Statistiques descriptives

Les ensembles étudiés sont appelés population.
Les éléments de la population sont appelés individus.
L'effectif de la population est le nombre de ces individus.
La population est étudiée selon un ou plusieurs caractères.

I Statistiques descriptives univariées

Une population d'effectif total n est observée statistiquement à travers un unique caractère quantitatif x. On suppose que les résultats obtenus pour les n individus sont $a_1, a_2, \cdots, a_n$ et les modalités ordonnées par ordre croissant sont $x_1 < x_2 < \ldots < x_p$.

1. Effectifs et fréquences

Définition 2 Pour tout $k \in [\![1,p]\!]$, on appelle

- effectif de la modalité x_k le nombre n_k d'individus pour lesquels le caractère vaut x_k.
- fréquence de la modalité x_k le nombre $f_k = \dfrac{\text{effectif } x_k}{\text{effectif total}} = \dfrac{n_k}{n}$ $(f_k \in [0,1])$

Il est d'usage de regrouper les effectifs (ou les fréquences) des modalités dans un tableau :

x	x_1	x_2	$\cdots$	x_p
effectif	n_1	n_2	$\cdots$	n_p

ou

x	x_1	x_2	$\cdots$	x_p
fréquence	f_1	f_2	$\cdots$	f_p

.

Remarque
- La somme des effectifs est égale à l'effectif total, c'est-à-dire $n = \sum_{i=1}^{p} n_i$.
- La somme des fréquences est égale à 1 (attention aux arrondis).

Exercice 1 On procède à l'analyse chimique de massifs granitiques de la chaîne des Pyrénées dont nous retenons les teneurs en silice.
On obtient les résultats suivants (pourcentages arrondis) :

67	65	70	72	72	71	72	75	71	74	76	76	76	78	75	76	74	72	76	75

Donner les effectifs et fréquences de la série statistique.

2. Représentation graphique

Pour représenter les effectifs et fréquences d'une série statistique, on place les modalités x_i en abscisse et les effectifs ou fréquences en ordonnée.
On utilise un diagramme en bâtons : pour chaque modalité x_j, on trace une trait de hauteur n_i ou f_i.

Pour des séries statistiques qui utilisent des classes, on utilise un histogramme : pour chaque classe, on trace un rectangle dont la largeur est égale à l'amplitude de l'intervalle correspondant et dont l'aire (et non la hauteur) est proportionnelle à n_i ou f_i.

On peut aussi représenter le polygone des fréquences cumulées : représentation graphique de la fonction constante par morceaux : $\forall i \in [\![1,p]\!]$, la fonction est égale à $\sum_{j=1}^{i} f_j$ sur l'intervalle $[x_i, x_{i+1}[$. On complète les segments horizontaux avec des traits verticaux qui les relient.

Exercice 2 Représenter le polygone des fréquences cumulées de la série statistique de l'exercice 1

3. Caractéristiques de position

Définition 3 Le mode d'une série statistique est une modalité dont l'effectif est maximal.

Graphiquement, cela correspond à une modalité pour laquelle le bâton est le plus haut.
Pour une série dont les modalités sont regroupées en classe, la classe modale est celle dont l'effectif divisé par l'amplitude de l'intervalle est maximal.

Exercice 3 Donner le mode de la série statistique étudiée dans lexercice 1

Définition 4 La moyenne de la série est le nombre réel :

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} a_k = \frac{n_1 x_1 + \cdots + n_p x_p}{n_1 + \cdots + n_p} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i.$$

Exercice 4 Donner la moyenne de la série statistique étudiée à l'exercice 1

Proposition 1 On définit une nouvelle série par $\forall k \in |[1,n]|,\ b_k = \lambda a_k + \mu$ alors $\bar{y} = \lambda\bar{x} + \mu$.

Définition 5 La médiane de la série, notée Q_2, est définie par :

- si n est impair, Q_2 est la modalité de l'individu du milieu numéroté $\frac{n+1}{2}$.
- si n est pair, Q_2 est la moyenne des modalités des individus numérotés $\frac{n}{2}$ et $\frac{n}{2}+1$.

