INSTITUT Mines-Télécom

Statistique pour ingénieur

Thème 0 : Statistique descriptive

F. Delacroix & M. Lecomte, 3 novembre 2016

Introduction générale

Ce cours porte sur les notions de statistique utiles pour l'ingénieur. Tout d'abord, **qu'est-ce que la statistique ?** Ce mot désigne à la fois un ensemble de données observées et les méthodes de recueil, de traitement et d'analyse de celles-ci. Par exemple, les relevés des nombres de pannes observées dans une unité de production constituent une statistique.

Les concepts développés en statistique sont utiles dans de nombreux domaines et font partie des connaissances de base de l'ingénieur, de l'économiste et du scientifique en général. Parmi les nombreuses applications dans l'industrie, on peut citer la fiabilité, le contrôle de qualité, la maîtrise statistique des procédés, *etc.*

Ce cours comprend *grosso modo* quatre parties déclinées en cinq thèmes :

- la **statistique descriptive** permet de décrire les données à l'aide de graphiques et de paramètres d'une façon compréhensible et utilisable. Cette partie sera principalement développée au sein du thème 0, objet du présent poly.
- Les concepts de base en **probabilités** sont essentiels pour modéliser efficacement les phénomènes étudiés en statistique. Ils seront traités dans le thème 1.
- La **statistique inférentielle** permet de faire des prévisions ou des généralisations à tout une population à partir d'échantillons. Elle sera développée dans les thèmes 2 et 3, respectivement dédiés à l'estimation et aux tests statistiques, et repose essentiellement sur la théorie des probabilités.
- La **régression linéaire**, développée dans le thème 4, permet d'étudier la relation existant entre deux variables. Elle met en place des modèles de prévision et des outils pour valider ceux-ci.

«La statistique» ou «les statistiques» ?

Au pluriel, «les statistiques» désignent des grandeurs (généralement numériques) que l'on calcule, ou que l'on est capable de calculer (selon une définition précise qui sera donnée lors du thème 2), sur un ensemble de données observées. *A contrario*, au singulier, «la statistique» est le nom de la science qui étudie ces grandeurs et propose des outils pour les concevoir.

Ce cours de «statistique pour ingénieur» a donc pour finalité d'initier à cette science permettant la mise en place de techniques rigoureuses pour l'étude de données en grands volumes (statistique descriptive) ou seulement partiellement connues (statistique inférentielle et régression linéaire).

Statistique : généralités

Dans ce premier thème, nous abordons les notions générales. Dans la démarche statistique, il s'agit, dans un premier temps, de décrire, présenter et résumer les données sous la forme de tableaux et de graphiques, puis de calculer certains paramètres (moyenne, écart-type, *etc.*) pour les caractériser. Nous entrons donc de plain-pied dans la statistique descriptive.

Table des matières

1 Vocabulaire de la statistique

En statistique, on utilise les notions de la théorie des ensembles avec un vocabulaire spécifique, comme résumé dans la table 1

Le terme «population» s'applique à des ensembles de toute nature : habitants d'une ville ou d'un pays, production d'une usine, entreprises d'un secteur donné, *etc.*

La collecte d'informations sur une population peut porter sur la totalité des individus : on parle alors d'**enquêtes exhaustives**. Dans le cas où l'effectif de la population est élevé, de telles enquêtes sont trop coûteuses voire impossibles à réaliser. On a alors recours aux enquêtes par sondage, qui portent sur une partie de la population qu'on nomme **échantillon**. Les observations faites sur l'échantillon peuvent alors, grâce aux outils de la

Vocabulaire ensembliste	Vocabulaire statistique
Ensemble	Population
Application	Variable ou caractère
Élément	Individu ou unité statistique
Sous-ensemble	Sous-population
Cardinal	Effectif

TABLE 1 – vocabulaire statistique courant

statistique inférentielle, s'étendre à toute la population, comme on le verra dans le thème 2 dédié à l'estimation.

L'effectif d'une population finie Ω est souvent noté $\operatorname{Card}\Omega$, plus rarement $|\Omega|$ ou $\#\Omega$.

Définition 1

*La **fréquence** d'une sous-population E de Ω est le rapport des effectifs de E et de Ω :*

$$f(E) = \frac{\operatorname{Card}(E)}{\operatorname{Card}(\Omega)} \in [0,1].$$

Cette fréquence est souvent exprimée sous forme de pourcentage.

La synthèse des données se fait très souvent sous la forme de tableaux, graphiques et de résumés numériques comme on va le voir dans les paragraphes suivants.

Exemple 1

Si l'on considère la population en 2007 des entreprises en France (hors micro-entreprise et hors secteur financier) et si l'on répartit celles-ci selon leur taille, on obtient la table 2.

Taille de l'entreprise	Effectif	Fréquence
PME Petites et Moyennes Entreprises	162 400	97,2%
ETI Entreprises de Taille Intermédiaire	4 510	2,7%
Grandes entreprises	219	0,1%
Total	167129	100%

TABLE 2 – données de l'exemple 1 (source : INSEE)

Les sous-populations issues du regroupement des données comme dans l'exemple 1 sont souvent nommées **classes**. Ce qui ressort de cet exemple est que 97,2% des entreprises en France en 2007 étaient des PME. Les fréquences calculées définissent la **loi empirique**, ou **distribution empirique**, de la variable étudiée, comme on le verra à la section 4.

2 Statistique et probabilités

La théorie des probabilités joue un rôle important en statistique car elle permet de modéliser certains phénomènes aléatoires, c'est-à-dire des expériences dont le résultat ne peut pas être prévu avec une totale certitude. Des développements spécifiques seront consacrés à cette théorie lors du thème 1.

Prenons quelques exemples pour illustrer la relation entre probabilités et statistique. L'intuition nous amène à penser que certains phénomènes obéissent à certaines lois. Par exemple, si on jette 6000 fois un dé bien équilibré, on s'attend à ce que le nombre d'apparitions de la face «6» soit voisin de 1000.

Autre exemple : une unité de production fabrique des tiges métalliques en grande quantité. On mesure les longueurs de 100 tiges choisies au hasard. On peut penser que les valeurs observées seront concentrées autour d'une certaine «valeur moyenne». Dans ce cas, on considère assez souvent que les valeurs observées se distribuent selon un certain modèle, une certaine loi, par exemple la **loi normale** (*cf.* figure 1). Cette hypothèse peut être confortée par un test d'ajustement (voir le thème 3 portant sur les tests statistiques).

