Statistique descriptive

Vincent Tariel

Université de la Nouvelle-Calédonie

09 Juillet 2020

Avant-propos

"Il y a trois sortes de mensonges : les mensonges, les sacrés mensonges et les statistiques." Mark Twain
"Selon les statistiques, une personne sur quatre serait folle. Si vos trois meilleurs amis ne le sont pas, alors c'est vous !" Rita Mae Brown
"Les lois de probabilités, si vraies en général, si fallacieuses en particulier." Edward Gibbon
"La mort d'un homme est une tragédie. La mort d'un million d'hommes est une statistique." Joseph Staline
"Les statistiques nous montrent que parmi ceux qui contractent l'habitude de manger, très peu survivent." Wallace Irwin.

Devinette probabiliste

- Ceci n'est pas un jeu de hasard, c'est un jeu sur le hasard.
- Ci-dessous sont portées deux séries de cent chiffres. L'une des deux séries est constituée à la main par quelqu'un à qui il a été demandé d'écrire une suite de 0 et de 1. L'autre correspond à une succession de tirages authentiques d'une pièce de monnaie. (face=0, pile=1).
- Série 1 :
  01011011100111101010101011010100010101011111001
  01010101110101000111100010100000111010101010001
  101000.
- Série 2 :
  01101101110100110000111100000000000111010001010
  11011000000101110001010000100001011111100011111
  110010.

Devinette probabiliste

- L'être humain est un piètre simulateur de hasard.
- On peut démontrer que la probabilité de présence d'une suite de cinq 0 ou 1 dans une liste de cent chiffres est supérieure à 97% or nous pensons, à tort qu'une suite trop longue d'un même chiffre ne fait pas très aléatoire. Cette remarque ne permet pas ici d'identifier les séries.
- Avec six chiffres identiques, la probabilité est de 80%. Seule la série 2 respecte ce critère.
- Seul le hasard peut permettre de donner une liste de onze chiffres identiques. . .

Conclusion : ne pas se fier à son intuition en statistiques mais aux mathématiques !

Statistique descriptive

Rôle :

- ressortir des propriétés de l'échantillon étudié,
- suggérer des hypothèses

Méthodes d'analyse de données :

- représentation des données
- classification pour réduire la taille de l'ensemble d'individus (i.e. regrouper ceux qui se ressemblent en « classes »)
- factorielles pour réduire le nombre de variables (i.e. analyse en composantes principales, analyse de correspondances)

Historique

Le mot “statistique” vient de l’allemand “Statistik”, qui, au milieu du XVIIe siècle, désigne l’analyse des données utiles à l’ ´Etat. Le traitement d’un grand nombre de données chiffrées qui sont triées, classées ou résumées correspond à ce que l’on appelle aujourd’hui “les statistiques” au pluriel. On les distingue de “la statistique”, au singulier, qui correspond à la modélisation de ces données, vues comme résultats d’expériences en présence d’aléa, et ‘a l’étude de cet aléa. On peut dater l’émergence de la statistique du début du XIXe siècle, avec l’étude de données provenant de l’astronomie sur les positions des planètes et leur trajectoire. En particulier, en 1805 Adrien-Marie Legendre (1752-1832) introduisit la méthode des moindres carrés pour estimer des coefficients à partir de données, et en 1809 Carl Friedrich Gauss (1777-1855), utilisant une modélisation des erreurs par la loi normale, retrouva en maximisant la densité de la loi normale des erreurs.

Sommaire

Sommaire

Définitions

Une étude statistique porte sur un ensemble d'objets ou de personnes appelé population.
En général, cet ensemble est trop grand pour que l'on interroge toutes les personnes. On se contente alors d'un certain nombre d'éléments de cet ensemble, appelés unités statistiques.
Pour chaque unité statistique, on pose une ou plusieurs questions qui correspondent à des caractères. Un caractère peut être

- qualitatif (la valeur est une catégorie) : couleur, profession, type de meuble, forme, etc.
- quantitatif (la valeur est un nombre) : nombre d'enfants, age, nombre de pièces, taille, poids, hauteur, largeur, salaire, etc.

