque de rejeter à tort H0

**Ce n'est pas :**

- Le risque de ne pas trouver un résultat dans un échantillon sachant que la tendance est vérifiée dans la population (dire que deux soupes sont salées à l'identique alors qu'en fait elles le sont différemment)
- La probabilité qu'un résultat soit vrai
- La preuve qu'une tendance est forte
- La probabilité de l'hypothèse nulle

**On exprime ce risque par une probabilité p, qui varie de 0 à 1 et peut être convertie en pourcentage. Le maximum admis est 0.05 (cette valeur est un usage, pas une décision scientifique)**

# Conclusion sur l'inférence statistique et les tests de significativité

## Pourquoi le seuil de 5 % ?

- **Inconvénient d'un seuil trop élevé :**

Supposons qu'on fixe le seuil à 20 %. Cela signifierait que dans une étude sur 5, on affirmerait des hypothèses (populations différentes, variables corrélées, etc.) qui sont fausses. Ce serait trop laxiste.

- **Inconvénient d'un seuil trop faible :**

Supposons qu'on fixe un seuil très faible, par exemple 0,1 % (0.001). On aurait une chance sur 1000 de rejeter à tort l'hypothèse nulle. Le problème, c'est que ce seuil sera rarement atteint. Le risque d'erreur de second espèce (non acceptation d'une hypothèse qui est vraie) devient trop élevé. On ne conclut plus grand-chose avec une valeur aussi conservatrice.

Dans certains cas, on peut cependant admettre des valeurs plus conservatrices (comme 1 % voir 0.5%), par exemple quand on veut démontrer l'effet d'un médicament qui par ailleurs présente des risques pour la santé (il faut être sûr que la balance bénéfices/risques soit avantageuse pour le patient).

# Probabilités, pourcentages, et significativité

## A vous de compléter

Proba.	%	Conclusion	Proba.	%	Conclusion	Proba.	%	Conclusion
p = .01	1 %	significatif	p=.23			p=.305		
p=.05	5 %	significatif	p=.45			p=.0634		
p=.15	15 %	non significatif	p=.0001			p=.00009		
P=.0507	5.07 %	non significatif (tendanciel)	p=.99			p < .03		

# Estimation de l'erreur standard

Pour une variable donnée (à échelle d'intervalle), on appelle **distribution d'échantillonnage** la distribution de ses valeurs pour tous les échantillons possibles. La moyenne des moyennes des différents échantillons est identique à la moyenne de l'ensemble : en revanche les variations de ces moyennes sont intéressantes car elle permettent d'évaluer la probabilité d'avoir un écart important par rapport à la vraie moyenne (celle de la population). On appelle l'**écart type** de ces moyennes l'**erreur standard** (ou *erreur type* ou *erreur standard de la moyenne*).

Quand on a un échantillon de taille  $n$ , pour une distribution proche de la normalité, on peut estimer l'erreur standard de la moyenne de la variable V par une formule simple :

$$\text{écart type}(V) / \text{racine carré }(n-1)$$

**Exercice :** Reprendre le fichier XLS de la distribution du QI entre bruns et blonds. Calculer l'erreur standard. Vérifier avec Jamovi qu'on trouve bien les mêmes valeurs.

Effectuer sur Jamovi une analyse descriptive de votre distribution, globalement, puis en séparant en deux groupes.

N.B. : Dans la plupart des logiciels (SPSS, Calc, Jamovi) les valeurs d'écart type sont *corrigées* (c'est-à-dire que la variance est calculée en divisant par  $n-1$  et non par  $n$ )

## 2.5 Quatre types d'objectifs pour l'analyse statistique

### Décrire, Inférer, Modéliser et Prédire

# Consomme-t-on plus de glace quand la température augmente ?

Pendant 20 jours consécutifs du premier au vingt août 2020, on a mesuré la température à Dublin à midi et on a recensé le nombre de glaces vendues dans 10 boutiques du centre ville ce même jour.

Table 6.1 Data for the number of ice-cream cones sold on days with different temperatures

Ice-cream cones sold	Temperature	Ice-cream cones sold	Temperature
1000	26	550	14
950	22	600	19
870	19	700	21
890	20	750	22
886	19	800	22
900	21	850	24
560	17	950	22
550	16	1050	26
400	12	1000	26
500	13	1000	26

# Approche descriptive

On se demande s'il y a une relation entre température et nombre de glaces vendues. Les jours les plus chauds vend-on plus de glaces ?

Une façon de répondre à la question est de représenter chaque jour sur un plan avec deux axes : le nombre de glaces vendues et la température : diagramme de dispersion.

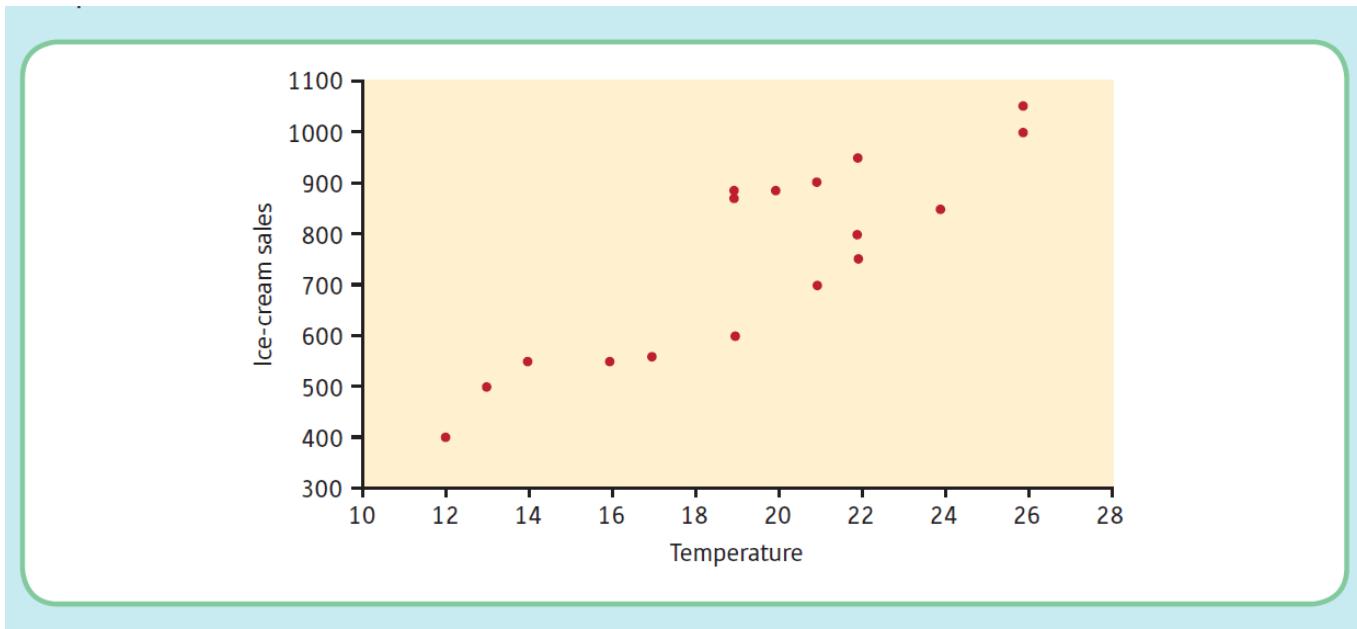


Diagramme de dispersion montrant le lien entre Température et vente de glaces

D'autres types de description ?

# Statistiques descriptives

On peut mesurer la corrélation linéaire observée visuellement, grâce au coefficient de Pearson.

Une façon de répondre est de calculer **un coefficient de corrélation r** associé à une valeur p qui estime le risque de d'obtenir une telle valeur par hasard.

Le coefficient de corrélation varie entre + 1 (les deux variables grandissent dans le même sens et sont parfaitement corrélées) et -1 (elles varient en sens inverses et sont parfaitement corrélées).

Matrice de corrélation

		A	B
A	r de Pearson	—	—
	valeur p	—	—
B	r de Pearson	0.893	—
	valeur p	< .001	—

Ici, la valeur p n'est pas à prendre en termes de statistiques inférentielles : elle quantifie le risque qu'il y a à estimer que nos deux variables sont liées. La significativité répond à la question : y a-t-il une corrélation linéaire entre A et B ?

# Statistiques inférentielles

Ces constats sont ils généralisables à l'ensemble de la population (p.ex. tous les marchands de glace de Dublin) ?

Le calcul de  $r$  peut être accompagné d'un intervalle de confiance, nous indiquant une borne inférieure et une borne supérieure entre lesquels le coefficient calculé sur l'ensemble de la population se situe probablement (avec une probabilité de 95%). On peut dans Jamovi afficher cet intervalle de confiance graphiquement, ainsi que dans le tableau contenant le calcul de  $r$ .

Matrice de corrélation

		A	B
A	$r$ de Pearson valeur p Borne sup de l'IC95% Borne inf de l'IC95%	— — — —	—
B	$r$ de Pearson valeur p Borne sup de l'IC95% Borne inf de l'IC95%	0.893 $< .001$ 0.957 0.745	— — — —

Ici on la borne inf. et sup. de l'intervalle de confiance qui correspondent à [0.745,0.957]. Ces valeurs sont calculées pour une probabilité de 95 % : cela signifie que le  $r$  de l'ensemble de la population ne soit pas dans cet intervalle est de 5 %.

# Statistiques inférentielles

L'intervalle de confiance de  $r$  se représente également graphiquement (ici avec Jamovi)

Matrice de corrélation

→

Coefficients de corrélation

Pearson  
 Spearman  
 Tau b de Kendall

Options additionnelles

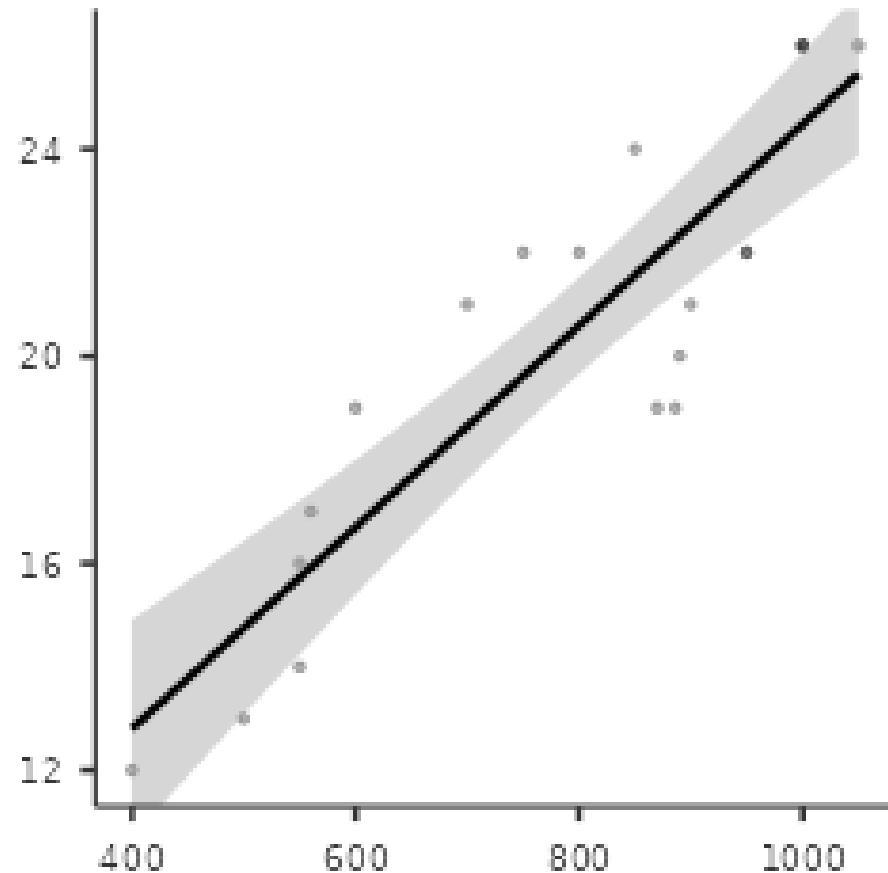
Rapport de significativité  
 Signaler les corrélations significatives  
 N  
 Intervalles de confiance  
Intervalle  %

Hypothèse

Corréle  
 Corréle positivement  
 Corréle négativement

Graphe

Matrice de corrélation  
 Densité des variables  
 Statistiques



# Approche en termes de statistiques prédictives ou de modélisation

On se demande s'il existe un modèle général de la relation entre température et nombre de glaces vendues

Une façon de répondre est de mettre en œuvre **une procédure de régression linéaire**. Cette procédure **calcule la droite qui s'ajuste le mieux au nuage de points** et la capacité de cette droite à rendre compte de tous les points.

## Régression linéaire

Mesures de l'ajustement du modèle

Modèle	R	R <sup>2</sup>
1	0.893	0.798

Coefficients du modèle - A

Prédicteur	Estimation	Erreur standard	t	p
Ordonnée à l'origine	-46.3	101.12	-0.457	0.653
B	41.0	4.87	8.421	< .001

Cette droite nous dit que lorsque la température augmente de 1 degré, le nombre de glaces vendues augmente d'environ 41 unités, suivant la formule :

$$\text{GlacesVendues} = 41 * T^\circ - 46,3$$

**L'erreur standard** ou erreur type est l'écart type de la **distribution d'échantillonnage** : elle permet d'estimer un intervalle de confiance pour les valeurs estimées ( $\pm 1.96 * \text{ES}$  c'est l'intervalle où se situe notre paramètre avec une probabilité de 95%) .

# Exercice

- Ouvrir glaces.xlsx. Rajouter une variable ordinaire indiquant trois types de température : frais, doux, chaud
- Importez sur Jamovi les données de glaces.xlsx et étudiez la corrélation entre les variables.
- Afficher différents diagrammes montrant le croisement de nos variables.
- Calculez la prédiction du nombre de glaces vendues pour des températures de 5°C et de 31°C. Indiquez à chaque fois un intervalle de confiance.

## 2.6 Analyses univariées, bivariées, multivariées

# Les analyses univariées ne concernent qu'une seule variable

Analyses univariées (exemples)	Mise en œuvre																																																																						
Calculer les paramètres de la distribution d'une variable quantitative	<p>Variables quanti d'un échantillon de 1000 comptes twitter</p> <table border="1" data-bbox="1088 345 2215 727"> <thead> <tr> <th colspan="7">Statistiques descriptives</th> </tr> <tr> <th></th> <th>N</th> <th>Plage</th> <th>Minimum</th> <th>Maximum</th> <th>Moyenne</th> <th>Erreur standard</th> </tr> <tr> <th></th> <th>Statistiques</th> <th>Statistiques</th> <th>Statistiques</th> <th>Statistiques</th> <th>Statistiques</th> <th></th> </tr> </thead> <tbody> <tr> <td>Revenu_par_unité</td> <td>1000</td> <td>20373,94359</td> <td>8779,256410</td> <td>29153,20000</td> <td>20091,19030</td> <td>118,0389597</td> </tr> <tr> <td>Densité_population</td> <td>1000</td> <td>72874,03800</td> <td>63,46200000</td> <td>72937,50000</td> <td>5793,633151</td> <td>306,5628398</td> </tr> <tr> <td>Proportion_proprio</td> <td>1000</td> <td>,4810000000</td> <td>,0500000000</td> <td>,5310000000</td> <td>,2280580000</td> <td>,0023215625</td> </tr> <tr> <td>Nombre_contact</td> <td>1000</td> <td>2231</td> <td>1</td> <td>2232</td> <td>59,10</td> <td>3,924</td> </tr> <tr> <td>Densité_réseau</td> <td>1000</td> <td>1,0000000000</td> <td>,0000000000</td> <td>1,0000000000</td> <td>,1255218582</td> <td>,0042683843</td> </tr> <tr> <td>Age</td> <td>799</td> <td>71,0</td> <td>14,0</td> <td>85,0</td> <td>27,621</td> <td>,3509</td> </tr> <tr> <td>N valide (liste)</td> <td>799</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	Statistiques descriptives								N	Plage	Minimum	Maximum	Moyenne	Erreur standard		Statistiques	Statistiques	Statistiques	Statistiques	Statistiques		Revenu_par_unité	1000	20373,94359	8779,256410	29153,20000	20091,19030	118,0389597	Densité_population	1000	72874,03800	63,46200000	72937,50000	5793,633151	306,5628398	Proportion_proprio	1000	,4810000000	,0500000000	,5310000000	,2280580000	,0023215625	Nombre_contact	1000	2231	1	2232	59,10	3,924	Densité_réseau	1000	1,0000000000	,0000000000	1,0000000000	,1255218582	,0042683843	Age	799	71,0	14,0	85,0	27,621	,3509	N valide (liste)	799					
Statistiques descriptives																																																																							
	N	Plage	Minimum	Maximum	Moyenne	Erreur standard																																																																	
	Statistiques	Statistiques	Statistiques	Statistiques	Statistiques																																																																		
Revenu_par_unité	1000	20373,94359	8779,256410	29153,20000	20091,19030	118,0389597																																																																	
Densité_population	1000	72874,03800	63,46200000	72937,50000	5793,633151	306,5628398																																																																	
Proportion_proprio	1000	,4810000000	,0500000000	,5310000000	,2280580000	,0023215625																																																																	
Nombre_contact	1000	2231	1	2232	59,10	3,924																																																																	
Densité_réseau	1000	1,0000000000	,0000000000	1,0000000000	,1255218582	,0042683843																																																																	
Age	799	71,0	14,0	85,0	27,621	,3509																																																																	
N valide (liste)	799																																																																						
Calculer la distribution d'une variable qualitative ou ordinaire	<p>Distribution des tranches d'âge dans l'échantillon de 1000 comptes twitter.</p> <p>Transformation de l'âge, variable continue, en tranches, variable ordinaire discrète :</p> <p>T1 &lt; 25, T2 de 25 à 34, T3 &gt; 34.</p> <table border="1" data-bbox="1498 842 2317 1188"> <thead> <tr> <th colspan="6">Tranche_age</th> </tr> <tr> <th></th> <th></th> <th>Fréquence</th> <th>Pourcentage</th> <th>Pourcentage valide</th> <th>Pourcentage cumulé</th> </tr> </thead> <tbody> <tr> <td rowspan="3">Valide</td> <td>T1</td> <td>489</td> <td>48,9</td> <td>61,2</td> <td>61,2</td> </tr> <tr> <td>T2</td> <td>165</td> <td>16,5</td> <td>20,7</td> <td>81,9</td> </tr> <tr> <td>T3</td> <td>145</td> <td>14,5</td> <td>18,1</td> <td>100,0</td> </tr> <tr> <td>Total</td> <td>799</td> <td>79,9</td> <td>100,0</td> <td></td> <td></td> </tr> <tr> <td>Manquant</td> <td>201</td> <td>20,1</td> <td></td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td>1000</td> <td>100,0</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	Tranche_age								Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé	Valide	T1	489	48,9	61,2	61,2	T2	165	16,5	20,7	81,9	T3	145	14,5	18,1	100,0	Total	799	79,9	100,0			Manquant	201	20,1				Total	1000	100,0																											
Tranche_age																																																																							
		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé																																																																		
Valide	T1	489	48,9	61,2	61,2																																																																		
	T2	165	16,5	20,7	81,9																																																																		
	T3	145	14,5	18,1	100,0																																																																		
Total	799	79,9	100,0																																																																				
Manquant	201	20,1																																																																					
Total	1000	100,0																																																																					
Tester si la moyenne d'une variable qualitative est significativement différente d'une valeur constante	Comparer la moyenne du maintien du « ne » de la négation dans les tweets écrits par les titulaires des 1000 comptes ( $m = 31,1\%$ ) au maximum qu'on a pu trouver dans des corpus d'interactions orales non-formelles (15.5%).																																																																						

# Les analyses bivariées concernent deux variables

## Analyses bivariées (exemples)

**Etudier la relation entre une variable qualitative (Statut Socio Economique - SSE) et une variable quantitative (taux de liaisons correctes)**

## Mise en œuvre au niveau descriptif (par des visualisations ou des tables)

Que remarquez vous ?

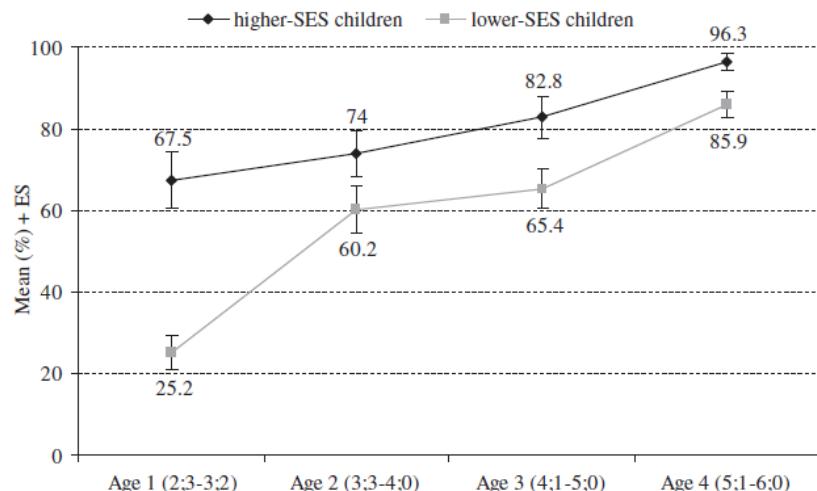
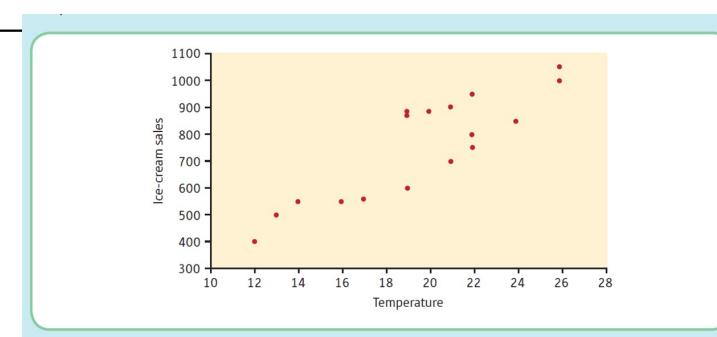


Fig. 1. Obligatory liaison: evolution with age of correct productions in higher and lower-SES children (mean percentages + standard errors).

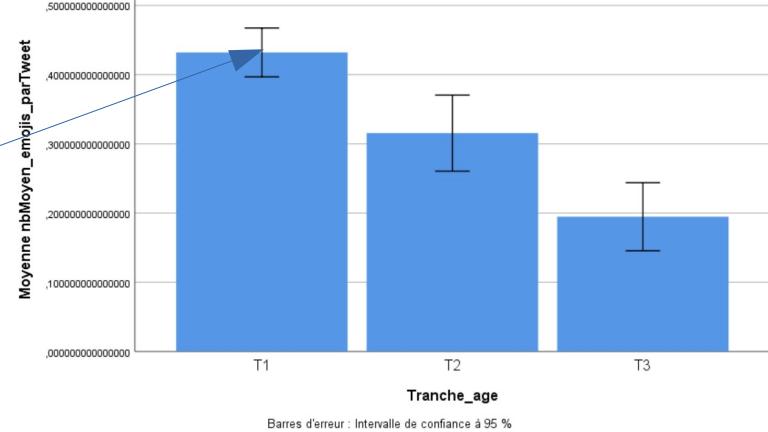
Chevrot, Nardy & Barbu, 2011

## Observer la relation (covariation) entre deux variables quantitatives

Voir diapo sur la consommation de glaces



# Les analyses bivariées concernent deux variables

Analyses bivariées (exemples)	Mise en œuvre au niveau descriptif (par des visualisations ou des tables)																	
Observer la relation (covariation) entre deux variables qualitatives	<p>Sous forme de <b>tableau croisé</b>. Ici le croisement du sexe et des tranches d'âge des utilisateurs de Twitter.</p> <p><b>Que remarquez vous ?</b></p>  <table border="1"><thead><tr><th rowspan="2">Tranche_age</th><th colspan="2">Sexe</th></tr><tr><th>f</th><th>h</th></tr><tr><th>Nombre</th><th>Nombre</th><th></th></tr></thead><tbody><tr><td>T1</td><td>291</td><td>198</td></tr><tr><td>T2</td><td>63</td><td>102</td></tr><tr><td>T3</td><td>46</td><td>99</td></tr></tbody></table>	Tranche_age	Sexe		f	h	Nombre	Nombre		T1	291	198	T2	63	102	T3	46	99
Tranche_age	Sexe																	
	f	h																
Nombre	Nombre																	
T1	291	198																
T2	63	102																
T3	46	99																
Observer la relation (covariation) entre une variable qualitative et une variable quantitative	<p>Sous forme de visualisation reliant le nombre moyen d'émoticônes par tweet aux tranches d'âge.</p> <p>Les <b>barres d'erreurs</b> représentent la fourchette dans laquelle chacune des 3 moyennes a 95% de chances de se situer.</p> <p><b>Que remarquez vous ?</b></p> <p><b>Observation généralisable ?</b></p>  <p>The bar chart displays the average number of emojis per tweet across three age groups: T1, T2, and T3. The y-axis represents the average number of emojis, ranging from 0.000,000,000,000,000 to 0,500,000,000,000,000. The x-axis shows the age groups T1, T2, and T3. The bars are blue, and each has a black error bar representing a 95% confidence interval. An arrow points to the top of the T1 bar. The chart includes a legend at the bottom stating "Barres d'erreur : Intervalle de confiance à 95 %".</p> <p>0,4 émoticônes par tweet signifie presque un émoticône tous les 2 tweets 0,3 : presque un émoticône tous les 3 tweets 0,2 : un émoticône tous les 5 tweets</p>																	

# Deux types principaux d'analyses multivariées

NB : Elles sont très nombreuses et très spécialisées selon les champs scientifiques. Par exemple, les économistes utilisent beaucoup la régression linéaire. Les psychologues expérimentalistes utilisaient beaucoup l'analyse de variance et maintenant les modèles mixtes. En textométrie on utilise souvent l'analyse factorielle des correspondances (AFC).

## Méthodes descriptives

« Les méthodes descriptives ont pour objectif d'aider à structurer et résumer un ensemble de données issues de plusieurs variables, sans privilégier particulièrement l'une de ces variables. Toutes les variables sont donc prises en compte au même niveau. Les traitements et représentations graphiques visent à apporter une vision globale la plus exacte possible de l'ensemble des données analysées, en minimisant la déperdition d'information.»

On voit plus tard si on a le temps

## Méthodes explicatives

« Les méthodes explicatives ont, quant à elles, pour objectif d'expliquer l'une des variables (dite dépendante) à l'aide de deux ou plusieurs variables explicatives (dites indépendantes)»

## 2.7 Analyses paramétriques vs. non paramétriques

# QUELLE EST LA DIFFÉRENCE ENTRE UN TEST PARAMÉTRIQUE ET UN TEST NON-PARAMÉTRIQUE ?

Quelle est la différence entre un test paramétrique et un test non-paramétrique ?

**Les tests paramétriques** se basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable. Par exemple, le test t de Student pour échantillons indépendants n'est fiable que si les données associées à chaque échantillon suivent une distribution normale et si les variances des échantillons sont homogènes.

**Les tests non-paramétriques** ne se basent pas sur des distributions statistiques. Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées.

<https://help.xlstat.com/s/article/guide-de-choix-de-test-statistique?language=fr>

On verra plus loin l'exemple d'un test paramétrique - test T de Student - qui permet de généraliser une différence de moyennes de l'échantillon à la population (en calculant une p-value).

La moyenne est une façon d'estimer la tendance centrale d'une distribution de valeur, c'est-à-dire trouver une seule valeur - celle de la moyenne - qui représente bien l'ensemble de la série.

Pour que ce test paramétrique fonctionne, il faut que la moyenne et l'écart-type représentent bien la distribution. Mais est-ce le cas de toutes les distributions ?

**Pour qu'un test paramétrique de comparaison de moyennes fonctionne, il faut que la moyenne représente bien la distribution.**

## **Un exemple extrême**

Analysons la distribution des fréquences absolues pour chaque vocabulaire d'un texte.