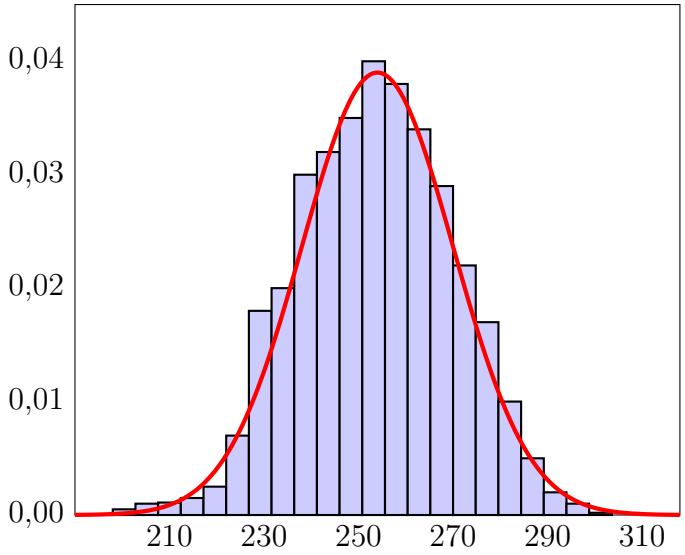

FIGURE 1 – Distribution empirique de 100 mesures de longueur

La théorie des probabilités permet de formaliser ces considérations un peu vagues. La statistique permet de confronter les modèles probabilistes avec la réalité observée afin de les valider ou de les invalider.

Les probabilités jouent aussi un rôle important dans la théorie de l'estimation qui a pour objet d'étendre les propriétés observées sur l'échantillon à toute la population. En effet, les échantillons d'individus sont la plupart du temps choisis **au hasard** au sein de la population. Par conséquent, les caractéristiques observées sur l'échantillon deviennent, par le biais de ce hasard, des variables aléatoires et la théorie des probabilités permet d'en étudier les propriétés.

Par exemple, le théorème central limite étudié au thème 1 permet d'établir que la moyenne d'une variable numérique mesurée sur n individus suit approximativement une **loi normale** pourvu que n soit suffisamment grand.

3 Variables ou caractères

En statistique, la population, généralement notée Ω, est un ensemble d'éléments définis sans ambiguïté, ces éléments étant appelés **individus**. La population constitue l'univers de référence lors d'une étude statistique.

Exemple 2

On se propose d'étudier les pièces produites en série dans une usine. On définit la population Ω comme l'ensemble de toutes les pièces produites pendant une certaine période.

Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées **variables** ou **caractères**. Certaines variables sont **qualitatives**, s'exprimant par l'appartenance à une catégorie : par exemple, dans le cadre de l'exemple 2, le caractère défectueux ou non d'une pièce.

D'autres variables sont **quantitatives** (ou : **numériques**). Par exemple la taille, le poids, le volume, la durée de vie sont des variables quantitatives. Une variable quantitative est qualifiée de **discrète** dans le cas où l'on observe un nombre fini ou infini dénombrable de valeurs.

Exemple 3

Le nombre de défauts observés sur une pièce produite dans un atelier est une variable discrète.

Étant donnée une variable discrète X, l'ensemble des valeurs (ou : modalités) prises par X est l'ensemble

$$X(\Omega) = \{x_1, x_2, \ldots, x_n \ldots\} = \{x_i, i \in \mathbb{N}^*\}.$$

Si on note n_i le nombre d'occurrences de x_i dans toute la population, et $n = \text{Card}(\Omega)$ la taille de la population, alors la fréquence correspondante est

$$f_i = \frac{n_i}{n}.$$

Exemple 4

La table 3 donne la répartition, selon les jours de la semaine, des 155 pannes observées dans une unité de production pendant une année.

Jour x_i	**Nombre de pannes** n_i	**Fréquence** f_i
lundi	45	29%
mardi	36	23%
mercredi	39	25%
jeudi	20	13%
vendredi	15	10%
Total	155	100%

TABLE 3 – répartition selon le jour de la semaine de 155 pannes observées dans l'exemple 4

On représente les effectifs ou les fréquences par des diagrammes adaptés :

— **Diagrammes en bâtons** (ou **à barres**) : l'effectif ou la fréquence correspondant à chaque valeur du caractère est représenté par la longueur d'un segment ou d'un rectangle de largeur constante. La représentation de plusieurs séries de données sur un même graphique peut se faire en empilant les barres comme à la figure 2.

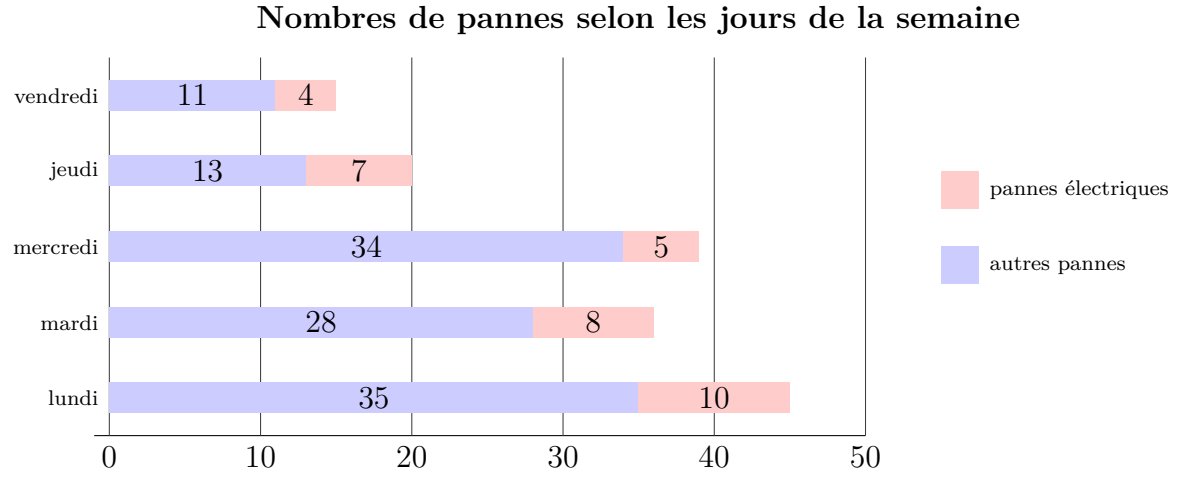

FIGURE 2 – Diagramme en bâtons (données de l'exemple 4)

