TD de Statistiques - Séance N°1
Statistiques Descriptives

Bibliographie

Blöss et Grosseti, Introduction aux méthodes statistiques en sociologie, PUF, Coll. Le Sociologue
P. Rateau, Méthode et statistique expérimentales en sciences humaines, Ellipses

Pourquoi faut-il étudier les statistiques ?

Les statistiques sont-elles utiles au sociologue ?
Les statistiques, il y a des calculatrices et des logiciels pour faire cela. Oui, mais ...

1 Introduction - Vocabulaire

On collecte des données. Sur qui ou sur quoi ? A propos de quoi ?

Sur qui ?

L'ensemble des "objets" étudié dans le cadre d'une étude statistique est appelé *population*. Souvent, la population est trop nombreuse pour être étudiée de façon exhaustive. On recueille alors des données relatives à un *échantillon* tiré de cette population.

Chaque objet de la population est appelé *individu statistique*, *unité statistique* ou *sujet*. Les logiciels utilisent souvent le terme *observation*.

A propos de quoi ?

On étudie une ou plusieurs caractéristiques présentes chez tous les individus de la population (données homogènes). Chacune des caractéristiques étudiées est appelée *attribut*, *caractère* ou *variable statistique*.

Exemple 1. Sur une population constituée par les élèves d'une classe, on peut étudier le genre (2 valeurs possibles), l'âge, le score obtenu à un test d'aptitude, etc.

Exemple 2. Dans le cadre d'une enquête, on note si le sujet interrogé habite en zone urbaine, en zone péri-urbaine ou en zone rurale : cela définit une variable (le lieu de résidence) avec trois valeurs possibles.

L'ensemble des valeurs prises par une variable (ou susceptibles de l'être) sont appelées *modalités* de cette variable.

Les modalités d'une variable doivent être *exhaustives* (la variable est définie pour chaque individu de la population) et *exclusives* (la variable ne possède qu'une seule valeur sur chaque individu).

Exemples de "variables" mal définies :
- On s'intéresse à l'âge des sujets au moment du mariage. Que fait-on des célibataires ? Des personnes remariées ?

1.1 Nature d'une variable statistique

Des variables telles que le genre, la couleur des yeux, la CSP sont des *variables nominales* ou *variables qualitatives*. Leur domaine de variation est une *échelle nominale*.

Des variables telles que le degré d'accord avec une affirmation sont des *variables ordinales*. Leur domaine de variation est une *échelle ordinale*. Par exemple, les échelles de Likert sont des échelles ordinales.

Des variables telles que le nombre d'enfants ou le score à un test sont des variables numériques. On distingue les variables numériques discrètes (le nombre d'enfants par exemple) et les variables numériques continues (la taille, le poids, le score à un test, etc).
Lorsqu'une variable est regroupée en classes, son domaine de variation constitue une échelle d'intervalles (on dit aussi échelle de rapports lorsque la grandeur est

2 Recueil et présentation d'un ensemble de données

Individus statistiques : $s_1, s_2, \ldots, s_n$
Variables étudiées : X, Y, ...

2.1 Tableau protocole :

Un tableau protocole croise les individus statistiques (lignes) et les variables étudiées (colonnes).
Exemple :

	Age	Genre	Cat. Prof.
s1	32	H	Ouvrier
s2	25	F	Cadre
s3	56	F	Prof. Interm.

2.2 Recensement ou tri à plat : tableau d'effectifs

Tableau indiquant pour chaque modalité, l'effectif correspondant (et éventuellement la fréquence).
Exemple :

CSP	Effectifs	Fréquences
0.Agric.	38	0,0406 ou 4,06%
1.Employés	45	0,0480 ou 4,80%
2.Patrons	75	0,0800 ou 8,00%
3.Cadres sup.	280	0,2988 ou 29,88%
...	...	
Total	937	1 ou 100%

Dans le cas d'une variable numérique comportant de nombreuses modalités, on peut procéder à un regroupement en classes ; par exemple : tailles de 40 étudiants

Classe	Effectifs
[155, 166[	8
[166, 170[	9
[170, 172[	7
[172, 176[	9
176, 190]	7
Total	40

2.3 Tableau de données chronologiques

On introduit une variable supplémentaire : le temps. Par exemple, on étudie à deux instants différents une même variable, définie sur deux populations caractérisées de façon analogue. Par exemple, CSP des parents sur une population d'étudiants de 1ère année en 1962 et 1970.

