Chapitre 11

Statistiques

11.1 Vocabulaire

La statistique descriptive est un moyen d'analyser certaines caractéristiques d'une population, appelées *caractères* ou *variables statistiques.*

Les éléments de la population sont appelés *individus.*

On distingue deux types de caractères :

Définition 11.1.1
On appelle variable quantitative toute variable qui prend des valeurs numériques ; on distingue

- *les variables quantitatives discrètes,*
- *les variables quantitatives continues,*

On appelle variable qualitative toutes les autres variables ; on distingue

- *les variables qualitatives nominales,*
- *les variables qualitatives ordinales,*

EXEMPLE

On peut prendre les exemples suivants :

- Variable quantitative discrète :
- Variable quantitative continue :
- Variable qualitative nominale :
- Variable quantitative ordinale :

Définition 11.1.2
On considère un caractère C. On appelle effectif *de C*

On appelle effectif total

On appelle fréquence *d'un caractère*

Si les caractères peuvent être ordonnés, on appelle effectif cumulé croissant de C

EXERCICE

Remplir le tableau suivant :

Mois	Jan	Fev	Mar	Avr	Mai	Jui	Jul	Aou	Sep	Oct	Nov	Dec
Effectif												
Fréquence												
ECC												

Dans le cas de variables quantitatives, on peut regrouper certains caractères pour former des classes.

EXERCICE

Remplir le tableau suivant :

Taille	$[130, 150]$	$]150, 170]$	$]170, 180]$	$]180, 200]$
Effectif				
Fréquence				

Pour représenter une série statistique, on peut utiliser plusieurs moyens.

Diagramme en bâtons Pour chaque caractère, on trace un rectangle fin de hauteur l'effectif ou la fréquence du caractère. On peut utiliser ce type de diagramme pour des valeurs qualitatives.

Pour la série des mois de naissances, on a donc le diagramme suivant :