— **Diagramme circulaire** : chaque valeur ou classe est représentée par un secteur angulaire d'un disque dont l'angle (et donc la surface) est proportionnel à sa fréquence : voir figure 3

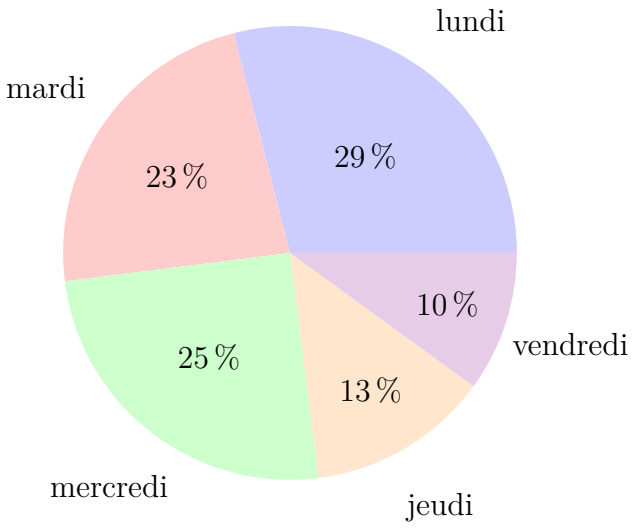

FIGURE 3 – Diagramme circulaire (données de l'exemple 4)

Un caractère est dit **continu** lorsque les valeurs qu'il prend constituent un intervalle de $\mathbb{R}$. Dans ce cas, il est fréquent de diviser la population en classes selon les intervalles de valeurs prises par le caractère. Ce procédé est parfois appelé **discrétisation** de la variable.

Exemple 5

Une entreprise fabrique des composants électroniques. La durée de vie X d'un composant est une variable continue.

Dans ce cas, on regroupe les valeurs observées en k classes d'extrémités e_0,e_1, . . . ,e_k ; et on note pour chaque classe $[e_{i-1},e_i[$ l'effectif n_i et la fréquence f_i, ainsi que les fréquences cumulées

$$F_i = \sum_{j=1}^{i} f_j.$$

On peut alors remarquer que F_i est la proportion d'individus pour lesquels $X < e_i$.

Exemple 6

Une entreprise fabrique des axes de roue dont le diamètre X est une variable continue. La table 4 donne la répartition en classes de 500 axes de roues selon leur diamètre (unité : dixième de millimètre).

Classe	Effectif	Fréquence (%)	Fréquence cumulée (%)
[244 ;248]	72	14,4	14,4
]248 ;250]	146	29,2	43,6
]250 ;252]	206	41,2	84,8
]252 ;254]	69	13,8	98,6
]254 ;258]	7	1,4	100
Total	500	100	

TABLE 4 – Valeurs observées du diamètre de 500 axes de roues (exemple 6)

On peut représenter cette série de données par un **histogramme** : chaque classe est représentée par un rectangle dont l'*aire* est proportionnelle à l'effectif, comme illustré à la figure 4.

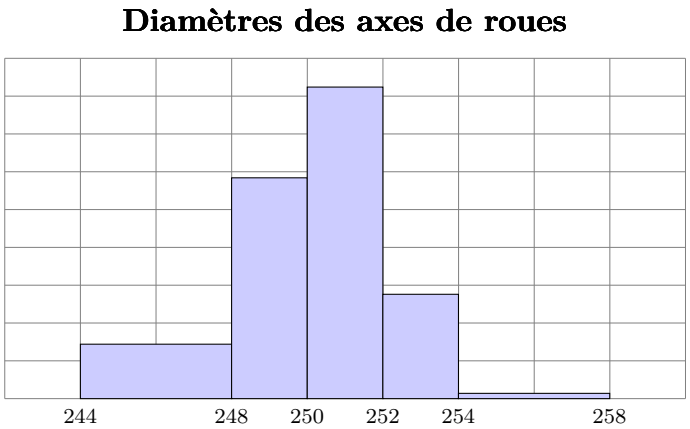

FIGURE 4 – Histogramme de distribution (données de l'exemple 6)

4 Loi d'une variable quantitative, fonction de répartition

La **loi**, ou **distribution empirique**, d'une variable X sur une population Ω est la donnée de la fréquence de chaque classe définie par la variable X.

— Si X est qualitative, ou quantitative discrète, sa loi est définie par la fréquence de chaque sous-population du type $\{X = x_i\} = \{\omega \in \Omega, X(\omega) = x_i\}$;

— si X est continue et si les valeurs possibles de X sont réparties en classes C_i, la loi empirique de X est la donnée de chaque fréquence des sous-populations $\{X \in C_i\} = \{\omega \in \Omega, X(\omega) \in C_i\}$.

Les exemples 4 et 6 illustrent ce concept, respectivement dans le cas d'une variable qualitative et d'une variable continue dont les valeurs sont regroupées en classes.

Dans le cas d'une variable *quantitative*, la fonction de répartition de X est un outil essentiel qui trouvera son prolongement en probabilités.

Définition 2

*La **fonction de répartition empirique de** X est la fonction, notée F_X, qui à $x \in \mathbb{R}$ associe la fréquence de la sous-population $\{X \leqslant x\}$:*

$$\begin{array}{rccc} F_X : & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & F_X(x) = \dfrac{\mathrm{Card}\{\omega \in \Omega, X(\omega) \leqslant x\}}{\mathrm{Card}\,\Omega}. \end{array}$$

On observera que la définition 2 définit toujours F_X sur $\mathbb{R}$ tout entier, même en des valeurs qui ne sont *a priori* pas des valeurs possibles de X (auquel cas la sous-population $\{X \leqslant x\}$ a tout de même un sens, et est éventuellement vide).

Une fonction de répartition empirique est toujours *croissante sur* $\mathbb{R}$, *de limites nulle en* $-\infty$ *et* 1 *en* $+\infty$. La population Ω étant finie, il s'agit en réalité toujours d'une *fonction en escaliers*. En pratique, pour une population «assimilable» à une population infinie et un caractère X continu, la fonction de répartition F_X est elle-même assimilée à une fonction continue croissante de 0 à 1.

La figure 5 illustre ceci sur les données de l'exemple 6. Bien entendu, comme on n'a que des données partielles, les points calculables de la fonction de répartition empirique sont «reliés» entre eux (interpolés) de façon plausible pour donner effectivement une fonction continue.

5 Grandeurs statistiques usuelles

Intéressons-nous à une variable quantitative X dont on possède, dans le cas discret, n valeurs notées $x_1, \ldots, x_n$. Si X est continue, on dispose couramment d'une discrétisation des données en k classes qui sont, en général, des intervalles de $\mathbb{R}$. On notera ces classes $[e_{i-1}, e_i[$.