Remarque La médiane se lit sur le polygone des fréquences cumulées : abscisse du point d'ordonnée $\frac{1}{2}$.

Exercice 5 Donner la médiane de la série statistique étudiée à l'exercice 1

4. Caractéristiques de dispersion

Définition 6 L'étendue de la série est le nombre réel $x_p - x_1$.

Définition 7

- La variance de la série est le nombre réel positif

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n}(a_k - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2.$$

- L'écart-type de la série est la racine carrée de sa variance : $\sigma_x = \sqrt{\sigma_x^2}$.
- Le coefficient de variation de la série est le nombre réel : $\frac{\sigma_x}{\bar{x}}$.

Proposition 2 Théorème de König-Huygens
En posant x^2 la série $(x_1^2, \cdots, x_n^2)$, on a $\sigma_x^2 = \overline{(x^2)} - (\bar{x})^2$

Exercice 6 Donner étendue, écart-type et coefficient de variation de la série statistique étudiée à l'exercice 1

Proposition 3 Soit $(x_1, \cdots, x_n)$ une série statistique, $(\lambda, \mu) \in \mathbb{R}^2$. On pose $\forall k \in |[1,n]|,\ y_k = \lambda y_k + \mu$.
On a alors $\sigma_y^2 = \lambda^2\sigma_x^2$ et $\sigma_y = |\lambda|\sigma_x$.

Remarques

- L'écart-type s'exprime dans la même unité que les valeurs de la série alors que la variance s'exprime dans cette unité au carré.
- Plus l'écart-type est faible, plus les valeurs de la série sont concentrées autour de sa moyenne.
- Décaler une série par une constante additive ($b_k = a_k + \mu$) ne modifie pas son écart-type
- Dilater une série par une constante multiplicative ($b_k = \lambda a_k$) multiplie son écart-type par la valeur absolue de cette constante.

Définition 8 Soit $(a_1, \cdots, a_n)$ une série de taille n, dont les modalités sont des nombres réels $x_1, \cdots, x_p$, d'effectifs $n_1, \cdots, n_p$ respectivement.

On suppose que l'on a rangé les modalités dans l'ordre croissant : $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_p$.

On sépare la population en quatre groupes d'effectifs égaux :

Le premier quartile, noté Q_1, est la modalité de l'individu situé au premier quart.

Le deuxième quartile, noté Q_2, est la médiane.

Le troisième quartile, noté Q_3, est la modalité de l'individu situé au troisième quart.

l'écart interquartile est la largeur de l'intervalle $[Q_1, Q_3]$.

Remarques • l'intervalle $[Q_1, Q_3]$ contient la moitié de la population

• l'écart interquartile est un indicateur de la dispersion des valeurs centrales de la série.

• Sur le pfc, les quartiles sont les abscisses des points d'ordonnées $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$.

Exercice 7 donner les quartiles et l'écart interquartile de la série statistique étudiée

Définition 9 On sépare la population en dix groupes d'effectifs égaux. Pour tout k de $|[1,9]|$,

Le *k-ième* décile, noté D_k, est la modalité de l'individu situé au *k-ième* dixième.

La largeur de l'intervalle $[D_1, D_9]$ est appelée l'écart interdécile.

Remarques • l'écart interdécile est aussi un indicateur de la dispersion de la série, il met de côté les valeurs extrèmes. Il est généralement préféré à l'étendue car il est moins sensibles aux valeurs extrèmes qui sont parfois non significatives.

• Sur le polygone des fréquences cumulées, les déciles sont les abscisses des points d'ordonnées $\frac{k}{10}$, $k \in |[1,9]|$.

Exercice 8 Donner les déciles et l'écart interdécile de la série statistique étudiée.

