BUT1 – OUTILS NUMÉRIQUES POUR LES STATISTIQUES DESCRIPTIVES

JOHAN LERAY

Table des matières

Notations. Pour a et b deux entiers relatifs tels que $a \leqslant b$, on note

$$[\![a, b]\!] \coloneqq [a, b] \cap \mathbb{Z} \ .$$

Par exemple, on a $[\![-2, 3]\!] = \{-2, -1, 0, 1, 2, 3\}$.

Remarque – Ce cours est très largement inspiré du très bon livre [Sap06] de Gilbert Saporta. Si le sujet vous intéresse, je vous en conseille la lecture.

Introduction

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents.

Définition 0.1 *(Population – Individu).* On appelle *population*, l'ensemble des objets que l'on observe, que l'on étudie, et qui possède des propriétés communes. Ces objets sont appelés des *individus* ou *unité statistiques.*
L'étude de tous les individus d'une population finie s'appelle *un recensement.* Lorsque l'on n'observe qu'une partie de la population, on parle de *sondage*, la partie de la population s'appelant *l'échantillon.*

Remarque 0.2 – Le terme de population est hérité des premières applications de la statistique à la démographie.

Définition 0.3 *(Variable).* Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées *variables* ou *caractères.* Ces variables peuvent être classées selon leur nature.

— Les *variables qualitatives* s'exprimant par l'appartenance à une *catégorie* ou *modalité* d'un ensemble fini.

— Les *variables quantitatives* ou numériques, qui s'expriment par des nombres réels sur lesquels les opérations arithmétiques courantes (somme, moyenne, ...) ont un sens. Certaines peuvent être *discrètes* ou *continues.* On pourra représenter une telle variable mathématiquement par une application ensembliste

$$\begin{array}{rccc} X\colon & P & \longrightarrow & \mathbb{R} \\ & i & \longmapsto & X(i) \end{array}$$

où P est l'ensemble qui représente la population.

Remarque 0.4 – Série statistique – Dans la pratique, on s'intéresse surtout à l'univers image $X(P)$ et on identifie souvent une variable statistique avec la liste des valeurs $X(i)$ prises par la variable. Dans ce cas, on parle de *série statistique.*

Exemple 0.5 – On peut vouloir faire des statistiques sur la population française. Dans ce cas, la population est « l'ensemble des Français », et chaque Français est un individu. Sur cette population, on pourra s'intéresser à différentes variables, comme par exemple

— le sexe, la couleur des cheveux, la catégorie socio-professionnelle, la ville de résidence,... ce sont des variables qualitatives ;

— la taille, le salaire, le nombre d'enfants, ... ce sont des variables quantitatives.

Dans la suite, on s'intéressera principalement à des variables quantitatives. Lorsque le nombre d'individus et le nombre de variables sont grands, on cherche à synthétiser cette masse d'informations sous une forme exploitable et compréhensible. Une première étape consiste à décrire séparément les résultats obtenus pour chaque variable : c'est la description *unidimensionnel.* La synthèse de ces données se fait en général sous forme de *tableaux*, de *graphiques* et de *résumés numériques.* Ce traitement des données est appelé *statistique descriptive.*

1 – Tableaux statistiques

Leur présentation diffère légèrement selon la nature des variables.

1.1. **Variables discrètes ou qualitatives.** On considère ici une variable X définie sur une population P, dont on ne connaît que n valeurs $x_1, \ldots, x_n$. Pour chaque valeur ou modalité x_i, on note n_i le nombre d'occurrences (ou effectif) de x_i dans l'échantillon. On a $\sum_{i=1}^{n} n_i = n$ et f_i la fréquence correspondante $f_i = \frac{n_i}{n}$.

Le tableau statistique se présente en général sous la forme suivante :

Table 1 – Tableau statistique

Modalité	Effectif	Fréquence
x_i	n_i	f_i

Le plus parlant est sûrement d'illustrer cela avec un exemple.

Exemple 1.1 – Diplômes en Loire-Atlantique I – On étudie le niveau d'étude de la population du département de la Loire-Atlantique. La population est donc l'ensemble des individus de plus de 16 ans recensés en 2017 en Loire–Atlantique. La variable statistique représentant le plus haut diplôme obtenu prend les valeurs suivantes :

> Aucun diplôme ; diplôme de niveau CEP ; diplôme de niveau BEPC ; diplôme de niveau CAP–BEP ; diplôme de niveau BAC ; diplôme universitaire du 1er cycle, BTS, DUT ; Diplôme universitaire du 2ème cycle.

On résume ces données dans le tableau 2.

Table 2 – Répartition des diplômes en Loire-Atlantique en 2017, pour les individus de plus de 16 ans
Source : Insee, Base historique des recensements de la population, exploitation complémentaire

Diplôme	Éffectif	Fréquence
Aucun diplôme	99 003	0.096
Niveau CEP	72 278	0.070
Niveau BEPC	48 056	0.046
Niveau CAP–BEP	274 194	0.267
Niveau BAC	182 371	0.178
1er cycle, BTS, DUT	132 610	0.129
2ème cycle	219 899	0.214
Total Population	1 028 411	

1.2. **Variables continues ou assimilées**. Dans le cas d'un très grand nombre de valeurs prises par notre variable X, et en particulier ci celle ci est continue, alors on regroupe les valeurs en k *classes* d'extrémités $e_0, e_1, \dots, e_k$. Les classes sont les intervalles $[e_i, e_{i+1}[$, et, pour chacune d'entre elles, on note l'effectif n_i, la fréquence f_i ainsi que la *fréquence cumulée*

$$F_i := \sum_{j=0}^{i} f_j \ ,$$

c'est-à-dire la proportion des individus pour lesquels $X < e_i$.

Exemple 1.2 – Durée de vie d'un composant électronique – Une entreprise veut mesurer la durée de vie sous tension du composant électronique qu'elle produit. Pour cela, elle teste 10000 composants. Pour stocker les 10000 résultats, on fait donc le choix de les rassembler par classe, que l'on compile dans le tableau 3. On y fait également apparaître les fréquences cumulées.

1.3. **Quelques représentations graphiques**.

