Chapitre 24 : Statistiques

Table des matières

1 Statistiques univariée

1.1 Données

Définition 1.

On va étudier une population constituée de N individus.
Pour chaque individu, on observe la valeur d'un caractère, noté X.

- un caractère quantitatif comme la taille, le poids, le nombre de naissance
- un caractère qualitatif comme la couleur des yeux, initiales, port de lunettes ou non.

On appelle série statistique la liste des valeurs de ce caractère pour la population étudiée $(X_1, \dots, X_N)$.
On appelle effectif total le nombre de données, c'est-à-dire N.

Exemple 2

- On relève les tailles en cm des nouveaux-nés dans un hôpital.
- On compte le nombre de pétales de pâquerettes dans le cadre d'un TIPE.

Remarque 3 L'objectif de l'étude statistique est de donner un résumé des informations présentes dans cette série statistique.

1.2 Modalités et fréquences

Définition 4.

Soit une série statistique $(X_1, \dots, X_N)$.

- On appelle modalités de la série les différentes valeurs prises par la série.
  On note p le nombre de valeurs différentes (avec $p \leq N$) et on les note $x_1, \dots, x_p$.
- Pour tout $k \in [\![1, p]\!]$, on appelle effectif de la modalité x_k le nombre d'individus pour lesquels le caractère est égal à x_k. On note cet effectif N_k.
- Pour tout $k \in [\![1, p]\!]$, on appelle fréquence de la modalité x_k le réel f_k de $[0, 1]$ défini par

$$f_k = \frac{\text{effectif de } x_k}{\text{effectif total}} = \frac{N_k}{N}$$

Remarque 5 Il est d'usage de regrouper les effectifs et les fréquences dans un tableau.

x	x_1	x_2	$\dots$	x_p
Effectifs	N_1	N_2	$\dots$	N_p
Fréquences	f_1	f_2	$\dots$	f_p

Remarque 6 On a toujours $\sum_{k=1}^{p} N_k = N$ et $\sum_{k=1}^{p} f_k = 1$.

Définition 7.

Soit une série statistique $(X_1, \dots, X_N)$ une série statistiques dont les modalités sont rangées dans l'ordre croissant : $x_1 < x_2 < \cdots < x_p$.

- Pour tout $k \in [\![1, p]\!]$, on appelle effectif cumulé de la modalité x_k le nombre d'individus pour lesquels le caractère est **inférieur ou égal** à x_k.
  Il vaut $N_1 + N_2 + \cdots + N_k$.
- Pour tout $k \in [\![1, p]\!]$, on appelle fréquence cumulé de la modalité x_k le réel de $[0, 1]$ défini par

$$\frac{\text{effectif cumulé de } x_k}{\text{effectif total}} = \frac{N_1 + N_2 + \cdots + N_k}{N} = f_1 + f_2 + \cdots + f_k$$

1.3 Regroupement par classe

Définition 8.

Soit une série statistique.
Lorsque les valeurs prises par le caractère sont proches, on peut avoir envie de les regrouper.
On compte par exemple le nombre de valeurs comprises dans différents intervalles $I_1, I_2, \dots, I_p$.
On parle de regroupement par classe.
On calcule alors des effectifs et des fréquences par classe.

Exemple 9 On a relevé les tailles en cm de nouveaux nés.

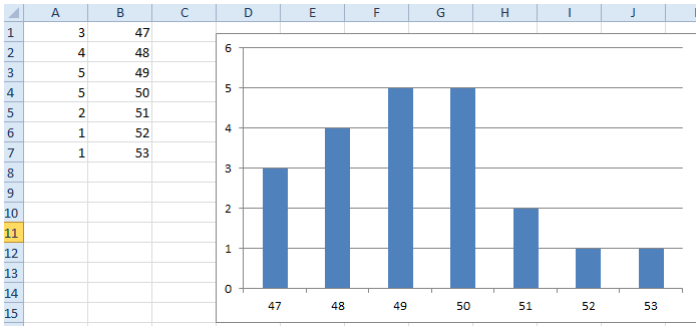
$n°$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
taille	47	50	52	48	53	49	48	48	50	50	41	49	49	47	51	49	50	48	50	51	49

1. Calculer les effectifs, les fréquences, les effectifs cumulés et les fréquences cumulées.
2. Effectuer un regroupement en 4 classes en considérant les intervalles suivants : $[46, 48[$, $[48, 50[$, $[50, 52[$ et $[52, 54[$.
   Calculer les effectifs, les fréquences, les effectifs cumulés et les fréquences cumulées.