CSP	1962	1990
Agric.	11	38
Sal. agric.	1	5
Patrons	37	75
Cadres. sup.	60	280
...	...	...
Total	197	894

L'*écart absolu* entre les effectifs des deux séries est (Effectif 1990) - (Effectif 1962).
L'*écart relatif*, ou *variation relative* est défini par :

$$\text{Ecart relatif} = \frac{\text{Effectif 1990} - \text{Effectif 1962}}{\text{Effectif 1962}}$$

Le *coefficient multiplicateur* est défini par :

$$\text{Coefficient multiplicateur} = 1 + \text{écart relatif}$$

Une grandeur qui augmente de z% est en fait multipliée par 1+z%.

Exemple :

CSP	1962	1990	Ecart absolu	Ecart relatif	Coef. multiplic.
Agric.	11	38	27	245%	3,45
Sal. agric.	1	5	4	400%	5
Patrons	37	75	38	103%	2,03
Cadres. sup.	60	280	220	367%	4,67
...	...	...	...		
Total	197	894	697	354%	4,54

Variations successives et taux de variation moyen.

Exemple : Le nombre de divorces a augmenté de la façon suivante :

- de 1980 à 1981 : 8,1%
- de 1981 à 1982 : 7,2%
- de 1982 à 1983 : 5,1%
- de 1983 à 1984 : 5,5%.

Quel est l'écart relatif entre 1980 et 1984 ?

On calcule le coefficient multiplicateur entre 1980 et 1984 :

$$C = 1{,}081 \times 1{,}072 \times 1{,}051 \times 1{,}055 = 1{,}285$$

L'écart relatif entre 1980 et 1984 est de 28,5%.

Taux de variation moyen.

On reprend l'exemple ci-dessus. Quel est le taux d'accroissement annuel moyen du nombre d'étudiants issus de la CSP Agriculteurs entre 1962 et 1990 ?

Coefficient d'accroissement entre 1962 et 1990 : 3,45

Coefficient annuel moyen entre 1962 et 1990 : $3{,}45^{1/28} = 1{,}045$

Taux moyen d'augmentation : 4,5%

2.4 Etude conjointe de deux variables ou tri croisé : tableau de contingence

On étudie conjointement deux variables nominales. On peut résumer les données à l'aide d'un tableau de contingence : les en-têtes de lignes et de colonnes sont les modalités des deux variables étudiées. Le corps du tableau donne l'effectif de chaque combinaison de modalités des deux variables.

Exemple : Genre et préférence pour un type de spectacle.

	Hommes	Femmes	Total
Comédie	90	150	240
Drame	50	90	140
Variétés	160	160	320
Total	300	400	700

Deux (voire trois) manières de calculer des fréquences :

Par ligne :

	Hommes	Femmes	Total
Comédie	37,5%	62,5%	100%
Drame	35,7%	64,3%	100%
Variétés	50%	50%	100%
Total	42,9%	57,1%	100%

Par colonne :

	Hommes	Femmes	Total
Comédie	30,0%	37,5%	34,3%
Drame	16,7%	22,5%	20,0%
Variétés	53,3%	40,0%	45,7%
Total	100%	100%	100%

Sur le total :

	Hommes	Femmes	Total
Comédie	12,9%	21,4%	34,3%
Drame	7,1%	12,9%	20%
Variétés	22,9%	22,9%	45,7%
Total	42,9%	57,1%	100%

Lequel de ces tableaux faut-il utiliser pour répondre à des questions telles que :
"Les hommes marquent-ils, plus que les femmes, une préférence pour les variétés ?"
"Un spectacle de variétés est organisé. A quel public doit-on s'attendre ?"