§ *Barres et camemberts.*

Pour des variables à modalités non-ordonnées, il existe plusieurs sortes de diagrammes. Les plus répandus sont :

— les *diagrammes en barres* (verticales ou horizontales) : les barres sont de longueurs proportionnelles aux fréquences des catégories, leur épaisseur étant sans importance.

TABLE 3 – Tableau de durée de vie du composant

Modalité	$[0,2[$	$[2,4[$	$[4,6[$	$[6,8[$	$[8,10[$	$[10,12[$	$[12,15[$	$[15,18[$	$[18,+\infty[$
Effectif	3088	2087	1496	1064	731	483	431	257	363
Fréquence	0.3088	0.2087	0.1496	0.1064	0.731	0.483	0.431	0.257	0.363
Fréq. cumul.	0.3088	0.5175	0.6671	0.7735	0.8466	0.8949	0.938	0.9637	1

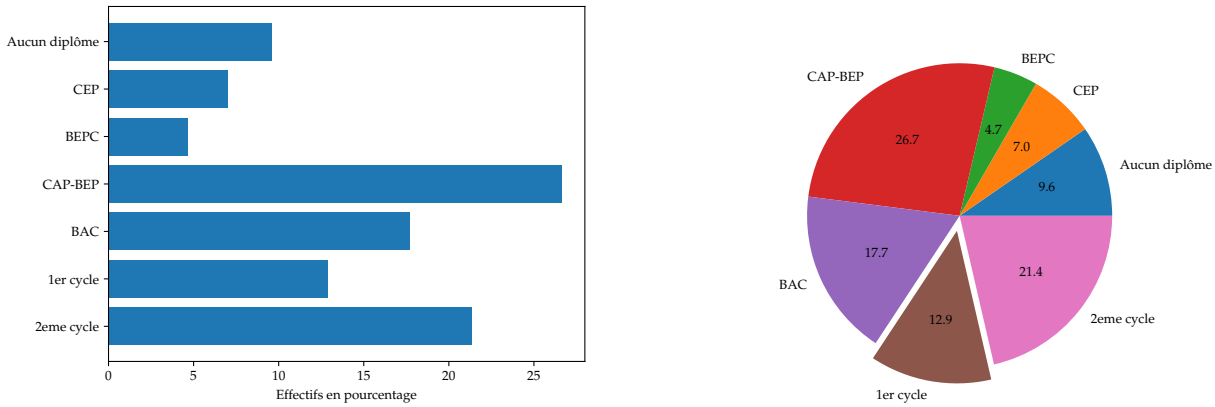
— les camemberts (ou *pie-chart* en anglais) : chaque catégorie est représentée par une portion de superficie proportionnelle à sa fréquence.

EXEMPLE 1.3 – DIPLÔMES EN LOIRE-ATLANTIQUE II – On représente les données de l'exemple 1.1 dans deux graphiques dans la Figure 1 :

— un diagramme en barres à gauche ;

— un diagramme en camembert à droite.

FIGURE 1 – Proportion des diplômes en Loire-Atlantique en 2017
Source : Insee, Base historique des recensements de la population, exploitation complémentaire



§ *Histogramme.* Analogue à la courbe de densité d'une variable aléatoire, un histogramme est un graphique à barres verticales accolées, obtenu après découpage en classes des observations d'une variable continue.

EXEMPLE 1.4 – DIPLÔMES EN LOIRE-ATLANTIQUE III – On compte le nombre de villes de Loire-Atlantique ayant un certain pourcentage de diplômés ayant le BAC comme dernier diplôme. On représente un certain nombre d'histogrammes correspondant à ces données à la figure 3.

2 – RÉSUMÉS NUMÉRIQUES

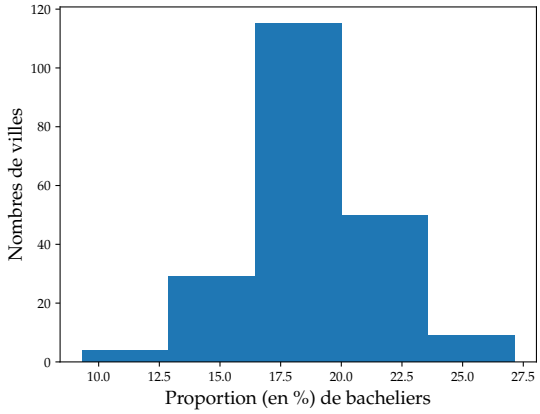
En général, il est indispensable de résumer une série d'observations par des indicateurs typiques.

2.1. **Caractéristiques de tendance centrale.** Il s'agit de définir une valeur c autour de laquelle se répartisse les observations. Les plus usitées sont la médiane et la moyenne empirique.

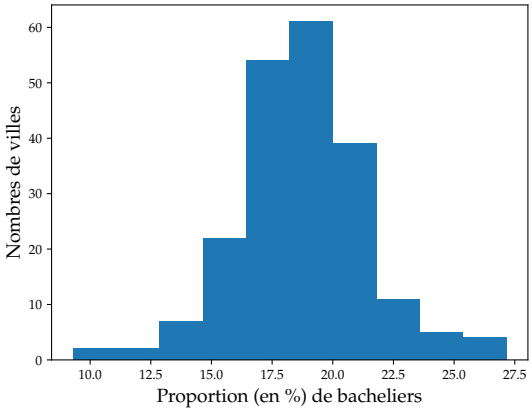
§ *Quantile, Médiane.* On commence par définir "l'équivalent statistique" de la fonction de répartition d'une variable aléatoire.