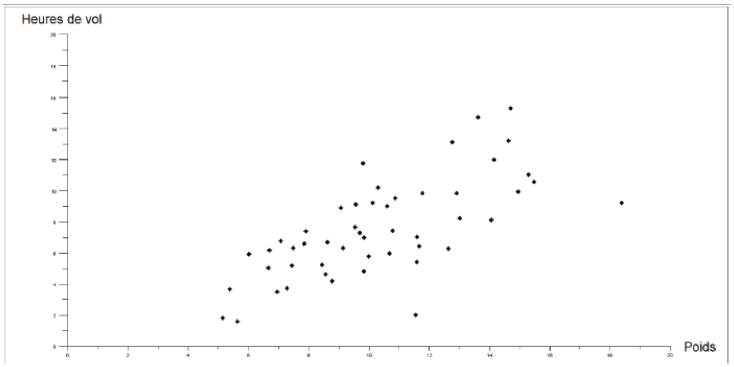
II Statistiques descriptives bivariées

On observe sur une population de taille n la donnée de deux caractères quantitatifs x et y. On dispose alors d'une série double (bivariée) $((x_1, y_1), \cdots, (x_n, y_n))$. L'objectif principal est de déterminer s'il existe une relation entre les deux caractères : est-ce que la valeur de x a une influence sur la valeur de y ?

1. Représentation d'une série statistique pour des caractères prenant des valeurs réelles

La série double est représentée par le nuage de points (x_i, y_i) pour $i \in |[1,n]|$.

Exemple Des traceurs ont permis de mesurer, pour 50 Colibris Serrirostris, la donnée de leur poids et le nombre moyen d'heures de vol quotidiennes. Les données sont représentées sur le graphique :



Exercice 9 Les archives d'un laboratoire de recherche ont fourni les périmètres du tronc et poids des arbres d'une réserve naturelle.

Périmètre	358	375	393	394	360	351	398	362	409	406	487	498
Poids	760	821	928	1009	766	726	1209	750	1036	1094	1635	1517

Périmètre	438	465	469	440	376	444	438	467	448	478	457	456
Poids	1197	1244	1495	1026	912	1398	1197	1613	1475	1571	1506	1458

Périmètre	389	405	405	392	327	395	427	385	404	416	479
Poids	944	1241	1023	1067	693	1085	1242	1017	1084	1151	1381

Représenter le nuage de points de la série statistique.

2. Point moyen du nuage seule caractéristique de position en statistique bivariée.

Définition 10 Le point moyen de la série double a pour coordonnées $G(\bar{x},\bar{y})$. G est donc
- l'isobarycentre du système de points $((x_1,y_1),\cdots,(x_n,y_n))$;
- le barycentre des modalités conjointes (a_j,b_k), munies des pondérations n_{jk}.

Il est pertinent de placer le point G sur le nuage de points.

Exercice 10 placer le point moyen de la série statistique étudiée à l'exercice 9.

3. Covariance de x et y

La covariance généralise la notion de variance au cas de deux caractères.

Définition 11 On appelle covariance de x avec y le nombre réel $s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$.

Proposition 4 Formule de König-Huygens $s_{x,y} = \overline{xy} - \bar{x}.\bar{y}$.

Preuve $s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\frac{1}{n}\sum_{i=1}^{n}y_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n}x_i + \bar{x}\bar{y}$
$= \overline{xy} - \bar{x}.\bar{y} - \bar{x}.\bar{y} + \bar{x}.\bar{y}.$

Remarques
- la covariance est la différence entre la moyenne du produit et le produit des moyennes.
- la variance d'un caractère x n'est autre que la covariance de x avec lui-même.

On vérifie en effet facilement que $s_{x,x} = \sigma_x^2$.

Interprétation de la covariance :
- Une covariance positive traduit le fait que $(x_i-\bar{x})$ et $(y_i-\bar{y})$ ont tendance à être de même signe. Ceci est d'autant plus marqué que $s_{x,y}$ est élevé.
- Une covariance négative traduit le fait que $(x_i-\bar{x})$ et $(y_i-\bar{y})$ ont tendance à être de signes opposés. Ceci est d'autant plus marqué que $s_{x,y}$ est fortement négatif.

C'est donc un indicateur qui décrit la tendance de la corrélation entre x et y :
elle dit si globalement les valeurs de x et de y varient dans le même sens ou dans des sens opposés.

Exercice 11 Calculer la covariance de la série statistique étudiée à l'exercice 9.