Le calcul de grandeurs caractéristiques liées à la distribution empirique de X permet souvent d'en résumer des informations essentielles. Présentons quelques unes de ces grandeurs, en distinguant les **paramètres de position** (également parfois appelés **paramètres de tendance centrale**) et les **paramètres de dispersion**.

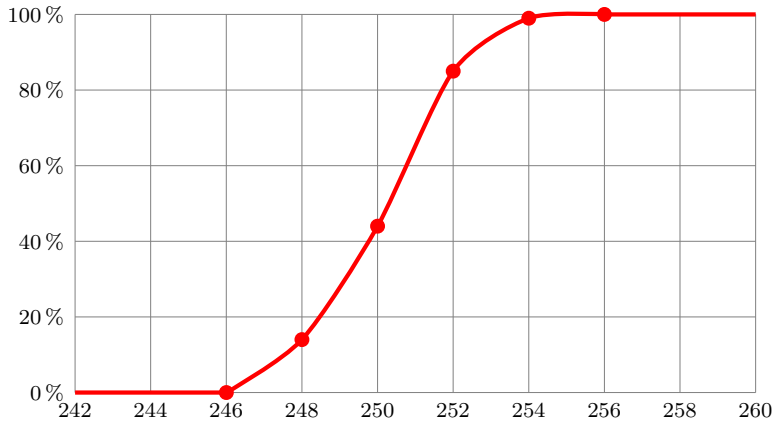

FIGURE 5 – Allure du graphe de la fonction de répartition de la variable de l'exemple 6

5.1 Paramètres de position

Définition 3 (*Moyenne*)

*La **moyenne** du caractère X est la quantité* $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

La moyenne est le paramètre le plus utilisé. Facile à calculer, elle a cependant l'inconvénient d'être très sensible au retrait ou à l'ajout de valeurs extrêmes ou «aberrantes». On dit que c'est une statistique *peu robuste*.

Exemple 7

Le tableau ci-dessous présente la série statistique donnant les primes annuelles (en k€) perçues par 11 cadres d'une PME.

Prime (k€)	5	4	3	6	7
Effectif	2	2	2	1	4

La prime moyenne touchée par les cadres est alors :

$$\overline{x} = \frac{1}{11}(2\times 5 + 2\times 4 + 2\times 3 + 1\times 6 + 4\times 7) = \frac{58}{11} \simeq 5{,}273.$$

La prime moyenne est donc de 5273€.

La proposition suivante, facile à démontrer, est une propriété très importante de la moyenne.

Proposition 1 (*Transformation affine de la moyenne*)

Si on effectue une transformation affine de la variable X, alors la moyenne $\overline{x}$ subit la même transformation affine.

Autrement dit, si $Y = a\,X + b$ avec a,$b \in \mathbb{R}$, alors

$$\overline{y} = a\,\overline{x} + b.$$

Dans le cas d'une variable continue, on fait généralement l'hypothèse que la répartition des observations est uniforme dans chaque classe. Alors la valeur moyenne des observations dans la classe $[e_{i-1},e_i[$ est $x_i = \frac{1}{2}(e_{i-1} + e_i)$.

Dans le cas où il y a k classes, on peut alors calculer la moyenne $\overline{x}$ sous la forme

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{n} f_i x_i$$

où n_i et f_i désignent respectivement l'effectif et la fréquence de la classe $[e_{i-1},e_i[$.

Définition 4

*Le **mode**, ou la **classe modale**, d'une distribution statistique est la valeur (ou la classe) du caractère qui correspond à la plus grande fréquence.*

Dans l'exemple 7, le mode est 7, correspondant à une prime annuelle de 7000€.

La moyenne est un paramètre peut robuste, c'est-à-dire qu'elle est sensible aux valeurs aberrantes (trop petites ou trop grandes). Pour éliminer le rôle des valeurs aberrantes, on définit un autre paramètre de position : la médiane. Intuitivement, il s'agit d'une valeur, notée M_e, qui partage la distribution statistique en deux sous-populations d'effectifs égaux. Plus précisément, on a la définition suivante.

Définition 5 (*Médiane*)

*On appelle **médiane** du caractère X tout nombre M_e tel que la fréquence de la sous-population $\{X \leqslant M_e\}$ est supérieure ou égale à $\frac{1}{2}$ et tel que la fréquence de la sous-population $\{X \geqslant M_e\}$ est elle aussi supérieure ou égale à $\frac{1}{2}$: en notant $f(\cdots)$ la fréquence,*

$$f(\{X \leqslant M_e\}) \geqslant \frac{1}{2} \quad \text{et} \quad f(\{X \geqslant M_e\}) \geqslant \frac{1}{2}.$$

En pratique, après avoir classé les observations dans le sens croissant, la médiane est la valeur de l'observation qui se trouve au rang $\frac{n+1}{2}$ si n est impair.

Si n est pair ($n = 2p$) on choisit, par convention, le milieu de l'intervalle $[x_p,x_{p+1}]$.

Reprenons les données de l'exemple 7, et classons-les par ordre croissant, chacune répétée un nombre de fois égal à son effectif :

$$\begin{array}{cccccccccccc} 3, & 3, & 4, & 4, & 5, & 5, & 6, & 7, & 7, & 7, & 7 \\ & & & & & \uparrow & & & & & \\ & & & & & M_e & & & & & \end{array}$$

On obtient une médiane M_e égale à 5, soit une prime médiane de 5000€.

Remarque

La médiane est plus robuste que la moyenne mais ses propriétés la rendent plus difficile à utiliser.

Définition 6

*Les **quartiles** Q_1, Q_2 et Q_3 sont, de manière analogue, des valeurs permettant de diviser la population en quatre sous-populations d'effectifs égaux, représentant chacune 25% de la population totale.*

Il existe des méthodes différentes pour obtenir les quartiles. Dans le cas d'une variable discrète, la méthode la plus courante consiste à déterminer les médianes de chacune des deux sous-populations délimitées par la médiane M_e pour obtenir les quartiles Q_1 et Q_3.

Dans le cas des données de l'exemple 7, on obtient les valeurs suivantes :

$$\begin{array}{cccccccccccc} 3, & 3, & 4, & 4, & 5, & 5, & 6, & 7, & 7, & 7, & 7 \\ & & \uparrow & & & \uparrow & & & \uparrow & & \\ & & Q_1 & & & M_e = Q_2 & & & Q_3 & & \end{array}$$

Le premier quartile est $Q_1 = 4$, ce qui signifie qu'au moins 25% des cadres de cette entreprise ont une prime inférieure ou égale à 4000€.