On note en général

- i un indice qui représente le numéro de l'unité statistique
- x_i la valeur du premier caractère mesuré sur l'unité statistique i
- y_i la valeur du deuxième caractère mesuré sur l'unité statistique i

Exemples

Sujet : notes de l'évaluation de l'enseignement statistique

- Population : licence DEG
- Unité statistique : étudiant
- Caractère : note de 0 à 20(quantitatif)

x_i est la note de l'étudiant d'indice i.

Sommaire

Tableaux de données

- **Liste**

  0 10 17 2 15 19 6 17 15 0 19 10 12 19 0 10 4 6 12 9 17

- **Effectif**

Valeurs	0	2	4	6	9	10	12	15	17	19
Effectifs	3	1	1	2	1	3	1	2	3	3

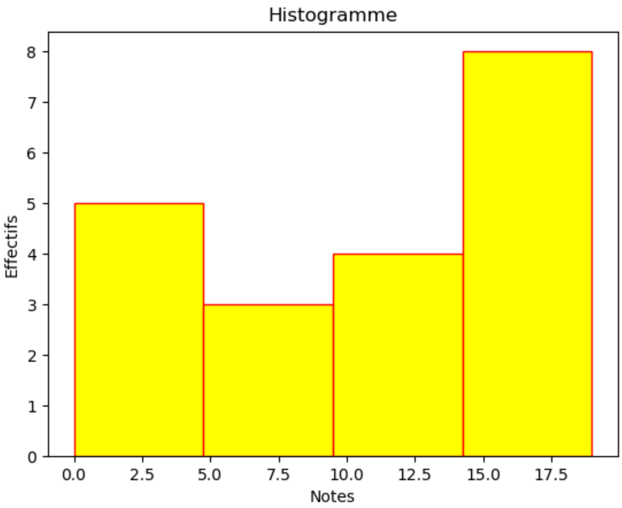
- **Fréquence**

Valeurs	0	2	4	6	9	10	12	15	17	19
Fréquence	0,15	0,05	0,05	0,1	0,05	0,15	0,05	0,1	0,15	0,15

- **classes et fréquences**

Valeurs	[0,5[	[5,10[	[10,15[	[15,20[
Fréquence	0,25	0,15	0,20	0,4

Histogramme



Un quart de la classe est en grande difficulté avec des notes <5 et de

Histogramme

Nombre de classes :
Le nombre de classes dépend du nombre de valeurs N dont on dispose.
Le nombre de classes K peut être déterminé par les formules suivantes :

$$K = \sqrt{N} \text{ ou } K = 1 + \log_2 N.$$

Taille de l'intervalle :
Généralement, dans le cadre d'une analyse de ce type, on utilise des classes de largeur identique.

Représentations de données : histogramme

Soit la fabrication de rations alimentaires, la pesée des rations avant emballage donne la série de mesures suivantes en kg :

0,547 0,563 0,532 0,521 0,514 0,547 0,578 0,532 0,552 0,526 0,534 0,560 0,502 0,503 0,516 0,565 0,532 0,574 0,521 0,523 0,542 0,539

0,543 0,548 0,565 0,569 0,574 0,596 0,547 0,578 0,532 0,552 0,554 0,596 0,529 0,555 0,559 0,503 0,499 0,526 0,551 0,589 0,588 0,568

0,564 0,568 0,556 0,523 0,526 0,579 0,551 0,584 0,551 0,512 0,536 0,567 0,512 0,553 0,534 0,559 0,498 0,567 0,589 0,579

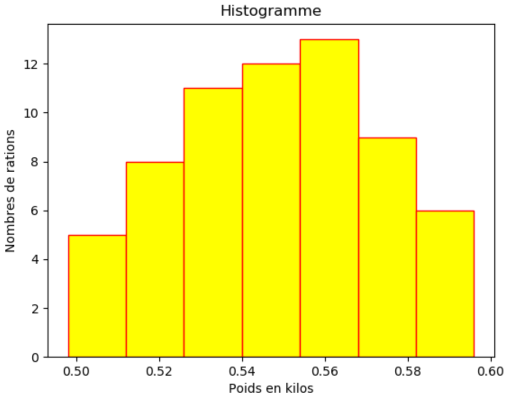
Les caractéristiques du relevé sont les suivantes :

1. Le nombre d'échantillons : N=64
2. L'étendue : w=0,098 kg
3. Valeur minimale : 0,498 kg
4. Valeur maximale : 0,596 kg

On en déduit les paramètres suivants pour l'histogramme :

1. Le nombre de classes est de 7 (en utilisant la formule avec le logarithme)
2. L'amplitude de classe est $0{,}098/7 = 0{,}014$ kg que l'on arrondit à 0,015 kg (résolution de la balance : 0,001 kg)
3. La valeur minimale de la première classe est de $0{,}498 - (0{,}001/2) = 0{,}4975$. Par souci de facilité pour l'interprétation, on peut arrondir cette valeur à 0,495 kg.