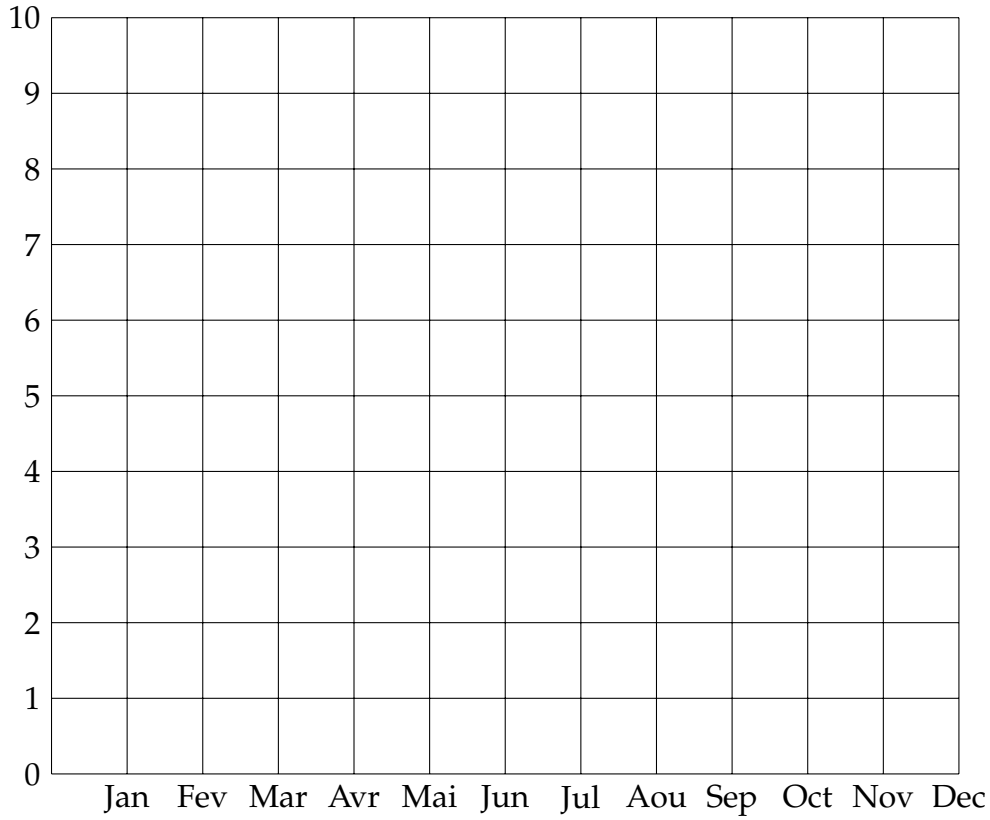
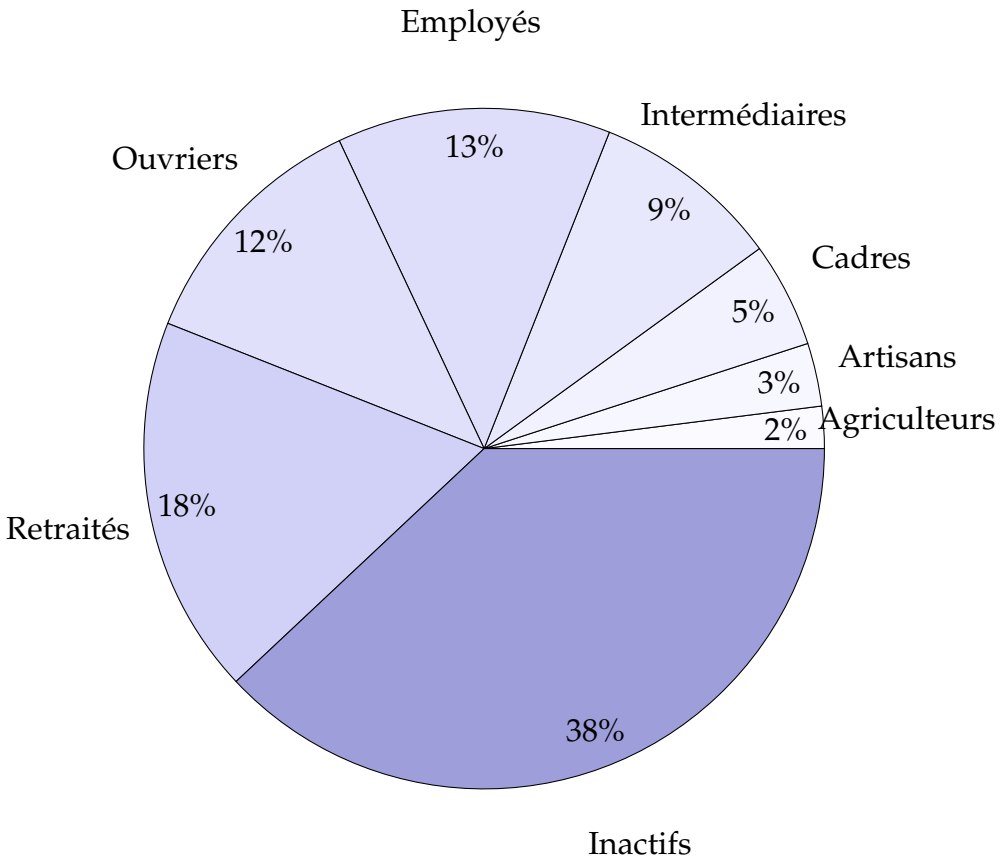