Remarque

*La **distance interquartile** $|Q_3 - Q_1|$ est un indicateur parfois utilisé pour mesurer la dispersion des données autour de la médiane.*

De la même façon, étant donné un entier $p \geqslant 2$, on peut définir les **quantiles**, qui sont les valeurs du caractère permettant de diviser la population en p sous-populations d'effectifs égaux. par exemple, les **déciles** d'une série statistique partagent la série en dix parties de même effectif. En pratique, seuls les premier et dernier déciles, respectivement notés D_1 et D_9, sont utilisés.

Remarque

Dans le cas d'une variable continue, on peut utiliser la fonction de répartition empirique F_X pour déterminer les quantiles. Par exemple, si la fonction F est continue et strictement croissante, alors elle réalise une bijection de $\mathbb{R}$ sur $]0,1[$, et on a

$$Q_1 = F_X^{-1}(0{,}25) \qquad M_e = F_X^{-1}(0{,}5) \qquad Q_3 = F_X^{-1}(0{,}75).$$

Dans la pratique, on obtient des valeurs approchées de ces quantités en procédant à une interpolation linéaire.

Une fois que les quartiles et déciles ont été calculés, nous pouvons représenter les données de façon synthétique à l'aide d'une **boîte à moustaches** (en anglais : **boxplot**). La partie centrale de la boîte est constituée d'un rectangle dont la longueur est la distance interquartile $|Q_3 - Q_1|$. Les «moustaches» sont des segments qui s'étendent de part et d'autre de la boîte jusqu'au premier décile D_1 pour la moustache inférieure, jusqu'au dernier décile D_9 pour la moustache supérieure. On dit alors que les *moustaches sont coupées*. Les moustaches non coupées, plus rares, consistent à aller jusqu'au minimum et au maximum de la distribution (qui peuvent être des valeurs aberrantes).

L'intérêt des boîtes à moustaches réside dans le fait qu'elles permettent de comparer aisément (visuellement) *plusieurs* séries de données comme dans l'exemple 8 ci-dessous.

Exemple 8

La table 5 regroupe des données relatives aux salaires annuels nets moyens dans l'industrie selon la catégorie socio-professionnelle en 2011. La figure 6 représente ces données aux fins de comparaison des différentes catégories socioprofessionnelles de ce secteur.

	Cadres[1]	Professions intermédiaires	Employés	Ouvriers	Ensemble
D1	28 900	19 370	13 380	14 650	15 290
Q1	35 560	22 810	15 400	16 710	18 110
Me	43 910	27 180	18 960	19 620	22 770
Q3	56 970	32 710	23 360	23 420	30 850
D9	76 870	39 210	28 540	27 920	43 790
D9/D1	2,7	2,0	2,1	1,9	2,9
Moy	50 600	28 670	20 310	20 780	27 450

Champ : France, salariés en EQTP du secteur privé et des entreprises publiques. Sont exclus les apprentis, les stagiaires, les bénéficiaires de contrats aidés et les salariés des particuliers-employeurs.
[1]Y compris chefs d'entreprises salariés

Source : INSEE, DADS, fichier semi-définitif (exploitation au 1/12)

TABLE 5 – Salaires nets annuels moyens dans l'industrie selon la CSP en 2011 (exemple 8)

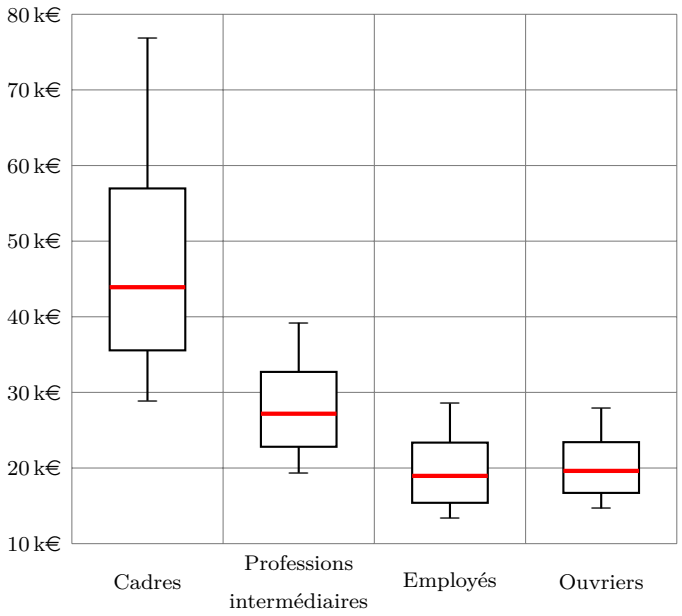

FIGURE 6 – Salaires annuels nets dans l'industrie par catégorie socio-professionnelle (données de l'exemple 8)

5.2 Paramètres de dispersion

Les caractéristiques de position ne suffisent pas en général pour résumer les données. Pour les compléter, on calcule des paramètres de dispersion qui rendent compte du plus ou moins grand «étalement» des valeurs observées.

Définition 7

*L'**étendue** (en anglais : **range**) est la différence entre les valeurs extrêmes du caractère :*

$$w = x_{max} - x_{min}.$$

L'étendue est bien entendu un paramètre grossier, peu robuste dans la mesure où sa sensibilité aux valeurs aberrantes est extrême.

Définition 8

*(1) La **variance** de la variable X est la quantité*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

qui représente donc la moyenne des carrés des écarts entre les observations et leur moyenne.

*(2) L'**écart-type** de X est la racine carrée σ de la variance.*

La variance joue, de part ses propriétés (que n'ont pas d'autres paramètres de dispersion) un rôle fondamental en statistique. En pratique, on la calcule souvent à l'aide de la relation suivante, qui se démontre en développant la formule de la définition 8.

Proposition 2 (*Formule usuelle de calcul de la variance*)

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

(moyenne des carrés moins carré de la moyenne).

L'une des propriétés importantes de la variance concerne l'effet d'une transformation affine (l'effet d'une telle transformation affine sur la moyenne a été mentionné à la proposition 1).

Proposition 3 (*Effet d'une transformation affine sur la variance*)

Si on effectue une transformation affine des données, alors la variance est multipliée par le carré du coefficient directeur de la transformation.