Histogramme



Sommaire

Moyenne

Moyenne

La moyenne d'une série statistique noté $\bar{x}$ est le quotient de la somme de toutes les valeurs de cette série par l'effectif total :

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

La moyenne exprime la valeur qu'aurait chacun si le partage était équitable.

Exemple des notes :

$$\bar{x} = \frac{0 + 10 + \cdots + 17}{20} = 10,35$$

Avec le tableau des effectifs, le calcul est :

$$\bar{x} = \frac{3 \times 0 + 1 \times 2 + \cdots + 3}{20} = 10,35$$

Variance et écart-type

Moyenne

L'écart d'une série statistique noté σ est la racine carrée de la variance, V, c'est-à-dire :

$$\sigma = \sqrt{V} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

La moyenne est une mesure de la dispersion des valeurs d'un échantillon statistique.

Exemple des notes :

$$\sigma = \frac{(0-10,35)^2 + (10-10,35)^2 + \cdots + (17-10,35)^2}{20} = 6,61$$

Avec le tableau des effectifs, le calcul est :

$$\sigma = \frac{3\times(0-10,35)^2 + 1\times(2-10,35)^2 + \cdots + 3\times(19-10,35)^2}{20} = 6,61$$

Comme l'écart type est de $6,61$, les notes sont très étalées.

Sommaire

Quantile

Quantile

Les quantiles sont les valeurs qui divisent un jeu de données en intervalles contenant le même nombre de données. Ainsi les quartiles sont les trois quantiles qui divisent un ensemble de données en quatre groupes de taille égale. La médiane quant à elle est le quantile qui sépare le jeu de données en deux groupes de taille égale.

Pour déterminer les quartile q_1 et q_3 et la médiane, on calcul le tableau des fréquences cumulées :

Valeurs	0	2	4	6	9	10	12	15	17	19
Fréquences	0,15	0,05	0,05	0,1	0,05	0,15	0,05	0,1	0,15	0,15
uences cumulées	0,15	0,2	0,25	0,35	0,4	0,55	0,6	0,7	0,85	0,1

Pour le premier quartile on cherche la valeur telle que la fréquence cumulée dépasse 0,25 pour la première fois, soit $q_1 = 4$.
Pour le troisième quartile, 0,75 pour la première fois, soit $q_3 = 17$.
Pour la médiane, 0,5 pour la première fois, soit $M = 10$.

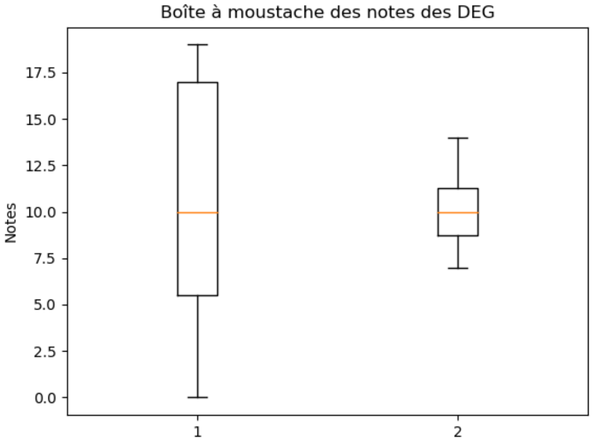
Boîte à moustache

Boîte à moustache

la boîte à moustaches est un moyen rapide de figurer le profil essentiel d'une série statistique quantitative en représentant quelques indicateurs de position du caractère étudié (médiane, quartiles, minimum, maximum ou déciles). Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes.

Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. Ce rectangle suffit pour le diagramme en boîte. On peut rajouter des segments aux extrémités menant jusqu'aux valeurs extrêmes, ou jusqu'aux premier et neuvième déciles (D_1/D_9).

Boite à moustache



Les notes de la première évaluation sont nettement plus hétérogènes que les notes de la seconde évaluation.