### **Exercice :**

- Analyser un texte avec Anatext et exporter le tableau donnant les fréquences de « Toutes les formes ».
- Importer le tableau au format CSV dans Jamovi.
- Analyser la variable « fréquence ». Au vu de ses paramètres (moyenne, min., max., mode, médiane), peut-on dire que la moyenne représente bien la distribution ?
- Afficher la distribution sous forme d'histogramme, et exporter le fichier en PDF.

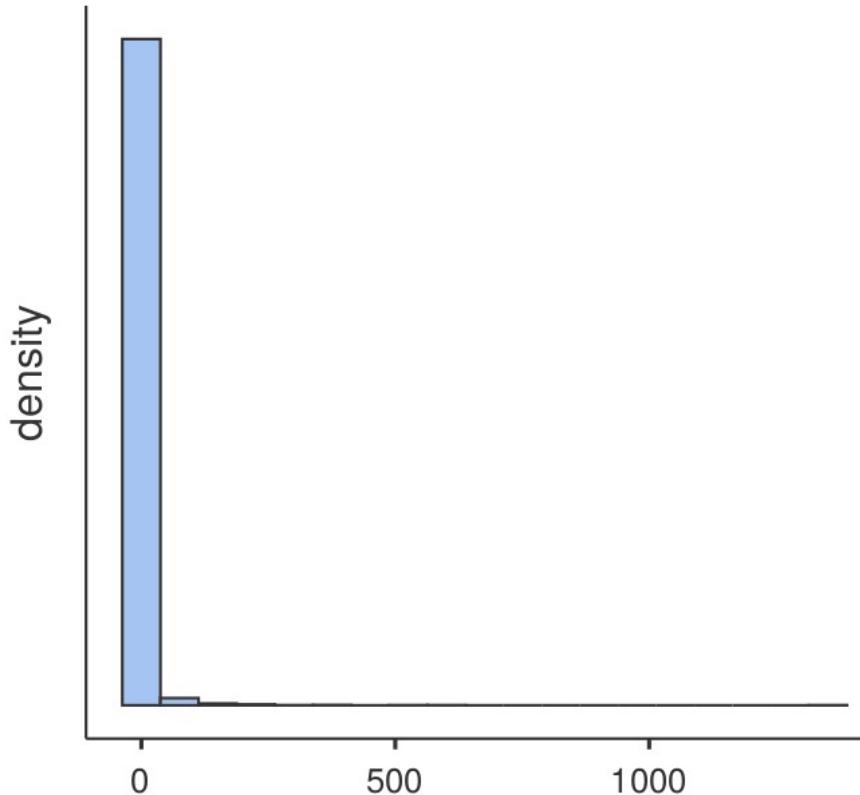
**Pour qu'un test paramétrique de comparaison de moyennes fonctionne, il faut que la moyenne représente bien la distribution.**

## Un exemple extrême

On voit que la distribution est très déséquilibrée du côté des hapax (à la fois mode et médiane). Le test de Shapiro-Wilk permet de rejeter l'hypothèse de normalité de manière significative.

Statistiques descriptives

	Fréquence
N	3760
Manquants	0
Moyenne	4.42
Médiane	1.00
Mode	1.00
Ecart-type	30.4
Minimum	1
Maximum	1353
W de Shapiro-Wilk	0.0746
Valeur p de Shapiro-Wilk	< .001



## **Pour qu'un test paramétrique de comparaison de moyennes fonctionne, il faut que la moyenne représente bien la distribution.**

### **Exemple d'une zone rurale où :**

- Il y a 99 fermiers avec un revenu annuel de 24 000 euros, avec de petites exploitations.
- Et 1 fermier qui a beaucoup investi, qui a 30 salariés et qui a racheté de nombreuses petites exploitations. Son revenu est de 1 000 000 d'euros.

La moyenne de la distribution des revenus est de : 33 760 euros

### **Cette moyenne ne représente pas la population de fermiers**

- Elle est supérieure aux revenus de 99% de fermiers
- Elle ne représente pas non plus le fermier très riche.

La mesure de tendance centrale qui représente le mieux la population est ici le mode, la valeur la plus représentée, qui est 24 000 et qui représente 99% des fermiers de la région.

# Quel est l'avantage des tests non paramétriques et quand les utiliser ?

Les tests paramétriques ont souvent des tests non-paramétriques équivalents. Vous trouverez une série de tests paramétriques courants ainsi que leurs équivalents non-paramétriques, lorsqu'ils existent, dans [cette grille](#).

## Quel est l'avantage d'utiliser un test non-paramétrique ?

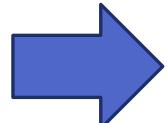
Les tests non-paramétriques sont plus **robustes** que les tests paramétriques. En d'autres termes, peuvent être utilisés dans un plus grand nombre de situations.

## Quel est l'avantage d'utiliser un test paramétrique ?

Les tests paramétriques sont, eux, plus **puissants** en général que leurs équivalents non-paramétriques. En d'autres termes, un test paramétrique sera plus apte à aboutir à un rejet de  $H_0$ , si ce rejet est justifié. La plupart du temps, la p-value calculée par un test paramétrique sera inférieure à la p-value calculée par un équivalent non-paramétrique exécuté sur les mêmes données.

<https://help.xlstat.com/s/article/guide-de-choix-de-test-statistique?>

En pratique, il faut utiliser des tests non paramétriques prioritairement dans les situations suivantes



- 1) Nombre limité d'observations (moins de 10)
- 2) La moyenne d'un des échantillons ne représente pas la tendance centrale
- 3) Les échantillons impliqués ont des dispersions inégales
- 4) Une des variables prend un nombre limité de valeurs (exemple des 100 fermiers dont les revenus ne prennent que deux valeurs)

## 2.8 Panorama général des méthodes statistiques

100

100

# Des centaines de méthodes statistiques...

Par exemple dans JASP, la plupart des tests sont proposés sous deux formes

- Classique : basée sur les fréquences observées
- Bayésienne : basée sur des connaissances a priori sur le résultat (Thomas Bayes, statisticien du 18<sup>ème</sup> siècle)

Sans formation spécifique, utiliser les statistiques fréquentielles.



échantillon_1000_2_infos (C:\Users\Jean-Pierre\Documents\Documents du DELL\Données et en cours)\2021a Tw			
	Descriptives	T-Tests	ANOVA
1	TwitterUID	Classical	th_tweets
2	1045965164	Independent Samples T-Test	8.9
3	1135411620	Paired Samples T-Test	7.4
4	778777092	One Sample T-Test	12
5	102049249		13
6	574387628		11.
7	2282446951	Bayesian	8.3
8	2799795709	Independent Samples T-Test	8.7
9	328960852	Paired Samples T-Test	20
10	1600678238	One Sample T-Test	13
11	0.66667		10
12	0.06667		12.07960199

Dans JASP ou SPSS, certaines méthodes sont basées sur l'IA (*Machine Learning ou Apprentissage machine*)

- Le système apprend des régularités sur une partie des données
- Il teste sa capacité à prédire ces régularités sur une autre partie des données.



Twitter samples 1000 comptes\9. 4 autres échantillons de 1000		
arn Bayes	Machine Learning	Meta-Analysis
	Regression	Results
	Boosting	Network
	K-Nearest Neighbors	Summary of N
	Random Forest	Number of n
	Regularized Linear	
	Classification	
	Boosting	
	K-Nearest Neighbors	
	Linear Discriminant	
	Random Forest	
	Clustering	
	Density-Based	
	Fuzzy C-Means	
	Hierarchical	
	K-Means	
	Random Forest	

Enfin de nombreuses approches statistiques sont spécifiques à un domaine

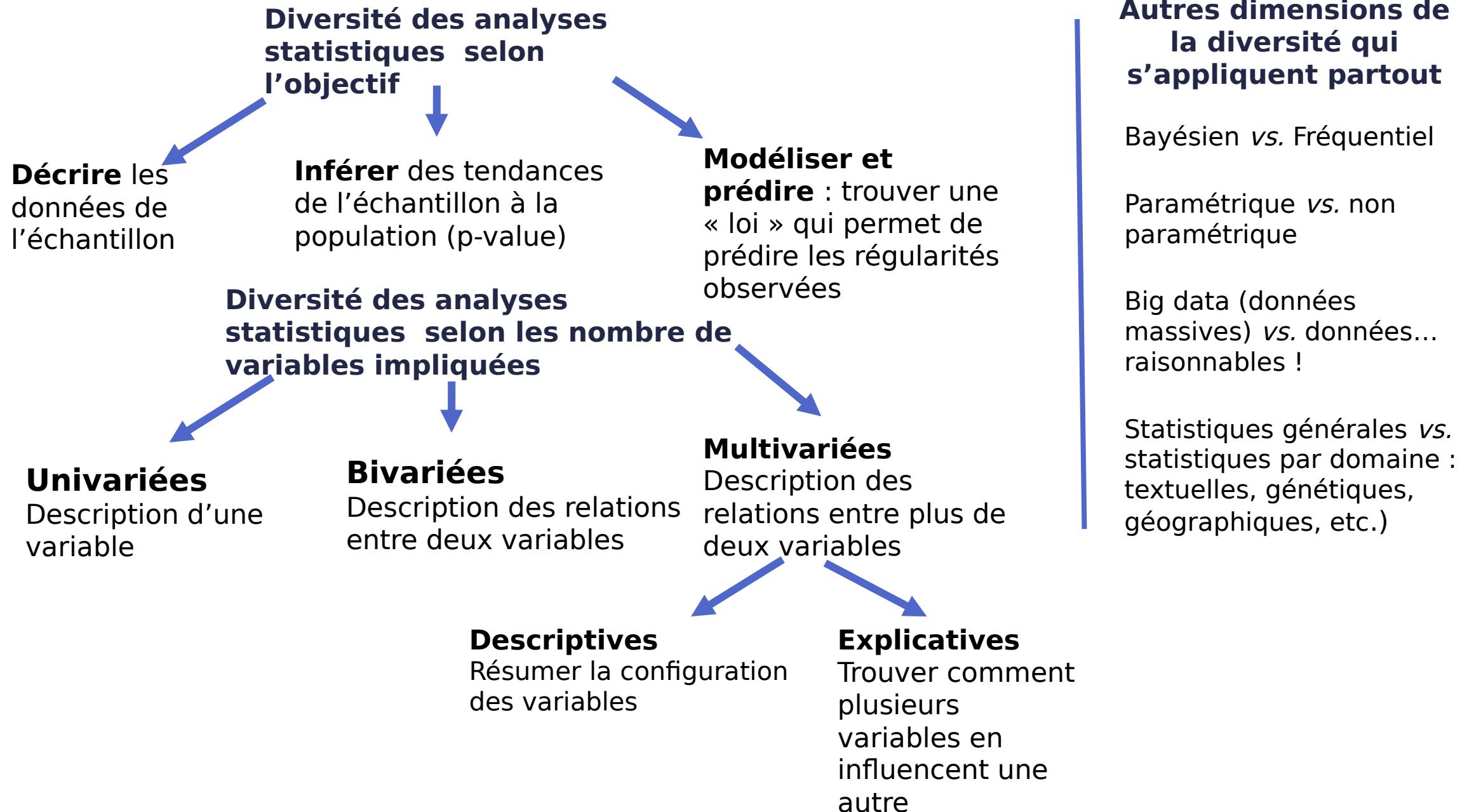


Par exemple, la **théorie des valeurs extrêmes** qui étudie le maxima de valeurs prises par les valeurs des variables.

- en hydrologie pour prévoir les crues
- en océanographie dans l'étude des vagues scélérates
- en finance pour prévoir les booms/ les krachs

[https://fr.wikipedia.org/wiki/Th%C3%A9orie\\_des\\_valeurs\\_extr%C3%A8mes](https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_valeurs_extr%C3%A8mes)

# La diversité des méthodes statistiques (esquisse)



# **3. analyses univariées de variables qualitatives et quantitatives**

## **3.1 Préparer ses données pour l'analyse statistique**

# Principes généraux

- 1/ On saisit dans Excel ou Calc sur LibreOffice car c'est le plus commode. On peut aussi saisir collaborativement sur un tableur en ligne (ex : Google sheet)
- 2/ On importe le fichier Excel ou autre directement à partir de Jamovi ou SPSS
- 3/ Les « objets » qu'on observe sont en ligne et les variables en colonne. Les noms des objets apparaissent dans la première colonne et les noms des variables dans la première ligne
- 4/ On laisse les cases des données manquantes vides (pas de NR, ni 0, ni ABS, etc.)
- 5/ Les noms des variables sont écrits sur une seule ligne et ne comportent ni espaces, ni accents
- 6/ On écrit à part un catalogue des variables précisant leur statut dans l'explication (explicative ou à expliquer) et leur nature (nominale, ordinaire, continue).

**Les « objets » qu'on observe sont en ligne et les variables en colonne. Les noms des « objets » apparaissent dans la première colonne et les noms des variables dans la première ligne**

**Exemple de la thèse de W.** : 12 chinois apprenants avancés du français traduisent 86 métaphores

1/ Pour la moitié d'entre eux : traduction du français vers le chinois, et pour l'autre moitié, traduction du chinois vers le français → **Variable explicative nominale 1 : sens de la traduction (zh → fr ou fr → zh)**

2/ Les métaphores sont construite dans un champs conceptuel → Colère = champ conceptuel de la couleur en Français : « une colère noire », « la colère a assombri son regard », « rouge de colère »

- La moitié des métaphores sont spécifiques à une langue : champs conceptuels différents en zh et en fr
- La moitié des métaphores sont non spécifiques : même champ conceptuel dans les deux langues

→ **Variable explicative nominale 2 : spécificité de la métaphore (spécifique ou non spécifique)**

**Les « objets » qu'on observe sont en ligne et les variables en colonne. Les noms des « objets » apparaissent dans la première colonne et les noms des variables dans la première ligne**

3/ Pour chaque traduction on code trois types de réponses. Et on compte chacune des réponses.

M-0 : la métaphore est traduite par une séquence sans métaphore

M-M : la métaphore est traduite par une métaphore du même champ conceptuel

M1-M2 : la métaphore est traduite par une métaphore d'un autre champ conceptuel

→ **Trois variables quantitatives à expliquer : M0, MM, M1M2**

**Comment présenter ces données dans un tableau ?**

**Les « objets » qu'on observe sont en ligne et les variables en colonne. Les noms des « objets » apparaissent dans la première colonne et les noms des variables dans la première ligne**

Est-ce que cette présentation convient ? Pourquoi ?

A	B	C	D	E	F	G
Groupe1 : vers chinois	Score M-0 sur les 18 metaphores spécifiques F-Ch	Score M-M sur les 18 metaphores spécifiques F-Ch	Score M1-M2 sur les 18 metaphores spécifiques F-Ch	Score M-0 sur les 29 metaphores non spécifiques F-Ch	Score M-M sur les 29 metaphores non spécifiques F-Ch	Score M1-M2 sur les 29 metaphores non spécifiques F-Ch
Part 1	0,78	0,00	0,22	0,31	0,55	0,14
Part 2	0,67	0,06	0,28	0,31	0,69	0,00
Part 3	0,67	0,00	0,33	0,28	0,59	0,14
Part 4	0,72	0,06	0,22	0,28	0,45	0,28
Part 5	0,83	0,00	0,17	0,24	0,52	0,24
Part 6	0,61	0,06	0,33	0,17	0,59	0,24
Groupe2 : vers français	Score M-0 sur les 14 metaphores spécifiques Ch-F	Score M-M sur les 14 metaphores spécifiques Ch-F	Score M1-M2 sur les 14 metaphores spécifiques Ch-F	Score M-0 sur les 26 metaphores non spécifiques Ch-F	Score M-M sur les 26 metaphores non spécifiques Ch-F	Score M1-M2 sur les 26 metaphores non spécifiques Ch-F
Part 1	0,79	0,00	0,21	0,35	0,46	0,19
Part 2	0,93	0,00	0,07	0,81	0,15	0,04
Part 3	0,86	0,00	0,14	0,50	0,31	0,19
Part 4	0,86	0,07	0,07	0,62	0,31	0,08
Part 5	0,79	0,07	0,14	0,42	0,42	0,15
Part 6	0,93	0,00	0,07	0,54	0,31	0,15

**Les « objets » qu'on observe sont en ligne et les variables en colonne.  
Les noms des « objets » apparaissent tous dans la première colonne et  
les noms des variables tous dans la première ligne**

**Est-ce que cette présentation convient ? Pourquoi ?**

Participants	Sens	M0 spe	MM spe	M1M2 spe	M0 nonspe	MM nonspe	M1M2 nonspe
Part 1	1	0,78	0,00	0,22	0,31	0,55	0,14
Part 2	1	0,67	0,06	0,28	0,31	0,69	0,00
Part 3	1	0,67	0,00	0,33	0,28	0,59	0,14
Part 4	1	0,72	0,06	0,22	0,28	0,45	0,28
Part 5	1	0,83	0,00	0,17	0,24	0,52	0,24
Part 6	1	0,61	0,06	0,33	0,17	0,59	0,24
Part 7	2	0,79	0,00				
Part 8	2	0,93	0,00				
Part 9	2	0,86	0,00				
Part 10	2	0,86	0,07				
Part 11	2	0,79	0,07				
Part 12	2	0,93	0,00				

NB - Ici une variable catégorielle est codé par des nombres. Ce n'est pas un problème car :

- On le signale au logiciel de traitement
- Il considère ces chiffres comme des étiquettes (comme un numéro INSEE, de sécurité social, etc.)

**Les « objets » qu'on observe sont en ligne et les variables en colonne.  
Les noms des « objets » apparaissent tous dans la première colonne et  
les noms des variables tous dans la première ligne**

Est-ce que cette autre présentation convient ? Quelle différence avec la première ? Pour quels objectifs ?

Métaphore	Sens trad.	Spécificité	Nombre de participants qui ont répondu M-0	Nombre de participants qui ont répondu M1-M2	Nombre de participants qui ont répondu M-M
1	1	2	0,50	0,33	0,17
2	1	1	0,33	0,67	0,00
3	1	2	0,00	0,00	1,00
4	1	2	0,00	0,67	0,33
5	1	1	0,83	0,17	0,00
6	1	2	0,67	0,17	0,17
7	1	1	0,67	0,00	0,33
8	1	1	0,67	0,33	0,00
9	1	2	0,17	0,67	0,17
10	1	1	1,00	0,00	0,00
11	1	2	0,33	0,50	0,17
12	1	2	0,50	0,00	0,50

# Les noms des variables sont écrits sur une seule ligne et ne comportent ni espaces, ni accents

Exemple de l'étude du réseau social construit par des apprenants sinophones et anglophone

	A	B	C	D	E	F	G
1	participants	sexes	origine	amis_masculins		amis_feminins	
2				meme orig.	autre_orig.	meme orig.	autre_orig.
3	Participant 1	M	ch	6	1	6	8
4	Participant 2	F	ch	1	2	8	1
5	Participant 3	M	ch	5	5	2	0
6	Participant 4	M	ch	5	1	1	0
7	Participant 5	F	ang	2	0	2	0
8	Participant 6	M	ang	0	1	3	0
9	Participant 7	F	ang	4	12	2	1
10	etc.						
11							

**Non** car les noms des variables sont sur deux lignes et le logiciel ne comprendra pas qu'il s'agit du nom des variables...

On considère que le nom des variables est sur la première ligne.

**Oui** car les noms des variables sont « mis à plat » sur une seule ligne

	A	B	C	D	E	F	G
1	participants	sexes	origine	masculins_meme	masculins_autre	feminins_meme	feminins_autre
2	Participant 1	M	ch	6	1	6	8
3	Participant 2	F	ch	1	2	8	1
4	Participant 3	M	ch	5	5	2	0
5	Participant 4	M	ch	5	1	1	0
6	Participant 5	F	ang	2	0	2	0
7	Participant 6	M	ang	0	1	3	0
8	Participant 7	F	ang	4	12	2	1
9	etc.						

## 3.2 Notion de distribution d'une variable nominale/qualitative/catégorielle

# Notion de distribution d'une variable

Lorsque l'on parle de « la distribution d'une variable dans la population », on se réfère à **la répartition des différentes valeurs possibles de cette variable entre les individus de la population**. Notons cependant que le concept de « distribution d'une variable » peut s'appliquer à n'importe quel ensemble de données, que ce soit une population ou un échantillon. On va voir dans ce chapitre à quoi ressemble une distribution pour les différents types de variables (...). Ceci se fera à l'aide de graphiques et de caractéristiques numériques. Une caractéristique numérique fameuse qui s'applique aux variables quantitatives est la moyenne arithmétique. Il s'agit d'un résumé des mesures individuelles. Pour les variables qualitatives, on remplacera la moyenne par des pourcentages, autre concept fameux de statistique descriptive.

Rousson, 2013

# Distribution d'une variable nominale (qualitative)

Voici un ensemble de données récoltées sur  $n = 158$  individus d'une population (fictive) du nord de l'Europe dont on a mesuré le groupe sanguin :

O	O	O	A	A	A	A	O	O	O	O	O	A	A	O
O	O	O	O	O	A	O	O	O	B	O	O	O	O	A
A	A	A	O	O	B	O	O	O	O	O	A	O	A	O
A	B	O	A	B	O	A	A	A	AB	O	A	O	O	O
A	AB	O	AB	O	O	B	A	A	A	O	A	O	B	O
O	O	O	A	O	A	A	A	O	O	A	O	A	A	O
O	AB	A	A	O	A	O	A	A	A	O	O	A	O	O
O	A	O	A	A	O	O	A	A	O	A	O	A	B	O
O	A	O	B	O	O	O	A	O	O	A	O	A	A	O
O	A	O	A	A	B	A	O	O	B	A	A	A	A	A

# Distribution d'une variable nominale

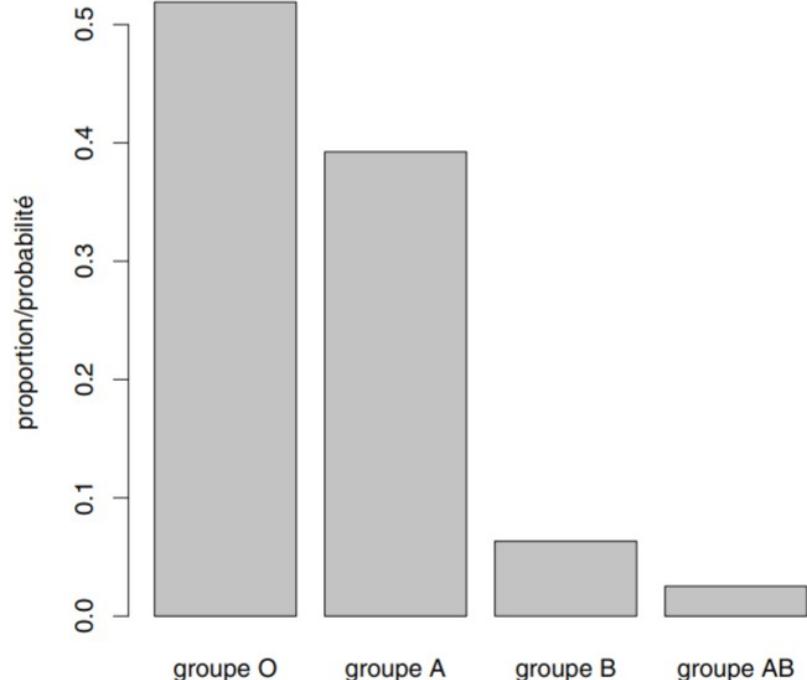


Figure 2.1 – Exemple de diagramme en barres.

## Représentation de la distribution de la variable nominale sous forme de diagramme

- Pour chacune des modalités/valeurs de la variable, on représente le nombre de réponses qui lui correspondent
- Ce nombre est ici représenté par des proportions (entre 0 et 1). On peut aussi donner directement le nombre de valeurs (nombre de O, de A, etc.) ou exprimer les proportions par des pourcentages.
- D'autres formes de représentations sont possibles, par exemple un diagramme en forme de « camembert »

# Distribution d'une variable nominale

groupe O	groupe A	groupe B	groupe AB	total
82	62	10	4	158
52 %	39 %	6 %	3 %	100 %

## Représentation de la distribution de la variable nominale sous forme de tableau

- Tableau représente à la fois les **effectifs** et les **pourcentages**
- Le total des pourcentages doit faire 100% et le total des effectifs doit faire l'effectif total de l'échantillon (moins les éventuelles non-réponses).

Rousson, 2013

Quels sont les avantages et inconvénients respectifs de la représentation en diagramme et de la représentation en tableau ?

# Distribution d'une variable qualitative/nominale

**Nous allons décrire les données du corpus du projet PhraseoRom, et examiner les distributions de plusieurs variables nominales**

Les résultats des analyses pourront être exportés en PDF ou HTML (HTML est conseillé pour plus de souplesse dans les copiers/coller ultérieur).

1. Afficher les métadonnées du corpus PhraseoRom en français (langue source) et les exporter au format .CSV
2. Importer les données dans Jamovi
3. Quelles sont les variables nominales qu'il peut être intéressant d'étudier ?
4. Vérifier que les variables sont correctement déclarées

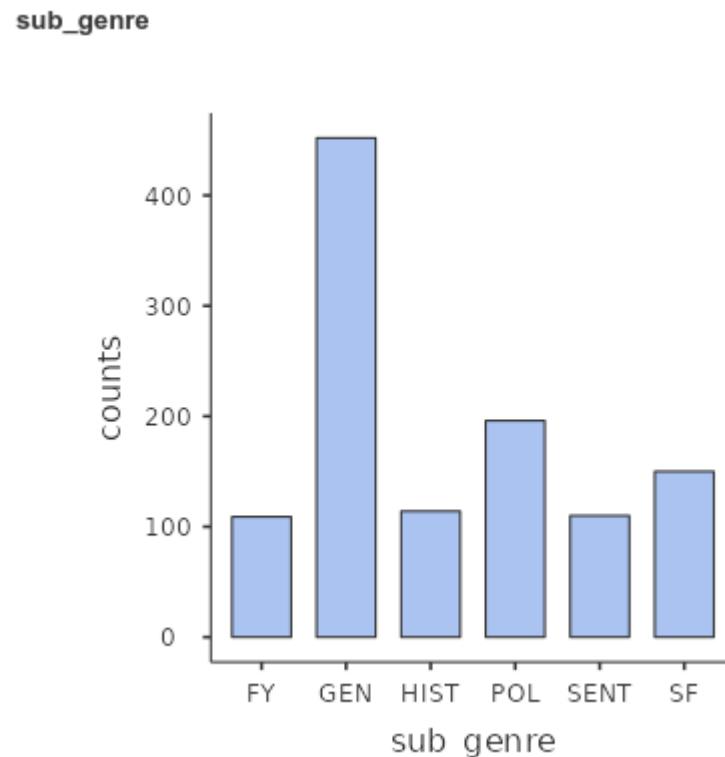
# Distribution d'une variable qualitative/nominale

**On peut par exemple étudier la distribution en sous-genres**

Cliquer sur l'onglet Analyse, puis choisir la variable « sous-genre ». Afficher le tableau des fréquences, puis le diagramme en barre correspondant.