Autrement dit, si on pose $Y = a\,X + b$ avec $a,b \in \mathbb{R}$, on a

$$\sigma_Y^2 = a^2\sigma_X^2.$$

Reprenons les données de l'exemple 7. On a vu que $\overline{x} \simeq 5{,}273$. On a donc, d'après la

proposition 2 :

$$\sigma^2 = \frac{1}{11}\left[2\times 5^2 + 2\times 4^2 + 2\times 3^2 + 1\times 6^2 + 4\times 7^2\right] - 5{,}273^2 \simeq 2{,}377$$

d'où un écart-type $\sigma \simeq 1{,}54\,k€ = 1540€$. Cet écart-type est une mesure de dispersion autour de la valeur moyenne.

Dans le cas d'une variable continue, on procède comme pour la moyenne, avec les centre des classes $x_i = \frac{1}{2}(e_{i-1} + e_i)$ comme représentants de celles-ci. On peut alors directement appliquer la définition 8 ou la proposition 2.

En guise de dernier paramètre de dispersion usuel, citons enfin la **distance interquartile**, déjà rencontrée à la section 5.1 :

$$I_Q = |Q_3 - Q_1|.$$

6 Distributions à deux caractères

Étudions maintenant une population de taille n selon deux variables X et Y, qui peuvent être qualitatives ou quantitatives, sans être nécessairement de même nature.

Exemple 9

On a relevé la taille et le poids d'une population constituée de 200 étudiants.

Exemple 10

Dans une ville moyenne, on a relevé, pour chaque logement proposé à la location, le type de logement (studio, F2,etc.) et le montant mensuel du loyer, en euros.

Si $X(\Omega)$ est fini, ses r modalités sont notées $x_1,\ldots,x_i,\ldots,x_r$. Si ses valeurs sont réparties en classes, celles-ci sont notées $C_1,\ldots,C_i,\ldots,C_r$.

De la même façon, si $Y(\Omega)$ est fini, on note $y_1,\ldots,y_j,\ldots,y_s$ ses éléments. Si les valeurs de Y sont réparties en classes, celles-ci sont notées $D_1,\ldots,D_j,\ldots,D_s$.

La répartition des n observations de ces variables sur la population Ω selon les modalités ou classes de X et Y se présente sous la forme d'un tableau à double entrée, appelé **tableau de contingence** :

X \ Y	D_1	$\cdots$	D_j	$\cdots$	D_s	**Total**
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
C_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.s}$	n

Dans ce tableau, n_{ij} désigne le nombre d'individus dont le caractère X observé appartient à la classe C_i et dont le caractère Y observé appartient à la classe D_j. On écrit donc

$$n_{ij} = \mathrm{Card}(C_i \cap D_j) \quad \text{et} \quad f_{ij} = \frac{n_{ij}}{n}.$$

Alors f_{ij} est la fréquence de $C_i \cap D_j$.

Définition 9

*La **loi conjointe du couple** (X,Y) est la donnée, pour chaque valeur de i et j, de la fréquence f_{ij}.*

On définit aussi :

— l'**effectif marginal en** X et la **fréquence marginale en** X de la classe C_i :

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij} \qquad f_{i\cdot} = \frac{n_{i\cdot}}{n} = \sum_{j=1}^{s} f_{ij}$$

— l'**effectif marginal en** Y et la **fréquence marginale en** Y de la classe D_j :

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \qquad f_{\cdot j} = \frac{n_{\cdot j}}{n} = \sum_{i=1}^{r} f_{ij}.$$

Définition 10

*La **loi conditionnelle de** Y **sachant** $X \in C_i$ est la donnée, pour tout $j \in \{1,\dots,s\}$, des fréquences relatives des classes D_j par rapport à C_i :*

$$f_{j/i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}.$$

*Les deux variables X et Y sont dites **indépendantes** si la loi conditionnelle de Y sachant $X \in C_i$ ne dépend pas de i.*

Dans le cas où X et Y sont indépendantes, on a alors

$$f_{j/i} = \frac{f_{ij}}{f_{i\cdot}} = f_{\cdot j} \quad \text{ou encore} \quad f_{ij} = f_{i\cdot} \times f_{\cdot j}.$$

On retrouvera toutes ces définitions dans la partie dédiée aux probabilités et aux variables aléatoires.

Exemple 11

Dans un groupe de 100 malades souffrant d'arthrose, certains ont pris un médicament et les autres un placebo. Tous pensaient prendre un médicament. Après un mois, on a demandé à chaque patient si le traitement avait été efficace. le tableau ci-dessous donne la répartition des réponses.

	Traitement efficace	*Traitement non efficace*
Médicament	*36*	*6*
Placebo	*28*	*30*

Les lois marginales des variables X, qui indique si un malade prend le médicament ou le placebo, et Y, qui indique si le traitement est perçu comme efficace ou non, sont données par les tableaux suivants :

x_i	Médicament	Placebo
$f_{i\cdot}$	42%	58%

y_j	Efficace	Non efficace
$f_{\cdot j}$	64%	36%

La loi conditionnelle de Y sachant qu'une personne prend le médicament est donnée par le tableau suivant.

y_j sachant $X = x_1$	Traitement efficace sachant que le médicament a été pris	Traitement inefficace sachant que le médicament a été pris
$f_{j/1}$	$\frac{36}{42} \simeq 86\%$	$\frac{6}{42} \simeq 14\%$

De même, la loi conditionnelle de Y sachant qu'une personne prend le placebo est donnée par :

y_j sachant $X = x_2$	Traitement efficace sachant que le placebo a été pris	Traitement inefficace sachant que le placebo a été pris
$f_{j/2}$	$\frac{28}{58} \simeq 48\%$	$\frac{30}{58} \simeq 52\%$

En comparant les lois conditionnelles $f_{j/1}$ et $f_{j/2}$, on est tenté d'affirmer que les variables X et Y ne sont pas indépendantes et conclure à l'efficacité du médicament. Il faut cependant être prudent car la population étudiée est un **échantillon** d'une population plus grande, celle de tous les malades. La preuve statistique de l'efficacité du médicament passera alors par un *test statistique* d'indépendance de X et Y qui prendra en compte l'*aléa statistique*. C'est l'objet du thème 3.

7 Cas de deux variables quantitatives

Étant données deux variables quantitatives X et Y, on souhaite étudier le lien éventuel entre ces deux variables : linéaire ou non, monotone ou non, *etc.*. On souhaite également évaluer «l'intensité» de cette liaison. Dans ce cadre, on considère que les deux variables sont symétriques, c'est-à-dire qu'on ne veut pas évaluer l'influence d'une variable sur l'autre. Ce dernier point est l'objet de la régression linéaire, qui sera étudiée au thème 4.