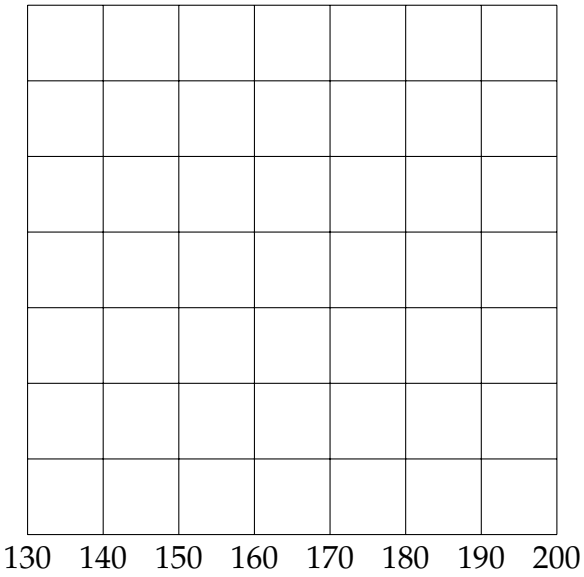
Diagramme circulaire Dans ce type de diagramme, chaque part du cercle a un angle proportionnel à la fréquence du caractère. On peut utiliser ce type de diagrammes pour représenter des variables qualitatives.

Voici par exemple la répartition de la population française en 1999.



Histogramme On utilise ce type de diagramme pour représenter des données regroupées par classes. L'aire de chaque rectangle doit être proportionnelle à la fréquence de chaque classe. On peut utiliser la formule

Pour le tableau des tailles :



11.2 Paramètres d'une série statistique

On distingue deux types de paramètres : les paramètres de position et les paramètres de dispersion.

Définition 11.2.1
On appelle mode *d'une série statistique*
.

Dans le cas de variables continues, on parle de

EXEMPLE

Considérons les mois de naissance : le mode est .
Pour les tailles, la classe modale est .

Définition 11.2.2
Pour une variable quantitative, la moyenne d'une série statistique $(x_1, \ldots, x_N)$ *avec effectifs* $(n_1, \ldots, n_N)$ *et fréquences* $(f_1, \ldots, f_N)$ *est donnée par, si* $n = \sum n_i$

$$\overline{x} =$$

Dans le cas de regroupement par classe, on prend pour x_i

EXEMPLE

Pour les deux séries de l'exemple précédent, les moyennes sont .

Proposition 11.2.3
La moyenne est linéaire : si $(x_1, \dots, x_N)$ est une série statistique avec effectifs $(n_1, \dots, l_N)$, alors la moyenne de $y = (ax_1 + b, \dots, ax_N + b)$ avec mêmes effectifs est

$$\overline{y} =$$

Proposition 11.2.4
Si S_1 et S_2 sont deux séries statistiques d'effectifs totaux N_1 et N_2, alors la moyenne de $S = S_1 \cup S_2$ est la moyenne de $(\overline{S_1}, \overline{S_2})$ d'effectifs (N_1, N_2) :

$$\overline{S} =$$

EXEMPLE

Dans une ville A de 100 000 habitants, les habitants ont en moyenne 30 ans, et dans une ville B de 80 000 habitants, les habitants ont en moyenne 28 ans.
Alors dans l'ensemble des deux villes, l'âge moyen est de

Définition 11.2.5
On appelle médiane d'une série statistique toute valeur m

MÉTHODE :

Pour calculer une médiane, on commence par ordonner la série dans l'ordre croissant.
Ensuite :

- s'il y a un nombre impair de valeurs,
- s'il y a un nombre pair de valeurs,

EXERCICE

Dans un bar, il y a 99 personnes, ayant tous 10 000 euros sur leur compte. Elon Musk, dont la fortune est estimée à 20 milliards d'euros, entre dans le bar.
Comparer la fortune moyenne, et la fortune médiane dans ce bar.

De la même façon, on peut choisir de séparer la série en plus petits intervalles.

Définition 11.2.6
Soit une série statistique.

- *Le premier quartile est*
- *Le troisième quartile est*
- *Le premier décile est*
- *Le neuvième décile est*

On peut aussi définir les quintiles, les percentiles, *etc.*

Passons maintenant aux paramètres de dispersion.

Définition 11.2.7
Soit $x = (x_1, \dots, x_N)$ une série statistique d'effectifs $(n_1, \dots, n_N)$. La variance de x est définie par, si $n = \sum n_i$

$$\mathrm{Var}(x) =$$

L'écart-type est donné par

$$\sigma_x =$$

Proposition 11.2.8 : Théorème de König [*]-Huygens [†]
On a

$$\mathrm{Var}(x) =$$

Démonstration. Il suffit de calculer :

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{k=1}^{N} n_k (x_k - \overline{x})^2$$

$$=$$

$$=$$

$$=$$

□

∗. Johann Samuel König, 1712 – 1757, Allemand
†. Christian Huygens, 1629 – 1695, Néerlandais

EXERCICE

Calculer la variance et l'écart-type des deux séries précédentes.