Fréquences de sub\_genre

sub_genre	Quantités	% du Total	% cumulés
FY	109	9.6 %	9.6 %
GEN	452	40.0 %	49.6 %
HIST	114	10.1 %	59.7 %
POL	196	17.3 %	77.0 %
SENT	110	9.7 %	86.7 %
SF	150	13.3 %	100.0 %



**Quel constate-t-on ?**

### 3.3 Distribution d'une variable continue/quantitative (échelle de rapport)

# Distribution d'une variable quantitative/continue

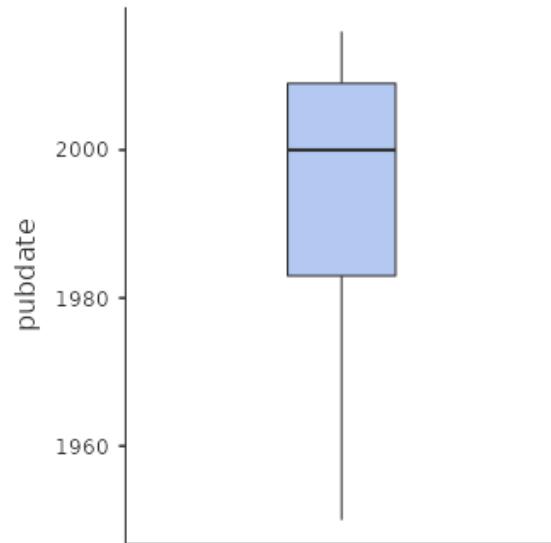
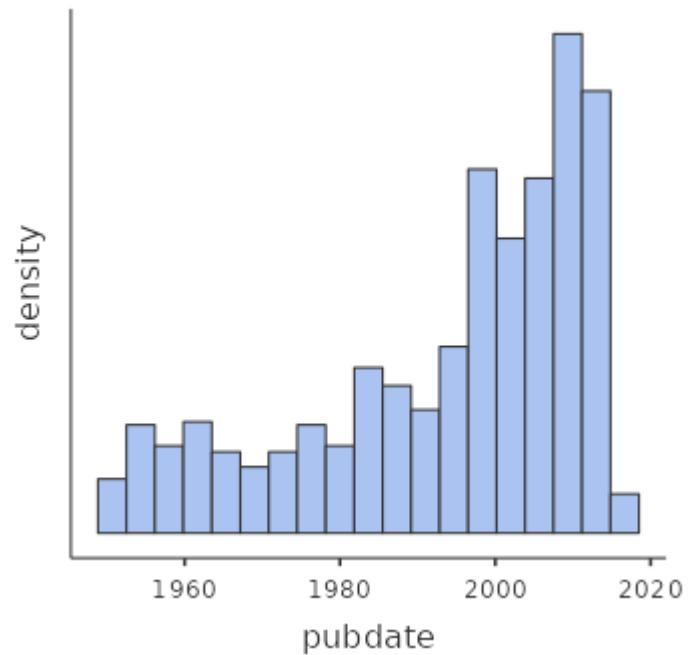
- Identifier les variables quantitatives dans les métadonnées ainsi extraites.
- Quelles statistiques et quelles représentations sont intéressantes pour ces données ?

# Distribution d'une variable quantitative/continue

- Exemple pour les dates :

Statistiques descriptives

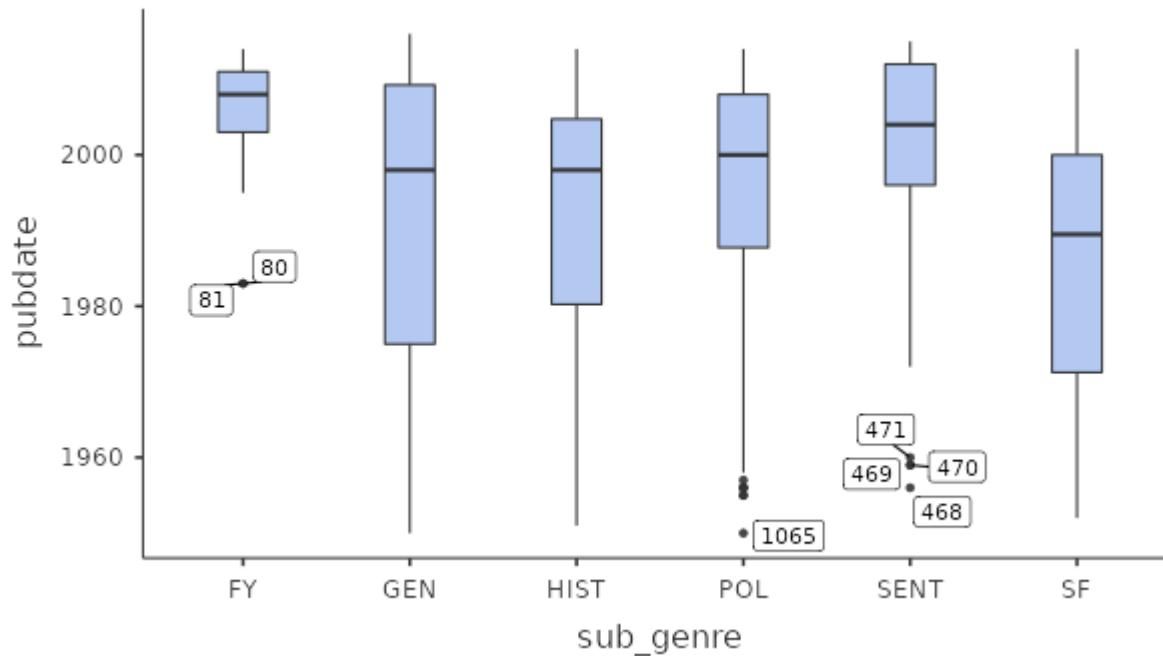
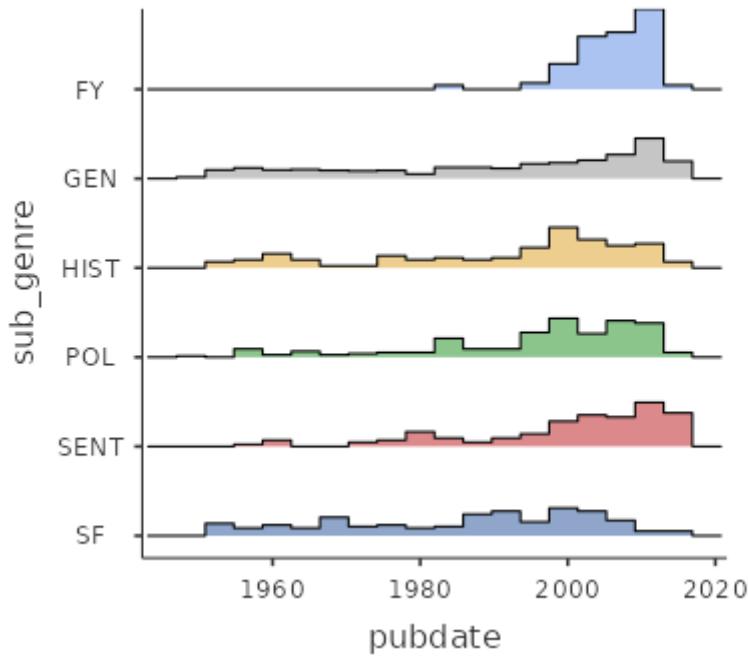
pubdate	
N	1131
Moyenne	1994
Moyenne de l'erreur-standard	0.532
Médiane	2000
Ecart-type	17.9
Minimum	1950
Maximum	2016



Croisez ces données avec les sous-genres. Affichez les diagrammes en barre. Qu'indiquent-ils ?

# Distribution d'une variable quantitative/continue

- Distributions des dates par genre :



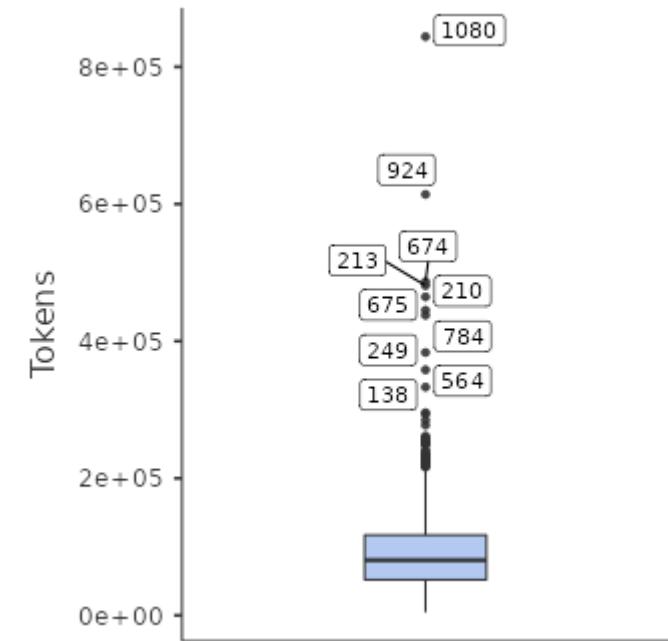
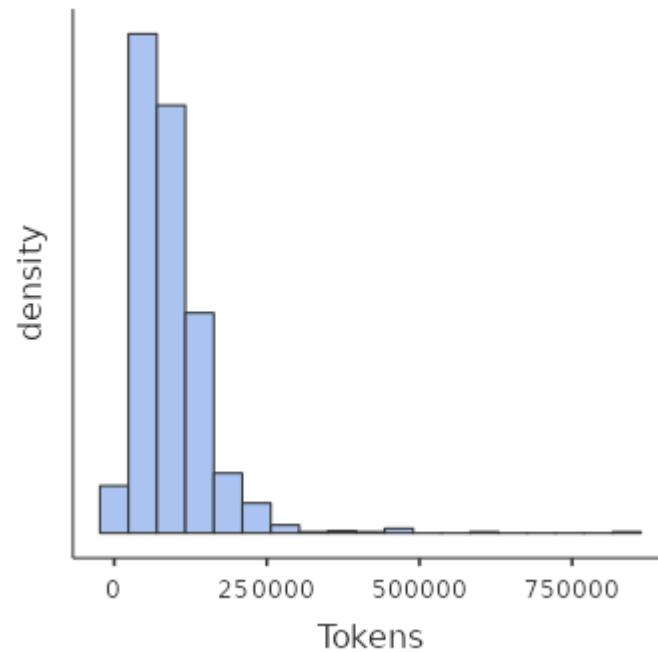
Que remarque-t-on ? a-t-on un corpus adapté à une étude diachronique ?

# Distribution d'une variable quantitative/continue

- Exemple pour les tailles en tokens :

Statistiques descriptives

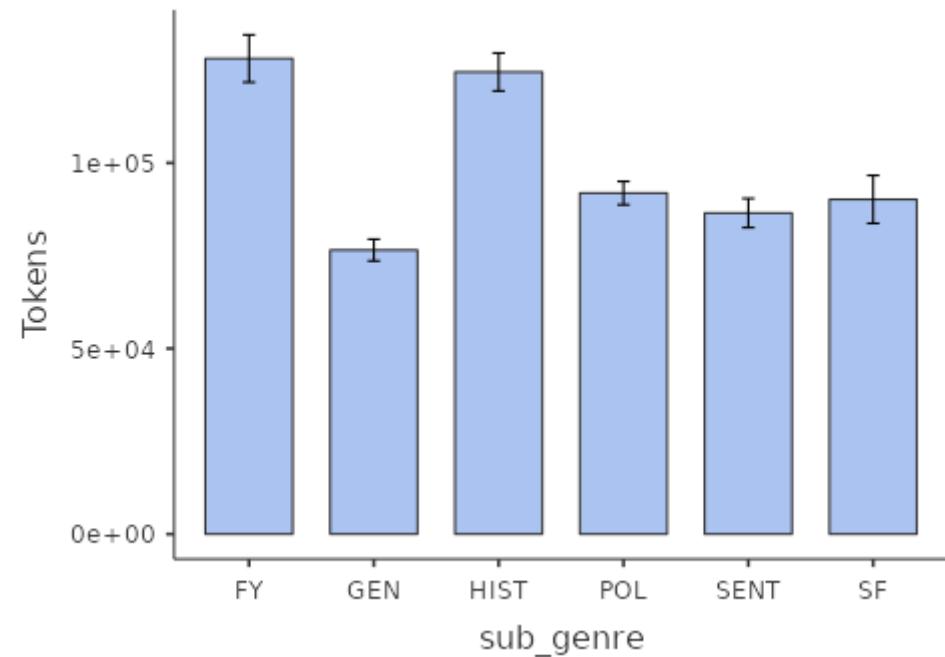
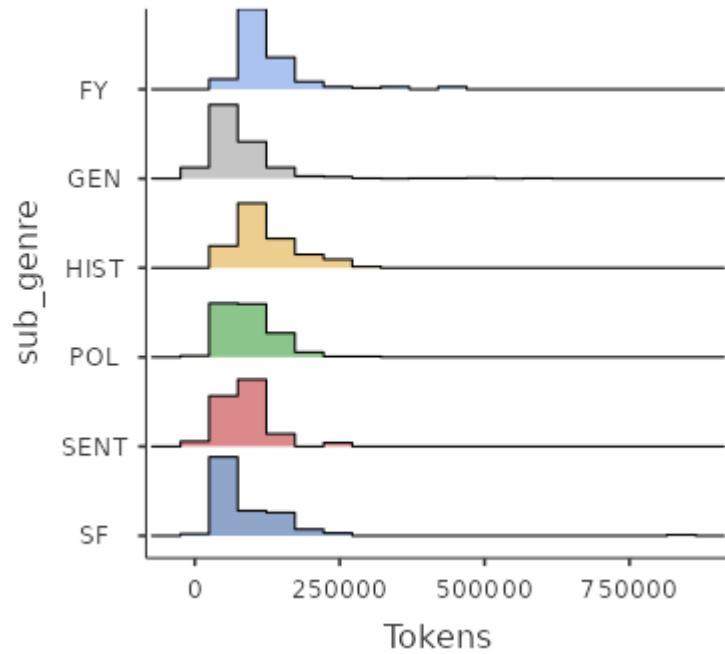
	Tokens
N	1131
Moyenne	91785
Moyenne de l'erreur-standard	1861
Médiane	80585
Ecart-type	62589
Minimum	4542
Maximum	844041
25-ième percentile	51965
50-ième percentile	80585
75-ième percentile	117146



Croisez ces données avec les sous-genres. Affichez les diagrammes en barre. Qu'indiquent-ils ?

# Distribution d'une variable quantitative/continue

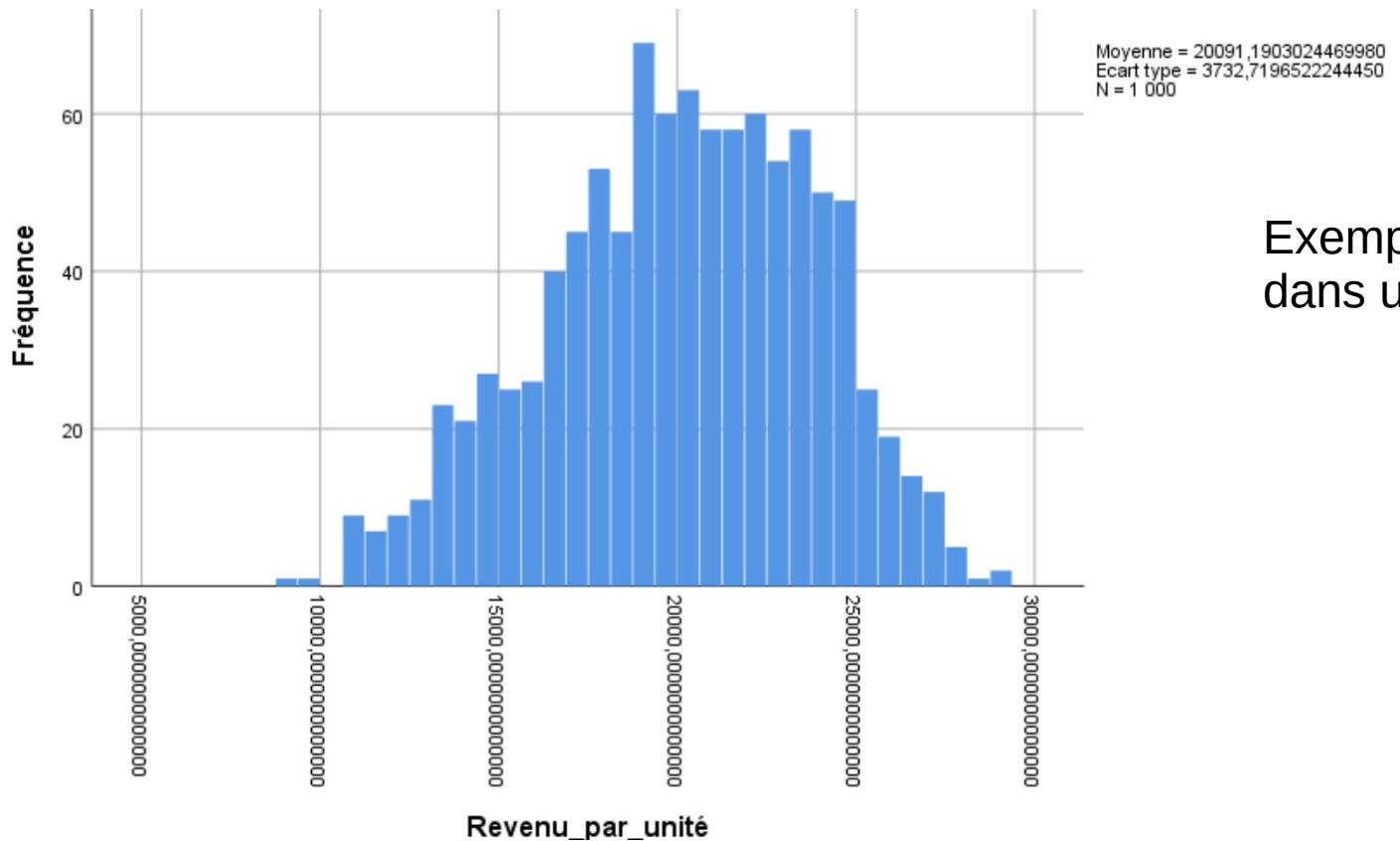
- On peut séparer les analyses en fonction du sous-genre



Cette fois le diagramme en barre n'indique pas des effectifs, mais des moyennes avec des barres d'erreur (+- erreur standard)

# Distribution d'une variable quantitative/continue

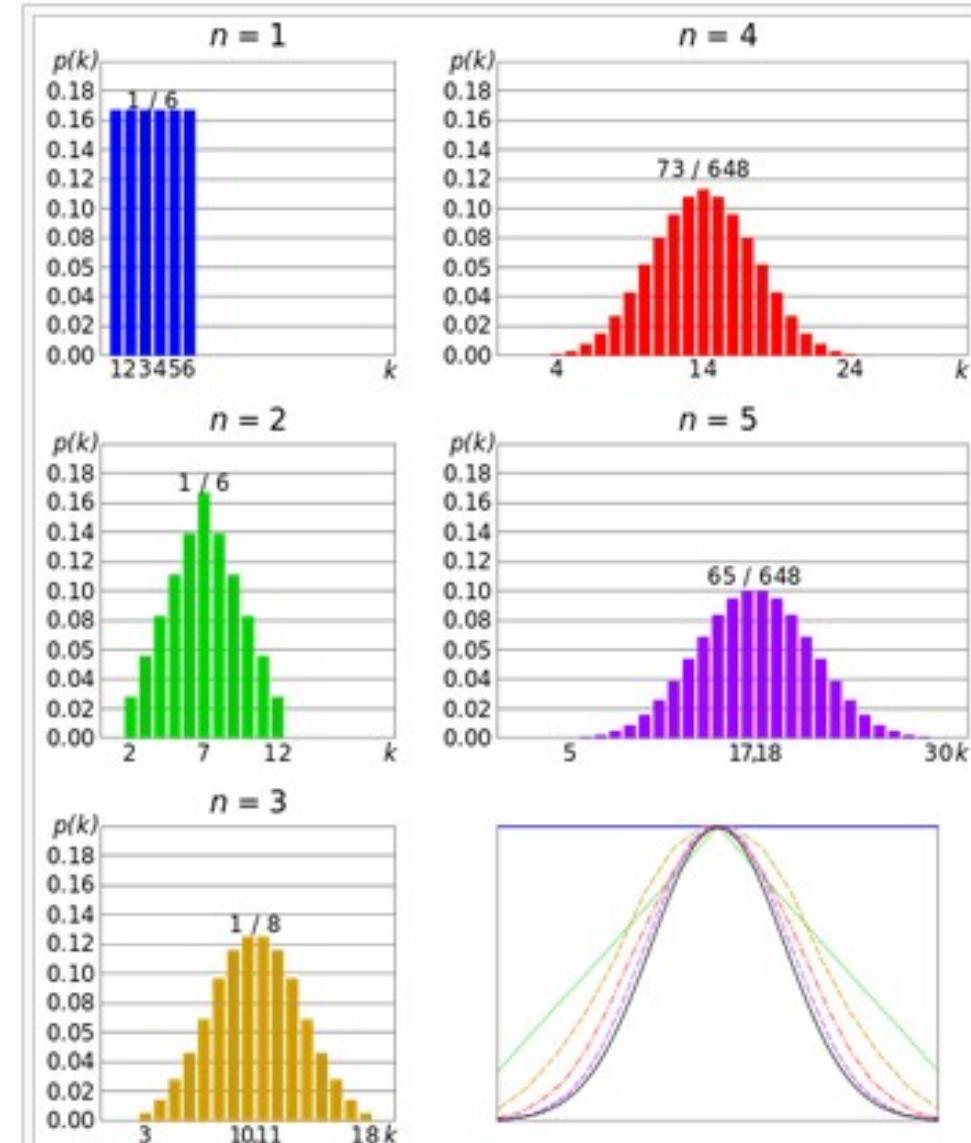
**Il est fréquent qu'une distribution aboutisse à une courbe en cloche symétrique autour de la moyenne.**



Exemple : distribution des revenus  
dans un échantillon de 1000 personnes

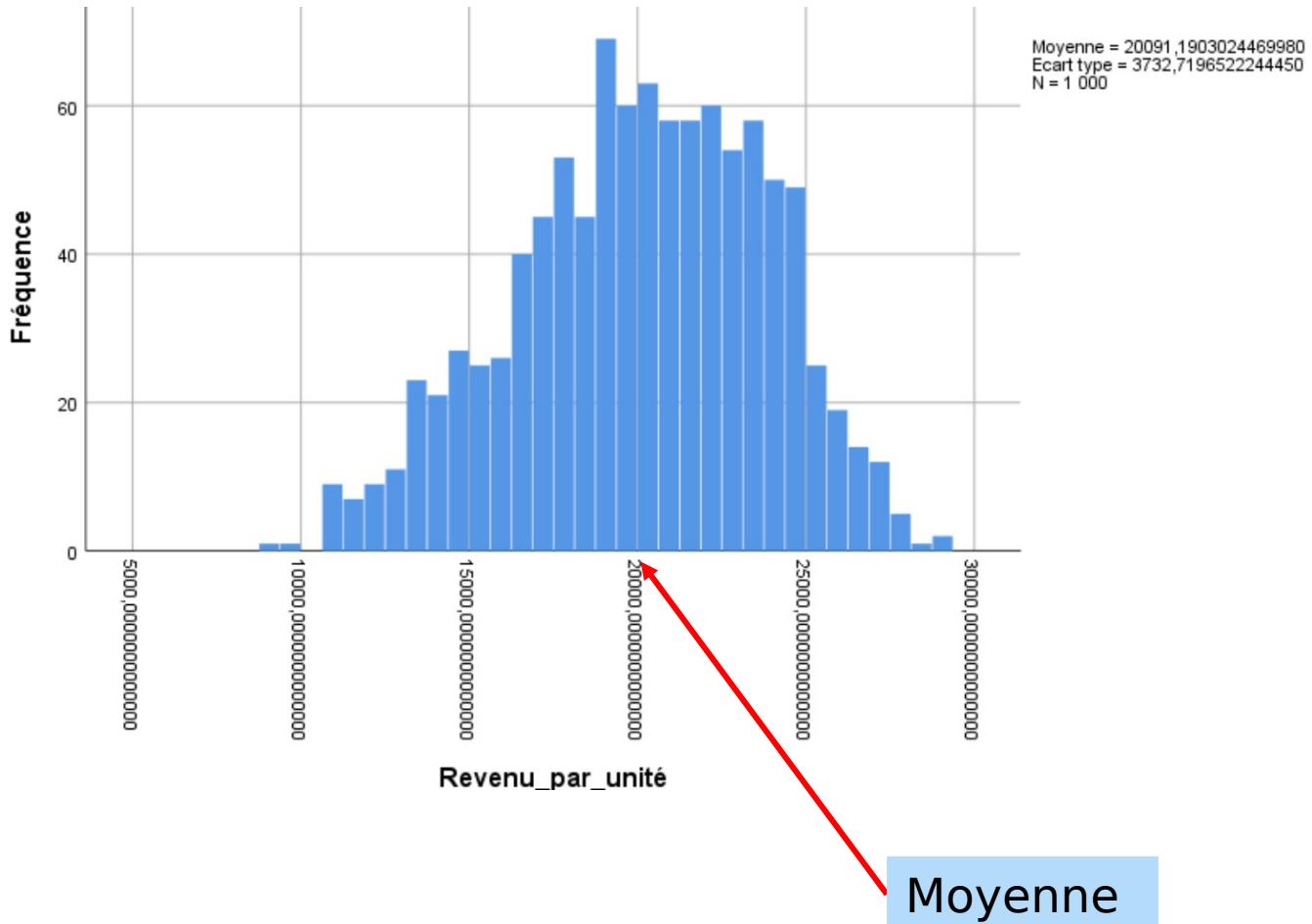
# Distribution d'une variable quantitative/continue

Ce type de distribution est dite « **normale** », et correspond à une fonction mathématique appelée « **gaussienne** ». Les distributions normales correspondent à l'addition de nombreux petits événements indépendants. P.ex. une série de lancer de dés ou de tirages à pile ou face tendent vers une loi de probabilité normale lorsque  $n$  est grand.



# Paramètres d'une distribution normale

Moyenne = 20091,1903024469980  
Ecart type = 3732,7196522244450  
N = 1 000



## 3 paramètres principaux

### N : la taille de l'échantillon

### $\mu$ : La moyenne

Une estimation de la **tendance centrale**. Les mesures de tendance centrale permettent de déterminer une valeur «typique» ou centrale autour de laquelle les données ont tendance à se rassembler.

### $\sigma$ : l'écart-type

Une estimation de la **dispersion**. Les indicateurs de dispersions permettent de mesurer comment les données se «répartissent» (par exemple, sont-elles «serrées» ou «étalées» autour de la moyenne).

# Approche intuitive de l'écart type (*standard deviation*)

Intuitivement, on peut concevoir l'écart type comme l'écart moyen des scores à la moyenne. Plus il est grand, plus les scores sont, en moyenne, « éloignés » de la moyenne et vice versa.