FIGURE 3 – Nombre de villes de Loire-Atlantique par proportion (en %) de bacheliers en 2017
Source : Insee, Base historique des recensements de la population, exploitation complémentaire

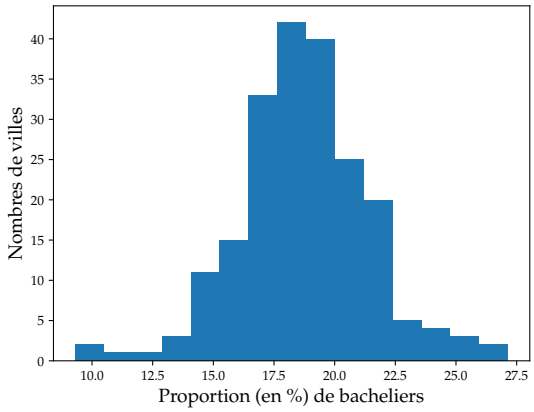
(A) 5 classes



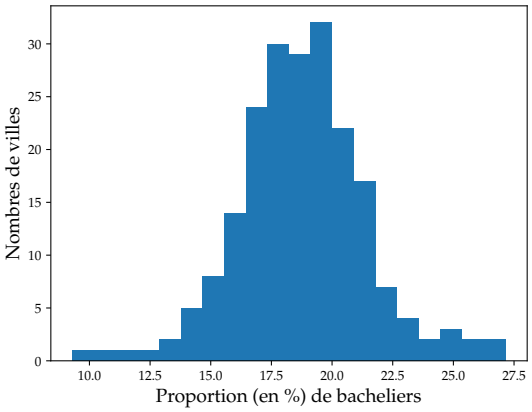
(B) 10 classes



(C) 15 classes



(D) 20 classes



DÉFINITION 2.1 *(Fonction de répartition empirique).* Soit $X : P \to \mathbb{R}$ une série statistique avec $\text{Card}(P) < +\infty$. La *fonction de répartition empirique de* X, que l'on note F_X, est la fonction :

$$\begin{array}{rccl} F_X : & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & \frac{1}{\text{Card}(P)} \text{Card}\left(\{i \in P \mid X(i) < x\}\right) \end{array}$$

REMARQUE 2.2 – C'est la fonction qui renvoie la proportion de valeurs inférieure à x dans l'ensemble $X(P)$ des valeurs prises par X.

DÉFINITION 2.3 *(Quantile).* Soient $p, q \in \mathbb{N}^*$ tels que $p < q$. Le *p-ième q-quantile* d'une série statistique X est "la" valeur $x_{(p/q)}$ telle que

$$F_X(x_{(p/q)}) = \frac{p}{q} .$$

Si $q = 4$, on parlera de *quartiles*, si $q = 10$, de *déciles* et si $q = 100$, de *centiles*.

REMARQUE 2.4 – En fait, cette définition n'est pas correcte, mais ce n'est pas très grave, il faut surtout retenir les définitions suivantes.

Définition 2.5 *(Médiane, Quartile).* Soit $X\colon P \to \mathbb{R}$ une série statistique avec $\mathrm{Card}(P) < +\infty$. La *médiane de* X est le 2-ème quartile de X. Si on note $X(P) = \{x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n\}$, alors

— si n est impair, alors la médiane de X est $x_{(n+1)/2}$;

— si n est pair, on prendra par convention pour médiane le nombre $\frac{x_{n/2}+x_{n/2+1}}{2}$.

Le *premier quartile* Q_1 de X est la plus petite valeur x telle que $F_X(x) \geqslant 0.25$. Le *troisième quartile* Q_3 de X est la plus petite valeur x telle que $F_X(x) \geqslant 0.75$.

Exemple 2.6 – Exemple jouet – On considère la série statistique X, résumée dans le tableau suivant.

Table 4 – Tableau de la série statistique X

Modalité X_i	0	1	2	3	4	6
Effectif n_i	26	23	12	36	2	1
Fréquence f_i	0.26	0.23	0.12	0.36	0.02	0.01

On a alors que la médiane de X est 2, le premier quartile Q_1 est 0 et le troisième quartile Q_3 est 3.

§ *Moyenne empirique.*

Définition 2.7 *(Moyenne empirique).* Soit $X\colon P \to \mathbb{R}$ une série statistique avec $\mathrm{Card}(P) < +\infty$. La *moyenne empirique de* X, que l'on note $\overline{X}$, est la valeur définie comme suit :

$$\overline{X} := \frac{1}{\mathrm{Card}(P)} \sum_{k \in P} X(k) \ .$$

Remarque 2.8 – On a trois manières de calculer la moyenne empirique d'une série statistique. On a les égalités suivantes :

$$\begin{aligned} \overline{X} &= \frac{1}{\mathrm{Card}(P)} \sum_{k \in P} X(k) && \text{calcul avec les données brutes} \\ &= \frac{1}{n} \sum_{i=1}^{n} n_i x_i && \text{effectifs par modalité} \\ &= \sum_{i=1}^{n} f_i x_i && \text{fréquence par modalité} \end{aligned}$$

On a vu que la moyenne empirique n'est bien définie que pour une variable discrète ; pour une variable continue, on utilise la notion de *moyenne approchée*

Définition 2.9 *(Moyenne approchée).* Soit X une variable (continue regroupée suivant les modalités $[x_1, x_2[$, $[x_2, x_3[, \ldots, [x_{n-1}, x_n]$. On pose, pour tout $i \in [\![1, n-1]\!]$,

$$c_i := \frac{x_i + x_{i+1}}{2}$$

le milieu de chaque modalité. La *moyenne approchée de* X, est définie par

$$\sum_{i=1}^{n} f_i c_i \ .$$

Attention. La moyenne approchée est seulement une approximation de la moyenne empirique !

Proposition 2.10 (Linéarité de la moyenne empirique). *Soient X et Y deux séries statistiques définies sur la même population P, avec* $\mathrm{Card}(P) < +\infty$*, et soient a et b deux nombres réels. On a*

$$\overline{aX + bY} = a\overline{X} + b\overline{Y} \ .$$

En particulier, on a $\overline{aX + b} = a\overline{X} + b$.