7.1 Nuage de points

Si les n observations de X et de Y sont connues individuellement, on commence par les visualiser en les représentant sous forme d'un **nuage de points**. Chaque point (x_i,y_i) est représenté dans un repère cartésien par un point M_i.

Exemple 12

On a relevé pour 10 *véhicules la masse, notée X (en kg), et la consommation de carburant, notée Y (en litres/100 km), dans des conditions normalisées. On a obtenu le tableau suivant.*

x_i	*1110*	*1140*	*1370*	*940*	*1400*	*1550*	*1330*	*1300*	*1670*	*1560*
y_i	*8,6*	*7,7*	*10,8*	*6,6*	*11,7*	*11,9*	*10,8*	*7,6*	*11,3*	*10,8*

En plaçant la masse en abscisse et la consommation en ordonnée, on obtient la figure 7.

Une analyse graphique du nuage de points est la première étape de la démarche statistique. En effet, un nuage de points de forme allongée avec des variables qui évoluent dans

le même sens comme dans l'exemple précédent conduit à penser qu'il s'agit d'une liaison linéaire positive (pour signifier une fonction croissante).

On peut aussi observer une liaison linéaire négative (fonction décroissante) ou l'absence de liaison linéaire comme sur les figures 8 et 9.

7.2 Covariance de deux variables quantitatives

Pour caractériser la liaison qui peut exister entre deux variables quantitatives, on peut calculer leur covariance.

Définition 11

*La **covariance** de deux variables quantitatives X et Y est*

$$\mathbb{C}\text{ov}\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

où les x_i (resp. y_i) désignent les valeurs prises par X (resp. y) sur les n individus de la population observée.

Remarques

1. *Comme pour la variance (cf. proposition 2), on a aussi*

$$\mathbb{C}\text{ov}\,(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}$$

   (moyenne des produits moins produit des moyennes).

2. *Dans le cas où les données sont regroupées dans un tableau de contingence, on a*

$$\mathbb{C}\text{ov}\,(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}\,(x_i - \overline{x})\,(y_j - \overline{y})\,.$$

La covariance permet de quantifier la liaison entre les deux variables X et Y : elle mesure leur «tendance à varier ensemble».

Si la covariance est positive, cela signifie que la variation de X et Y se fait dans le même sens. Si elle est négative, leurs variations se font en sens contraire. On parle respectivement de **liaison positive** et de **liaison négative**.

En guise d'exemple de calcul, reprenons et complétons les données de l'exemple 12 :

											Total
x_i	1110	1140	1370	940	1400	1550	1330	1300	1670	1560	13370
y_i	8,6	7,7	10,8	6,6	11,7	11,9	10,8	7,6	11,3	10,8	97,8
$x_i y_i$	9546	8778	14796	6204	16380	18445	14364	9880	18871	16848	134112

Alors :

$$\mathbb{C}\text{ov}\,(X,Y) = \frac{134112}{10} - \frac{13370}{10} \times \frac{97{,}8}{10} = 335{,}3.$$

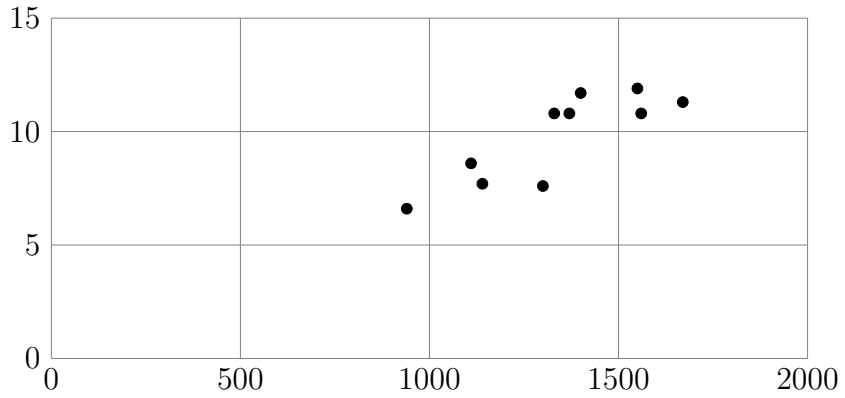

FIGURE 7 – Nuage de points de l'exemple 12

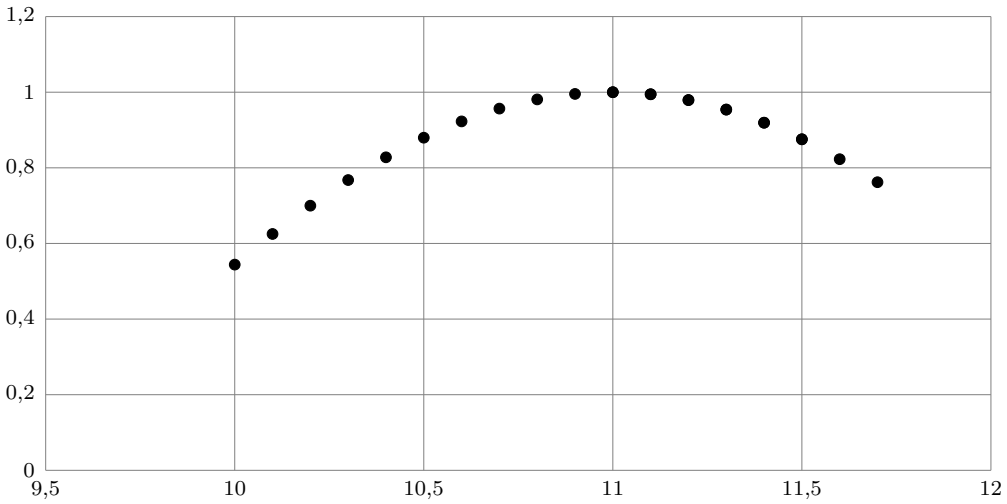

FIGURE 8 – Exemple de nuage de points

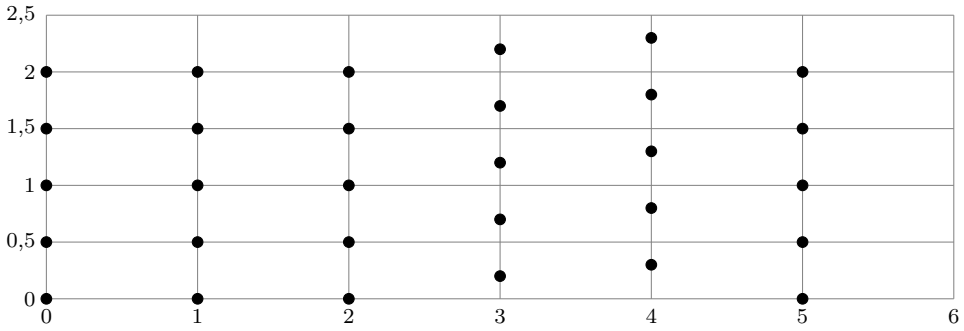

FIGURE 9 – Exemple de nuage de points

On constate ici que la liaison entre masse et consommation est positive, ce qui n'est pas étonnant !