Moyenne = 20091,1903024469980  
Ecart type = 3732,7196522244450  
N = 1 000

Moyenne = ,228058000000000  
Ecart type = ,073414251261511  
N = 1 000

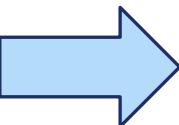
Moyenne = 5793,633151000001000  
Ecart type = 9694,368198279586000  
N = 1 000

Moyenne = 59,1  
Ecart type = 124,082  
N = 1 000

Moyenne = 27,62  
Ecart type = 9,919  
N = 799

On ne peut comparer des écarts types pour des grandeurs différentes. Le problème est que l'écart type dépend de la mesure et des objets mesurés

- il est plus grand si on mesure des éléphant que des souris en cm
- il sera plus grand si on mesure des éléphants en cm plutôt qu'en mètres



Pour comparer la dispersion de deux distributions, on calcule le **coefficient de variation** (CV): l'écart type divisé par la moyenne

- on exprime l'écart type comme une proportion de la moyenne
- on « standardise » ainsi la mesure de dispersion

# Paramètres décrivant la dispersion d'une distribution

Nom du paramètre	Définition	Exemple de calcul pour la distribution suivante <b>5,6 - 5,5 - 5,3 et 5,2</b> <b>Moyenne = 5,4</b>
Etendue ou plage	Ecart entre la valeur maximum et la valeur minimum	$5,6 - 5,2 = 0,4$
Variance	Moyenne du carré des écarts à la moyenne	$0,1/4 = 0,25$ si variance de la population $0,1/3 = 0,33$ si variance d'un échantillon <b>(variance corrigée=estimateur sans biais)</b>
Ecart type	Racine carré de la variance	$\sigma=0,5$ si population $s=0,56$ si échantillon (N.B. la notation est différente)
Ecart moyen	Moyenne de la valeur absolue des écarts à la moyenne	$ 5,6-5,4 + 5,5-5,4 + 5,3-5,4 +  5,2-5,4  = 0,2 + 0,1 + 0,1 + 0,2 = 0,6$ . Et $0,6/4 = 0,15$

## Calcul de la variance

- Carré des écarts à la moyenne

Valeur	Différence à la moyenne	Carré des différences
5,6	$5,6 - 5,4 = 0,2$	$0,2 \times 0,2 = 0,04$
5,5	$5,5 - 5,4 = 0,1$	$0,1 \times 0,1 = 0,01$
5,3	$5,3 - 5,4 = -0,1$	$-0,1 \times -0,1 = 0,01$
5,2	$5,2 - 5,4 = -0,2$	$-0,2 \times -0,2 = 0,04$

- Somme des carrés des écart à la moyenne

$$0,04 + 0,01 + 0,01 + 0,04 = 0,1$$

- On divise cette somme par le nombre de valeurs, et on obtient la variance

$$0,1 / 4 = 0,25$$

- si on dispose de toute la population :

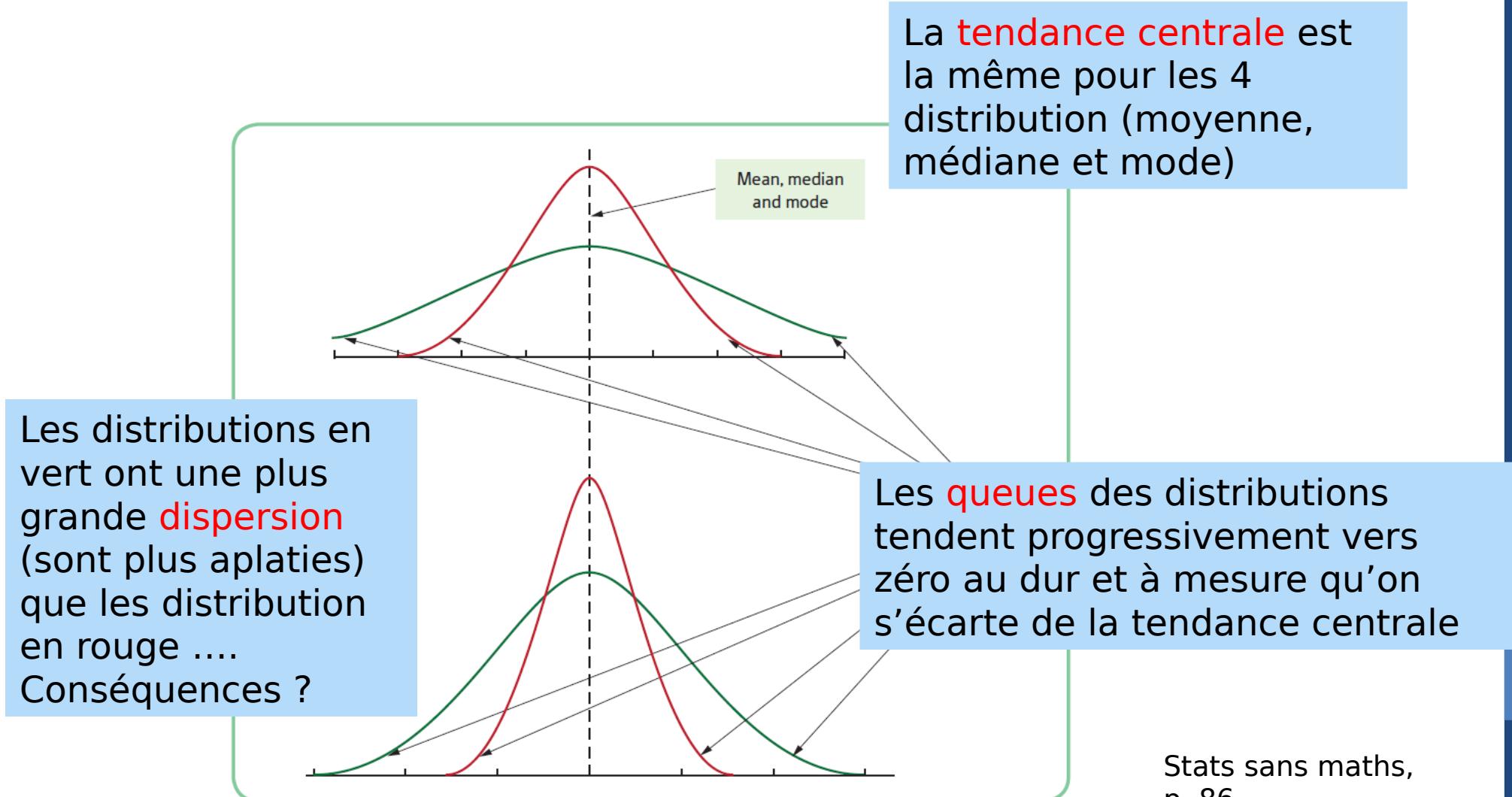
variance= somme des carrés divisée par nombre de valeurs

- si on ne dispose que d'un échantillon, on calcule son estimateur sans biais :

variance = somme des carrés divisée par nombre de valeurs -1

# Comparaison de distributions normales avec des écarts types différents

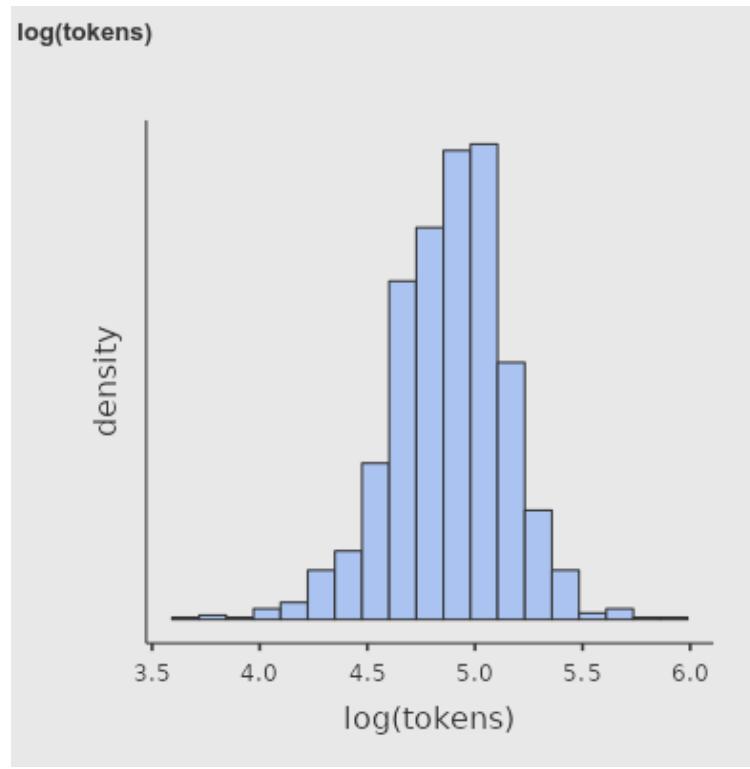
Quatre distributions normales avec des tendances centrales identiques mais des dispersions différentes



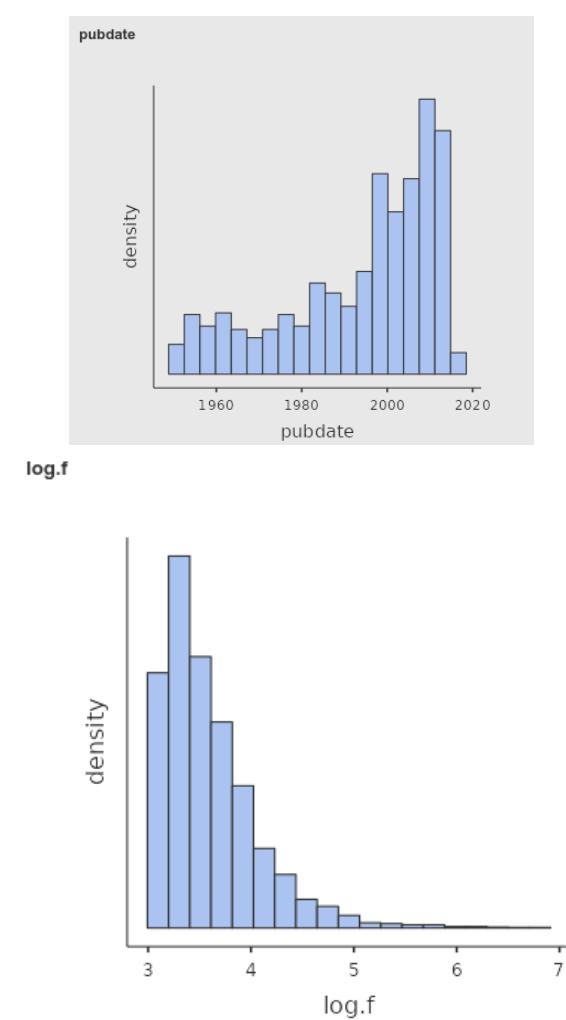
- Représenter la distribution du logarithme du nombre de tokens de chaque roman. Que constate-t-on ?
- Représenter la distribution des fréquences lexicales pour le corpus Phraseorom\_fr\_fr
- Représenter la distribution du logarithme des fréquences lexicales
- classer les différentes courbes obtenues

# 1. Peut-on classer les distributions en deux ou trois catégories distinctes selon leur forme générale ?

Catégorie 1

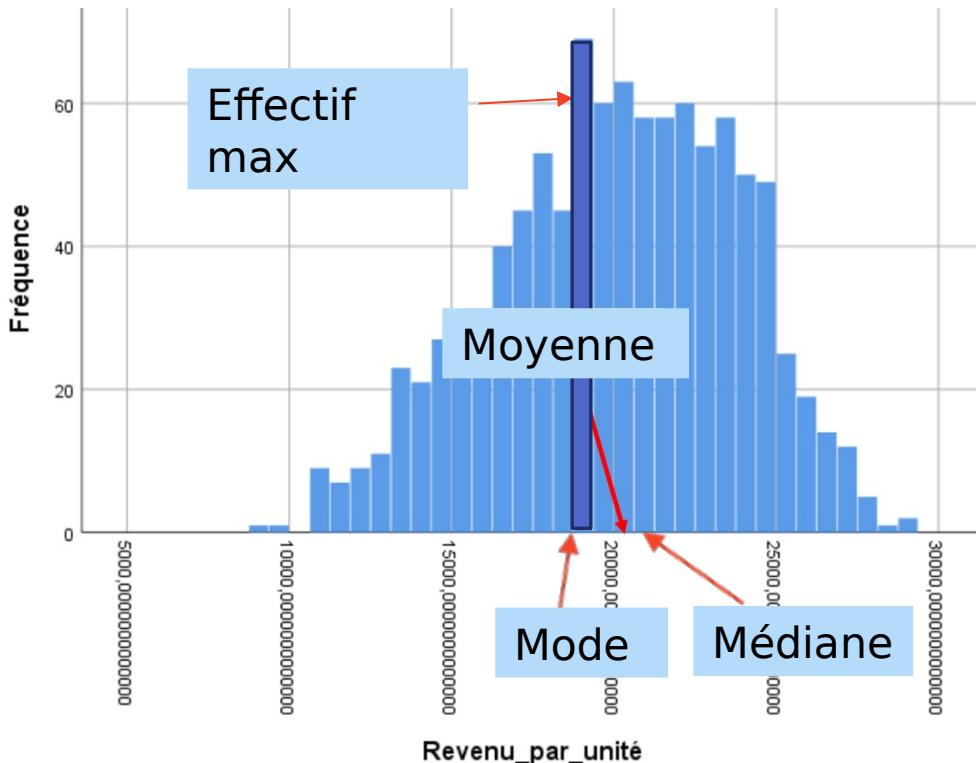


Catégorie 2



# Comment décrire ces catégories ?

## Paramètres décrivant la tendance centrale d'une distribution



### Remarque

Si une distribution suit parfaitement une loi normale, alors les valeurs de la moyenne, du mode et de la médiane sont identiques

Qu'est-ce qui dans la forme de la distribution du revenu suggère qu'elle ne suit-elle pas parfaitement une loi normale ?

**Moyenne**  
Ici **20091 €**

### Mode

Valeur dont l'effectif est le plus élevé  
Dans l'intervalle [15000+6x625, 15000+7x625]  
Soit **entre 18750 et 19375**

### Médiane

Valeur qui coupe l'échantillon en deux partie égales (il y a autant de valeurs au-dessous de la médiane qu'au-dessus)

### Estimation de la médiane

1. On classe les 1000 valeurs de revenu par ordre croissant.
2. Comme 1000 est un nombre pair on n'a pas de valeur qui soit au milieu
3. Mais la valeur au rang 500 vaut 20255 et la valeur au rang 501 vaut 20268
4. On choisit la médiane au milieu de l'intervalle entre ces deux valeurs :  $(20268 + 20255)/2 = 20261,5$

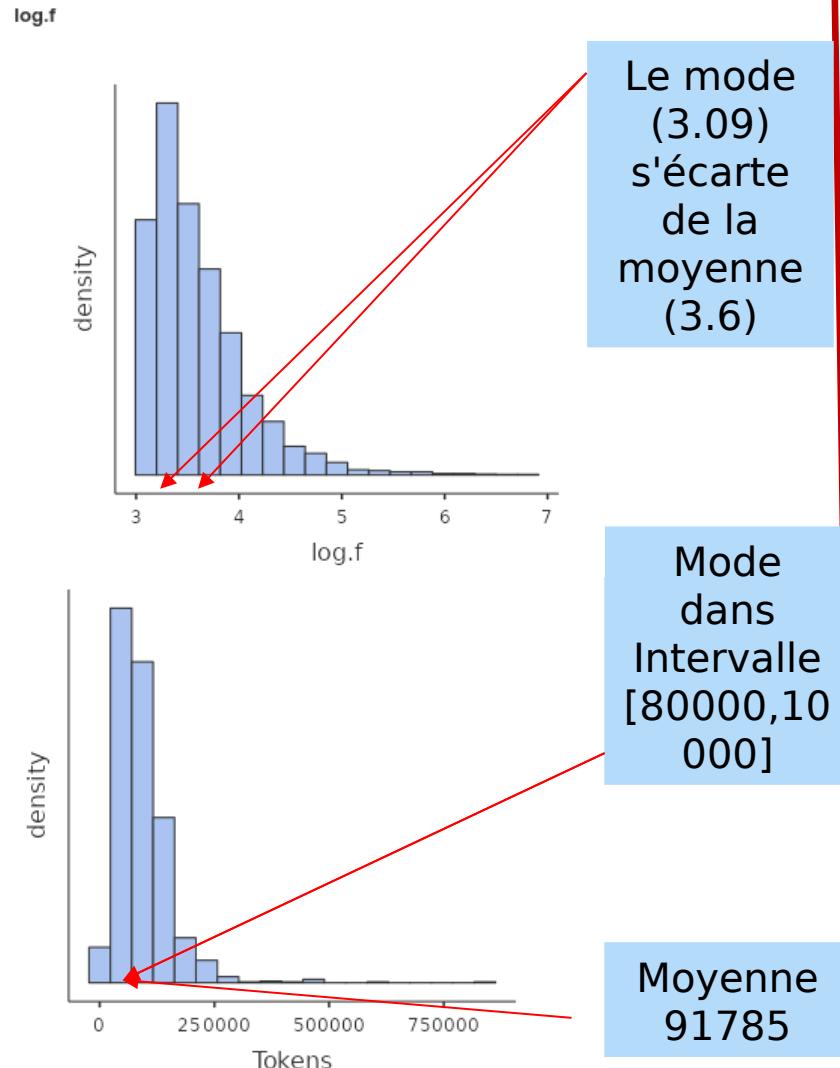
Donc sur le graphe, dans la barre qui va de 20165 à 20330 [20000+1x165, 20000+2x165]

Rang du revenu	Revenu_par_unité	Revenu_médiane
499	20254	Rev_inf
500	20255	Rev_inf
501	20268	Rev_sup
502	20271	Rev_sup

5. Si on avait un nombre impair de valeurs – par exemple 1001- alors la médiane serait la valeur du rang du milieu – par exemple dans ce cas 501

# Comment décrire ces catégories ?

## Catégorie 2



**La distribution des fréquences lexicales est asymétrique et ne suit pas une loi normale**

- Elle appartient à la catégorie des distributions asymétriques.
- Avec un mode à gauche.
- Et une queue sur la droite, très allongée si on considère les fréquences en valeur absolue (mots outils). Il faut passer au logarithme pour ramasser l'axe des x et avoir une vision complète de la répartition.

**En termes plus parlant**

- Un grand nombre d'unités lexicales ont une faible fréquence alors qu'un petit nombre ont une fréquence élevée
- Concernant la taille en tokens un grand nombre de romans sont répartis autour de la moyenne, mais on a une queue à droite assez étendue (quelques romans très longs, mais très peu de romans très courts).

**Attention**

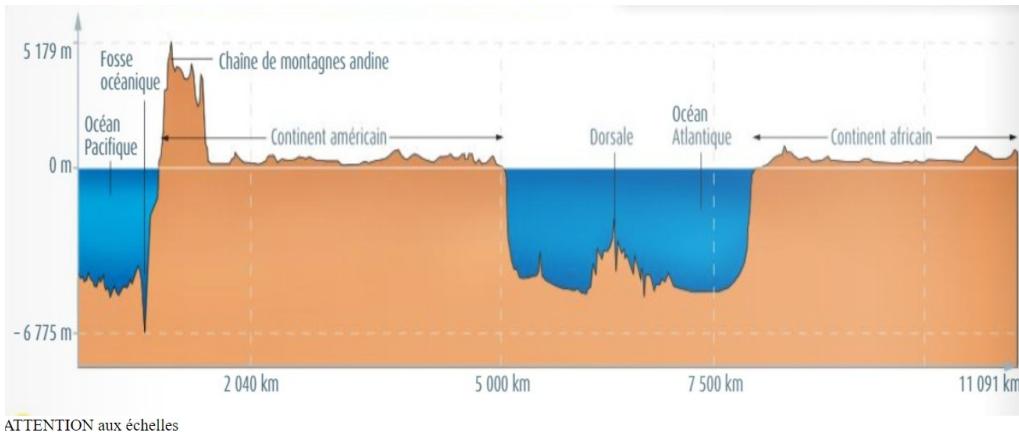
Si une distribution est très asymétrique (comme ici  $\log.f$ ), la moyenne n'est pas une bonne mesure de la tendance centrale.

# Autres types de distribution

## Un autre type de distribution qui ne suit pas une loi normale : la distribution bimodale

- La distribution présente deux modes
- ces deux modes sont égaux ou pas

## Un exemple connu en géophysique : la répartition des altitudes terrestres

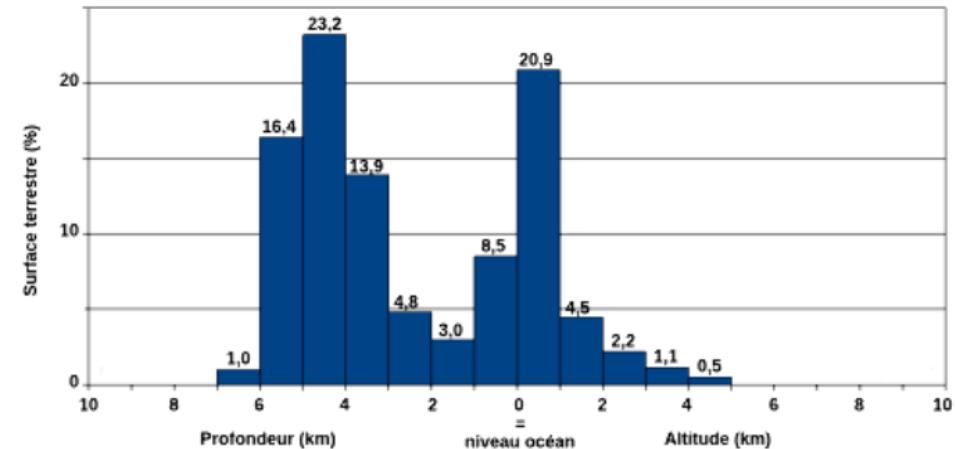


ATTENTION aux échelles

[http://nfabien-svt.fr/courslycee2019/1ere\\_spe2019/01theme1b/chap01.htm](http://nfabien-svt.fr/courslycee2019/1ere_spe2019/01theme1b/chap01.htm)

**Quels phénomènes humains, de nature sociale ou cognitive, sont-ils susceptibles d'avoir une distribution bimodale ? cf. Stats sans maths, p. 97**

Figure 2. La distribution des altitudes sur Terre



Ce diagramme (représente) la distribution des altitudes en montrant le pourcentage de la surface terrestre solide occupé par des terrains d'altitudes données. Les altitudes sont regroupées par tranches de 1 kilomètre. (Les) tranches extrêmes intègrent les valeurs maximales. La tranche 4-5 km d'altitude comprend (...) les terrains allant de 4 km à 8850 m (maximum pour l'Everest), et la tranche 6\_-7 km de profondeur les planchers océaniques allant de 6 à 11 km (maximum dans la fosse des Mariannes). La répartition des altitudes montre une distribution bimodale, les deux modes (pics de fréquence) étant 4-5 km de profondeur (océans) et 0-1 km d'altitude (continents).

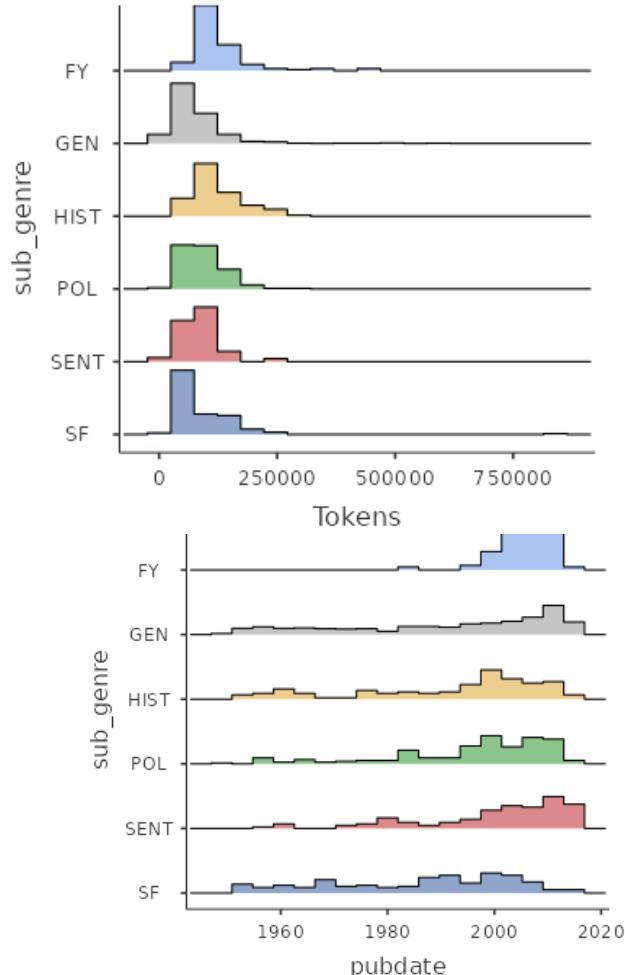
<https://planet-terre.ens-lyon.fr/article/distribution-bimodale-altitudes-tectonique-plaques.xml>

## Comparer les dispersions de différentes distributions

On peut comparer directement les écarts types pour chaque sous-genre, tant au niveau de la distribution des tailles que des années

Statistiques descriptives

	sub_genre	Tokens	pubdate
Ecart-type	FY	66588	5.59
	GEN	62189	19.8
	HIST	54513	17.8
	POL	43974	15.0
	SENT	41142	14.2
	SF	79072	17.1



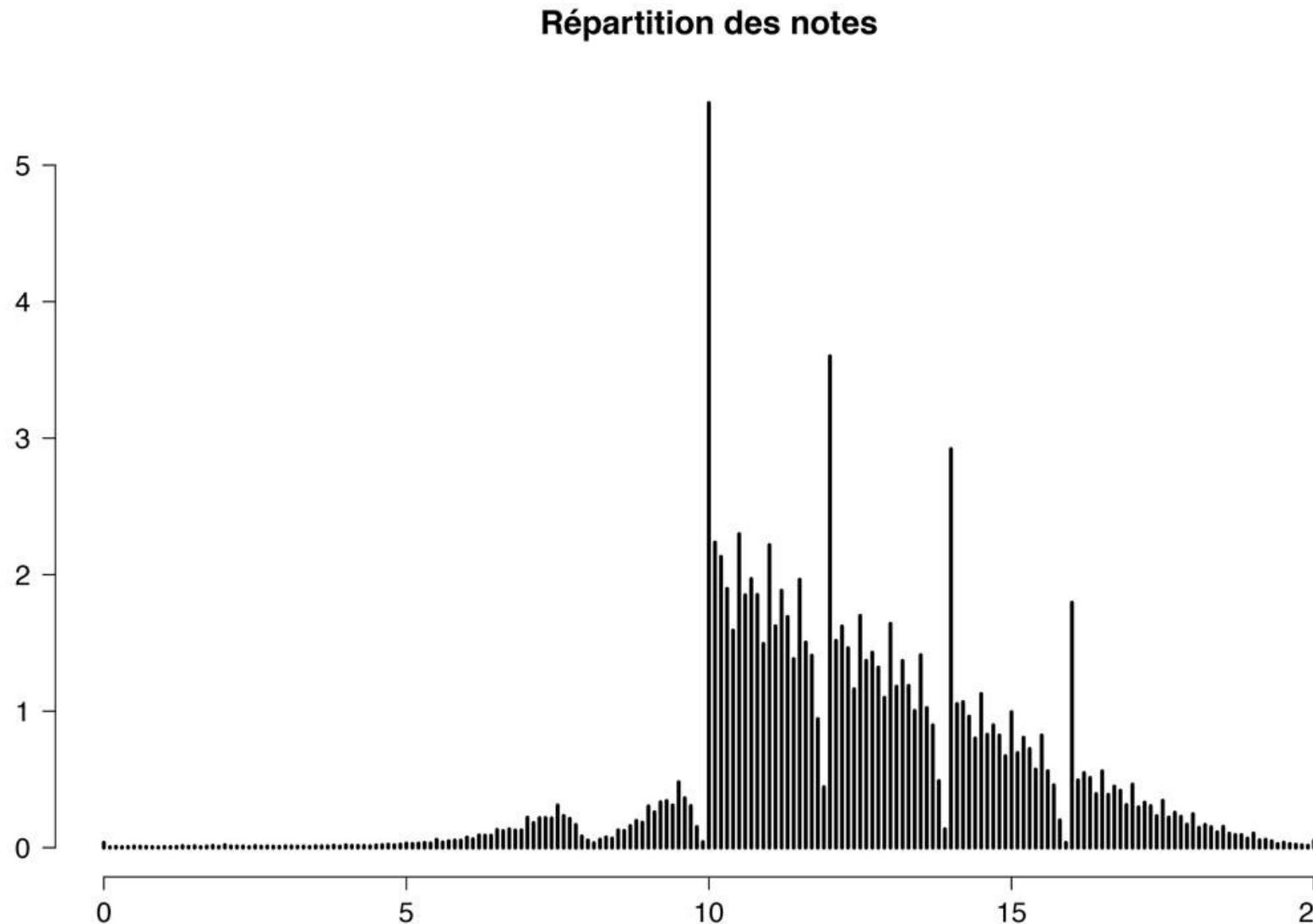
Peut-on aussi comparer les dispersions des dates et des tokens pour un même sous-genre ?  
**Effectuez le calcul des coeff. de variations et comparez. Que constatez-vous ?**

## Comparer les dispersions de différentes distributions

Le coefficient de variation CV n'a pas de sens pour la variable *Années*, car ce n'est pas une échelle de rapport (l'origine est arbitraire).