Sommaire

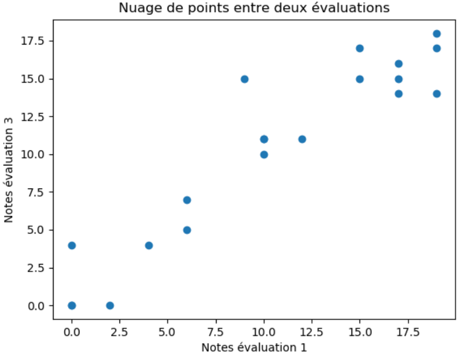
Sommaire

Nuage de points

Nuage de points

Le nuage de points, aussi appelé diagramme de dispersion, est la représentation graphique d'une série statistique à deux variables. Il permet d'observer la relation entre ces deux variables. Il est possible d'effectuer un ajustement de ce nuage de points par une courbe afin d'effectuer des prévisions.

Notes 1	Notes 2	Notes 3
0	8	4
10	10	10
17	11	15
2	12	0
15	8	17
19	9	14
6	13	5
17	14	16
15	9	15
0	7	0
19	8	17
10	7	11
19	12	18
0	13	0
10	9	11
4	10	4
6	10	7
12	10	11
9	10	15
17	10	14



Sommaire

Corrélation et causalité

- Coluche : " 1/3 des accidents de la route étant dus à des conducteurs alcooliques, qu'est ce qu'on attend pour punir les 2/3 de conducteurs sobres responsables de la majorité des accidents ? "
- "Une étude anglaise a prouvé que les gens habitant près de pylônes à haute tension étaient significativement plus souvent malades que le reste de la population. Est-ce la faute du courant électrique ? Ce n'est pas évident parce qu'une autre étude a révélé que les habitants sous les pylônes étaient en moyenne plus pauvres ; et on sait les liens santé-pauvreté... À elle seule, cette étude ne permet pas de conclure."

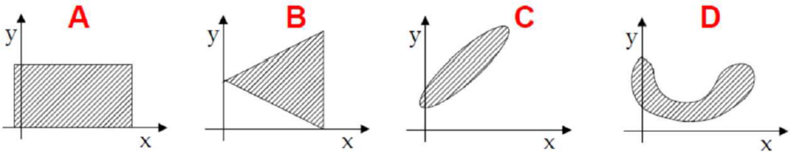
L'objectif est de déterminer un *corrélation* entre deux variables X et Y, étudiées sur le même échantillon. Dans certains cas, cette liaison peut être considérée a priori comme causale, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité.

Corrélation

Corrélation

La corrélation entre deux variables aléatoires (X, Y) est une notion de liaison qui contredit leur indépendance.
A $X = x$ fixé, la moyenne $\bar{Y}$ est une fonction de x.
Cette corrélation est très souvent réduite à la corrélation linéaire entre variables quantitatives, c'est-à-dire l'ajustement d'une variable par rapport à l'autre par une relation affine obtenue par régression linéaire.

Corrélation



- 1 : corrélation non linéaire
- 2 : absence de liaison en moyenne mais pas en dispersion
- 3 : corrélation linéaire
- 4 : absence de liaison

Coefficient de corrélation

Partant d'un échantillon $\{(x_i, y_i) \,|\, 1 \leq i \leq N\}$ de réalisations indépendantes de deux variables X et Y, le coefficient de corrélation est donné par

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

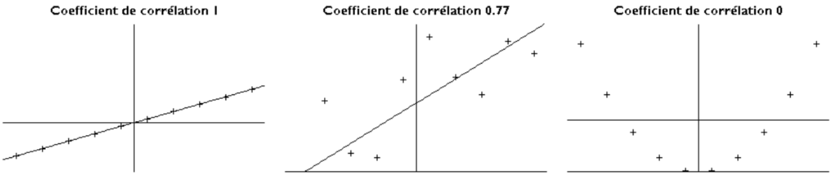
avec

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$\sigma_X = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\sigma_Y = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2}$$

Coefficient de corrélation



X = a*Y (corrélation linéaire)

Nuage de point

$Y = X^2$: Y est complètement déterminée par X (X et Y ne sont pas indépendants), mais leur corrélation vaut 0