Proposition 4 (*Propriétés de la covariance*)

(1) La covariance est symétrique : $\mathbb{C}\text{ov}(X,Y) = \mathbb{C}\text{ov}(Y,X)$.

(2) Covariance de X avec elle-même : $\mathbb{C}\text{ov}(X,Y) = \mathbb{V}(X)$.

(3) Transformation affine : $\mathbb{C}\text{ov}(a\,X + b, c\,Y + d) = a\,c\,\mathbb{C}\text{ov}(X,Y)$.

(4) Variance d'une somme : $\mathbb{V}(X+Y) = \mathbb{V}(X) + 2\mathbb{C}\text{ov}(X,Y) + \mathbb{V}(Y)$.

(5) Inégalité de Cauchy-Schwarz : $|\mathbb{C}\text{ov}(X,Y)| \leqslant \sigma(X)\,\sigma(Y)$ *avec égalité si et seulement si il existe une relation affine entre X et Y :* $Y = a\,X + b$ *ou* $X = c\,Y + d$.

(6) Cas de variables indépendantes : si X et Y sont indépendantes, leur covariance est nulle. La réciproque est fausse !

Lorsque deux variables ont une covariance nulle, on dit qu'elles sont **décorrélées**. L'assertion (6) de la proposition 4 montre que des variables indépendantes sont décorrélées, mais *attention* : il est possible de trouver des variables décorrélées qui ne sont pas indépendantes.

7.3 Coefficient de corrélation linéaire

Le coefficient de corrélation de deux variables X et Y correspond à une normalisation de leur covariance par le produit des écart-types de X et Y. Il mesure le degré de **dépendance linéaire** de X et Y. Il suppose bien sûr que X et Y ne soient pas des constantes (donc que leurs écart-types ne soient pas nuls).

Définition 12 (*Coefficient de corrélation linéaire*)

Le coefficient de corrélation linéaire de X et Y est la quantité

$$r(X,Y) = \frac{\mathbb{C}\text{ov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

Ses propriétés découlent directement de celles de la covariance explicitées à la proposition 4.

Proposition 5 (*Propriétés de $r(X,Y)$*)

(1) On a toujours $-1 \leqslant r(X,Y) \leqslant 1$;

(2) corrélation linéaire parfaite : $r(X,Y) = \pm 1$ *si et seulement si il existe une relation affine entre X et Y ;*

(3) si X et Y sont indépendantes, alors $r(X,Y) = 0$, *la réciproque étant* fausse.

Remarques

1. *Si on a* $|r(X,Y)| \geqslant 0{,}8$, *on dit que les variables X et Y sont* ***fortement corrélées***.

2. *L'existence d'une corrélation, même forte, entre deux variables statistiques ne permet pas de mettre en évidence une relation de cause à effet.*

Avec les données de l'exemple 12, on trouve

$$\begin{cases} \sigma(X) \simeq 214{,}57 \\ \sigma(Y) \simeq 1{,}85 \end{cases} \quad \text{donc} \quad r(X,Y) = \frac{\mathbb{C}\text{ov}\,(X,Y)}{\sigma(X)\sigma(Y)} \simeq \frac{335{,}34}{214{,}57 \times 1{,}85} \simeq 0{,}845.$$

On peut donc dire que la masse d'un véhicule et sa consommation de carburant sont des variables fortement corrélées.

Exercices

Exercice 1 : Quelques calculs de statistique descriptive

Dans une entreprise, on a recensé les salariés par tranche d'âge et par sexe. Les résultats sont donnés dans le tableau ci-dessous.

Tranche d'âge	Hommes	Femmes
Moins de 20 ans	32	51
20 — 30	1309	2118
30 — 40	1902	3025
40 — 50	1730	2330
50 — 60	1468	1624
Plus de 60 ans	114	131

1. Quelles sont les caractéristiques étudiées ? Préciser s'il s'agit de caractères discrets ou continus.

2. Quelle est la proportion de salariés dans les tranches d'âge inférieures ou égales à 40 ans ? Mêmes questions pour les hommes et femmes séparément. Que peut-on en conclure ?

3. Déterminer l'âge moyen, l'âge médian, les quartiles et l'écart-type pour les hommes. Mêmes questions pour les femmes.

4. Comparer les deux sous-populations (hommes et femmes) à l'aide de boîtes à moustaches.

Exercice 2 : Étude d'une corrélation

On a relevé la taille (X, exprimée en cm) et le poids (Y, exprimé en kg) d'une population humaine donnée. Les résultats sont regroupés en classes et les effectifs conjoints notés dans le tableau de contingence suivant.

X \ Y	]50,60]	]60,70]	]70,80]	]80,90]
]150,155]	24	11	2	0
]155,160]	22	27	10	1
]160,165]	13	30	14	3
]165,170]	3	6	15	7
]170,180]	0	2	3	7

1. Déterminer les lois marginales.

2. En choisissant les centres des classes comme représentants, calculer :
- la taille moyenne de cette population,
- son poids moyen,
- les écart-types correspondants,
- la covariance de X et de Y,
- le coefficient de corrélation linéaire.

3. Déterminer la loi conditionnelle de Y sachant $\{150 < X \leqslant 155\}$. Calculer la moyenne conditionnelle de Y sachant sachant $\{150 < X \leqslant 155\}$.

4. Mêmes questions avec les autres classes de la variable X.

5. Représenter graphiquement les points de coordonnées $\left(x_i, \overline{y_i | X \in C_i}\right)$ où :
- C_i désigne l'une des classes de taille,
- x_i est le centre de la classe C_i,
- $\overline{y_i | X \in C_i}$ est la moyenne conditionnelle de Y sachant $\{X \in C_i\}$.

Construire une courbe de régression de Y en X, c'est-à-dire une courbe passant par les points précédemment représentés. Conclure.