Proposition 11.2.9
Soient $x = (x_1, \dots, x_N)$ une série d'effectifs $(n_1, \dots, n_N)$, et $y = ax + b$ de mêmes effectifs. Alors

$$\mathrm{Var}(y) =$$

NOTA

Attention, la variance n'est donc pas linéaire. En particulier, $\mathrm{Var}(x + y) \neq \mathrm{Var}(x) + \mathrm{Var}(y)$.

Définition 11.2.10
On appelle étendue *d'une série statistique*

On appelle intervalle interquartile *(resp. interdécile) l'intervalle (resp.), et* distance interquartile *(resp. distance interdécile) le nombre (resp.).*

11.3 Statistiques bivariées

Définition 11.3.1
Une série statistique double est une série statistique portant sur deux variables d'une même population. Les données sont donc des couples de valeurs.

EXEMPLE

Dans toute cette partie, on considèrera l'exemple suivant, qui indique, pour chaque pays, la consommation annuelle de chocolat moyenne par habitant en kg, et le nombre de prix Nobels obtenus pour 10 millions d'habitants :

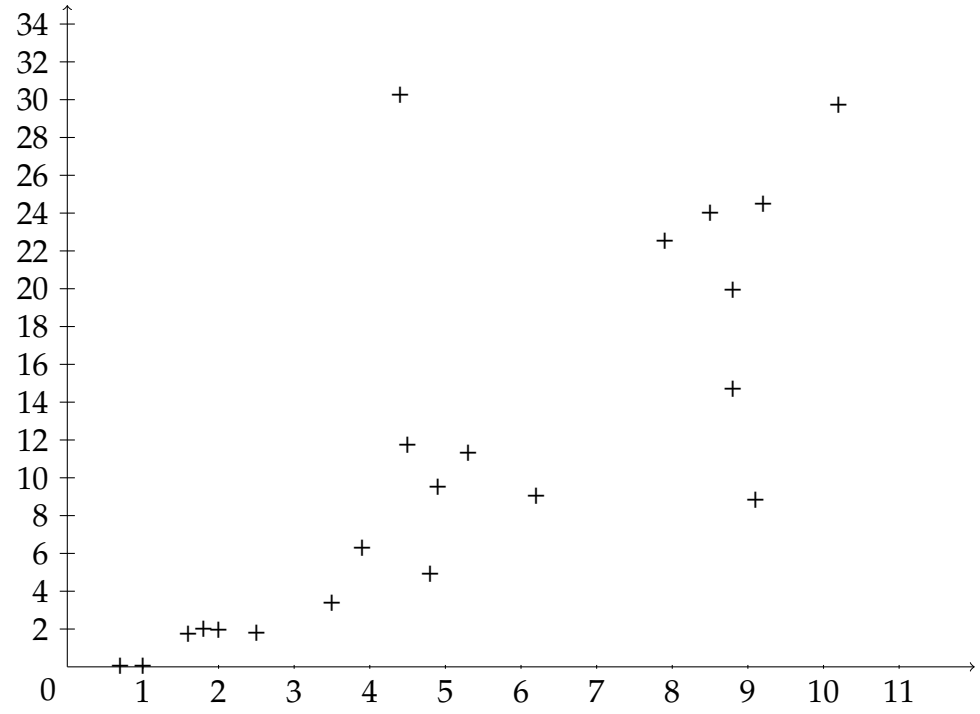
Pays	Chocolat	Nobel
Australie	4.8	4.908
Begique	9.1	8.85
Brésil	1	0.048
Chine	0.7	0.064
Danemark	7.9	22.516
Finlande	6.2	9.053
France	4.9	9.541
Grèce	2.5	1.792
Irlande	8.8	14.701
Italie	3.5	3.369

Japon	1.8	2.039
Canada	3.9	6.28
Pays-Bas	4.5	11.74
Norvège	9.2	24.503
Autriche	8.5	24.04
Portugal	2	1.936
Suède	4.4	30.27
Suisse	10.2	29.728
Espagne	1.6	1.726
USA	5.3	11.34
Royaume-Uni	8.8	19.945

Pour représenter une série double, on utilise un *nuage de points* : un caractère sera représenté sur l'axe des abcisses, et l'autre sur l'axe des ordonnées.