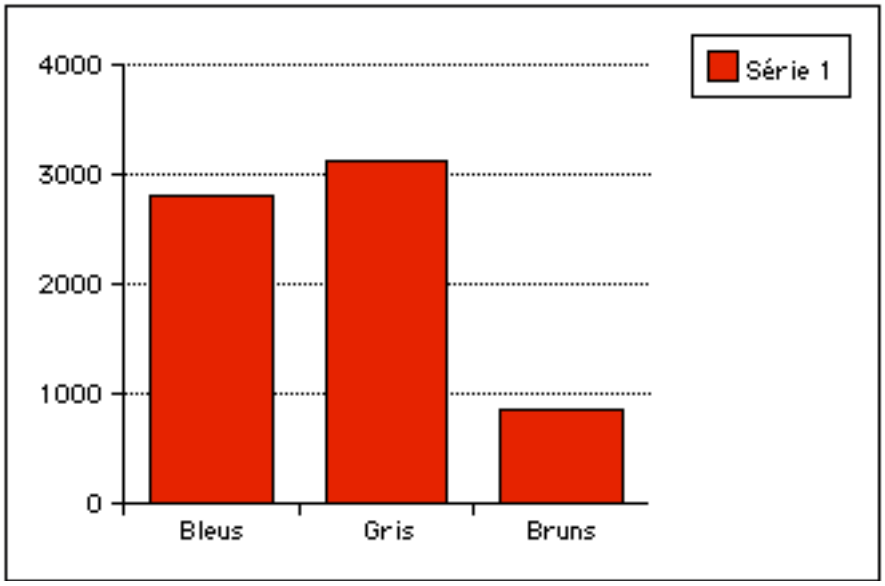
3 Représentations graphiques

Une variable nominale peut être représentée par un *diagramme à bandes* ou un *diagramme circulaire* (ou semi-circulaire). Exemple : couleur des yeux de 6800 sujets.

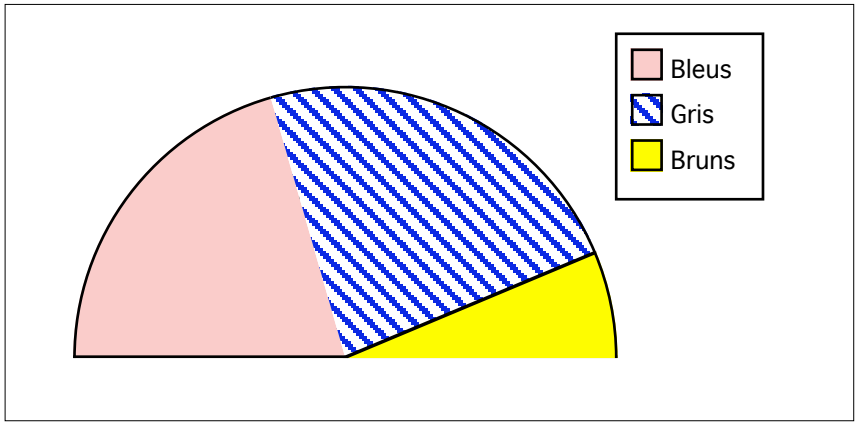
Modalités	Effectifs	Fréquences	Fréq en %

Bleus	2811	0,41	41%
Gris	3132	0,46	46%
Bruns	857	0,13	13%
Total	6800	1	100%

Diagrammes à bandes



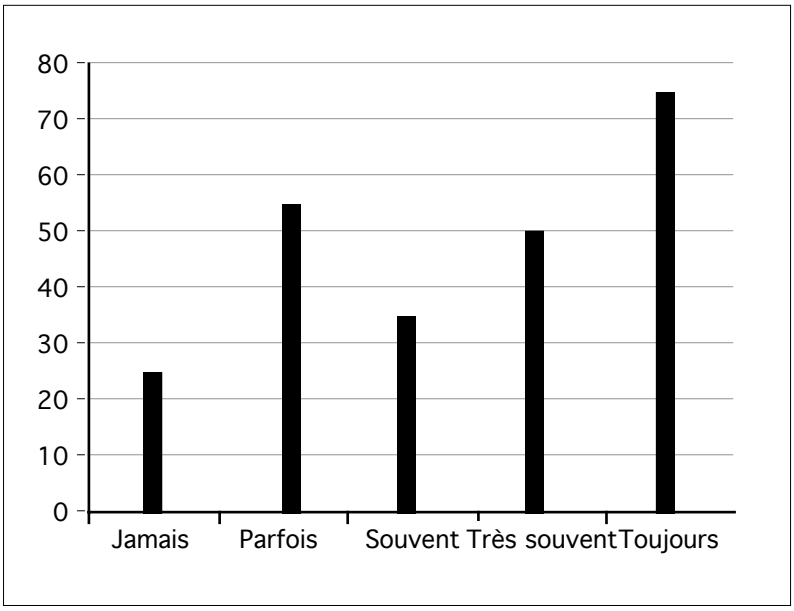
Diagrammes circulaires ou semi-circulaires



Une variable ordinale, ou une variable numérique discrète peut être représentée à l'aide d'un *diagramme en bâtons*.

Exemple : On a demandé à 240 sujets s'ils fermaient à clef la porte de leur appartement.