1.4 Représentation des données

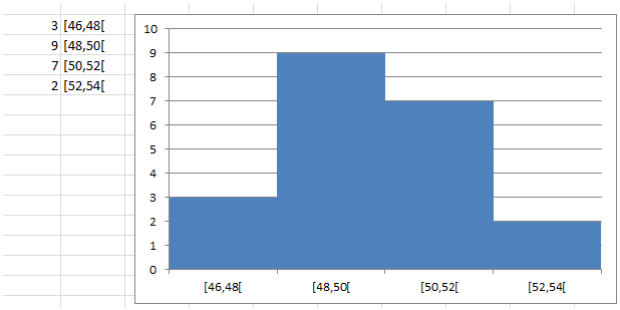
1.4.1 Diagramme en bâtons

Pour représenter les données, on peut associer à chaque modalité de la série statistique une barre de hauteur l'effectif de cette modalité.



1.4.2 Histogramme

Dans le cas d'un regroupement par classe, on peut vouloir que la largeur de chaque bâton soit égale à la longueur de l'intervalle.
On parle alors d'histogramme.



1.5 Caractéristiques de position

Définition 10.

Soit une série statistique $(X_1, \dots, X_N)$ dont les modalités et les effectifs sont données par le tableau suivant :

x	x_1	x_2	$\dots$	x_p
Effectifs	N_1	N_2	$\dots$	N_p

- On appelle mode de la série une modalité de la série dont l'effectif est le plus élevé. (Il peut y en voir plusieurs).
- On appelle moyenne de la série le réel défini par

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{p} N_k x_k = \frac{N_1 x_1 + \dots + N_p x_p}{N_1 + \dots + N_p}$$

- On appelle médiane de la série le réel m_X tel que
  - la moitié de l'effectif a un caractère dont la valeur est inférieure à m_X.
  - la moitié de l'effectif a un caractère dont la valeur est supérieure à m_X.

Exemple 11 Si on reprend le tableau des tailles des nouveaux nés, l'effectif total est 21.
La moyenne vaut 49,28571.
La médiane vaut 49. C'est la modalité pour laquelle l'effectif cumulé dépasse 10.5.

Remarque 12 La moyenne n'est, en général, pas égale à la moyenne.
Elle est moins sensible aux valeurs extrêmes.

Exemple 13 Pour leur TIPE, des élèves de BCPST cueillent 16 pâquerettes et comptent leurs pétales. Ils obtient les données suivantes : 41, 50, 48, 45, 38, 42, 51, 43, 49, 50, 52, 42, 44, 51, 48, 48.

1. Déterminer un mode de la série.
2. Calculer la moyenne et la médiane de cette série.
3. Un des élèves, cueille une dernière pâquerette et compte 100 pétales...
   Donner la nouvelle moyenne et la nouvelle médiane.
   Que remarque-t-on ?

1.6 Caractéristiques de dispersion

Définition 14.

Soit une série statistique $(X_1, \dots, X_N)$ dont les modalités et les effectifs sont données par le tableau suivant :

x	x_1	x_2	$\dots$	x_p
Effectifs	N_1	N_2	$\dots$	N_p

- On appelle variance empirique de la série le réel défini par

$$V_x = \frac{1}{N}\sum_{k=1}^{N}(X_k - \overline{x})^2 = \frac{1}{N}\sum_{k=1}^{p} N_k\,(x_k - \overline{x})^2$$

- On appelle écart-type de la série le réel défini par $s_X = \sqrt{V_X}$.

Remarque 15 La variance et l'écart-type sont des réels positifs.

Théorème 16.

Soit une série statistique $(X_1, \dots, X_N)$.

$$V_X = \frac{1}{N}\left(\sum_{k=1}^{N} X_k^2\right) - (\overline{x})^2 = \overline{x^2} - (\overline{x})^2$$

Définition 17.

Soit une série statistique $(X_1, \dots, X_N)$.

- On appelle quartiles de la série trois réels $Q_1 \leq Q_2 \leq Q_3$ tels que
  - Un quart de l'effectif a un caractère dont la valeur est inférieure à Q_1.
  - $Q_2 = m_X$, la médiane.
  - Trois quarts de l'effectif ont un caractère dont la valeur est inférieure à Q_3.
- On appelle déciles de la série neuf réels $D_1 \leq D_2 \leq \dots \leq D_9$ tels que
  - 1/10 de l'effectif a un caractère dont la valeur est inférieure à D_1.
  - 2/10 de l'effectif a un caractère dont la valeur est inférieure à D_2.
  - $\dots$
  - 9/10 de l'effectif ont un caractère dont la valeur est inférieure à D_9.

Remarque 18 Avec les quartiles, on partage la série en quatre groupes. Avec les déciles, en 10 groupes.