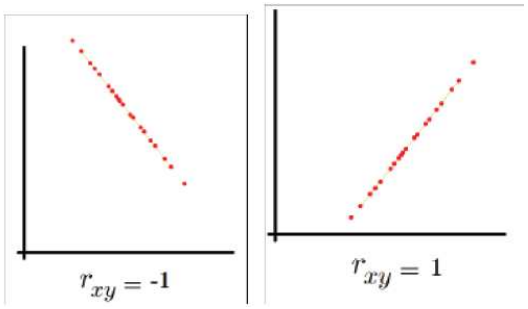
4. Coefficient de corrélation de x et y

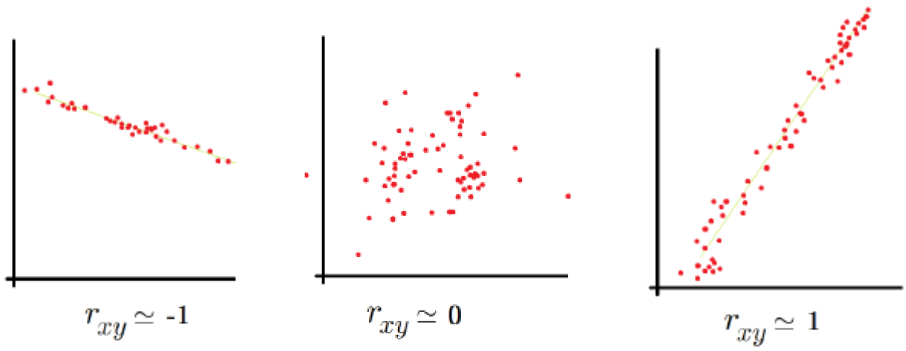
Définition 12 On appelle coefficient de corrélation de x et y le nombre réel $r_{x,y} = \dfrac{s_{x,y}}{\sigma_x \sigma_y}$.

Remarques • $r_{x,y}$ n'est pas défini si $\sigma_x = 0$ ou $\sigma_y = 0$. (si l'un des deux caractères est constant).
• $r_{x,y}$ est sans unité. Il ne dépend pas de l'unité choisie pour exprimer les caractères x et y.

Proposition 5 • $r_{x,y} \in [-1, 1]$.
• $r_{x,y} = \pm 1 \Leftrightarrow \exists (a,b) \in \mathbb{R}^2 \mid y = ax + b \Leftrightarrow y$ est une fonction affine de x.

Remarques Si $r_{x,y} = 1$ alors les points sont alignés sur une droite de pente positive.
Si $r_{x,y} = -1$ alors les points sont alignés sur une droite de pente négative.
Si $|r_{x,y}|$ est proche de 1 alors
le nuage de points est presque rectiligne
les séries x et y sont fortement corrélées.
Inversement, si $|r_{x,y}|$ est proche de 0 alors
le nuage de points est très dispersé
les séries x et y ne sont que très peu corrélées :





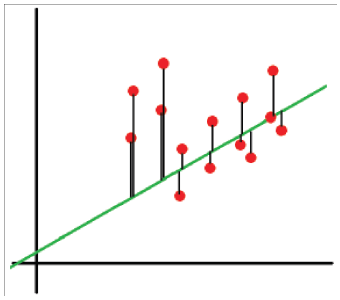
Attention !
Une corrélation entre x et y ne signifie pas forcément un lien de causalité.
Il se peut que les deux caractères dépendent fortement d'un troisième.
Par exemple, il existe certainement une corrélation forte entre le nombre de médecins en France et l'émission de CO2 au Brésil depuis 1950 mais on ne peut pas en conclure que l'émission de CO2 au Brésil est la cause du nombre de médecins en France.
Cependant la corrélation traduit parfois un vrai lien de causalité, par exemple, la consommation de cigarettes est la cause de la diminution de la capacité pulmonaire.
En général la statistique ne démontre pas la causalité, elle permet seulement de la détecter.

Exercice 12 Calculer le coéfficient de corrélation de la série statistique étudiée à l'exercice 9.

5. Ajustement affine

Lorsque $|r_{x,y}|$ est proche de 1, le nuage de points est presque aligné.
Il est donc naturel de chercher l'équation d'une droite qui approche au mieux le nuage de points.
On parle d'ajustement affine ou de régression linéaire.