	Coefficient de variation	
	Tokens	Années
FY	0,519674715532177	0,002786640079761
GEN	0,812566963702407	0,009939759036145
HIST	0,437834321237531	0,008935742971888
POL	0,478540025247029	0,00751503006012
SENT	0,475547592902965	0,007096451774113
SF	0,876736630852986	0,008559919436052

## Interprétez la distribution des moyennes finales au baccalauréat en France



### Indications

- On réussit le bac avec une moyenne de 10
- Les mentions sont à 12 (assez bien), 14 (bien), 16 (très bien)
- Les jurys de bac ont le droit d'attribuer des « points jury » pour favoriser la réussite ou l'obtention d'une mention.

### Statistiques des taux de réussite au bac.

<https://www.letudiant.fr/bac/bac-2021-93-8-de-reussite-apres-les-rattrapages.html>

# Mise au point générale sur Décrire et Interpréter

**Règle générale sur description, analyse et interprétation : le travail d'interprétation doit être fait après le travail de description et d'analyse des résultats**

1. Décrire et analyser les tendances dans les résultats : ici des formes de distribution mais aussi dans d'autres cas, des tableaux à double entrée, des différence de moyenne, etc.
2. Interpréter les tendances par rapport aux questions de recherche, à des sources externes, à une théorie, à un modèle.

## Étapes du cheminement des données



# Distribution d'une variable quantitative/continue : erreur standard

1. Comment est calculée la mesure Erreur standard qui est affichée dans la colonne à droite de la moyenne ?
2. Recalculez l'erreur standard de la variable revenu\_par\_unite « à la main »

Statistiques descriptives								
N Statistiques	Plage Statistiques	Minimum Statistiques	Maximum Statistiques	Moyenne Statistiques	Erreur standard	Ecart type Statistiques	Variance Statistiques	
Revenu_par_unité	1000	20373,944	8779,2564	29153,20000	20091,19	118,03896	3732,71965	13933196,002
Densité_population	1000	72874,038	63,462000	72937,50000	5793,633	306,56284	9694,36820	93980774,764
Proportion_proprio	1000	,48100000	,050000000	,5310000000	,2280580	,00232156	,073414251	,005
Nombre_contact	1000	2231	1	2232	59,10	3,924	124,082	15396,390

$$ES(M) = \frac{\sigma}{\sqrt{n}} \cong \frac{s}{\sqrt{n-1}}$$

Quand on a des distributions normales, l'**erreur d'échantillonnage** réellement commise est **inférieure à l'erreur standard** dans environ **deux tiers des cas** et supérieure dans le tiers restant de cas.

# Distribution d'une variable quantitative/continue : erreur standard

1. Comment est calculée la mesure Erreur standard qui est affichée dans la colonne à droite de la moyenne ?
2. Recalculez l'erreur standard de la variable revenu\_par\_unite « à la main »

Statistiques descriptives								
N Statistiques	Plage Statistiques	Minimum Statistiques	Maximum Statistiques	Moyenne Statistiques	Erreur standard	Ecart type Statistiques	Variance Statistiques	
Revenu_par_unité	1000	20373,944	8779,2564	29153,20000	20091,19	118,03896	3732,71965	13933196,002
Densité_population	1000	72874,038	63,462000	72937,50000	5793,633	306,56284	9694,36820	93980774,764
Proportion_proprio	1000	,48100000	,050000000	,5310000000	,2280580	,00232156	,073414251	,005
Nombre_contact	1000	2231	1	2232	59,10	3,924	124,082	15396,390

$$ES(M) = \frac{\sigma}{\sqrt{n}} \cong \frac{s}{\sqrt{n-1}}$$

Quand on a des distributions normales, l'**erreur d'échantillonnage** réellement commise est **inférieure à l'erreur standard** dans environ **deux tiers des cas** et supérieure dans le tiers restant de cas.

## Calcul

$$N = 1000$$

$$\text{Racine carré de } 1000 = 31,62$$

$$\text{Ecart type} = 3732,72$$

$$\text{SEM} = 3732,72 / 31,62 = \mathbf{118,04}$$

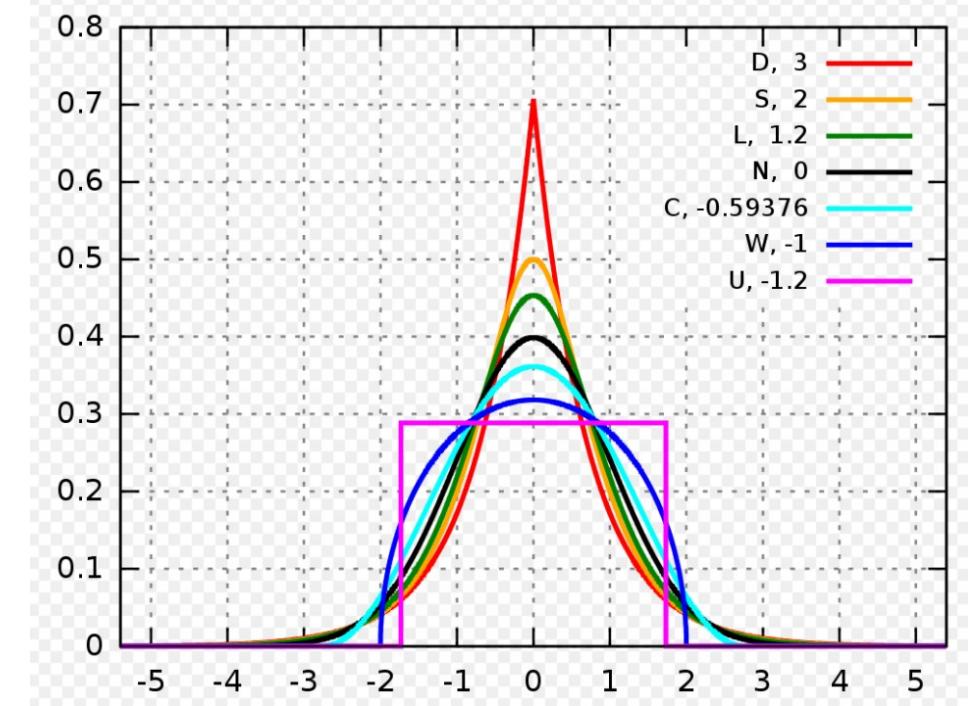
# Distribution d'une variable quantitative/continue : skewness et Kurtosis

**Kurtosis** : mesure de l'**aplatissement** d'une distribution

Mesure de l'étendue du regroupement des observations autour d'un point central. **Dans le cas d'une distribution normale, la valeur de la statistique kurtosis est égale à zéro.** Un kurtosis positif indique que par rapport à une distribution normale, les observations sont plus regroupées au centre et présentent des extrémités plus fines (...). Un kurtosis négatif indique que les observations sont moins regroupées au centre et présentent des extrémités plus épaisses atteignant les valeurs extrêmes de la distribution.

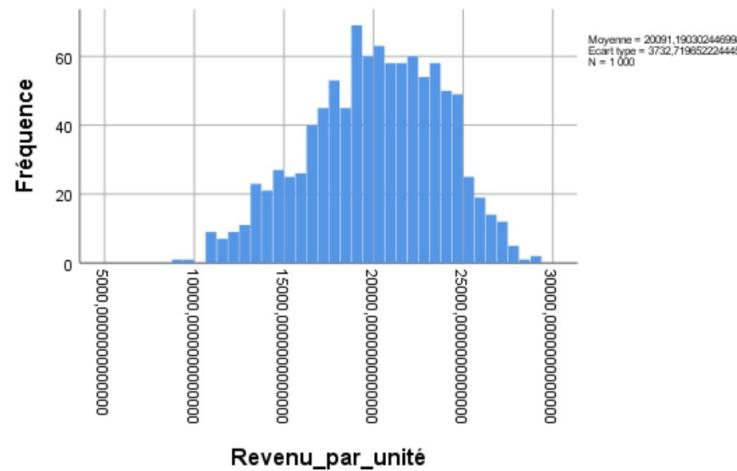
## Une façon intuitive de se représenter les choses

Un Kurtosis proche de zéro correspond à une distribution normale (courbe noire). Un Kurtosis positif est comme un « pincement de la distribution » (courbes verte, orage et rouge). Un kurtosis négatif est un « rassemblement au centre (en tapant dessus et sur les côtés !) » (courbes bleue ciel, bleu marine et violette). Les courbes avec kurtosis positif sont dites **leptocurtique**, les courbes aplatis sont dites **platycurtiques** et les courbes voisines de la normale sont dites **mésocurtiques**.

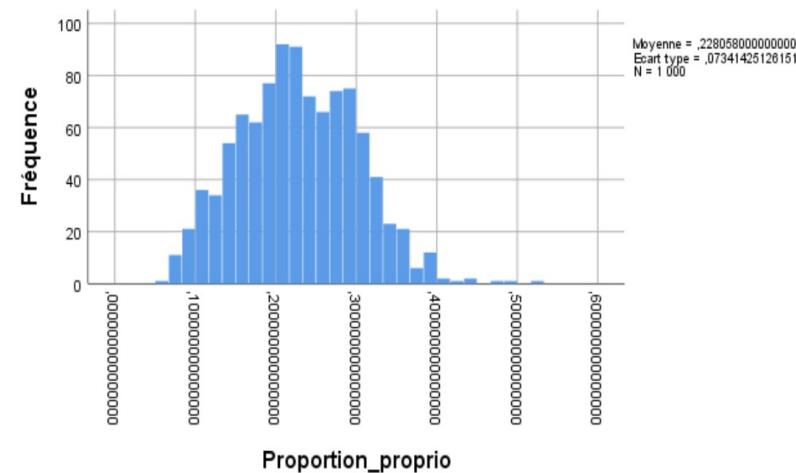


<https://fr.wikipedia.org/wiki/Kurtosis>

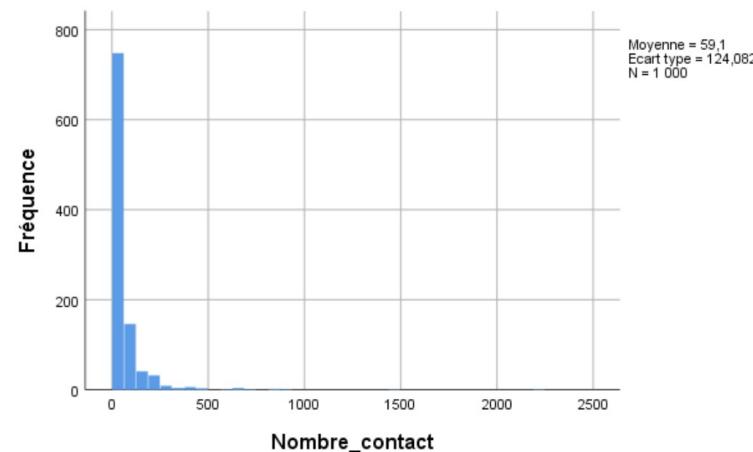
## Kurtosis : mesure de l'aplatissement d'une distribution



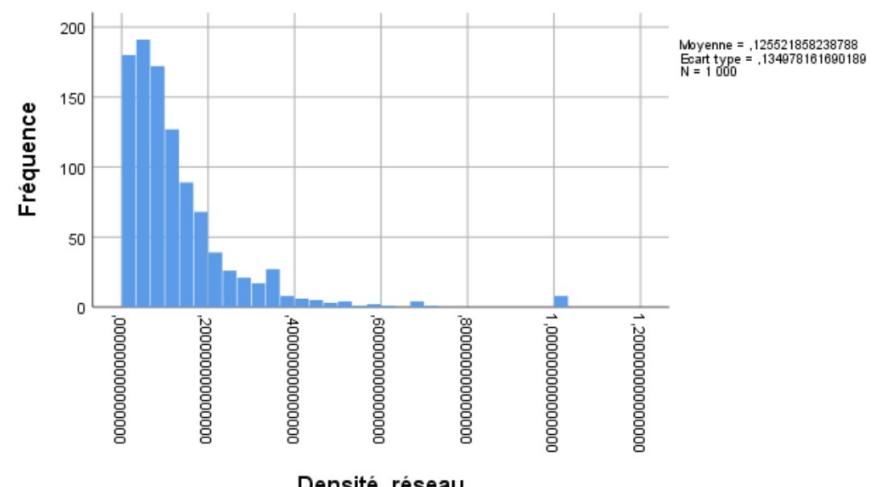
Kurtosis = - 0,438. Assez proche de loi normale



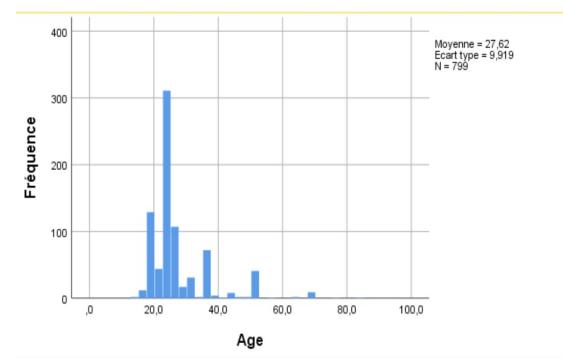
Kurtosis = - 0,1. Proche loi normale



Kurtosis = 117,79. Très regroupée

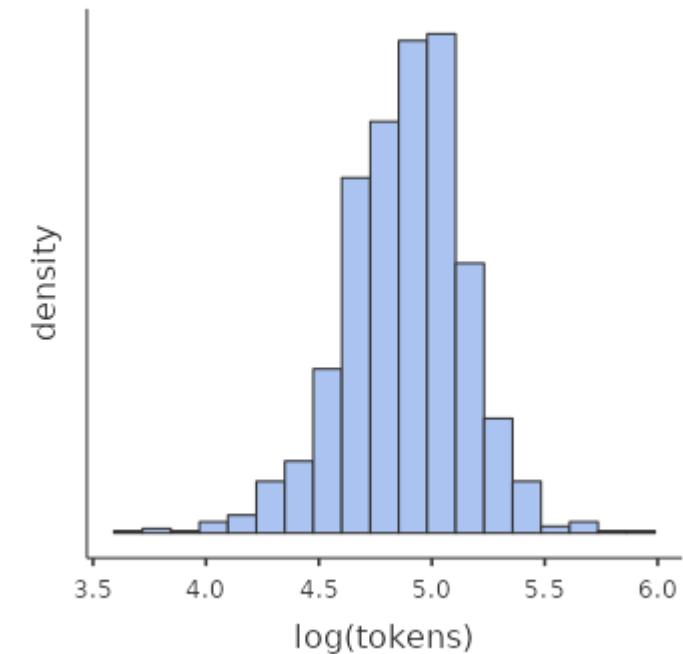
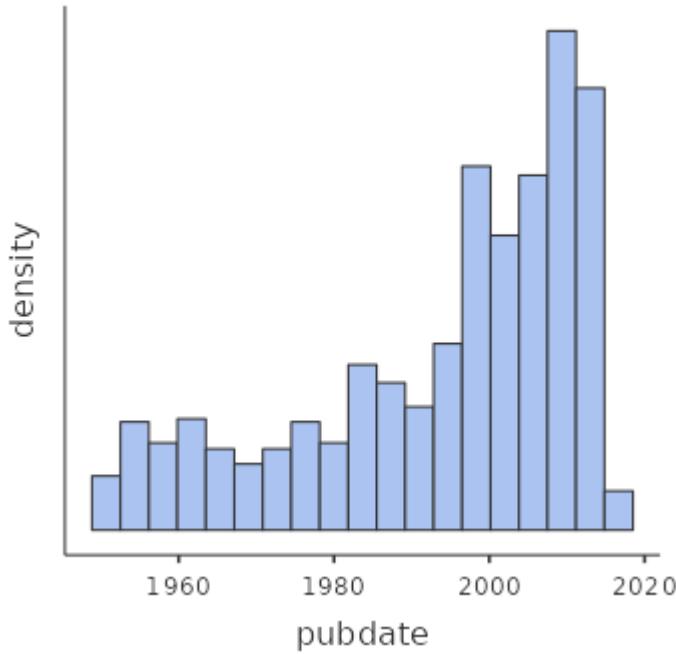
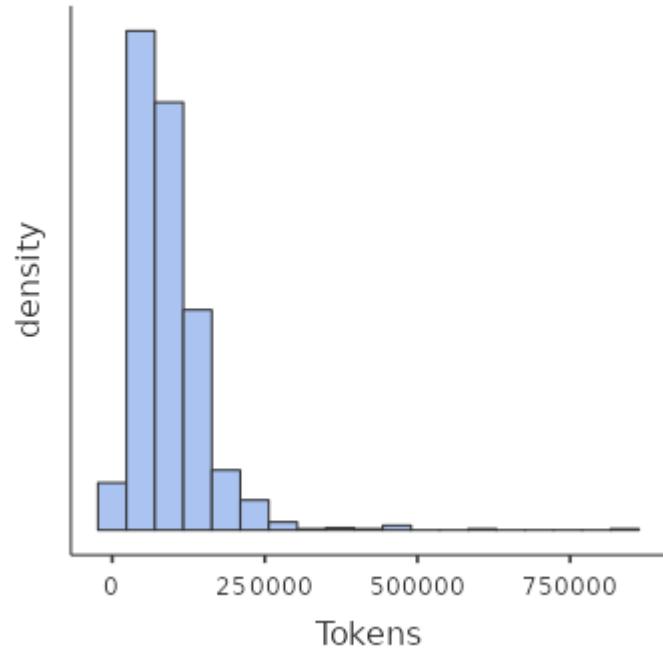


Kurtosis = 14,28. Regroupé



Kurtosis = 5,78.  
Plus regroupée  
qu'une loi normale

## Kurtosis : mesure de l'**aplatissement** d'une distribution



### Statistiques descriptives

	Tokens	pubdate	log(tokens)
Kurtosis	27.9	-0.350	1.14
Kurtosis de l'erreur-standard	0.145	0.145	0.145

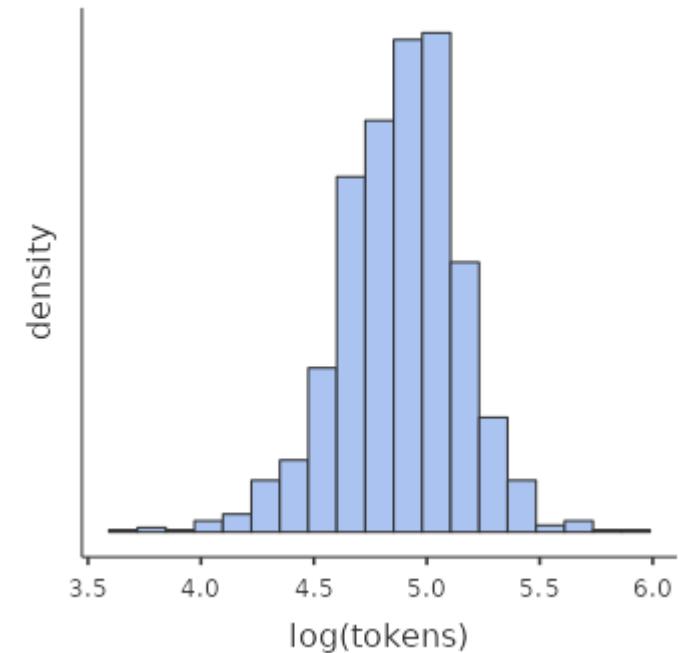
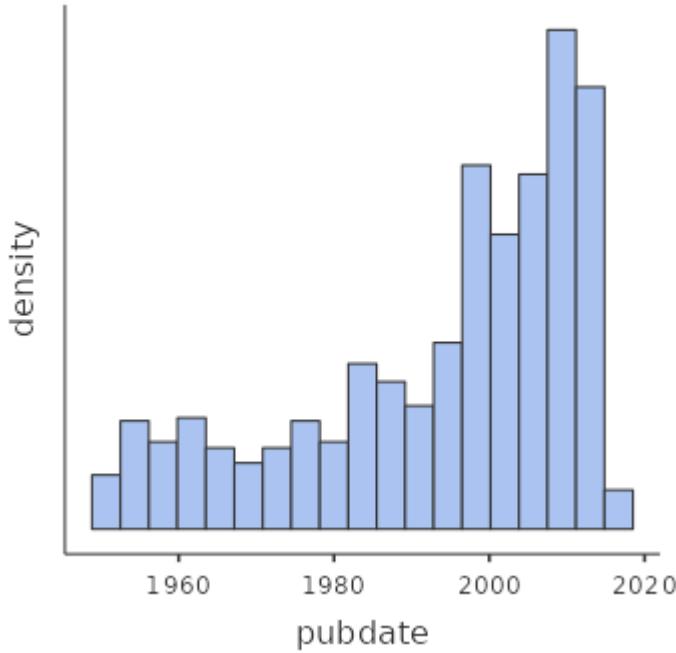
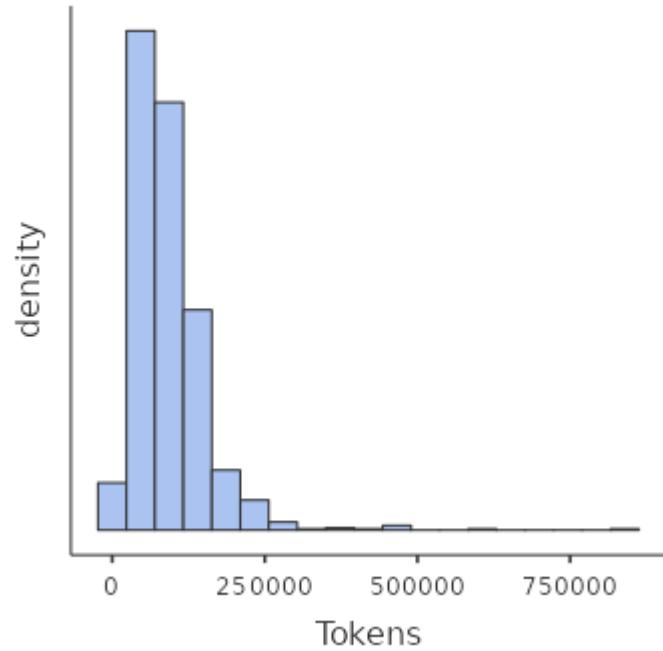
# Skewness : mesure de l'asymétrie d'une distribution

**La distribution normale est symétrique et a une valeur d'asymétrie égale à 0.** Une distribution caractérisée par **une importante asymétrie positive présente une partie droite plus allongée.** Une distribution caractérisée par une **importante asymétrie négative présente une extrémité gauche plus allongée.** Pour simplifier, **une valeur d'asymétrie deux fois supérieure à l'erreur standard correspond à une absence de symétrie.**

Ici on a un critère précis pour dire qu'une distribution n'est pas symétrique. Et donc non normale : valeur absolue du score de skewness deux fois supérieur à l'erreur standard de la mesure d'asymétrie

	Skewness		Deux fois l'erreur standard de l'indice	Conclusion
	Statistiques	Erreur standard		
Revenu_par_unité	-,277	,077	0,154	Dissymétrique à gauche
Densité_population	2,723	,077	0,154	Dissymétrique à droite
Proportion_proprio	,213	,077	0,154	Dissymétrique à droite
Nombre_contact	8,722	,077	0,154	Dissymétrique à droite
Densité_réseau	3,044	,077	0,154	Dissymétrique à droite
Age	2,246	,086	0,172	Dissymétrique à droite

## Skewness : mesure de l'asymétrie d'une distribution



### Statistiques descriptives

	Tokens	pubdate	log(tokens)
Coefficient d'asymétrie	3.62	-0.890	-0.353
Asymétrie de l'erreur-standard	0.0727	0.0727	0.0727

## En bref

### Paramètres pour décrire la distribution d'une variable quantitative

Paramètres de tendance centrale	Moyenne Médiane Mode
Paramètres de dispersion usuels	Etendue ou plage Variance Ecart type Ecart moyen Coefficient de variation (écart type standardisé par la moyenne)
Paramètres de forme	Aplatissement (Kurtosis) Asymétrie (Skewness)

5. Etudier la  
relation entre deux  
variables ou entre  
une variable et  
une constante

## 5.1 Vue générale

# Comparer deux mesures vs comparer une mesure et une constante

Exemple : 185 enfants âgés de 2 à 6 ans

4 tranches d'âge, 2 sexes, 2 milieux sociaux

Tranche d'âge	Milieu social	Effectif		
		Filles	Garçons	Total
2-3 ans (2;3-3;2)	parents cadres	10	11	21
	parents ouvriers	7	10	17
3-4 ans (3;3-4;0)	parents cadres	13	12	25
	parents ouvriers	12	8	20
4-5 ans (4;1-5;0)	parents cadres	19	8	27
	parents ouvriers	10	17	27
5-6 ans (5;1-6;0)	parents cadres	9	16	25
	parents ouvriers	13	10	23

Nardy, 2008

## Tâche

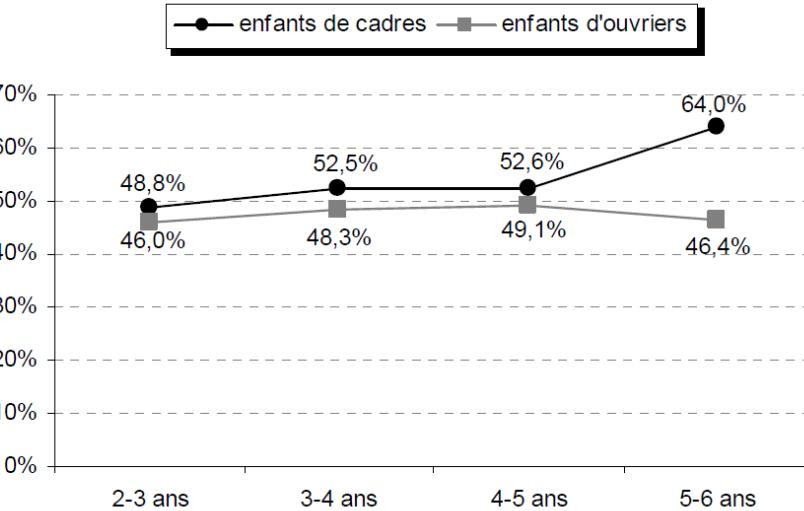
Deux marionnettes disent un énoncé avec un site de liaison facultative.

- une marionnette prononce la liaison : c'est un gros-Z-ours
- une marionnette ne la prononce pas : c'est un gro/ours

*L'enfant doit montrer du doigt « qui a bien parlé ».*

## Calcul de scores de réponses conformes à la norme

- Pour chaque enfant on calcule la proportion de réponses conformes (montrer la marionnette qui fait la liaison)
- On fait la moyenne des proportions par âge et milieu



## Deux questions

1/ Est-ce que les jugements des deux milieux sociaux sont différents ? **comparer deux mesures**

**Quelles variables met-on alors en relation ?**

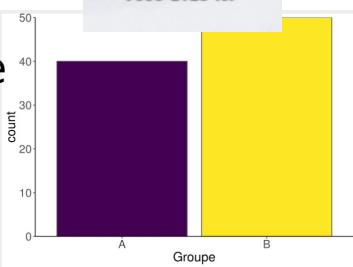
2/ Est-ce que les jugement pour chaque milieu et à chaque âge sont différents de réponses au hasard : **comparer une mesure à la valeur de hasard**

**Quelle est la valeur constante de hasard ?**

# Quand on étudie la relation entre deux variables, quels types de test d'inférence statistique utiliser pour quelle question de recherche ?

Les tests d'inférence statistiques permettent d'estimer le risque d'inférer un résultat d'un échantillon à une population et de décider si on « prend le risque » (si  $p \leq 0.05$  ou 5 %)

Un résultat ? Mais de quel type ?



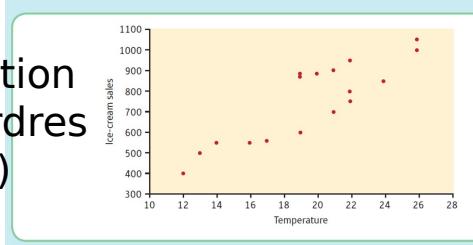
Une différence de moyennes  
(T de Student)

- Le revenu moyen est-il plus élevé chez les hommes que chez les femmes ?
- Le rythme cardiaque est-il en moyenne plus bas pendant le sommeil que pendant la veille ?