Démonstration. Soient X et Y deux séries statistiques et oient a et b deux réels. On a

$$\begin{aligned}
\overline{aX + bY} &= \frac{1}{\mathrm{Card}(P)} \sum_{k \in P} (aX(k) + bY(k)) \\
&= \frac{1}{\mathrm{Card}(P)} \left(a \sum_{k \in P} X(k) + b \sum_{k \in P} Y(k) \right) \\
&= a \left(\frac{1}{\mathrm{Card}(P)} \sum_{k \in P} X(k) \right) + b \left(\frac{1}{\mathrm{Card}(P)} \sum_{k \in P} Y(k) \right) \\
&= a\overline{X} + b\overline{Y}
\end{aligned}$$

□

2.2. **Caractéristiques de dispersion : variance et écart-type.**

Définition 2.11 *(Variance empirique et écart-type).* Soit X une série statistique de moyenne $\mu = \overline{X}$ définie sur une population P, avec $\mathrm{Card}(P) < +\infty$. On appelle *variance empirique de X* la valeur

$$\mathbb{V}(X) \coloneqq \overline{(X - \mu)^2} = \frac{1}{\mathrm{Card}(P)} \sum_{k \in P} \left(X(k) - \overline{X} \right)^2$$

et *l'écart-type de X* est la valeur

$$\sigma_X \coloneqq \sqrt{\mathbb{V}(X)} \ .$$

Théorème 2.12 – Théorème de Koenin

Soit X une série statistique définie sur une population P avec $\mathrm{Card}(P) < +\infty$*, alors*

$$\mathbb{V}(X) = \overline{X^2} - \overline{X}^2 = \frac{1}{\mathrm{Card}(P)} \sum_{k \in P} (X(k))^2 - \left(\overline{X} \right)^2 \ .$$

Démonstration. Soit X une telle variable statistique, et on note $\mu = \overline{X}$ sa moyenne empirique. Comme $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$, en utilisant la linéarité de la moyenne (cf. Proposition 2.10), on a alors que

$$\begin{aligned}
\mathbb{V}(X) &= \overline{(X - \mu)^2} \\
&= \overline{X^2 - 2\mu X + \mu^2} \\
&= \overline{X^2} - 2\mu\overline{X} + \mu^2 \\
&= \overline{X^2} - 2\overline{X} \cdot \overline{X} + \overline{X}^2 \\
&= \overline{X^2} - \overline{X}^2 \ .
\end{aligned}$$

□

PROPOSITION 2.13 (Deux indicateurs de dispersion). *Soit X une série statistique.*

1. *L'*inter-quartile *est la quantité $Q_3 - Q_1$: c'est la largeur de l'intervalle $[Q_1, Q_3]$ qui contient plus de 50% des données.*
2. *L'écart-type $\sigma_X := \sqrt{\mathbb{V}(X)}$ nous permet de définir l'intervalle $[\overline{X}-\sigma_X, \overline{X}+\sigma_X]$ où sont concentrées environ $2/3$ des données.*

2.3. **Un diagramme pour les gouverner tous : la boîte à moustache ou box-plot.** Ce type de diagramme, introduit par J.W. Tukey, est une représentation synthétique extrêmement efficace des principales caractéristiques d'une variable numérique.

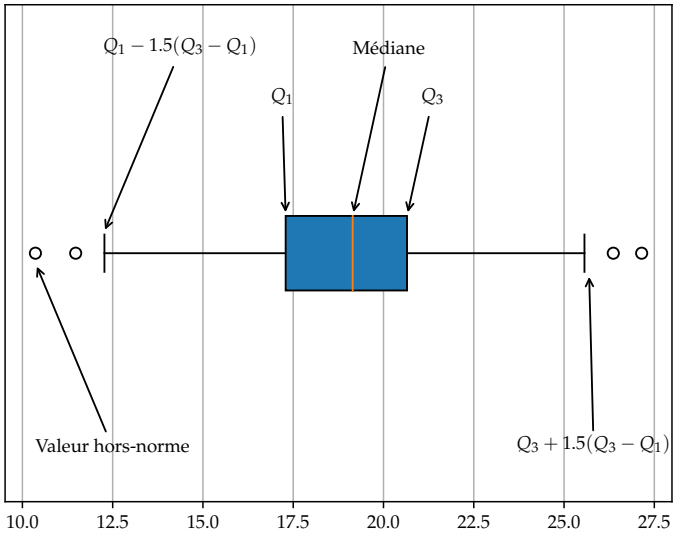
La boîte correspond à la partie centrale de la distribution : la moitié des valeurs comprises entre le premier et le troisième quartile Q_1 et Q_3. Les moustaches s'étendent de part et d'autres de la boîte jusqu'aux valeurs suivantes :

— à gauche jusqu'à $Q_1 - 1.5(Q_3 - Q_1)$ s'il existe des valeurs encore plus petites, sinon jusqu'à la valeur minimale ;

— à droite jusqu'à $Q_3 + 1.5(Q_3 - Q_1)$ s'il existe des valeurs encore plus grandes, sinon jusqu'à la valeur maximale.

Les valeurs au-delà des moustaches repérées par des points, sont des valeurs hors normes éventuellement suspectes ou aberrantes, mais pas nécessairement.

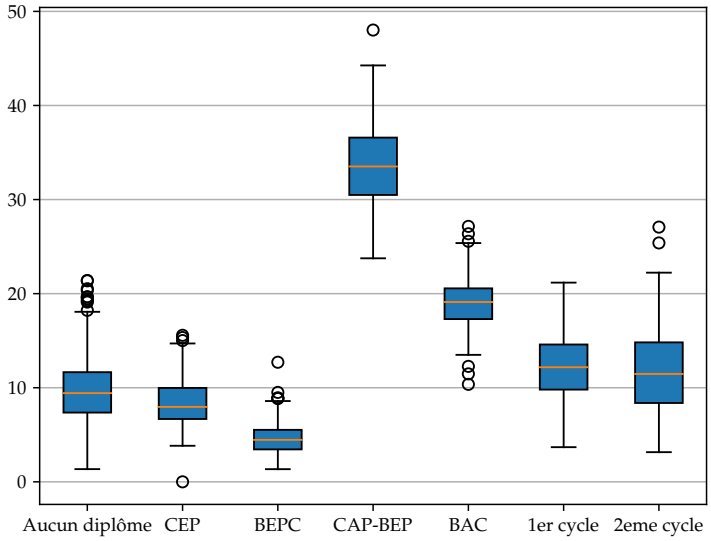
FIGURE 5 – Exemple de boîte à moustache : proportion (en %) de bacheliers dans les villes de 500 à 4500 diplômés en 2017
Source : Insee, Base historique des recensements de la population, exploitation complémentaire



EXEMPLE 2.14 – DIPLÔMES EN LOIRE-ATLANTIQUE IV – On représente dans la figure 6, la proportion de chaque classe de diplôme dans chaque ville de Loire-Atlantique ayant entre 1000 et 3000 diplômés.