La méthode utilisée pour déterminer cette droite est la méthode dite des moindres carrés.
La qualité de l'approximation du nuage de points par une droite d'équation $y = ax + b$ est mesurée par la somme des carrés des écarts verticaux des points à la droite, autrement dit, par la quantité $S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$:



Définition 13 On appelle droite de régression de y par rapport à x la droite d'équation $y = ax + b$ où a et b sont des réels qui minimalisent la somme $S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$.

Proposition 6 La droite de régression de y par rapport à x est la droite passant par le point moyen $G(\bar{x},\bar{y})$ et de coefficient directeur $\frac{s_{x,y}}{\sigma_x^2}$.
Elle a donc pour équation $y = ax + b$ avec $a = \frac{s_{x,y}}{\sigma_x^2}$ et $b = \bar{y} - a\bar{x}$.

Preuve On recherche le minimum de la fonction $(a,b) \mapsto S(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$
On recherche les points singuliers c'est-à-dire tels que $\frac{\partial S}{\partial a}(a,b) = \frac{\partial S}{\partial b}(a,b) = 0$.

Or $\begin{cases} \frac{\partial S}{\partial a}(a,b) = -2\sum_{i=1}^{n} x_i(y_i - (ax_i + b)) = -2\left(\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i\right) \\ \frac{\partial S}{\partial b}(a,b) = -2\sum_{i=1}^{n}(y_i - (ax_i + b)) = -2\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i + nb\right) \end{cases}$.

On a alors , en divisant par n et simplifiant par -2 $\begin{cases} \frac{\partial S}{\partial a}(a,b) = 0 \\ \frac{\partial S}{\partial b}(a,b) = 0 \end{cases} \Leftrightarrow \begin{cases} a\overline{x^2} + b\bar{x} = \overline{xy} \\ a\bar{x} + b = \bar{y} \end{cases}$.

$\underset{L_1 \leftarrow L_1 - \bar{x}L_2}{\Leftrightarrow} \begin{cases} a\left(\overline{x^2} - \bar{x}^2\right) = \overline{xy} - \bar{x}\bar{y} \\ a\bar{x} + b = \bar{y} \end{cases}$. On a bien $a\sigma_x^2 = s_{x,y}$ et $b = \bar{y} - a\bar{x}$.
On admettra que ce point est bien celui qui minimalise la somme $S(a,b)$.

Remarque La droite de régression existe toujours, mais lorsque $|r_{x,y}|$ n'est pas proche de 1 elle n'a pas d'intérêt : elle passe à travers le nuage sans en donner l'allure.

Exercice 13 Déterminer et représenter la droite de régression de la série statistique étudiée à l'exercice 9.

Remarque **Attention** !

L'ajustement affine par les moindres carrés n'est pas adapté pour des données issues de mesures physiques avec des incertitudes variables.

Dans ce contexte on préfèrera utiliser la méthode du χ^2 qui introduit des pondérations donnant plus d'importance aux mesures qui ont faible incertitude dans le calcul des coefficients a et b.

6. Changement de variable : anamorphose

Il peut arriver que le nuage de points suggère l'existence d'une relation entre x et y, mais pas affine : les points se répartissent à peu près autour d'une certaine courbe qui n'est pas une droite. Il existe d'autres types d'ajustement qui utilisent un ajustement affine via un changement de variable.

Exemples **1) ajustement exponentiel**

On étudie l'existence d'une relation de la forme $y = ae^{bx}$ avec $(a,b) \in \mathbb{R}^2$.

En posant $Y = \ln y$ et $\beta = \ln a$, on obtient une relation affine $Y = \beta + bx$.

2) ajustement selon une fonction puissance

On étudie l'existence d'une relation de la forme $y = ax^b$ avec $(a,b) \in \mathbb{R}^2$.

En posant $Y = \ln y$, $X = \ln x$ et $\beta = \ln a$ on obtient une relation affine $Y = \beta + bX$.

3) ajustement logarithmique

On étudie l'existence d'une relation de la forme $y = a\ln x + b$ avec $(a,b) \in \mathbb{R}^2$.

En posant $X = \ln x$ on obtient une relation affine $Y = aX + b$.