EXEMPLE

Pour l'exemple précédent



Définition 11.3.2
Le point moyen *d'une série double* $(a, b) = ((a_1, b_1), \ldots, (a_N, b_N))$ *est le point*

EXERCICE

Calculer le point moyen de la série précédente, et le placer sur le nuage de points.

Définition 11.3.3
La covariance d'une série double est donnée par

$$\mathrm{Cov}(a, b) =$$

Proposition 11.3.4 : Formule de Koenig-Huygens
On a

$$\mathrm{Cov}(a, b) =$$

où

NOTA

Le signe de la covariance permet de voir l'évolution des deux caractères : si la covariance est positive, les deux caractères ont tendance à varier dans le même sens ; sinon, ils varient plutôt en sens inverses.

Définition 11.3.5
On appelle coefficient de corrélation linéaire *la quantité*

$$\rho(a, b) =$$

Proposition 11.3.6
On a toujours $-1 \leqslant \rho(a, b)m \leqslant 1$..

NOTA

Le coefficient de corrélation linéaire mesure l'"alignement" des points du nuage : si $|\rho| = 1$, a et b ont une relation linéaire ; sinon, on ne peut rien dire (les caractères peuvent n'avoir aucun lien, ou un lien non linéaire).

EXERCICE

Calculer le coefficient de corrélation linéaire de l'exemple.

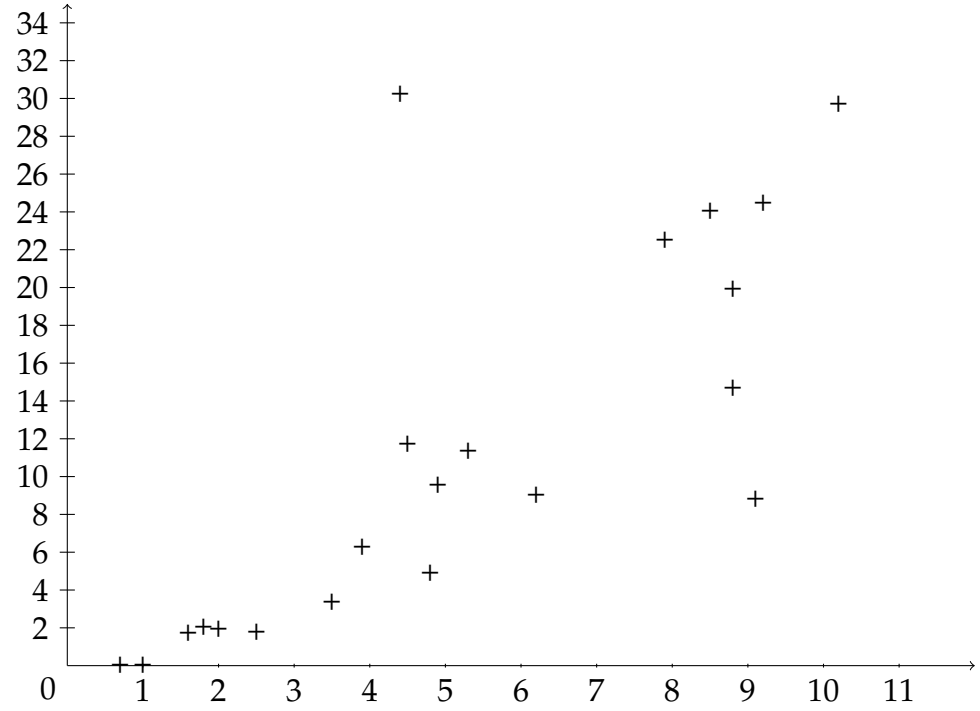
On souhaite maintenant savoir, dans le cas d'un coefficient de corrélation proche de 1 ou -1, l'équation de la droite qui approxime au mieux nos données.