	Jamais	Parfois	Souvent	Très souvent	Toujours
Effectifs	25	55	35	50	75

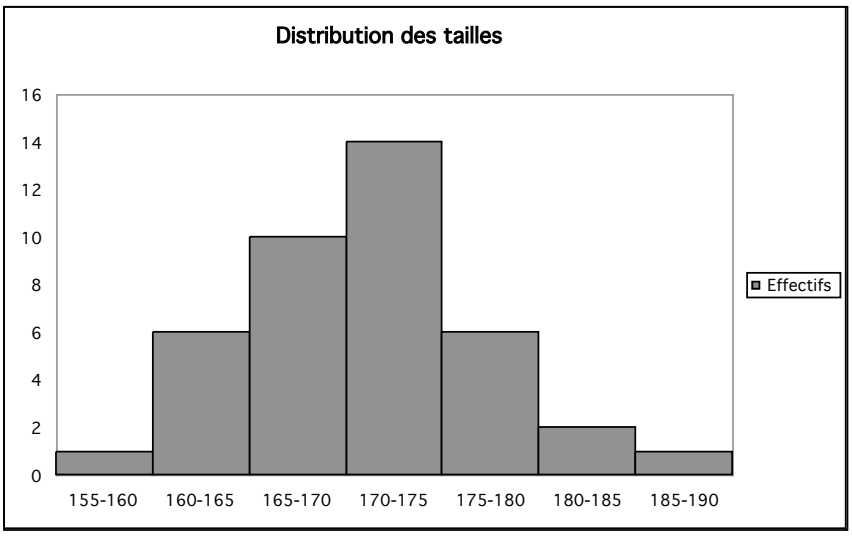


Une variable numérique regroupée en classes est généralement représentée à l'aide d'un *histogramme*.

Un histogramme est formé de rectangles adjacents :
- dont la base est proportionnelle à l'amplitude de la classe
- dont l'aire est proportionnelle à l'effectif de la classe

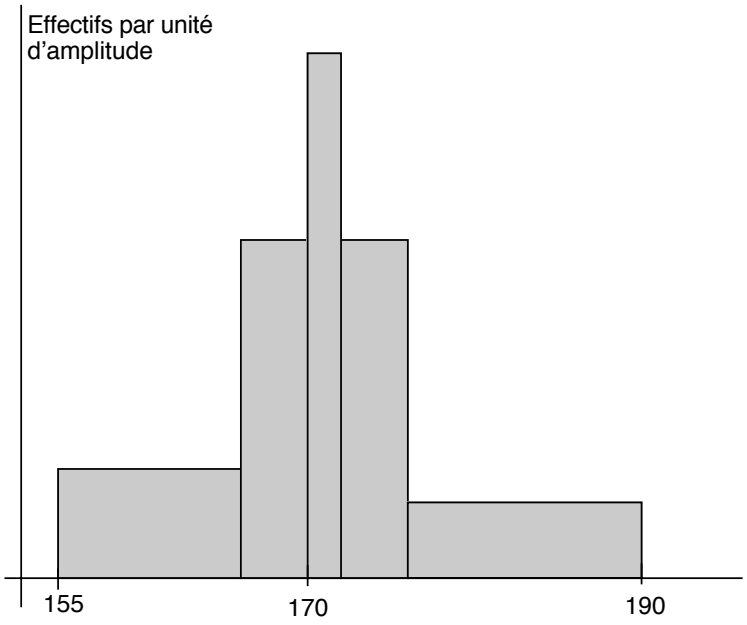
Classes de même amplitude : les hauteurs des rectangles sont alors proportionnelles aux effectifs.

Classe	Eff.
[155,160[	1
[160,165[	6
[165,170[	10
[170,175[	14
[175,180[	6
[180,185[	2
[185,190[	1



Classes d'amplitudes différentes : les hauteurs des rectangles représentent les effectifs par unité d'amplitude ou *densités* des différentes classes :

Classe	Eff.	Amplitude	Densité
[155,166[	8	11	0.73
[166,170[	9	4	2.25
[170,172[	7	2	3.5
[172,176[	9	4	2.25
[176,190]	7	14	0.5



4 Caractéristiques de position

4.1 Mode, classe modale

Mode d'une série statistique (nominale, ordinale ou numérique) : modalité correspondant à l'effectif le plus élevé.
N.B. Une série statistique peut admettre plusieurs modes.

Classe modale d'une série statistique regroupée en classes : classe qui a la plus forte densité.
N.B. : c'est la classe correspondant au rectangle de hauteur maximale dans l'histogramme.