Exemple 19 Une classe est partagée en deux demi-groupes. Leurs notes eu premier devoir d'informatique sont les suivantes/
Groupe 1 : 11,14,13,8,16,8,13,14,13,15,11,12
Groupe 2 : 6,4,7,19,20,18,17,4,15,15,16,7

1. Calculer la moyenne de chaque groupe.
2. Calculer la variance et l'écart-type de chaque groupe.
3. Représenter le diagramme en bâtons pour chacun des groupes.
4. Déterminer les quartiles des deux groupes.

1.7 Diagramme en boîte

Les 5 grandeurs minimum, premier quartile, médiane, troisième quartile et maximum peuvent être regroupés dans un diagramme en boîte (à l'échelle).

2 Statistique bivariée

2.1 Données

Dans cette partie, on observe toujours une population de taille N.
On observe cette fois deux caractères X et Y pour chaque individu.
Les données sont N couples de $\mathbb{R}^2$ notés $(X_1, Y_1), \dots, (X_N, Y_N)$.

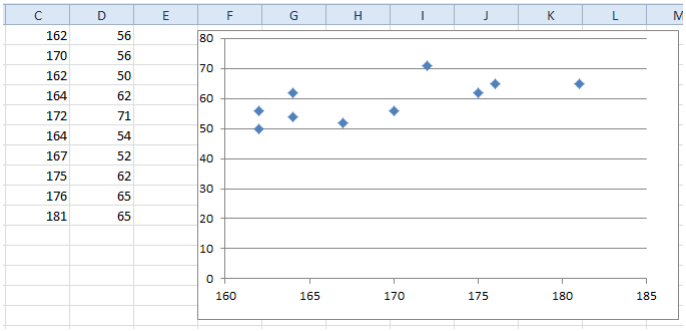
On cherche à déterminer le lien possible entre les caractères X et Y.

Exemple 20 On a relevé taille et poids de 10 personnes.

Taille (en cm)	162	170	162	164	172	164	167	175	176	181
Poids (en kg)	56	56	50	62	71	54	52	62	65	65

2.2 Représentation des données

Les données sont présentées sous forme de nuage de points.
Cette représentation permet de conjecturer une relation entre les deux caractères, par exemple $Y = aX + b$.



2.3 Caractéristiques de position et de dispersion

Définition 21.

Soit $(X_1, Y_1), \dots, (X_N, Y_N)$ une série statistique pour deux caractères.
On appelle point moyen le point de coordonnées $(\overline{x}, \overline{y})$.

Définition 22.

Soit $(X_1, Y_1), \ldots, (X_N, Y_N)$ une série statistique pour deux caractères.

- On appelle covariance empirique le réel défini par

$$c_{X,Y} = \frac{1}{N}\sum_{k=1}^{N}(X_i - \overline{x})(Y_i - \overline{y})$$

- On appelle coefficient de corrélation le réel défini par

$$r_{X,Y} = \frac{c_{X,Y}}{s_X . s_Y}$$

Théorème 23.

Soit $(X_1, Y_1), \ldots, (X_N, Y_N)$ une série statistique pour deux caractères.

$$c_{X,Y} = \frac{1}{N}\left(\sum_{k=1}^{N} X_k Y_k\right) - \overline{x}.\overline{y}$$

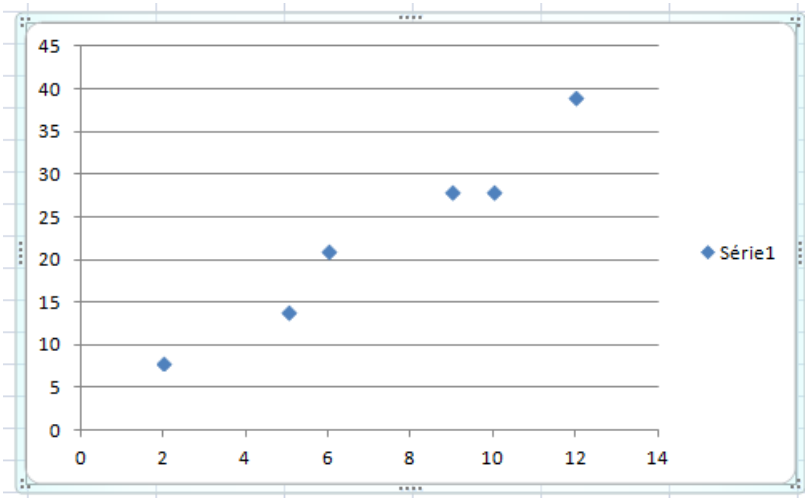
2.4 Ajustement affine

On va appliquer cette méthode aux données suivantes

X	2	5	6	12	10	9
Y	8	14	21	39	28	28

.

Voici leur représentation sous forme de nuage de points.



Il semble y avoir une relation affine entre X et Y. On cherche deux réels a et b tels que $Y = aX + b$.