FIGURE 6 – Proportion (en %) des diplômes dans les villes de Loire-Atlantique ayant entre 500 et 4500 diplômés en 2017

Source : Insee, Base historique des recensements de la population, exploitation complémentaire



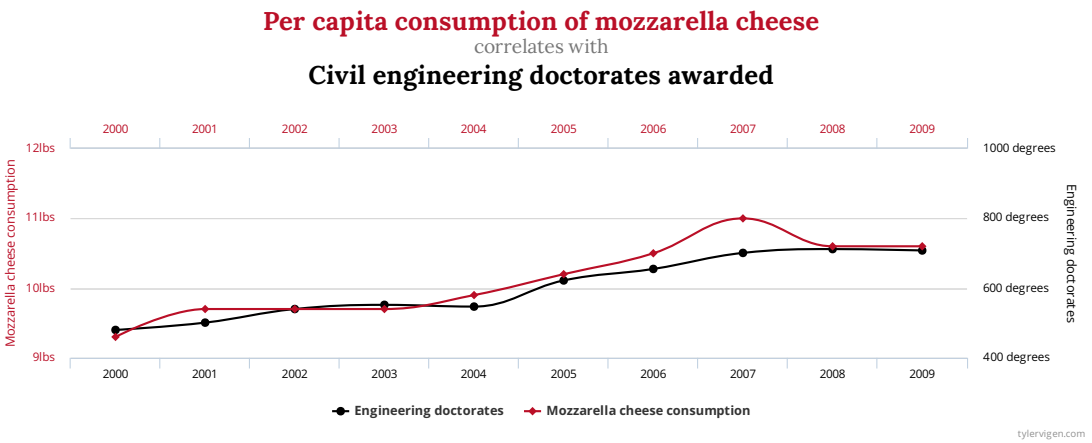
3 – Un peu de statistiques bidimensionnelles

On pourra consulter [Mol] ou [Sap06, Chapitre 16], dont cette sous-section est très fortement inspirée.

Les statistiques bivarirées permettent d'établir des liens entre différentes caractères sur une population. Ici, nous n'étudierons que les caractères quantitatifs (i.e. qui sont des nombres).

Attention. Les statistiques permettent de constater *une corrélation* entre plusieurs caractères, mais elles n'ont pas de légitimité pour discerner les causes de corrélation. On peut constater des corrélations entre de nombreux caractères ; on pourra par exemple consulter le site suivant https://www.tylervigen.com/spurious-correlations d'où la Figure 7 est issue.

Figure 7 – Corrélation fallacieuse



3.1. **Indicateurs et représentation graphique.**

Définition 3.1 *(Modalité conjointe).* Soient X et Y deux caractères quantitatifs sur une population, tels que X puisse prendre les valeurs $x_1, \dots, x_p$ et Y puisse prendre les valeurs $y_1, \dots, y_q$.

1. Un couple (x_i, y_j) s'appelle *une modalité conjointe* des caractères X et Y.
2. *L'effectif conjoint* de la modalité (x_i, y_j) est le nombre d'individus vérifiant simultanément les caractères $X = x_i$ et $Y = y_j$.
3. *L'effectif marginal* de x_i est le nombre d'individus vérifiant simultanément les caractères $X = x_i$ (indépendamment de la valeur du caractère Y).

Exemple 3.2 – Consommation des ménages (I) – Dans un certain nombre de cas, notamment si l'effectif de l'échantillon est petit, on résumera les données dans un tableau à deux lignes, où chaque colonne correspond à un individu de l'échantillon. En Table 6, on a un tableau indiquant le PIB et les dépenses de consommation des ménages en France de 1990 à 2002 en euros constants (base 1995). Ici, les individus de mon échantillon sont les différentes années et l'évaluation de chacun des caractères est donnée par une ligne du tableau. Cela est une représentation plus compacte de l'information qui évite de stocker les données dans un grand tableau rempli de 0 avec des 1 sur la diagonale.

TABLE 5 – Modalités des échantillons

X \ Y	y_1	y_2	$\cdots$	y_q	eff. marginal de x_i
x_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,q}$	$n_{1,*}$
x_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,q}$	$n_{2,*}$
$\vdots$					
x_i	$n_{i,1}$	$n_{i,2}$	$\cdots$	$n_{i,q}$	$n_{i,*} \coloneqq \sum_{j=1}^{q} n_{i,j}$
$\vdots$					
x_p	$n_{p,1}$	$n_{p,2}$	$\cdots$	$n_{p,q}$	$n_{k,*}$
eff. marginal de y_j	$n_{*,1}$	$n_{*,2}$		$n_{*,r}$	n

TABLE 6 – PIB français et dépenses de consommation des ménages en France de 1990 à 2002 en milliards euros (euros constants, base 1995)
Source : Insee

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
PIB	1121.0	1132.2	1149.1	1138.9	1162.4	1181.8	1194.9	1217.6	1259.1	1299.5	1348.8	1377.1	1393.7
Dépense	627.5	631.7	637.5	633.7	641.2	649.0	657.3	658.2	680.7	702.6	721.2	740.1	748.9

Plutôt que de s'intéresser aux données brutes des effectifs, on peut également renseigner les proportions des différentes modalités dans l'échantillon : cela est donné par la notion de fréquence.

DÉFINITION 3.3 *(Fréquence conjointe).* Soient X et Y deux caractères quantitatifs sur une population, tels que X puisse prendre les valeurs $x_1, \dots, x_p$ et Y puisse prendre les valeurs $y_1, \dots, y_q$.