Une association entre des classifications  
(analyse de tableaux  
(Chi-deux ou  $\chi^2$ )

	Alcool	Sobre
Crime violent	19 70,3%	5 21,7%
Non-violent	8 29,6%	18 78,3%

- Les crimes violent sont-ils plutôt commis par des personnes alcoolisées ?



- La consommation de cigarettes (nbre de cig./jour) augmente-t-elle avec le revenu net mensuel (en euros /mois) ?

La possibilité de poser une des ces trois questions dépend du type de variable : on ne peut pas calculer une moyenne sur des variables « étiquettes » comme le sexe (femmes/homme) ou ordinales (titulaire du bac / d'une licence / d'un master)

# Types de problèmes traités par les tests de significativité et types de variables

	Différence de moyennes	Association entre des classifications	Association entre des ordres
Variables impliquées	- une variable quantitative et une variable nominale	- plusieurs variables nominales	- Plusieurs variables impliquant une relation d'ordre, donc ordinaires ou quantitatives.
Exemples	<ul style="list-style-type: none"> <li>- La moyenne au bac (variable quantitative) est-elle différente chez les filles et les garçons (variable nominale) ?</li> <li>- Le rythme cardiaque (variable quantitative) est-il plus lent la nuit que le jour (variable nominale : nuit vs.jour) ?</li> </ul>	<ul style="list-style-type: none"> <li>- La proportion de fumeurs et de non fumeurs est-elle différente chez les hommes et les femmes ? (fumeur/ non fumeur et sexe : deux variables nominales)</li> </ul>	<p><u>Quantitative / quantitative</u></p> <ul style="list-style-type: none"> <li>- Le revenu mensuel (en euros) augmente-t-il avec le nombre d'années d'études ?</li> </ul> <p><u>Quantitative / ordinaire</u></p> <ul style="list-style-type: none"> <li>- Le revenu mensuel (en euros) est-il lié à l'orientation politique à droite (sur une échelle LFI, PS, LREM, RN) ?</li> </ul> <p><u>Ordinale / ordinaire</u></p> <p>L'orientation politique à droite (échelle LFI, PS, LREM, RN) est-elle liée au niveau d'étude (pré-bac, bac à L3, supérieur à L3) ?</p>

## Trois caractéristiques d'une relation entre deux variables

La significativité de la relation	Peut-on généraliser de l'échantillon à la population ? (calcul p-value)
La question de la relation causale : son existence et son sens	<p>« la consommation d'alcool est (associée à) deux (types) de troubles psychiatriques/psychologiques (...): plus grande anxiété/propension à la dépression et moindre capacité à contrôler ses impulsions ».</p> <p><a href="https://www.addictaide.fr/impact-des-consommations-dalcool-sur-la-personnalite-boire-fait-devenir-anxieux-impulsif-et-moins-agreable/">https://www.addictaide.fr/impact-des-consommations-dalcool-sur-la-personnalite-boire-fait-devenir-anxieux-impulsif-et-moins-agreable/</a></p> <ul style="list-style-type: none"><li>- L'alcool provoque les troubles</li><li>- L'existence préalable de troubles incite à la consommation d'alcool</li><li>- Alcool et troubles s'entretiennent mutuellement</li><li>- Pas de lien causal entre les deux mais existence d'un troisième facteur qui les relie : le fait d'avoir traversé de nombreuses épreuves favorise les troubles psychologiques et la consommation d'alcool (effet « cigogne »)</li></ul>
La question de la force du lien	<p>Comparer</p> <ul style="list-style-type: none"><li>- Le risque de mourir d'un cancer du poumon en fumant plus de 25 cigarettes par jour est 36 fois plus élevé chez les hommes en comparaison à des non-fumeurs</li><li>- Le risque de mourir d'un cancer du poumon en fumant plus de 25 cigarettes par jour est 12 fois plus élevé chez les hommes en comparaison à des non-fumeurs</li></ul> <p>Dans lequel de ces deux cas le lien entre consommation du tabac et cancer est-il le plus fort ? (plus d'effet du premier sur le second)</p>

# Pour info

## Health consequences of smoking 1–4 cigarettes per day

K Bjartveit, A Tverdal

Tobacco Control 2005;14:315–320. doi: 10.1136/tc.2005.011932

**Objectives:** To determine the risk in men and women smoking 1–4 cigarettes per day of dying from specified smoking related diseases and from any cause.

**Design:** Prospective study.

**Setting:** Oslo city and three counties in Norway.

**Participants:** 23 521 men and 19 201 women, aged 35–49 years, screened for cardiovascular disease risk factors in the mid 1970s and followed throughout 2002.

**Outcomes:** Absolute mortality and relative risks adjusted for confounding variables, of dying from ischaemic heart disease, all cancer, lung cancer, and from all causes.

**Results:** Adjusted relative risk (95% confidence interval) in smokers of 1–4 cigarettes per day, with never smokers as reference, of dying from ischaemic heart disease was 2.74 (2.07 to 3.61) in men and 2.94 (1.75 to 4.95) in women. The corresponding figures for all cancer were 1.08 (0.78 to 1.49) and 1.14 (0.84 to 1.55), for lung cancer 2.79 (0.94 to 8.28) and 5.03 (1.81 to 13.98), and for any cause 1.57 (1.33 to 1.85) and 1.47 (1.19 to 1.82).

**Conclusions:** In both sexes, smoking 1–4 cigarettes per day was associated with a significantly higher risk of dying from ischaemic heart disease and from all causes, and from lung cancer in women. Smoking control policymakers and health educators should emphasise more strongly that light smokers also endanger their health.

See end of article for authors' affiliations

Correspondence to:  
Dr Kjell Bjartveit, Fridtjof  
Nansens vei 24 B, N-0369  
Oslo, Norway (home  
address);  
kjell.bjartveit@chello.no

Received 23 March 2005  
Accepted 20 July 2005

**Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.**

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
<b>Males</b>							
<b>All causes</b>							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
<b>Ischaemic heart disease</b>							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
<b>All cancer</b>							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
<b>Lung cancer</b>							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
<b>Females</b>							
<b>All causes</b>							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
<b>Ischaemic heart disease</b>							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
<b>All cancer</b>							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
<b>Lung cancer</b>							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

\*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

†Adjusted for age.

‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

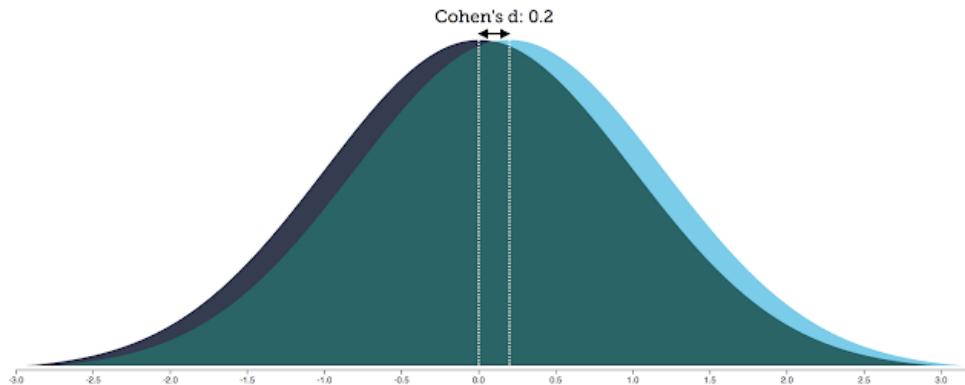
# Attention : significativité $\neq$ force d'une relation

- Une différence *significative* entre deux groupes (correspondant au croisement entre une variable nominale et une variable quantitative) n'implique pas que cette différence soit *importante*.
- Pour mesurer la force d'une relation entre ces deux variables, on calcule la **taille d'effet** ( $d$  de Cohen) :

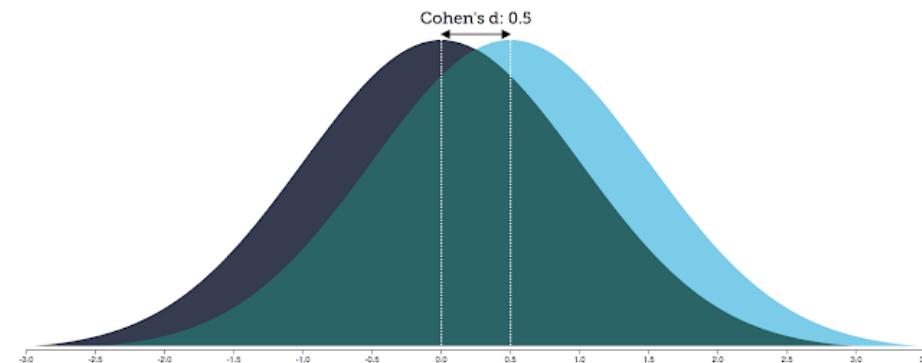
$$\delta = \frac{\mu_2 - \mu_1}{\sigma}$$

# Taille de l'effet

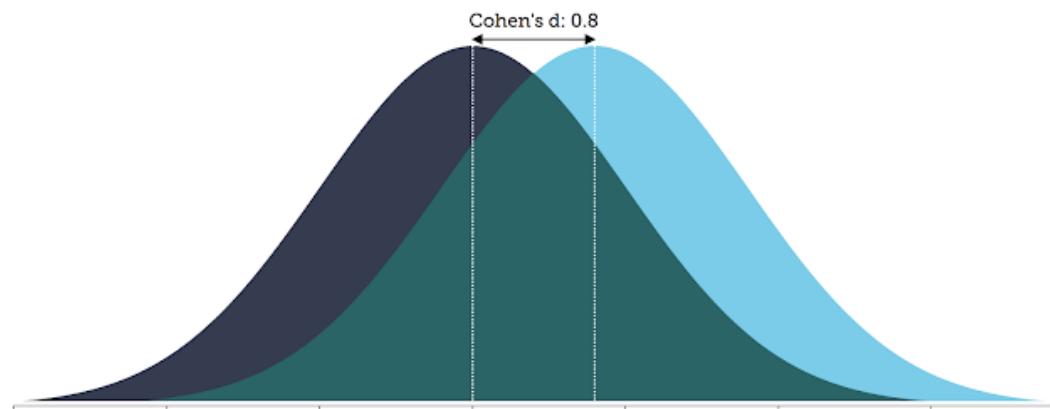
- Taille de l'effet



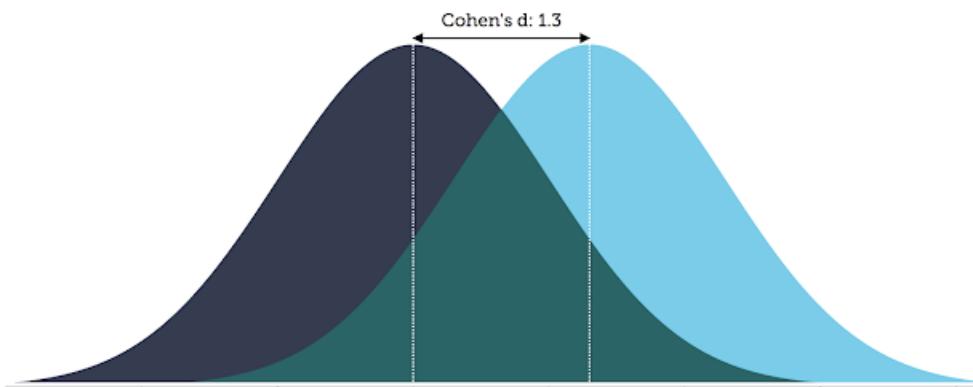
$d=0,2$  petit effet (85 % de superposition)



$d=0,5$  effet moyen (68 % de superposition)



$d=0,8$  effet fort (53 % de superposition)



$d=1,3$  effet très fort

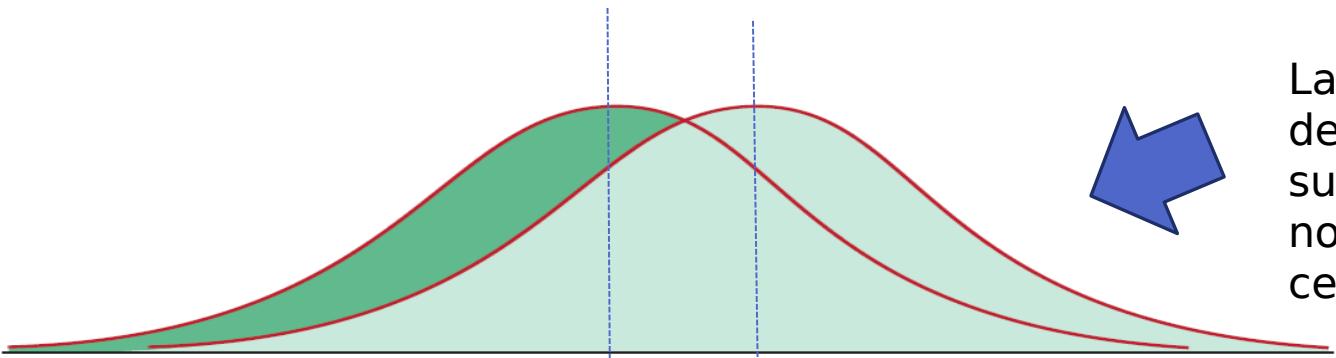
# Taille de l'effet

- Quelques exemples (tirés de *Stats sans maths*, p. 286) :
  - $d=2$  : différence de taille entre hommes et femmes.
  - $d=1$  : différence de niveau scolaire entre le CM1 et le CM2
  - $d=0,9$  : différence hommes-femmes dans certaines tâches de rotation mentale dans l'espace (à l'avantage des hommes)
  - $d=0,33$  : différence hommes-femmes dans certaines tâches de fluence verbale (à l'avantage des femmes)
  - $d=0,26$  : effet de la menace de stéréotype sur les performances.  
Il s'agit de l'effet négatif que le rappel d'un stéréotype (p.ex. « attention les filles/les garçons sont généralement mauvais.e.s à ce test ») peut avoir sur les performances d'un test scolaire.

## 5.2 Comparer les tendances centrales de deux distributions en comparant leur moyennes : le test T de student

# Comparer deux moyennes : le test de Student

## Logique du T de Student



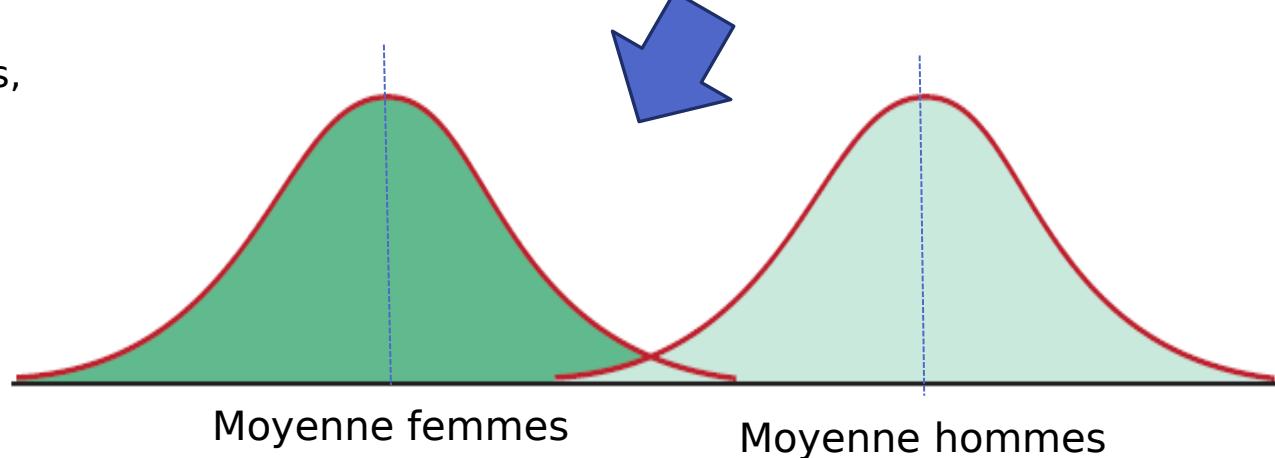
Le T de Student est un rapport entre la différence des moyennes **entre les groupes** et la dispersion **dans** les groupes (plus précisément, c'est le rapport entre la différence des moyennes, et **l'erreur standard**).

Si ce rapport est grand, alors les tendances centrales sont différentes et les valeurs se superposent peu. On estime  $p$  en calculant la probabilité d'obtenir une telle valeur de T dans une population où il n'y a pas de différence.

Distribution fictive des salaire des deux sexes : **femmes en vert** et **hommes en bleu**

La moyenne des femmes est inférieure à celle des hommes mais les distributions sont suffisamment dispersées pour que de nombreuses femmes gagnent plus que certains hommes

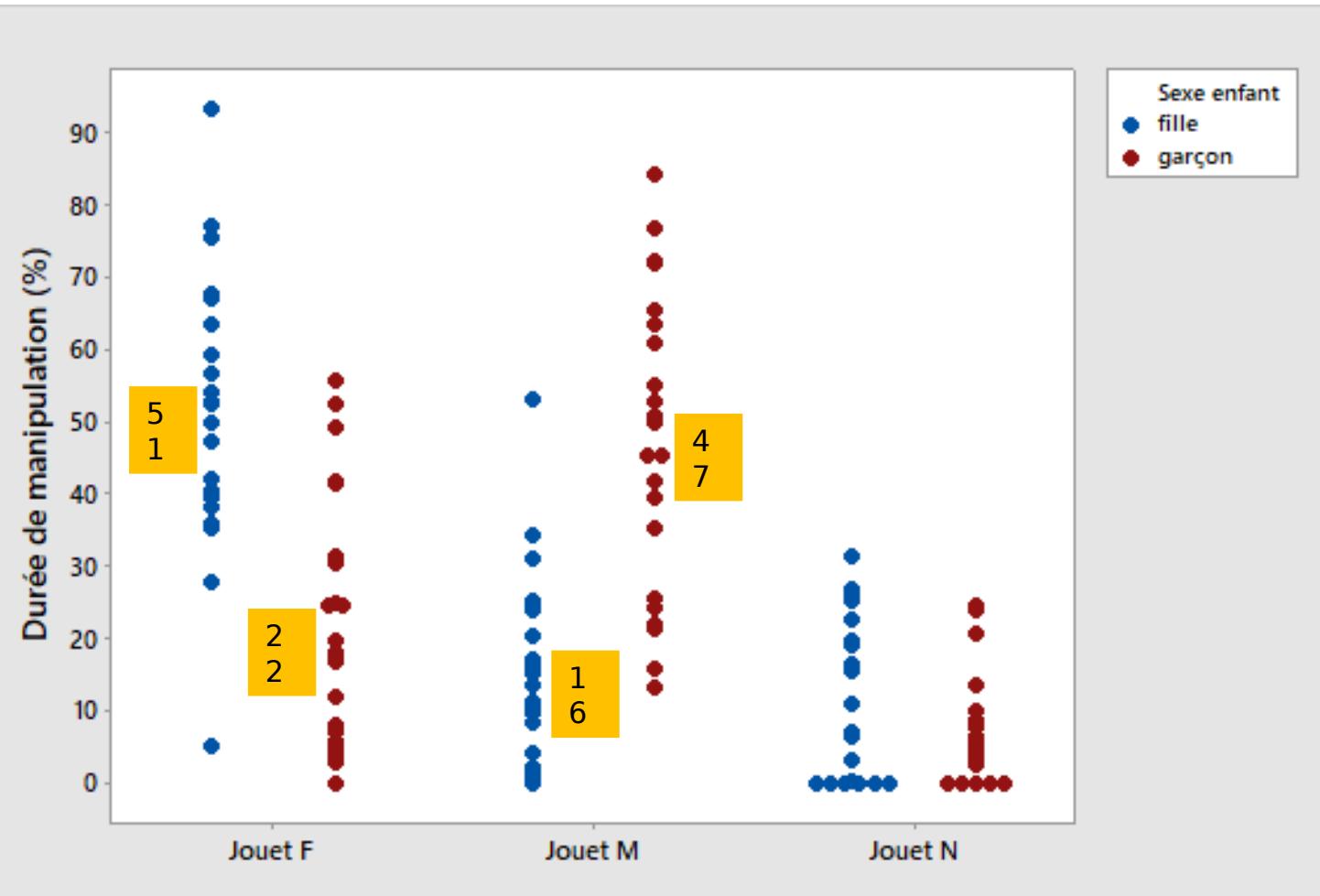
La moyenne des femmes est inférieures à celle des hommes mais les distributions sont suffisamment « resserrées » pour que très peu de femmes gagnent plus que les hommes



# Comparer deux moyennes : le test de Student

## Logique du T de Student

Exemple sur les jeux générés chez des enfants de 3 ans : thèse de Julie Bardet



12 filles et 12 garçons jouent deux fois en compagnie de leurs parents avec : des jouets stéréotypés en genre : maisons de poupée / garage et voiture des jouets neutres

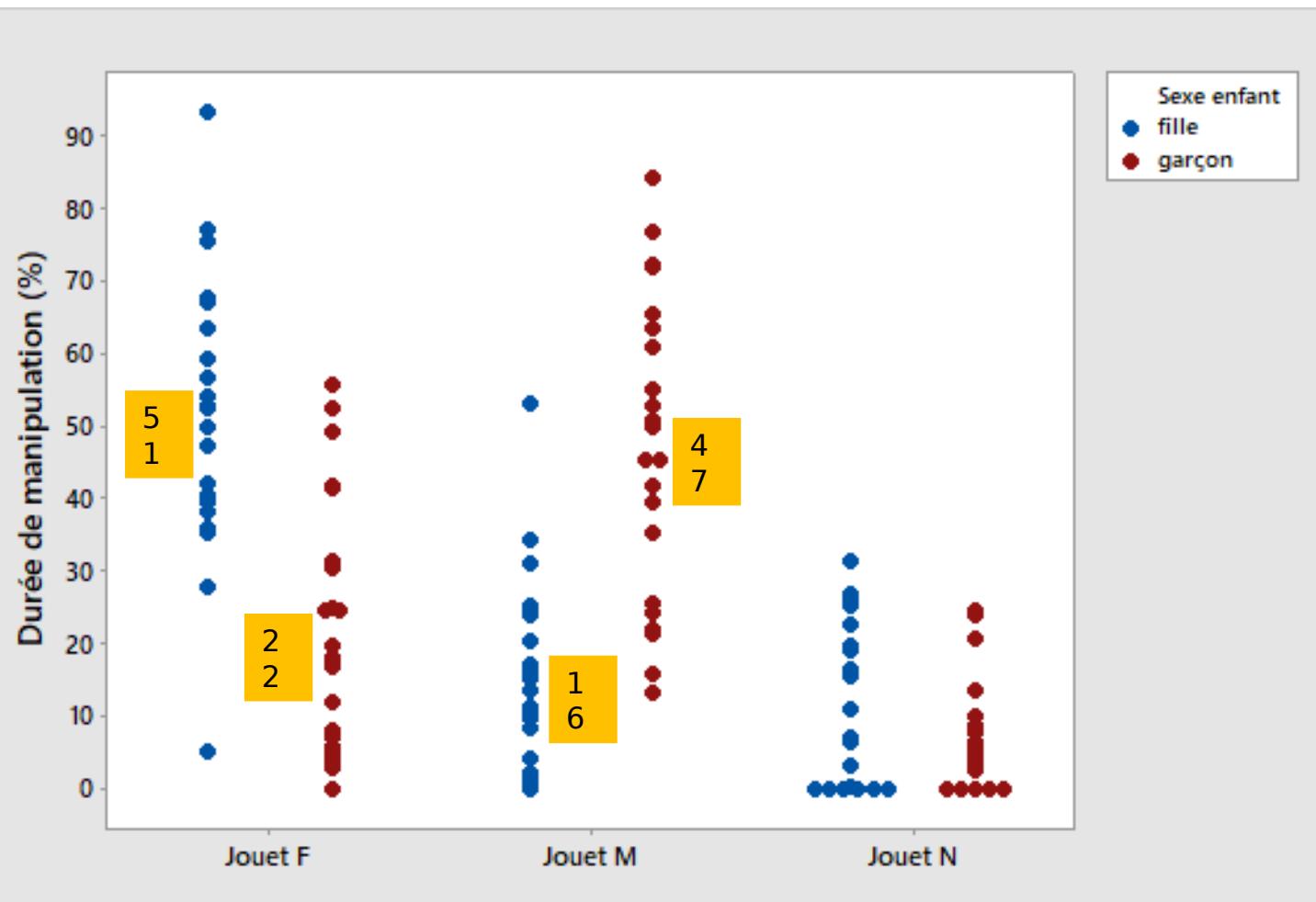
Le graphique ci-contre montre la proportion de temps que chaque enfant de chaque sexe ont passé à jouer avec chaque type de jouets : données individuelles et moyennes .

Chaque sexe joue globalement davantage que l'autre sexe avec les jouets correspondant aux stéréotypes attendus.

Mais pour savoir si on peut généraliser ces différences dans l'échantillon à la population des filles et garçons de 3 ans, il faut faire un test associant une *p-value* à cette différence de moyennes.

# Comparer deux moyennes : le test de Student

## Logique du T de Student



Exemple sur les jeux générés chez des enfants de 3 ans : thèse de Julie Bardet

Quelles moyennes peut-on comparer ?

Quels doutes font planer la visualisation des données individuelles ?

Le T de Student nous permet de savoir si on peut conclure à une différence entre les tendances centrales des groupes malgré les variations entre les individus.

# Différents types de comparaisons de moyennes

## Différents types de comparaisons de moyennes et de Test de Student

Comparer <b>deux groupes</b>	Test T pour échantillons indépendants	Comparer salaire des hommes et des femmes Comparer le temps de portage des jouets « féminins » chez les garçons et les filles
Comparer des échantillons appariés  1. Le même échantillon qui obtient deux séries de scores 2. Deux échantillons dans lesquels on peut appairer les scores	Test T pour échantillons appariés	1. Comparer le temps de portage des jouets « féminin » et « masculin » chez les filles. 2. Comparer le temps de portage des jouets « féminins » dans deux échantillons de jumeaux (garçons et filles)
Comparer une moyenne à une valeur constante	Test T pour échantillon unique	Comparer la moyenne des salaires des usagers de Twitter à la moyenne nationale.
Comparer plus de deux groupes ou plus de deux échantillons appariés	- Tests T répétés par paire : comparer A à B, B à C, A à C, etc. - ou ANOVA	Comparer les moyennes des utilisateurs de Twitter entre les 3 groupes d'âge

# Comparer une moyenne à une valeur constante

TABLEAU I Temps réel – maintien du <i>ne</i> et l'année de provenance des corpus			
Corpus	Nombre de <i>ne</i> maintenus sur possibles	... en pourcentages	
Corpus Péretz-Juillard (1972-74)	38/241	15.8 %	
Corpus Hansen (1989-1993)	67/727	9.2 %	Les corpus des années 1990 ensemble :
Corpus Malderez (1992-93)	42/602	7.0 %	109 / 1329 : 8.2 %

Voici le pourcentage de négation standard, où le *ne* est maintenu dans différents corpus d'échanges oraux **quotidiens du 20eme siècle**

Hansen, A., & Malderez, I. (2004). Le *ne* de négation en région parisienne: Une étude en temps réel. *Langage & Société*, 107, 5-30.

Dans les écrits standard (livres, journaux) ce taux est de 100%

## Question de recherche

Est-ce que du point de vue de l'usage du *ne*, l'écrit utilisé sur Twitter est plus proche de l'oral ou de l'écrit ?

- Est-ce que la moyenne des pourcentages de *ne* standard sur Twitter est différente de 100 % ?
- Est-ce que la moyenne des pourcentages de *ne* standard sur Twitter est différente de 15,8 % (valeur la plus haute des corpus oraux)?