Définition 11.3.7
On appelle droite d'ajustement affine *ou* droite de régression linéaire *d'une série statistique double la droite d'équation* $y = mx + p$*, où*

$m =$ *et* $p =$

EXERCICE

Calculer la droite de régression linéaire de l'exemple, et la tracer sur le nuage de points.



Proposition 11.3.8
La droite de régression linéaire est l'unique droite minimisant la somme des carrés des écarts verticaux, i.e. *minimisant la quantité*

$$\sum_{k=1}^{N} (b_i - ma_i - p)^2.$$

NOTA

Attention, des points presques alignés indiquent une corrélation, mais absolument pas une causalité. On renvoie au sophisme *Cum hoc ergo propter hoc*.

On pourra consulter le site Spurious Correlations pour s'en convaincre.

11.4 Exercices

Exercice 1. Vous devez vous faire opérer pour des calculs rénaux, et devez pour cela choisir entre deux traitements : le traitement A et le traitement B.

a) Un médecin vous fournit les données suivantes :

Nombre de patients	Traitement A	Traitement B
Total	350	350
Soignés	273	289

Quel traitement choisissez-vous ?

b) Un second médecin vous donne les données suivantes, selon la taille des calculs des patients

Petits calculs	Traitement A	Traitement B
Total	87	270
Soignés	81	234

Gros calculs	Traitement A	Traitement B
Total	263	80
Soignés	192	55

Quel traitement choissisez-vous ?

Exercice 2. La distance de freinage d'une véhicule est donnée par le tableau suivant :

Vitesse	40	50	60	70	80	90	100	110
Distance	8	12	18	24	32	40	48	58

a) Représenter cette série par un nuage de points.

b) Calculer le coefficient de corrélation.

c) En utilisant la méthode de moindre carrés, déterminer l'équation de la droite de régression linéaire de la distance en fonction de la vitesse.

d) Estimer la distance pour un véhicule roulant à 120km/h.

Exercice 3. Montrer que pour une série statistique $x = (x_1, \dots, x_N)$, si $\mathrm{Var}(x) = 0$, alors la série statistique est constante.

Exercice 4. On veut dans cet exercice retrouver les valeurs de la droite de régression linéaire, et prouver qu'ils minimisent bien les carrés des écarts verticaux.

On considère donc deux séries $x = (x_1, \dots, x_N)$ et $y = (y_1, \dots, y_N)$, x étant non constante. On définit la fonction

$$f(a,b) = \frac{1}{n}\sum_{k=1}^{N}(ax_k + b - y_k)^2.$$

a) Montrer que pour tous a, b

$$f(a,b) = a^2\overline{x^2} + 2ab\overline{x} + b^2 - 2a\overline{xy} - 2b\overline{y} + \overline{y^2}.$$

b) Calculer les dérivées partielles de f.

c) Résoudre le système

$$\begin{cases} \dfrac{\partial f}{\partial a}(a,b) = 0 \\ \dfrac{\partial f}{\partial b}(a,b) = 0 \end{cases}$$

On note A et B les solutions.

d) Pour tout a, on note f_a la fonction définie par $f_a(b) = f(a,b)$. Montrer que pour tout a, on a

$$f_a(b) \geqslant f_a(\overline{y} - a\overline{x}).$$

e) En étudiant la fonction $\varphi : a \mapsto f_a(\overline{y} - a\overline{x})$, montrer que pour tout a,

$$\varphi(a) \geqslant \varphi(A).$$

f) En déduire que pour tous a, b,

$$f(a,b) \geqslant f(A,B).$$

g) Conclure.

Exercice 5. Montrer que le point moyen d'une série double est toujours sur la droite de régression linéaire.