4.2 Médiane

Variable ordinale ou numérique comportant *N* observations.

Les individus sont classés par valeurs croissantes de la variable. La médiane est la valeur du caractère observée sur l'individu "médian", à savoir :

- Si *N* est impair, la médiane est la modalité observée sur l'individu de rang $\frac{N+1}{2}$.

- Si N est pair et si le caractère est numérique, la médiane estla moyenne des modalités observées sur les individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$.

4.3 Moyenne arithmétique

On considère une variable numérique.

- Calcul à partir d'un tableau protocole : une variable X, de valeurs (x_i) est définie sur une population d'effectif total N. La moyenne est donnée par :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Calcul à partir d'un tableau d'effectifs : une variable X, de modalités $a_1, a_2, ..., a_k$, avec les effectifs $n_1, n_2, ... n_k$ est définie sur une population d'effectif total N. La moyenne est donnée par :

$$\mu = \frac{1}{N}\sum_{i=1}^{k} n_i a_i$$

Remarque : la moyenne peut aussi être calculée à l'aide des fréquences (f_i) :

$$\mu = \sum_{i=1}^{k} f_i a_i$$

Exemple :

Mod. a_i	Effect. n_i	$n_i\, a_i$
0	5	0
1	18	18
2	32	64
3	29	87
4	14	56
5	2	10
Total	100	235

On obtient : $\mu = 2{,}35$.

Remarque. Cas d'une variable répartie en classes : on considère que la masse de chaque classe est concentrée au centre $c_i = \frac{a_i + a_{i+1}}{2}$ de la classe.

5 Caractéristiques de dispersion

5.1 Etendue

$x_1, x_2, ... , x_n$: valeurs observées d'une variable statistique numérique X.

$x_{max} = \text{Max}(x_1, x_2, ... , x_n)$
$x_{min} = \text{Min}(x_1, x_2, ... , x_n)$

L'étendue de la variable est :

$$e = x_{max} - x_{min}$$

5.1.1 Quartiles

Soit une série statistique numérique de médiane M.

Premier quartile Q_1 : médiane de la série des observations strictement inférieures à M.

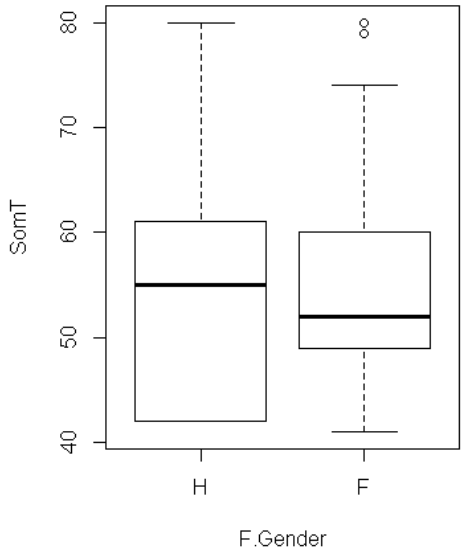
Deuxième quartile Q_2 : médiane M de la série complète.

Troisième quartile Q_3 : médiane de la série des observations strictement supérieures à M.

L'*écart interquartile* est défini par :

$$I_q = Q_3 - Q_1$$

Représentation graphique permettant de visualiser l'étendue et les quartiles : boîte à moustaches.



Généralisation : déciles, centiles...

5.1.2 Variance et écart type

Définition : La variance est la moyenne des carrés des écarts à la moyenne.

- Calcul à partir d'un tableau protocole

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Calcul à partir d'un tableau d'effectifs

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{k} n_i(a_i - \mu)^2$$

L'écart type est donné par : $\sigma = \sqrt{\text{Var}}$.

Calcul pratique

"Moyenne des carrés moins carré de la moyenne"

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

$$\text{Var} = \frac{1}{N}\sum_{i=1}^{k} n_i a_i^2 - \mu^2$$

Remarques
Cas d'une variable répartie en classes : utiliser les centres de classes.
Unités, effet d'un changement d'origine ou d'unités.

Organisation des calculs

Mod.	Effect.	$n_i\, a_i$	$n_i a_i^2$
0	5	0	0
1	18	18	18
2	32	64	128
3	29	87	261
4	14	56	224
5	2	10	50
Total	100	235	681

On obtient : $\mu = 2{,}35$; $\text{Var} = 6{,}81 - 2{,}35^2 = 1{,}29$; $\sigma = 1{,}13$.