# Comparer une moyenne à une valeur constante : le pourcentage de ne est-il inférieur à 100% ?

Données\_Twitter\_1000

Variables Données Analyse Editer

Exploration Tests t ANOVA Régression Fréquences Facteur

Test t

Test t pour échantillons indépendants  
Test t pour échantillons appariés  
Test t pour un échantillon

Variables dépendantes  
pct\_réalisation\_ne

Identifiant  
Revenu\_par\_unité  
Densité\_population  
Proportion\_proprio  
Nombre\_contact  
Densité\_reseau  
Age  
nbNEG

Tests

Student  
 Facteur de Bayes  
A priori 0.707  
 Rangs signés de Wilcoxon

Hypothèse

Valeur testée 1  
 ≠ de la valeur testée  
 ≥ à la valeur testée  
 < à la valeur testée

Statistiques additionnelles

Différence moyenne  
 Intervalle de confiance 95 %  
 Taille de l'effet  
 Intervalle de confiance 95 %  
 Statistiques descriptives  
 Graphes descriptifs

Vérifications des hypothèses

Test de normalité  
 Graphe Q-Q

Résultats

### Test t pour un échantillon

Test t pour un échantillon

	Statistique	ddl	p	Différence moyenne
pct_réalisation_ne	t de Student	-76.5	979	< .001
Note. H <sub>a</sub> μ < 1				

Test de normalité (Shapiro-Wilk)

	W	p
pct_réalisation_ne	0.874	< .001

Note. Une valeur p faible suggère une violation de la condition de normalité

Statistiques descriptives

	N	Moyenne	Médiane	Ecart-type	Erreur standard
pct_réalisation_ne	980	0.312	0.198	0.282	0.00899

La variable « pourcentage » est exprimée comme une proportion entre 0 et 1. La valeur de 100% est donc bien 1.

Valeur de t = -76  
ddl = 979  
significativité :  
p<.001

Moyenne des pourcentages de maintien du ne dans les tweets (31,2%)

Significativité unilatérale et non bilatérale. Pourquoi ?  
Qu'est-ce qui change ?

<https://www.edimark.fr/Front/frontpost/getfiles/6655.pdf>

# Comparer une moyenne à une valeur constante : le pourcentage de *ne* est-il supérieur à 15,8% (0.158)?

## Test t pour un échantillon

Test t pour un échantillon

		Statistique	ddl	p	Différence moyenne
pct_réalisation_ne	t de Student	17.1	979	< .001	0.154

Note.  $H_a \mu > 0.158$

Statistiques descriptives

	N	Moyenne	Médiane	Ecart-type	Erreur standard
pct_réalisation_ne	980	0.312	0.198	0.282	0.00899

## Même résultat

La valeur moyenne de 31,2 % dans les *tweets* est significativement supérieure à la valeur la plus haute dans les trois corpus oraux.

## Comment rédiger ce type de résultats ?

Dans l'échantillon de 1000 comptes Twitter, le taux moyen de réalisation standard du *ne* de négation ( $m = 31,2\%$ ) est significativement inférieur à la valeur de 100% de réalisation ( $T_{(979)} = 76$ ;  $p < 0.001$ ) et significativement supérieur à la valeur de 15,8% ( $T_{(979)} = 17.1$ ;  $p < 0.001$ ).

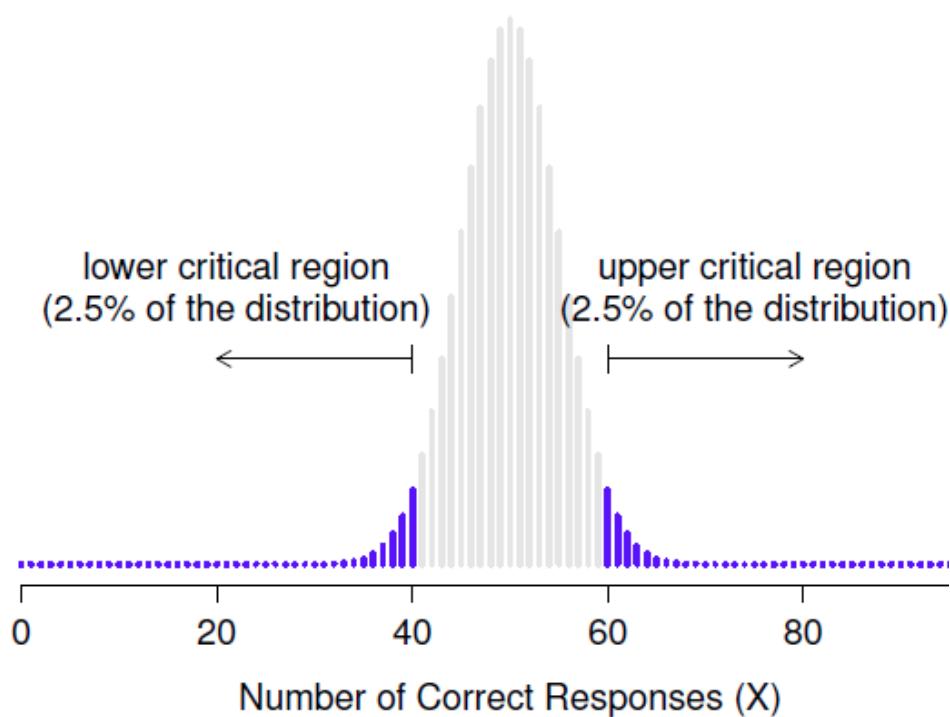
## On peut rendre la différence plus « parlante »

La réalisation moyenne de *ne* dans les *tweets* se situe donc entre sa valeur dans l'écrit standard et sa valeur dans un échange oral familier.

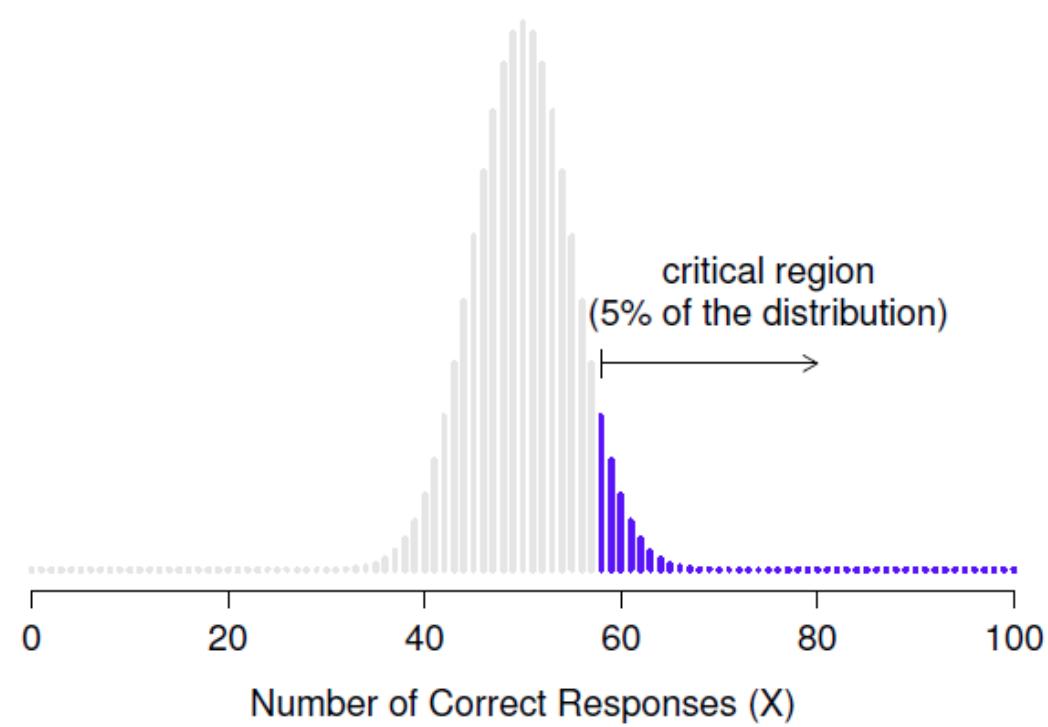
**Ensuite on peut interpréter, donner du sens par rapport à une hypothèse, une théorie...**

## Test unilatéral vs bilatéral

Critical Regions for a Two-Sided Test



Critical Region for a One-Sided Test



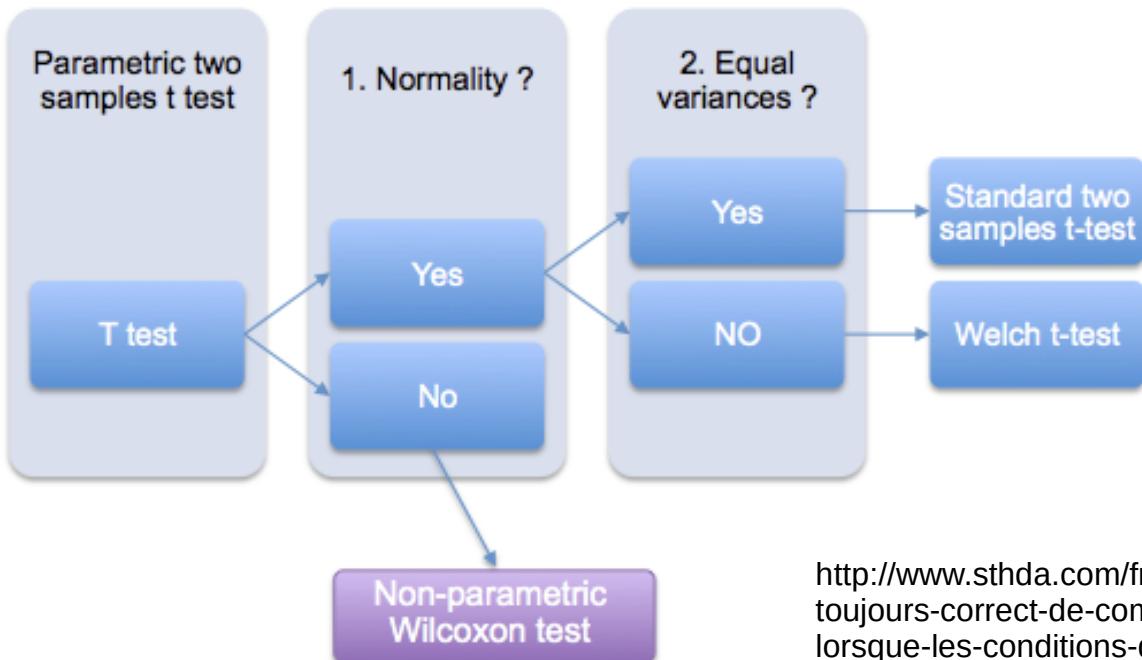
Suivant le type de test, les 5 % de chance d'obtenir une certaine observation sous l'hypothèse nulle se répartissent différemment par rapport à la courbe de Gauss. Un test unilatéral est donc plus puissant quand il peut s'appliquer (on divise  $p$  par 2)

# Comparer des moyennes : conditions d'application des différents tests

## Le T-test s'applique dans des conditions de normalité

- Lorsque les variances des deux groupes sont proches (test de Levene) on peut appliquer le test Standard (dans le cas d'une comparaison de deux groupes)
- Sinon on applique la variante de Welch

Si on ne peut supposer la normalité, on pourra appliquer un test non paramétrique tel que le test U (aussi appelé test de Wilcoxon)



<http://www.sthda.com/french/wiki/test-de-student-est-il-toujours-correct-de-comparer-des-moyennes#que-faire-lorsque-les-conditions-dapplication-du-test-de-student-ne-sont-pas-remplies>

# Le test t standard peut-il s'appliquer dans le cas précédent ?

Dans le cas précédent, les conditions de normalité étaient elles vérifiées ?

Test de normalité (Shapiro-Wilk)		
	W	p
pct_réalisation_ne	0.874	< .001

Note. Une valeur p faible suggère une violation de la condition de normalité

Non. Il faut donc appliquer plutôt le test de Wilcoxon, qui confirme la significativité

## Test t pour un échantillon

	Statistique	p	Différence moyenne
pct_réalisation_ne	W de Wilcoxon	< .001	0.131

Note.  $H_a \mu > 0.158$

## **Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)**

**Dans un corpus d'une centaine de traductions de textes de Heine, Keats et Leopardi, on s'intéresse à différents phénomènes traductionnels et stylistiques :**

- allongement traductionnel
- nombre d'omissions
- longueur moyenne des phrases
- fréquence et position des adjectifs épithètes
- fréquence des ponctuations
- fréquence de la conjonction « et » par rapport aux fréquences des conjonctions équivalentes en allemand, italien et anglais

Ouvrir le fichier *Données corpus traduit HDR.Pascale.xlsx*. Quelles sont à votre avis les variables explicatives et les variables expliquées ?

# **Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)**

## **Première étude - Fréquence des adjectifs**

- dans l'onglet « Les adj épithètes », rajouter 3 colonnes « Keats », « Leopardi » et « Heine », afin d'en faire des variables binaires indépendantes.
- importer dans Jamovi les données de cet onglet (passer par un export CSV)
- examiner la distribution des fréquences : que concluez-vous ?
- tester l'hypothèse selon laquelle la fréquence des adjectifs est supérieure chez Keats. Vérifiez les conditions d'application du test.
- que concluez-vous ?

# Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)

## Première étude - Fréquence des adjectifs

On constate un effet pour la fréquence des adjectifs.

Les traductions de Keats contiennent donc en moyenne significativement plus d'adjectifs.

Concernant la variable « Taux de postpositions » on ne constate aucun effet, en revanche.

### Test t pour échantillons indépendants

Test t pour échantillons indépendants

		Statistique	ddl	p
Fréquence (pour 100 mots)	t de Student	-8.241	94.0	< .001
Taux de postposition (%)	t de Student	-0.451	94.0	0.653

### Hypothèses

Test de normalité (Shapiro-Wilk)

	W	p
Fréquence (pour 100 mots)	0.994	0.929
Taux de postposition (%)	0.962	0.007

Note. Une valeur p faible suggère une violation de la condition de normalité

Test d'homogénéité des variances (test de Levene)

	F	ddl	ddl2	p
Fréquence (pour 100 mots)	3.65	1	94	0.059
Taux de postposition (%)	2.20	1	94	0.141

Note. Une valeur p faible suggère une violation de la condition d'égalité des variances

## **Deuxième étude - Fréquence des omissions**

- dans l'onglet « Omissions », rajouter 3 colonnes « Vers libres », « Vers métriques » et « Prose », afin d'en faire des variables binaires séparées.
- importer dans Jamovi les données de cet onglet (passer par un export CSV)
- examiner la distribution des fréquences d'omission, en fonction des différentes formes : quelle hypothèse peut-on émettre ?
- tester les hypothèses selon laquelle la forme (vers libres, vers métriques et prose) détermine la fréquence des omissions. Peut-on appliquer le test T ? a-t-on des variances homogènes ? Peut-on appliquer le test de Welch ? L'hypothèse de normalité est-elle vérifiée ? Appliquez le test U de Mann-Whitney (aussi appelé test de Wilcoxon).
- que concluez-vous ?

# Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques

UFR D'Angoulême

## Deuxième étude - Fréquence des omissions entre « vers libre » et le reste

- On ne peut appliquer de test paramétrique, les conditions de normalité n'étant pas respectées.
- Le test U de Mann-Whitney permet cependant d'affirmer que la différence de moyenne de 5 entre le groupe « vers métrique » et le reste est significatif.

Test t pour échantillons indépendants

		Statistique	ddl	p	Différence moyenne	Différence d'erreur standard
Nb d'omissions	t de Student	-7.80 <sup>a</sup>	94.0	< .001	-4.82	0.618
	t de Welch	-6.83	46.0	< .001	-4.82	0.705
	U de Mann-Whitney	356		< .001	-5.00	

<sup>a</sup> Le test de Levene est significatif ( $p < 0.05$ ), suggérant une violation de la condition d'égalité des variances

### Hypothèses

Test de normalité (Shapiro-Wilk)

	W	p
Nb d'omissions	0.946	< .001

Note. Une valeur p faible suggère une violation de la condition de normalité

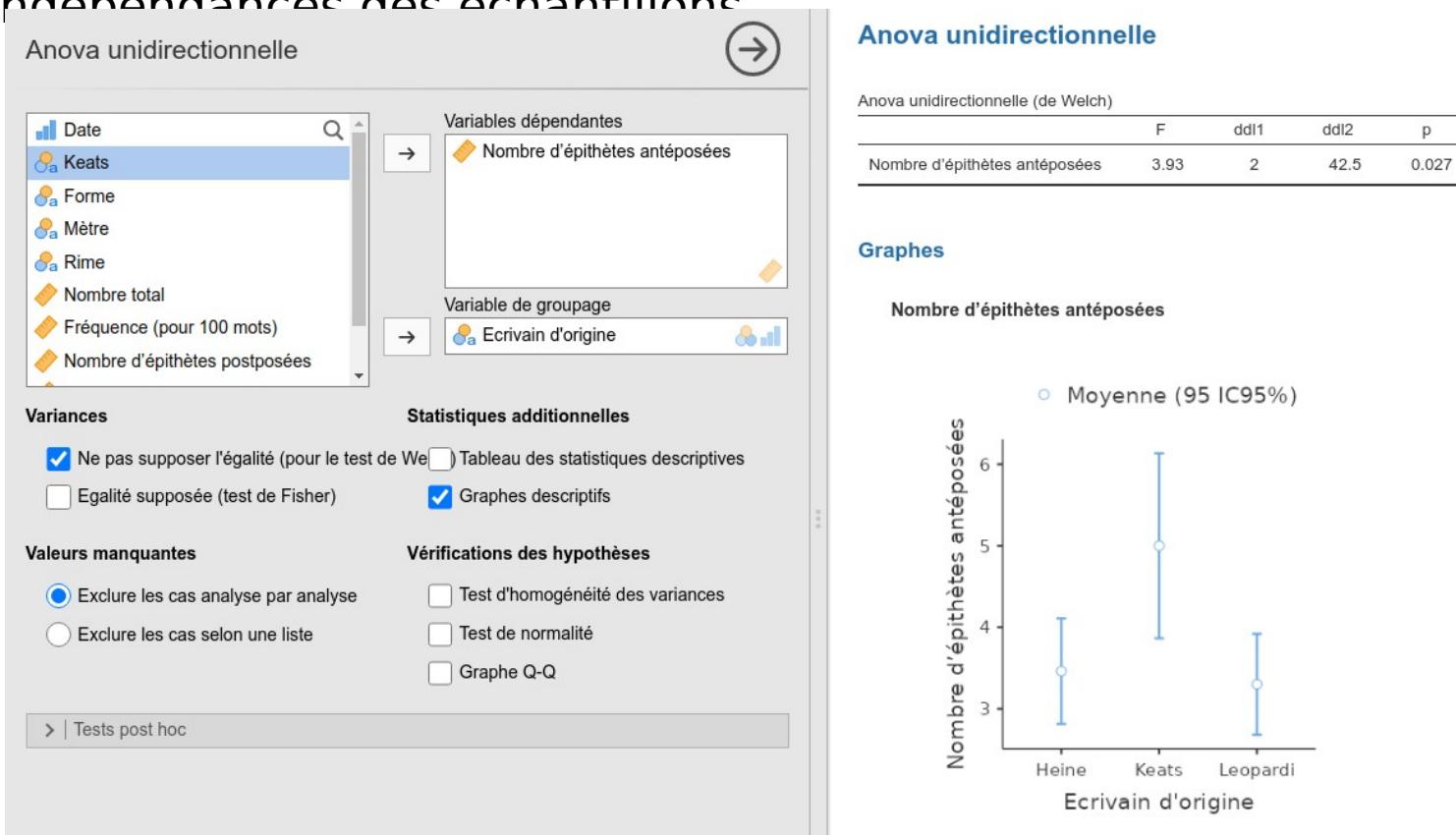
Test d'homogénéité des variances (test de Levene)

	F	ddl	ddl2	p
Nb d'omissions	42.6	1	94	< .001

Note. Une valeur p faible suggère une violation de la condition d'égalité des variances

# Comparer plus de deux groupes : Vers libres vs Vers métriques vs Prose

Quand on compare plus de deux moyennes on utilise en principe l'**analyse de variance ou ANOVA**. Dans sa version simple (**ANOVA unidirectionnelle**), on ne prend en compte qu'un facteur (une seule variable explicative). C'est une généralisation du test T à 3 groupes ou plus. Elle possède les mêmes conditions d'applications : distribution normale, homogénéité des variances, indépendances des échantillons.



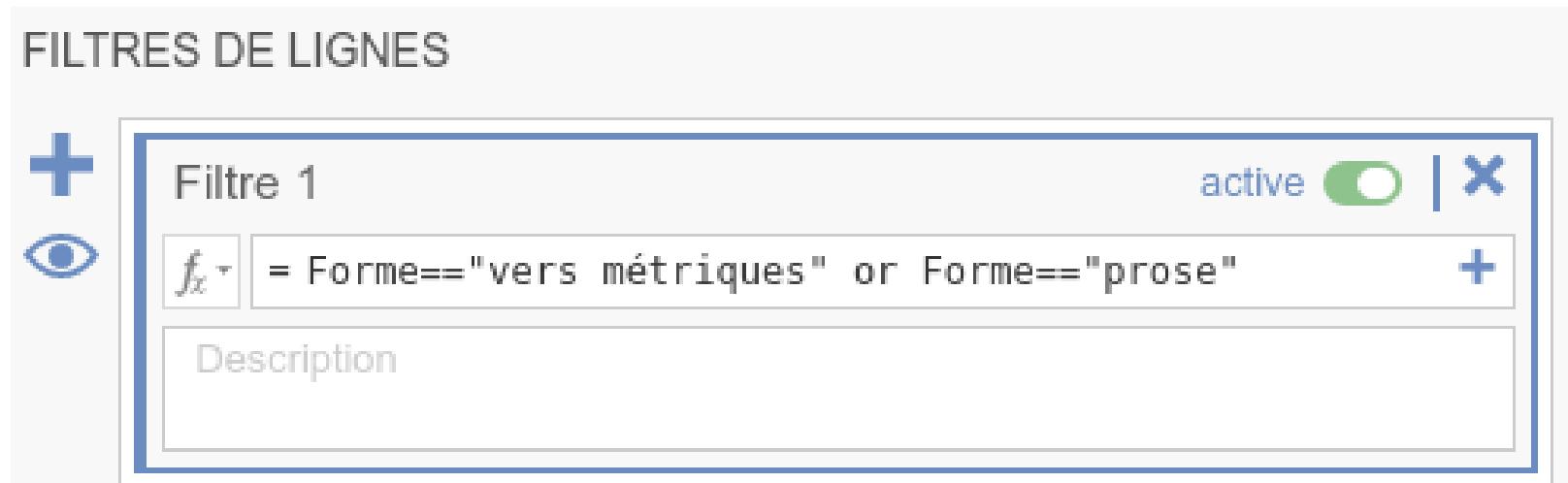
→ Si le test permet d'identifier un effet sur au moins un des groupes, il n'indique pas sur quel groupe. **Pour comparer précisément les moyennes des groupes G1, G2 et G3, nous allons comparer G1 à G2, puis G1 à G3, puis G2 à G3.**

# Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques

(HDR, P. Roux)

## Troisième étude - Fréquence des omissions, entre « Prose » et « Vers métriques »

Pour comparer nos deux groupes, il faut filtrer les données afin de supprimer les lignes correspondant à « vers libres » : pour ce faire on ajoute un filtre (dans une syntaxe assez proche de celle d'Excel). De la sorte pas la peine de réimporter un nouveau tableau de données :



# Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)

## Troisième étude - Fréquence des omissions, entre « Prose » et « Vers métriques »

Dès lors, on peut lancer un test de Welch unilatéral (hypothèse moy(Vers métrique) > moy(Prose)) sur la variable « Omissions » en groupant nos données en fonction de la variable « Forme ». La différence apparaît comme forte et significative :

### Test t pour échantillons indépendants

#### Test t pour échantillons indépendants

		Statistique	ddl	p	Différence moyenne	Différence d'erreur standard
Nb d'omissions	t de Welch	-6.60	51.5	< .001	-4.92	0.746

Note.  $H_0: \mu_{\text{prose}} = \mu_{\text{vers métriques}}$

### Hypothèses

#### Test de normalité (Shapiro-Wilk)

	W	p
Nb d'omissions	0.961	0.066

Note. Une valeur p faible suggère une violation de la condition de normalité

# Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)

## Troisième étude - Fréquence des omissions, entre «Vers libres» et «Vers métriques»

La même démarche permet de conclure que  $\text{moy}(\text{Vers libres}) < \text{moy}(\text{Vers métriques})$  :

### Test t pour échantillons indépendants

Test t pour échantillons indépendants

	Statistique	ddl	p	Différence moyenne	Différence d'erreur standard
Nb d'omissions	t de Welch	-6.61	49.9	< .001	-4.78

Note.  $H_0: \mu_{\text{vers libres}} \leq \mu_{\text{vers métriques}}$

### Hypothèses

Test de normalité (Shapiro-Wilk)

	W	p
Nb d'omissions	0.960	0.013

Note. Une valeur p faible suggère une violation de la condition de normalité

## **Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)**

### **Troisième étude - Fréquence des omissions, entre «Vers libres » et « prose »**

Concernant cette dernière comparaison, on ne peut conclure à un effet (les moyennes étant très proches) :

#### **Test t pour échantillons indépendants**

Test t pour échantillons indépendants

		Statistique	ddl	p	Différence moyenne	Différence d'erreur standard
Nb d'omissions	t de Welch	-0.369	35.1	0.714	-0.150	0.406

## **Comparer les moyennes de deux groupes: une étude sur un corpus de traductions poétiques (HDR, P. Roux)**

### **Quatrième étude - Indice d'originalité lexicale**

Dans un autre jeu de données, on s'intéresse au lien entre forme (prose, vers libre, vers métrique) et l'originalité du lexique. Des indices sont calculés pour mesurer la « conformité lexicale » (lexnorm\_mots ou lexnorm\_mots conv), c'est-à-dire le fait qu'une traduction partage une plus ou moins grande proportion de son lexique avec les autres traductions du même texte. Un autre indice a été calculé pour quantifier l'originalité : lexorig\_normalisé.

On émet les hypothèses suivantes :

- 1/ Les traductions en vers métriques sont moins conformes à la norme / plus originales que celles en prose et celles en vers libres.
- 2/ Les traductions en vers libres sont moins conformes / plus originales que les traductions en prose.

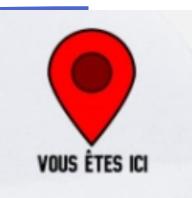
**Testez ces hypothèses à partir des données issues du fichier : [Données corpus traduit\\_HDR.Pascale.v2.enrichie.xlsx](#). Quel est l'indice qui montre un meilleur effet ?**

5.3 Chercher un lien entre deux variables nominales en examinant si la répartition des effectifs dans un tableau est différente du hasard

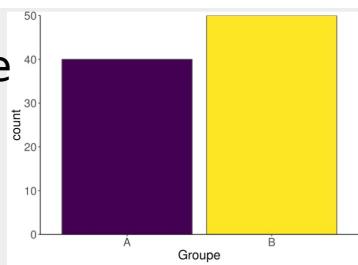
# Quand on étudie la relation entre deux variables, quels types de test d'inférence statistique utiliser pour quelle question de recherche ?

Les tests d'inférence statistiques permettent d'estimer le risque d'inférer un résultat d'un échantillon à une population et de décider si on « prend le risque » (si  $p \leq 0.05$  ou 5 %)

Un résultat ? Mais de quel type ?



Une différence de moyennes  
(T de Student)



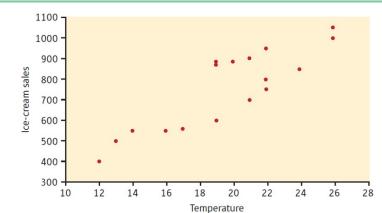
- Le revenu moyen est-il plus élevé chez les hommes que chez les femmes ?
- Le rythme cardiaque est-il en moyenne plus bas pendant le sommeil que pendant la veille ?