Régression linéaire

Dans le cadre d'un modèle linéaire simple, on peut représenter graphiquement la relation entre x et y à travers un nuage de points. L'estimation du modèle linéaire permet de tracer la droite de régression, d'équation $y = \beta_0 + \beta_1 x$. Le paramètre β_0 représente l'ordonnée à l'origine et β_1 le coefficient directeur de la droite. On exprime β_0 et β_1 ainsi :

$$\beta_1 = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$$

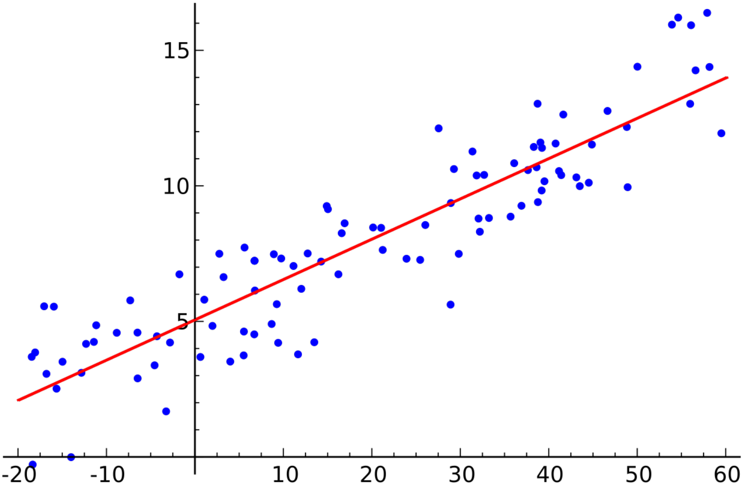
$$\beta_0 = E(Y) - \beta_1 E(X)$$

Partant d'un échantillon $\{(x_i, y_i) \,|\, 1 \leq i \leq n\}$, on a :

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}$$

Régression linéaire

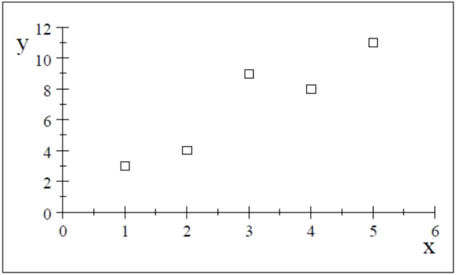


Régression linéaire

Exemple.
On considère la série double statistique suivante :

x_i	2	3	5	1	4
y_i	4	9	11	3	8

Le nuage de points correspondant est représenté sur le graphe ci-dessous.



Nuage de points

Régression linéaire

La droite de regression de y en x a pour équation : $y = \hat{a}x + \hat{b}$, avec $\hat{a} = \dfrac{cov(x,y)}{s_x^2}$ et $\hat{b} = \overline{y} - \hat{a}\overline{x}$.

x_i	y_i	x_iy_i	x_i^2
2	4	8	4
3	9	27	9
5	11	55	25
1	3	3	1
4	8	32	16
15	35	125	55

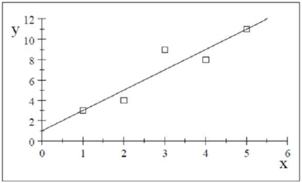
On a $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{5} \times 15 = 3$, $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{5} \times 35 = 7$,

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \overline{x}\overline{y} = \frac{1}{5} \times 125 - 3 \times 7 = 4,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2 = \frac{1}{5} \times 55 - (3)^2 = 2.$$

On en déduit que $\hat{a} = \dfrac{cov(x,y)}{s_x^2} = \dfrac{4}{2} = 2$ et $\hat{b} = \overline{y} - \hat{a}\overline{x} = 7 - 2 \times 3 = 1$.

La droite de regression de y en x a donc pour équation : $y = 2x + 1$.



Nuage de points et droite de régression de y

Régression linéaire

Exercice : On cherche à étudier la relation entre le nombre d'enfants d'un couple et son salaire. On dispose de la série bidimensionnelle suivantes :

Salaire en euros (Y)	Nombre d'enfants (X)
510	4
590	3
900	2
1420	1
2000	0
600	5
850	6
1300	7
2200	8

a) Calculer le coefficient de corrélation linéaire entre ces deux variables statistiques. Conclusion ?

b) Un expert en démographie affirme que les deux caractéristiques sont indépendantes. Qu'en pensez-vous ?