1. *La fréquence conjointe* de la modalité (x_i, y_j) est donnée par
$$f_{i,j} = \frac{\text{effectif conjoint de la modalité } (x_i, y_j)}{\text{effectif total}} .$$
2. *La fréquence marginale* de x_i est donnée par
$$f_{i,*} = \frac{\text{effectif marginal de la modalité } x_i}{\text{effectif total}} .$$

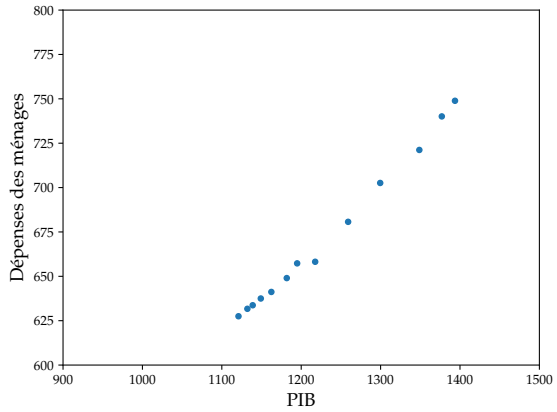
On peut représenter ce type de données par un nuage de points. On illustre cela par un exemple.

EXEMPLE 3.4 – CONSOMMATION DES MÉNAGES (II) – En Table 6, on a un tableau indiquant le PIB et les dépenses de consommation des ménages en France de 1990 à 2002 en euros constants (base 1995). On représente ces données dans la Figure 8. On peut constater un relatif alignement des points.

3.2. **Régression linéaire**.

Étant donné deux caractères quantitatifs évalués sur un échantillon d'une population, une question naturelle se pose : existe-t-il un lien entre ces deux caractères ? Dans cette section, on présente une manière d'exhiber une potentielle relation de linéarité entre deux caractères, c'est-à-dire qu'étant donné deux caractères X et Y, existe-t-il deux nombres α et β tels qu'on ait l'approximation $Y \approx \alpha X + \beta$? Et si oui, est-ce qu'une telle approximation est pertinente dans mon expérience statistique ?

FIGURE 8 – Représentation graphique des dépenses des ménages (en milliards d'euros) en fonction du PIB (en milliards d'euros)



§ *Point moyen.* On généralise la notion de moyenne qui existe en une dimension.

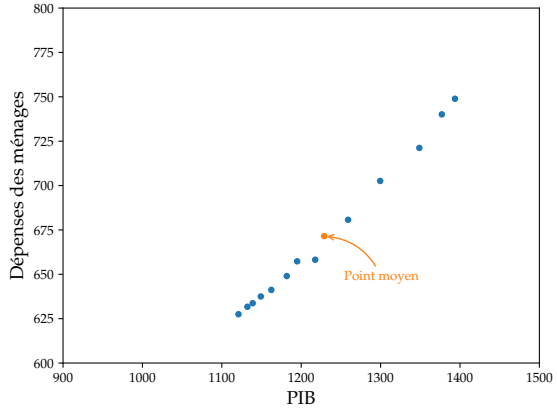
DÉFINITION 3.5 *(Point moyen).* Le *point moyen* d'une série statistique pour les caractères (X, Y) est le point de coordonnées

$$G = (\overline{X}, \overline{Y}),$$

où $\overline{X}$ et $\overline{Y}$ sont les moyennes empiriques de X et Y.

EXEMPLE 3.6 – CONSOMMATION DES MÉNAGES (III) – On calcule le point moyen pour la série statistique donnée en Table 6. On calcule le PIB moyen $P_{\text{moy}} \approx 1228.9$ et la dépense moyenne $D_{\text{moy}} \approx 671.5$: on peut alors placer le point moyen G de coordonnées $(P_{\text{moy}}, D_{\text{moy}})$ sur le graphique, comme sur la Figure 9.

FIGURE 9 – Représentation graphique des dépenses des ménages (en milliards d'euros) en fonction du PIB (en milliards d'euros) avec le point moyen



On voit sur la représentation graphique que les points ont l'air aligné. Nous allons à présent déterminer la droite "la plus proche" (en un certain sens que nous allons expliciter par la suite) du nuage de points.

§ *Coefficient de corrélation.*

Définition 3.7 *(Covariance empirique).* La *covariance empirique* de deux caractères (X, Y) d'une série statistique est définie par

$$\sigma_{X,Y} := \overline{(X - \overline{X})(Y - \overline{Y})} = \sum_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \frac{n_{i,j}}{N}(x_i - \overline{X})(y_j - \overline{Y})$$

où $\overline{X}$ et $\overline{Y}$ sont les moyennes empiriques de X et Y, $n_{i,j}$ est l'effectif conjoint de la modalité (x_i, y_j) et N est l'effectif total.

En pratique, on n'utilise jamais la définition de la covariance pour la calculer, mais la formule suivante.

Proposition 3.8 (Formule de Koenig–Huygens). *La covariance empirique de deux caractères (X, Y) d'une série statistique satisfait l'égalité suivante*

$$\sigma_{X,Y} = \overline{XY} - \overline{X} \cdot \overline{Y} \ .$$

De plus, on a

$$\sigma_{X,X} = \sigma_X^2 (= \mathbb{V}(X)) \ .$$

Démonstration. On a les égalités suivantes

$$\begin{aligned}
\sigma_{X,Y} :=& \overline{(X - \overline{X})(Y - \overline{Y})} \\
=& \sum_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \frac{n_{i,j}}{N}(x_i - \overline{X})(y_j - \overline{Y}) \\
=& \sum_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \frac{n_{i,j}}{N}(x_i y_j - \overline{X} y_j - \overline{Y} x_i + \overline{X} \cdot \overline{Y}) \\
=& \sum_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \frac{n_{i,j}}{N} x_i y_j - \overline{X} \cdot \overline{Y} - \overline{Y} \cdot \overline{X} + \overline{X} \cdot \overline{Y}) \\
=& \overline{XY} - \overline{X} \cdot \overline{Y}
\end{aligned}$$

d'où le résultat. On a directement par le théorème 2.12 que $\sigma_{X,X} = \overline{XX} - \overline{X} \cdot \overline{X} = \mathbb{V}(X) = \sigma_X^2$. □

Exemple 3.9 – Considérons les deux caractères suivants

X	-3	-1	1	2	3
Y	4	2	6	3	2

et calculons leur covariance. On a

$$\overline{X} = \frac{-3 + (-1) + 1 + 2 + 3}{5} = 0.4, \ \overline{Y} = \frac{4 + 2 + 6 + 3 + 2}{5} = 3.4$$

et

$$\overline{XY} = \frac{-3 \times 4 + (-1) \times 2 + 1 \times 6 + 2 \times 3 + 3 \times 2}{5} = 0.8$$

donc

$$\sigma_{X,Y} = \overline{XY} - \overline{X} \cdot \overline{Y} = 0.8 - 3.4 \times 0.4 = -0.56.$$

Définition 3.10 *(Coefficient de correlation linéaire).* Pour deux caractères (X, Y) d'écart-type $\sigma_X \neq 0$ et $\sigma_Y \neq 0$, on définit *le coefficient de corrélation* par

$$\rho_{X,Y} \coloneqq \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \ .$$

Proposition 3.11. *Le coefficient de corrélation est toujours compris dans l'intervalle* $[-1, 1]$.