Une association entre des classifications  
(analyse de tableaux  
(Chi-deux ou  $\chi^2$ )

	Alcool	Sobre
Crime violent	19 70,3%	5 21,7%
Non-violent	8 29,6%	18 78,3%

- Les crimes violent sont-ils plutôt commis par des personnes alcoolisées ?

Une association entre des ordres  
(Corrélation)



- La consommation de cigarettes (nbre de cig./jour) augmente-t-elle avec le revenu net mensuel (en euros /mois) ?

La possibilité de poser une des ces trois questions dépend du type de variable : on ne peut pas calculer une moyenne sur des variables « étiquettes » comme le sexe (femmes/homme) ou ordinaires (titulaire du bac / d'une licence / d'un master)

# La répartition des effectifs dans un tableau et le test de chi-deux

On demande à 100 doctorant LLSH (50 femmes et 50 hommes) si ils sont pour ou contre la peine de mort. On examine la relation entre sexe (variable nominale 1) et opinion pour/contre la peine de mort. On représente le résultat dans ce tableau.

Sexe de l'étudiant	Opinion face à la peine de mort		Total
	Contre	Pour	
Femme			50
Homme			50
			183

# La répartition des effectifs dans un tableau et le test de chi-deux

Voici une répartition possible des réponses : en **effectifs** et en **pourcentages**

Sexe de l'étudiant		Opinion face à la peine de mort		Total
		Contre	Pour	
Femme	FE	30 (50%)	20 (50%)	50 (100%)
Homme	FE	30 (50%)	20 (50%)	50 (100%)
Total		60	40	100

1. Comment sont calculés les totaux marginaux en bas du tableau ?
2. Pourquoi les totaux marginaux à droite du tableau sont-ils imposés et fixes ?
3. Comment sont calculés les pourcentages dans le tableau ? Et comment aurait-on pu les calculer autrement ?

4. Et surtout : que nous dit cette répartition sur la relation entre sexe et opinion sur la peine de mort ?

# La répartition des effectifs dans un tableau et le test de chi-deux

Cette répartition est celle qu'on attend si il n'y a aucun lien entre sexe et opinion

Sexe de l'étudiant		Opinion face à la peine de mort		Total
		Contre	Pour	
Femme	FE	30 (50%)	20 (50%)	50 (100%)
Homme	FE	30 (50%)	20 (50%)	50 (100%)
Total		60	40	100

Chez les hommes et chez les femmes les proportions de pour et de contre sont identiques : 60 % de Contre (30/50) et 40% de Pour (20/50)

Chez les Pour et chez les Contre les proportions d'hommes et de Femmes sont identiques : la moitié d'homme et la moitié de femme

Dans cette répartition, les effectifs dans les cases reproduisent à l'identique les mêmes proportions d'hommes et de femmes, de Pour et de Contre, que dans les totaux marginaux.

Cette répartition est dite « théorique » et les effectifs dans les cases sont dits « théoriques » ou « attendus ». Elle est celle qu'on attend si il n'y a aucun lien entre Sexe et Opinion.

Quand on connaît les valeurs marginales, on peut toujours remplir le tableau avec des valeurs attendues

# La répartition des effectifs dans un tableau et le test de chi-deux

On suppose maintenant que l'on a un échantillon déséquilibré dans sa répartition H/F : Calculer la répartition théorique du tableau suivant. Quelle formule appliquer ?

Sexe de l'étudiant	Opinion face à la peine de mort		Total
	Contre	Pour	
Femme			80
Homme			20
		60	40

$$FT \text{ (fréquence théorique)} = \text{tot\_Ligne} * \text{tot\_Colonne} / N$$

# La répartition des effectifs dans un tableau et le test de chi-deux

Voici maintenant la répartition observée « réellement » (en gras) à la suite du sondage sur un effectif équilibré. Elle est ici comparée à la répartition théorique.

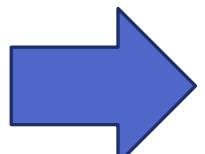
Sexe de l'étudiant		Opinion face à la peine de mort		Total
		Contre	Pour	
Femme	FT	30 (50%)	20 (50%)	<b>50 (100%)</b>
	FO	<b>36 (72%)</b>	14 (28%)	
Homme	FT	30 (50%)	20 (50%)	<b>50 (100%)</b>
	FO	24 (48%)	26 (52%)	
Total		<b>60</b>	<b>40</b>	<b>100</b>

72% des femmes sont contre la peine de mort. Mais seulement 48% des hommes.

Parmi les réponses Contre, il y a une majorité de femmes ( $36/60 = 60\%$ ). Parmi les réponses Pour, il y a une majorité d'hommes ( $26/40 = 65\%$ )

**Le chi-deux calcule la distance entre les effectifs observés et les effectifs théoriques (pas de relation entre les variables). Plus cette distance est grande, plus la valeur du chi-deux est grande et plus il y a des chances que la *p-value* associée soit significative (i.e. faible)**

Tests $\chi^2$	Valeur	ddl	p
$\chi^2$	6.00	1	0.014
N	100		



La *p-value* associée au chi-deux est inférieure au seuil usuel de 0.05. On peut donc prendre le risque de généraliser ces résultats.

# La répartition des effectifs dans un tableau et le test de chi-deux

## Formule du chi-deux

$$\chi^2 = \sum_{ij} \frac{(eff.\cdot obs_{ij} - eff.\cdot theo_{ij})^2}{eff.\cdot theo_{ij}}$$

**Attention : plus un chi-deux est élevé et plus il y a des chances qu'il soit significatif. Toutefois la valeur du chi-deux ne permet pas de trancher sur la force de la relation**

Le chi-deux en B vaut deux fois le chi-deux en A.

Mais la taille d'effet est identique. Pourquoi ?

		Tableau X. Lien entre le sexe des répondants et l'attitude face à la peine de mort	
		Hommes	Femmes
Pour	30	20	50
	20	30	50
	50	50	100

La valeur du Chi carré est de 4,0, p<0,05

		Tableau X. Lien entre le sexe des répondants et l'attitude face à la peine de mort	
		Hommes	Femmes
Pour	60	40	100
	40	60	100
	100	100	200

La valeur du Chi carré est de 8,0, p<0,01, soit exactement le double.

Que pensez-vous de ce tableau ?

		Hommes	Femmes	Total
Pour	0	50	50	
	50	0	50	
	50	50	100	

# La répartition des effectifs dans un tableau et le test de chi-deux

Ce test du chi-deux est dit « **test du chi-2 d'indépendance** ». Il permet de rejeter l'hypothèse d'indépendance entre 2 variables ayant 2 ou plus modalités.

**Quelles sont ses conditions d'application ? pour une table  $n \times m$  le test s'applique si au moins 25 % des valeurs théoriques sont supérieures ou égales à 5.**

Pour le cas des tables  $2 \times 2$ , si cette condition est violée, on peut recourir au **test exact de Fischer** (qui indique une valeur p).

Si on a  $n \times m$  modalités, les degrés de liberté sont :  $ddl = (n-1) * (m-1)$

N.B. : il est possible d'appliquer le test de chi2 sur une seule variable, afin de tester si la répartition des effectifs sur plusieurs modalités est uniforme (ou bien correspond à une répartition a priori) : il s'agit du **test de chi2 « de conformité » ou de « qualité d'ajustement »**.

# **La répartition des effectifs dans un tableau et le test de chi-deux**

Exercice : effectuer le test sur Jamovi avec les données ci-dessus.

- Quelles sont les variables ?
- Comment intégrer les effectifs de manière efficace ?

# La répartition des effectifs dans un tableau et le test de chi-deux

Exercice : effectuer le test sur Jamovi avec les données ci-dessus.

Sexe	Opinion	Effectif
H	Pour	26
H	Contre	24
F	Pour	14
F	Contre	36

Tables de contingence

Sexe	Opinion		Total
	Pour	Contre	
H	26	24	50
F	14	36	50
Total	40	60	100

Tests  $\chi^2$

	Valeur	ddl	p
$\chi^2$	6.00	1	0.014
Correction de continuité du $\chi^2$	5.04	1	0.025
N	100		

## Chi<sup>2</sup> Corrigé de Yates

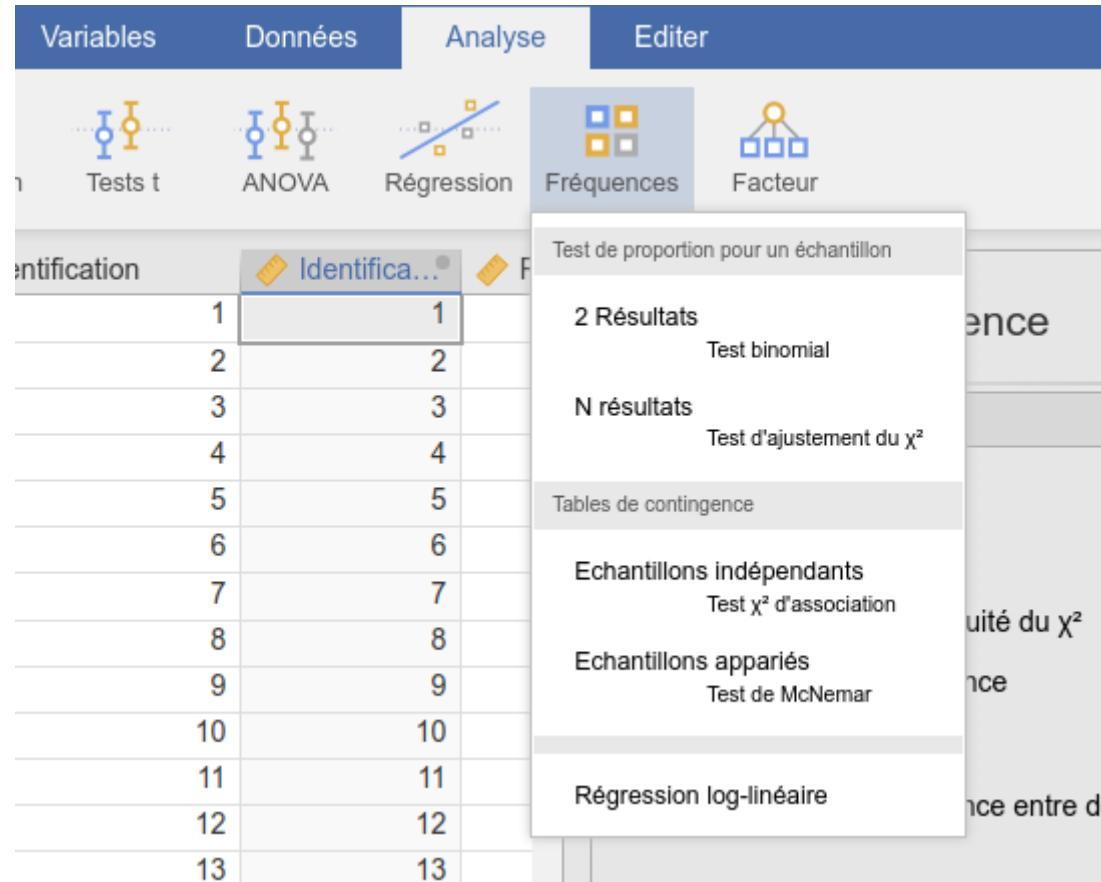
L'approximation de la statistique du Chi<sup>2</sup> pour les petits tableaux 2 x 2 peut être améliorée en réduisant de 0,5 la valeur absolue des différences entre les effectifs théoriques et les effectifs observés avant de l'élever au carré (correction de Yates). Cette correction, qui rend l'estimation plus conservatrice est souvent appliquée lorsque la table ne contient que peu d'effectifs observés, et que les effectifs théoriques deviennent inférieurs à 10 (pour une présentation plus approfondie de cette correction, voir Conover, 1974 ; Everitt, 1977 ; Hays, 1988; Kendall & Stuart, 1979 et Mantel, 1974).

<https://www.statsoft.fr/concepts-statistiques/glossaire/c/chi-corrigé.html>

# Le test chi-deux d'indépendance (entre deux variables)

## 1. Analyse du croisement des variables Sexe et Revenu\_médiane sur les données des 1000 comptes de Twitter.

- Ouvrir le fichier *Données\_Twitter\_1000.sav*
- Lancer l'analyse avec Jamovi : menu Fréquences / Echantillons indépendants



# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les effectifs

# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les totaux par ligne et par colonne

# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les pourcentages  
par ligne

# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les pourcentages  
par colonne

# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les pourcentages  
par colonne

# Le test chi-deux d'indépendance (entre deux variables)

Tables de contingence

		Revenu_médiane		
Sexe		Rev_sup	Rev_inf	Total
f	Observé	214	222	436
	% par ligne	49.1 %	50.9 %	100.0 %
	% par colonne	49.1 %	50.1 %	49.6 %
	% du total	24.3 %	25.3 %	49.6 %
h	Observé	222	221	443
	% par ligne	50.1 %	49.9 %	100.0 %
	% par colonne	50.9 %	49.9 %	50.4 %
	% du total	25.3 %	25.1 %	50.4 %
Total	Observé	436	443	879
	% par ligne	49.6 %	50.4 %	100.0 %
	% par colonne	100.0 %	100.0 %	100.0 %
	% du total	49.6 %	50.4 %	100.0 %

Les pourcentages  
par colonne

# Le test chi-deux d'indépendance (entre deux variables)

Tests  $\chi^2$

	Valeur	ddl	p
$\chi^2$	0.0933	1	0.760
Correction de continuité du $\chi^2$	0.0566	1	0.812
Test exact de Fisher			0.787
N	879		

Si tableaux  $2 \times 2$  et effectifs réduits ( $<10$ ) utiliser le chi-deux corrigé

Si 25 % (ou plus) des fréquences théoriques est inférieure à 5, utiliser le Test exact de Fisher.

Que conclure de ce test ?

# Le test chi-deux d'indépendance (entre deux variables)

Tests  $\chi^2$

Le chi-deux vaut 0.0933

	Valeur	ddl	p
$\chi^2$	0.0933	1	0.760
Correction de continuité du $\chi^2$	0.0566	1	0.812
Test exact de Fisher			0.787
N	879		

La p-value associé vaut 76 % (0.76)

**Pas de lien significatif entre les deux variables Sexe et Revenu\_Mediane**

## Rédaction possible

« Pour les compte de Twitter, on n'observe pas de relation significative entre le fait d'avoir un revenu supérieur ou inférieur à la médiane et celui d'être une femme ou un homme (Chi-carré = 0.093 (1ddl), p = 0.76, N=879) »

# Le test chi-deux d'indépendance (entre deux variables)

Tests  $\chi^2$

	Valeur	ddl	p
$\chi^2$	6.00	1	0.014
Correction de continuité du $\chi^2$	5.04	1	0.025
Ratio de vraisemblance	6.07	1	0.014
N	100		

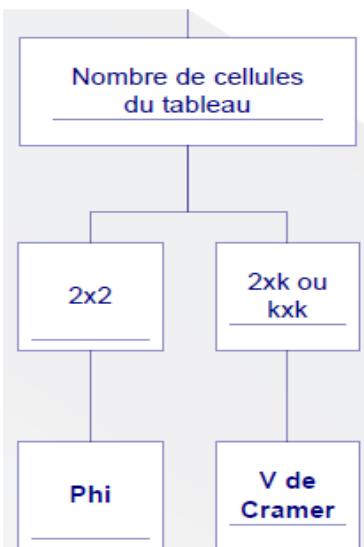
significativité

Nominal

	Valeur
Coefficient de contingence	0.238
Coefficient Phi	0.245
V de Cramer	0.245

Taille d'effet  
(force du  
lien)

Choix  
entre Phi  
et V



Dans le cadre du cours, nous utiliserons 2 indicateurs de force pour ce type de test. Les 2 varient de 0 à 1 (plus ça tend vers 1 plus le lien est fort entre les 2 variables)

- V. de Cramer
- Phi

Valeur	Force du lien statistique
0	Absence de relation
Entre 0,05 et 0,10	Très faible
Entre 0,10 et 0,20	Faible
Entre 0,20 et 0,40	Modérée
Entre 0,40 et 0,80	Forte
Entre 0,80 et 1	Louche (Colinéarité)



En bref, pas de relation significative entre le fait d'être parmi la moitié la plus pauvre ou la plus riche des 1000 utilisateurs des comptes Twitter et le fait d'être une femme ou un homme. Et même si le lien était significatif, il serait « très faible ».

# Le test chi-deux d'indépendance (entre deux variables)

- Autre utilisation du test : mesure d'association dans la cooccurrence

	Présence Mot2	Absence Mot2	Total
Présence Mot1	$f_{12}$	$f_1 - f_{12}$	$f_1$
Absence Mot1	$f_2 - f_{12}$	$N - f_1 - f_2 + f_{12}$	$N - f_1$
Total	$f_2$	$N - f_2$	$N$

- $f_{12}$  : co-fréquence (nombre d'événements où mot1 et mot2 cooccurrent)
- $f_1$  (resp.  $f_2$ ) : fréquence de mot 1 (resp. de mot2)
- $N$  : nombre total d'événements (p.ex. nombre d'empans, nombre de phrases, nombre de relations syntaxiques, etc.)

# Le test chi-deux d'indépendance (entre deux variables)

- Chercher sur le Lexicoscope, dans le corpus PhraseoAnticipation, les statistiques de cooccurrences de *délit* avec *un*, et *délit* avec *flagrant*
- Remplir les deux tableaux de contingence, et faire un test de chi2

	Présence <i>délit</i>	Absence <i>délit</i>	Total
Présence <i>flagrant</i>			
Absence <i>flagrant</i>			
Total			
Qu'en concluez-vous ?			

	Présence <i>délit</i>	Absence <i>délit</i>	Total
Présence <i>un</i>			
Absence <i>un</i>			
Total			

# Le test chi-deux d'indépendance (entre deux variables)

- D'autres tests existent pour ce type d'association :
  - rapport de vraisemblance ( $\rightarrow$  log likelihood ratio)
  - z-score
  - t-score
  - information mutuelle spécifique
  - odd ratio
  - ...
- Ces tests ont des caractéristiques particulières :
  - certains mettent l'accent sur l'association des événements fréquents (rapport de vraisemblance)
  - d'autres sur les événements rares (information mutuelle spécifique)
- Certains tests sont paramétriques (t-score, z-score) et d'autres non (chi2, loglike, etc.)

# Comparaison d'une répartition observée avec une répartition de référence : le chi-deux de conformité (entre une variable et une constante)

Question de recherche : la répartition des âges dans notre échantillon Twitter est-elle la même que dans la population française ?

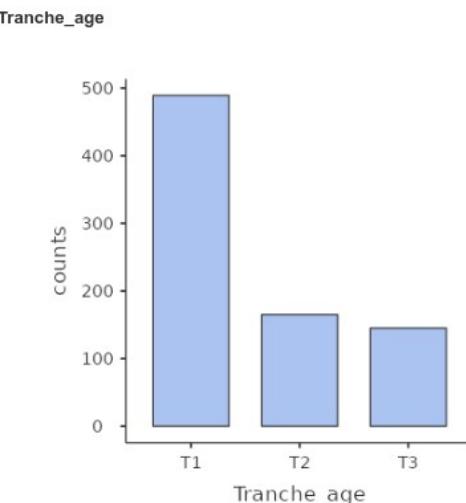
## Dans notre échantillon

(Analyses > Exploration > Statistiques descriptives)

### Fréquences

Fréquences de Tranche_age			
Tranche_age	Quantités	% du Total	% cumulés
T1	489	61.2 %	61.2 %
T2	165	20.7 %	81.9 %
T3	145	18.1 %	100.0 %

### Graphes



### Avertissement :

Une version provisoire de ce chiffre-clé a été diffusée le 19 janvier 2021. Il a été actualisé le 29 mars 2021 à partir des données d'état civil de l'ensemble de l'année 2020.

Effectifs  Proportions

! TÉLÉCHARGER  
(XLSX, 6 Ko)

### Population par sexe et groupe d'âges en 2021 : effectifs

Groupe d'âges	Femmes	Hommes	Ensemble
Moins de 15 ans	5 856 900	6 105 989	11 962 889
15-19 ans	2 024 642	2 136 473	4 161 115
20-24 ans	1 883 418	1 943 616	3 827 034
25-29 ans	1 869 943	1 829 074	3 699 017
30-34 ans	2 089 420	1 976 672	4 066 092
35-39 ans	2 167 585	2 045 970	4 213 555
40-44 ans	2 109 573	2 030 367	4 139 940
45-49 ans	2 241 600	2 191 233	4 432 833
50-54 ans	2 248 475	2 175 511	4 423 986
55-59 ans	2 255 010	2 135 166	4 390 176
60-64 ans	2 157 194	1 968 499	4 125 693
65-69 ans	2 058 594	1 816 440	3 875 034
70-74 ans	1 977 748	1 702 561	3 680 309
75 ans ou plus	3 906 839	2 502 729	6 409 568
Ensemble	34 846 941	32 560 300	67 407 241



Téléchargement  
du tableau dans  
Excel/Calc

205

### Dans la population française

<https://www.insee.fr/fr/statistiques/2381474>

# Comparaison d'une répartition observée avec une répartition de référence : le chi-deux de conformité (entre une variable et une constante)

Question de recherche : la répartition des âges dans notre échantillon Twitter est-elle la même que dans la population française ?

Population par sexe et groupe d'âges en 2021 : effectifs

Groupe d'âges	Femmes	Hommes	Ensemble	
Moins de 15 ans	5 856 900	6 105 989	11 962 889	
15-19 ans	2 024 642	2 136 473	4 161 115	7 988 149
20-24 ans	1 883 418	1 943 616	3 827 034	
25-29 ans	1 869 943	1 829 074	3 699 017	7 765 109
30-34 ans	2 089 420	1 976 672	4 066 092	
35-39 ans	2 167 585	2 045 970	4 213 555	
40-44 ans	2 109 573	2 030 367	4 139 940	
45-49 ans	2 241 600	2 191 233	4 432 833	
50-54 ans	2 248 475	2 175 511	4 423 986	39 691 094
55-59 ans	2 255 010	2 135 166	4 390 176	
60-64 ans	2 157 194	1 968 499	4 125 693	
65-69 ans	2 058 594	1 816 440	3 875 034	
70-74 ans	1 977 748	1 702 561	3 680 309	
75 ans ou plus	3 906 839	2 502 729	6 409 568	

ATTENTION OUBLI DES MOINS DE 15 ANS

Dans Excel, on fait trois additions pour retomber sur les 3 catégories qu'on a distinguées dans Twitter

- Inf. à 25
- De 25 à 34
- Sup. à 34

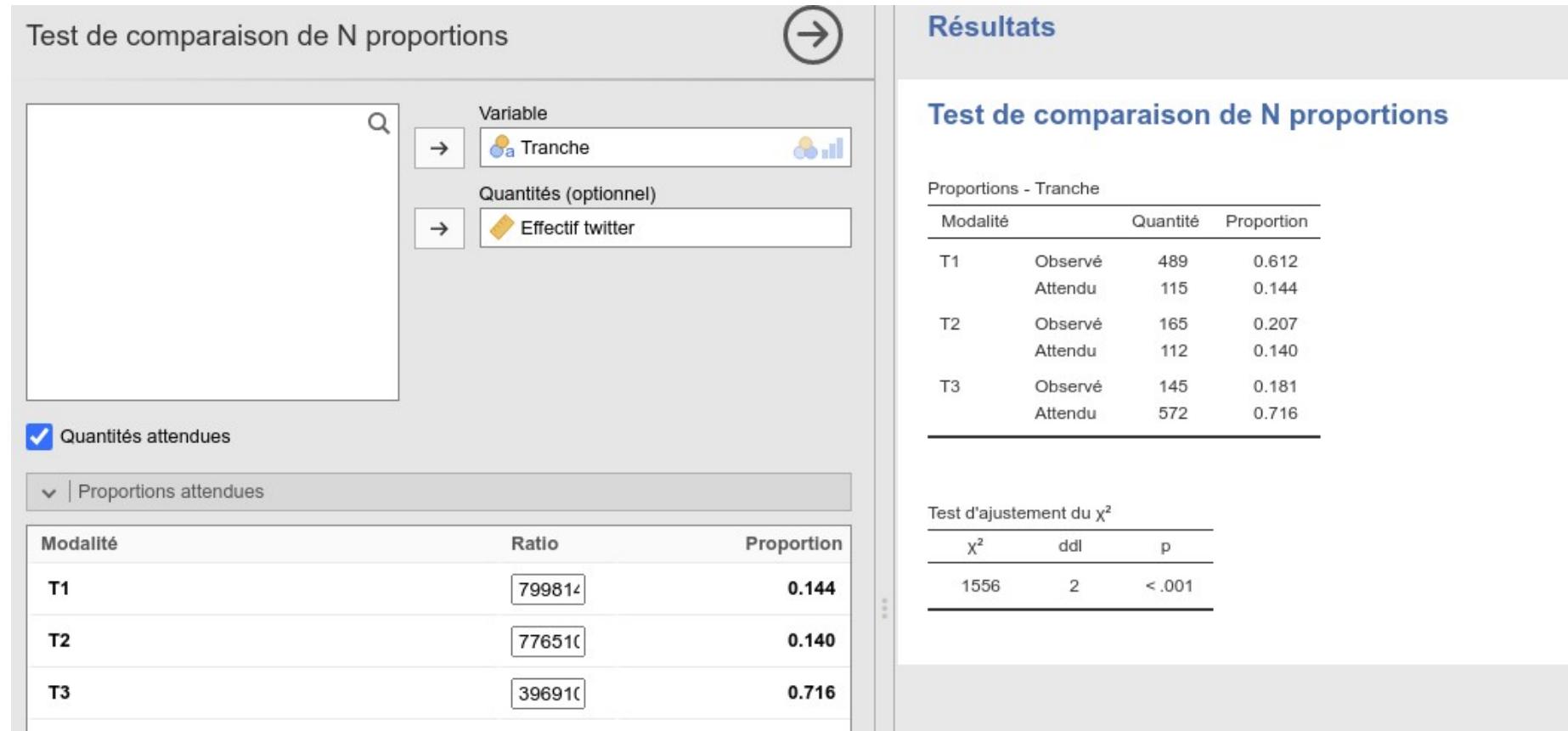
Maintenant on peut mettre les deux distributions dans le même tableau et les comparer pour répondre à la question.

	Inf. à 25	De 25 à 34	Sup. à 34
Twitter	489	165	145
Insee	7 998 149	7 765 109	39 691 094

Pensez-vous à vue de nez que la répartition des âges est la même sur Twitter et dans la population française ?

# Comparaison d'une répartition observée avec une répartition de référence : le chi-deux de conformité (entre une variable et une constante)

Question de recherche : la répartition des âges dans notre échantillon Twitter est-elle la même que dans la population française ?



Le test confirme ce que l'intuition nous indiquait : la répartition n'est pas du tout conforme à la population prise dans son ensemble ( $\chi^2=1556$ ,  $ddl=2$ ,  $p<0.01$ )

## Comparaison d'une répartition observée avec une répartition de référence: le chi-deux de conformité (entre une variable et une constante)

### Exemple de rédaction

« *La répartition des utilisateurs de Twitter dans les tranches d'âge est significativement différente de la répartition de la population française, telle que décrite dans les données INSEE (lien en bas de page) (chi-deux = 1556, ddl = 2, p < 0.001). Plus précisément, les utilisateurs de moins de 25 ans sont les plus nombreux sur Twitter, alors que ce sont les personnes de plus de 34 ans qui sont majoritaires dans la population.* »

# Synthèse

209

209

# Notions importantes du cours

- Notion de variable, nature des variables
- Variable explicative et variable expliquée
- Paramètres de tendance centrale vs paramètres de dispersion
- Différence statistique descriptives vs inférentielles
- $P_0$  et  $P_1$
- Notion de significativité et valeur p (risque alpha, risque bêta)
- Distribution d'échantillonnage, erreur standard
- Coefficient de variation
- Distribution normale et tests de normalité
- Différence entre significativité et taille d'effet
- Les différents tests et conditions d'application
  - test de Student et tests alternatifs
  - test du Chi2 et tests alternatifs
  - test de Pearson