On va choisir a et b pour minimiser l'erreur commise par notre hypothèse $\varepsilon(a,b) = \sum_{k=1}^{6}(y_k - ax_k - b)^2$.

Théorème 24.

Soit $(X_1, Y_1), \ldots, (X_N, Y_N)$ une série statistique pour deux caractères.

Les paramètres qui minimisent l'erreur $\varepsilon(a,b) = \sum_{k=1}^{6}(y_k - ax_k - b)^2$ sont

$$\hat{a} = \frac{c_{X,Y}}{V_X} \text{ et } \hat{b} = \overline{y} - \hat{a}\overline{x}$$

2.5 D'autres modèles

La relation entre X et Y peut être autre que affine. Par exemple,

- $Y = aX^2 + b$.
- $Y = \dfrac{1}{aX+b}$.
- $Y = b\mathbf{e}^{aX}$

Dans tous les cas, on se ramènera à rechercher un ajustement affine entre deux variables, qui pourront être d'autres que X et Y.

- $Y = aX^2 + b$.
  On appliquera la méthode précédente à Y et X^2.
- $Y = \dfrac{1}{aX+b} \Leftrightarrow \dfrac{1}{Y} = aX + b$.
  On appliquera la méthode précédente à $\dfrac{1}{Y}$ et X.
- $Y = b\mathbf{e}^{aX} \Leftrightarrow \ln(Y) = aX + \ln(b)$.
  On appliquera la méthode précédente à $\ln(Y)$ et X.

Exemple 25 Loi SPAR. (extrait du livre Mathématiques, Méthodes et Exercices. A.Bégyn, R.Leroy,G.Connau)

Plus une région est vaste, plus le nombre d'espèces y vivant est grand. Pour modéliser mathématiquement ce phénomène (et mesurer ce qu'on appelle la biodiversité), les scientifiques utilisent régulièrement la loi SPAR("species-area relationship"). Elle stipule que si S représente la surface de la région à étudier et N le nombre d'espèces présentes dans cette région, alors on a

$$N = CS^{\alpha}$$

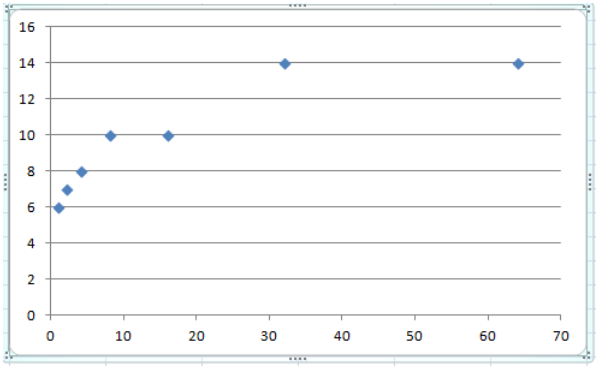
où C et α sont des constantes à ajuster selon la région étudiée.

On demande à des élèves de vérifier cette loi pour les plantes dans une prairie.
Les données récoltées sont résumées dans le tableau suivant

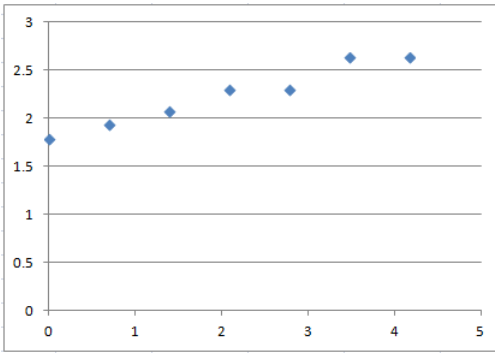
Surface S (en cm^2)	1	2	4	8	16	32	64
Nombre d'espèces N	6	7	8	10	10	14	14

1. A quelles variables peut-on appliquer la régression linéaire (l'ajustement affine) ?
2. Tracer le nuage de points pour ces nouvelles variables.
3. Déterminer une approximation de C et de α.

1. On commence par tracer les données brutes.



   Si la loi SPAR est correcte, on a $N = CS^{\alpha}$ ou encore $\ln(N) = \ln(C) + \alpha.\ln(S)$.
   Donc, on va appliquer la régression linéaire aux variables $X = \ln(S)$ et $Y = \ln(N)$.
2. On obtient alors le nuage de point suivant



3. Le théorème 24 nous dit que la meilleure valeur pour minimiser l'erreur sont

$$\hat{\alpha} = \frac{c_{\ln(S),\ln(N)}}{V_{\ln(S)}} = 0.206 \quad \widehat{\ln(C)} = \overline{\ln(N)} - \hat{\alpha}.\overline{\ln(S)} = 1.803$$

. Finalement, $\alpha \simeq 0.206$ et $C \simeq \mathbf{e}^{1.803} \simeq 6.07$.