Remarque 3.12 – Interprétation – Lorsque le coefficient de corrélation vaut ± 1, cela veut dire que X et Y sont alignés sur une droite d'équation $y = ax+b$. Par extension, lorsque le coefficient de corrélation se rapproche de ± 1, les points ont tendance à s'aligner. On dit alors que les caractères X et Y sont *corrélés linéairement.* Le résultat précédent provient de la nature géométrique de la covariance (c'est une sorte de produit scalaire).

Attention. La corrélation de deux caractères n'indique pas un lien de causalité entre eux (comme on l'a déjà évoqué en début de section 3).

§ *Droite de régression.*

Définition 3.13 *(Droite de régression).* La *droite de régression* d'un nuage de points pour les caractères (X, Y) est la droite passant par le point moyen et de coefficient directeur $\sigma_{X,Y}/(\sigma_X^2)$. Elle a pour équation

$$y = \frac{\sigma_{X,Y}}{\sigma_X^2} x + \left(\overline{Y} - \frac{\sigma_{X,Y}}{\sigma_X^2} \overline{X} \right) \ .$$

Théorème 3.14 – Régression linéaire

La droite de régression est l'unique droite d'équation $y = ax + b$ qui minimise la somme des carrées des distances verticales entre la droite et les points du nuage. Autrement dit, les nombres $a = \frac{\sigma_{X,Y}}{\sigma_X^2}$ et $b = \left(\overline{Y} - \frac{\sigma_{X,Y}}{\sigma_X^2} \overline{X} \right)$ minimisent la grandeur

$$\sum_{k=1}^{n} (y_k - (ax_k + b))^2 \ ,$$

où n est le nombre de points (x_k, y_k) du nuage de points.

Remarque 3.15 – On remarquera que lorsque l'on applique une régression linéaire, les deux variables X et Y ne jouent pas le même rôle. On a choisi ici d'exprimer Y en fonction de X. On dira que X est la variable explicative et que Y est la variable à expliquer.

Idée de preuve, pour les plus courageuses et courageux. Commençons par rappeler que si l'on considère une fonction $f : \mathbb{R} \to \mathbb{R}$ dérivable, les points critiques de f sont les points où sa dérivée s'annule. Ainsi, si f admet un minimum en $\alpha \in \mathbb{R}$, on a que $f'(\alpha) = 0$. Ainsi, pour trouver les extremums de f, on commence par chercher les points où la dérivée de f s'annule, puis on trouve les extremums parmi ces points-là. On va utiliser ici une stratégie similaire mais avec un niveau de généralisation en plus : on va devoir considérer une fonction à deux variables.

Soient (X, Y) deux caractères qui prennent respectivement les valeurs $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$. On considère la fonction à deux variables suivante

$$\begin{array}{rccc} F: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (a,b) & \longmapsto & \sum_{k=1}^{n} (y_k - ax_k - b)^2 . \end{array}$$

On cherche les nombres a et b telle que la quantité $F(a,b)$ soit minimal. On va montrer que cette fonction n'admet qu'un unique point critique. Un point critique d'une fonction à deux variables est un point pour lequel *son gradient* s'annule.

Comme F possède deux variables, on peut dériver F par rapport à l'une ou à l'autre. Le gradient de F, noté ∇F, est le vecteur qui contient ces deux dérivées. On peut l'interpréter comme la généralisation de la dérivée. La première composante de ce vecteur correspond à la dérivée par rapport à la variable a et la seconde, la dérivée par rapport à la variable b.

$$\begin{aligned} \nabla F(a,b) &= \left(\frac{\partial F}{\partial a}(a,b), \frac{\partial F}{\partial b}(a,b) \right) \\ &= \left(-2 \sum_{k=1}^{n} x_k (y_k - ax_k - b), -2 \sum_{k=1}^{n} (y_k - ax_k - b) \right). \end{aligned}$$

Ainsi, le gradient de F est nul, c'est-à-dire $\nabla F(a,b) = (0,0)$ si et seulement si on a

$$\begin{cases} \sum_{k=1}^{n} x_k (y_k - ax_k - b) &= 0, \\ \sum_{k=1}^{n} (y_k - ax_k - b) &= 0. \end{cases}$$

La deuxième équation de ce système est équivalente à $b = \overline{Y} - a\overline{X}$. En substituant b par $\overline{Y} - a\overline{X}$ dans la première équation, on obtient l'équation

$$\sum_{k=1}^{n} x_k \left(y_k - ax_k - \overline{Y} + a\overline{X} \right) = 0,$$

ce qui nous donne finalement que $\nabla F(a,b) = (0,0)$ si et seulement si

$$\begin{cases} \displaystyle\sum_{k=1}^{n} x_k (y_k - \overline{Y}) - a \sum_{k=1}^{n} x_k (x_k - \overline{X}) = 0 \\ b = \overline{Y} - a\overline{X} \end{cases}$$

qui est équivalent au système

$$\begin{cases} \displaystyle\sum_{k=1}^{n} \left(x_k - \overline{X} \right) \left(y_k - \overline{Y} \right) - a \sum_{k=1}^{n} \left(x_k - \overline{X} \right)^2 = -\overline{X} \left(\sum_{k=1}^{n} y_k - \overline{Y} - a(x_k - \overline{X}) \right) \\ b = \overline{Y} - a\overline{X} \end{cases}.$$

Finalement, comme le terme de droite de la première équation est nul car $\frac{\partial F}{\partial b}(a,b) = 0$ par hypothèse, on a donc que $\nabla F(a,b) = (0,0)$ si et seulement si

$$\begin{cases} \displaystyle\sum_{k=1}^{n} \left(x_k - \overline{X} \right) \left(y_k - \overline{Y} \right) = a \sum_{k=1}^{n} \left(x_k - \overline{X} \right)^2 \\ b = \overline{Y} - a\overline{X} \end{cases}.$$

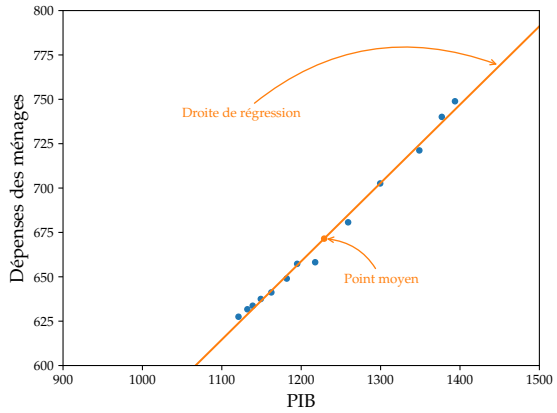
Ainsi, on montre que le couple (a, b), avec $a = \frac{\sigma_{X,Y}}{\sigma_X^2}$ et $b = \left(\overline{Y} - \frac{\sigma_{X,Y}}{\sigma_X^2}\overline{X}\right)$, est l'unique point critique de F. Il ne nous reste plus qu'à montrer que ce point critique est un minimum pour F, ce que l'on admettra ici. □

Exemple 3.16 – Consommation des ménages (IV) – On reprend la série statistique donnée en Table 6. On a

$$\frac{\sigma_{X,Y}}{\sigma_X^2} \approx 0.44 \quad \text{et} \quad \overline{Y} - \frac{\sigma_{X,Y}}{\sigma_X^2}\overline{X} \approx 129.14 \ .$$

On peut alors tracer la droite de régression, comme en Figure 10.

Figure 10 – Représentation graphique des dépenses des ménages (en milliards d'euros) en fonction du PIB (en milliards d'euros) et sa droite de régression



3.3. **Interprétation géométrique de la régression linéaire.** Cette section a pour but de donner une explication géométrique à différentes quantités introduites dans les sections précédentes. On considère une série statistique à deux variables (X, Y) avec $X = (x_1, \dots, x_n)$ et $Y(y_1, \dots, y_n)$ deux vecteurs de $\mathbb{R}^n$. On note également $\mathbb{1} := (1, \dots, 1) \in \mathbb{R}^n$. On "rappelle" que l'espace $\mathbb{R}^n$ est muni d'un produit scalaire

$$\langle X, Y \rangle := \sum_{k=1}^{n} x_k y_k,$$

qui nous permet de définir la norme associée $\|X\| := \sqrt{\langle X, X \rangle}$ (ce nombre correspond à la longueur du vecteur X), et qu'elle nous permet de définir la *distance euclidienne de* $\mathbb{R}^n$; ainsi, la distance entre X et Y sera donnée par $\|X - Y\|$.

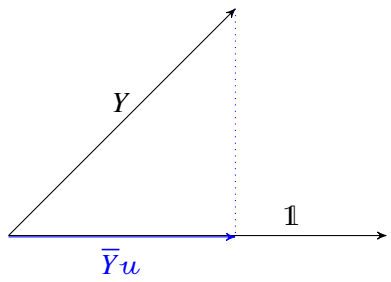
Commençons par donner une interprétation géométrique de la moyenne. Considérons le vecteur $Y \in \mathbb{R}^n$. Sa projection orthogonale pour le produit scalaire précédemment rappelé sur la droite vectorielle de vecteur directeur $\mathbb{1}$ est donné par

$$\frac{\langle Y, \mathbb{1} \rangle}{\|\mathbb{1}\|^2}\mathbb{1} = \frac{\sum_{k=1}^{n} y_k}{n^2}\mathbb{1} = \overline{Y}u,$$

où u est le vecteur unitaire colinéaire et de même sens que $\mathbb{1}$. On donne une représentation de cette projection orthogonale en Figure 11.

À présent, on va donner une interprétation de la régression linéaire. On va considérer le plan vectoriel de $\mathbb{R}^n$ engendré par $\mathbb{1}$ et X (en général, X et $\mathbb{1}$ ne sont pas colinéaires), que l'on note $\mathcal{P}(X)$. Ainsi, pour tout vecteur v de $\mathcal{P}(X)$, il existe deux nombres α et β tels que $v = \alpha\mathbb{1} + \beta X$.

FIGURE 11 – Représentation de la projection orthogonale de Y sur $\mathbb{1}$



On considère $\widetilde{Y}$, le projeté orthogonal de Y sur ce plan $\mathcal{P}(X)$, qui possède donc des coordonnées a et b tels que

$$\widetilde{Y} = aX + b\mathbb{1}.$$

De plus, on sait que ce vecteur est l'unique vecteur de $\mathcal{P}(X)$ qui minimise la distance avec Y, c'est-à-dire que

$$\|Y - \widetilde{Y}\| = \min_{v \in \mathcal{P}(X)} \|Y - v\|.$$

Or, par définition de la norme euclidienne, déterminer $\widetilde{Y}$ revient à déterminer le couple $(a, b) \in \mathbb{R}^2$ tel que la quantité

$$\|Y - \widetilde{Y}\| = \sum_{k=1}^{n} (y_k - (ax_k + b))^2$$

soit minimale. On retrouve bien le problème que l'on a traité dans la preuve du théorème 3.14.

RÉFÉRENCES

[Mol] M. Molin. Cours BCPST – Statistiques. https://molin-mathematiques.fr/sources/cours90.php.

[Sap06] G. Saporta. *Probabilités, analyse des données et statistique.* Editions Technip, 2006.

Email address: johan.leray@univ-nantes.fr

DÉPARTEMENT D'INFORMATIQUE – IUT DE NANTES