Eléments de Statistique descriptive

Jean VAILLANT
Mars 2015

Table des matières

1 Terminologie

La **statistique** est le domaine des mathématiques qui étudie les outils de recueil, de traitement et d'interprétation des données. La statistique mathématique s'appuie fortement sur la théorie des probabilités et développe des outils théoriques, tandis que la statistique appliquée s'attache à proposer des méthodologies dans divers domaines scientifiques (biologie, sciences médicales, sismologie, agronomie, économie, sciences sociales,...). La statistique désigne donc la science du recueil, du traitement et de l'interprétation des données. Notons que l'utilisation du nom au pluriel (**statistiques**) correspond à des données obtenues par certains type de calcul, par exemple : *revenu moyen, revenu médian, taux de chômage.*

La **statistique descriptive** est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer des données nombreuses et variées.

Il faut d'abord préciser l'ensemble étudié, appelé **population statistique**, dont les éléments sont des **individus**, ou **unités statistiques**. Il est fréquent qu'on ne puisse observer toute la population statistique, pour des raisons techniques ou budgétaires. On effectue alors une observation partielle de cette population à travers un **échantillon** qui est, par définition, un sous-ensemble de la population statistique. Il existe différentes procédures pour choisir un échantillon. On parle de **procédure d'échantillonnage**. Les plus courantes sont l'échantillonnage aléatoire simple et l'échantillonnage aléatoire stratifié. Pour le premier, tous les échantillons de même taille ont les mêmes chances d'être sélectionnés. Pour le second, la population statistique est divisée en strates (disjointes et relativement homogènes), et dans chacune de ces strates, un échantillonnage aléatoire simple est appliqué et ceci indépendamment d'une strate à l'autre.

La **statistique inférentielle** est l'ensemble des méthodes permettant, à partir d'un échantillon, d'estimer des paramètres d'une population statistique et/ou de tester des hypothèses sur cette population. A l'inverse de la statistique descriptive, la statistique inférentielle fait appel à la théorie des probabilités à travers les notions de précision statistique et de risque d'erreur décisionnel.

Notons qu'un individu statistique n'est pas forcément un individu biologique ni même un objet matériel. Ainsi, on peut s'intéresser à l'ensemble des accidents de la route survenus dans une région au cours d'une période donnée. L'individu statistique est alors l'accident, qui est une occurrence donc immatériel. Voici quelques exemples de population statistique :

1. Ensemble des collèges d'une académie. Pour chaque collège, on peut

s'intéresser au taux de passage en seconde, au nombre d'élèves, à la présence ou pas d'une cuisine scolaire, à la commune d'implantation, au numéro de département.

2. Ensemble des parents d'élève d'un lycée. On s'intéresse à leur opinion sur un projet éducatif selon leur profession, leur revenu, leur statut marital, le nombre d'enfants scolarisés, la distance domicile-lycée, le moyen de locomotion.
3. Ensemble des incidents de violence remontés à un rectorat au cours de l'année scolaire 2013-2014. Pour chaque incident, l'établissement concerné indique : le statut du principal acteur (élève, personnel de sécurité, personnel enseignant, personnel administratif ou technique), le type violence (physique et/ou verbale), le nombre de protagonistes, lieu (intérieur, extérieur de l'enceinte de l'établissement), le nombre de blessés.
4. Ensemble des élèves de CM2 d'une région. L'ARS (Agence Régionale de Santé) désire étudier le comportement alimentaire chez certains jeunes et ses conséquences sur l'obésité et autres risques sanitaires. Les enquêteurs notent le poids, la hauteur, l'âge, tour de taille, tour de hanche, le sexe, la commune de résidence, le nombre de sports pratiqués, la fréquence de prise de petit-déjeuner, la taille de fratrie, régularité de consommation de divers produits.

.

Chaque individu statistique est donc décrit par un ou plusieurs traits distinctifs ou grandeurs physiques le caractérisant. On les appelle **variables statistiques**.

Une **variable statistique** (ou **caractère statistique**) est donc ce qui est observé ou mesuré sur un individu statistique.

Quand on observe une variable statistique sur un nombre n d'individus statistiques, on obtient une suite $x_1, x_2, \cdots, x_n$ où x_i est la modalité ou valeur observée sur le ième individu. Cette suite est appelée **série statistique**. On parle de **série statistique simple (ou univariée)**. Le nombre n est la **taille (ou longueur) de la série**. Si on observe sur chaque individu deux variables, on a alors une suite $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ appelée **série statistique double (ou bivariée)**. D'une façon générale, si sur chaque individu statistique, il est observé un nombre de variables k (supérieur à 2), on dit que la série statistique est **multivariée**.

La statistique descriptive concernant une seule variable statistique est appelée **statistique descriptive univariée (ou unidimensionnelle)**. La statistique descriptive concernant plusieurs variables statistiques est dite **statistique descriptive multivariée (ou multidimensionnelle)**. Cette

dernière permet la description des caractères observés sur des individus et des liens éventuels entre ces caractères.

Une variable peut être :

1) **quantitative** : elle concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité, et sur lesquelles les opérations arithmétiques (addition, multiplication, etc,...) ont un sens.
La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre. Une **variable quantitative discrète** ne peut prendre que des valeurs isolées. Ces valeurs sont en nombre fini ou dénombrable. Le cas le plus répandu est celui où les valeurs possibles sont des nombres entiers naturels : nombre d'insectes sur une plante ; nombre de descendants dans une portée ; nombre de fruits dans un arbre ; taille de fratrie, effectif d'un établissement. Une **variable quantitative continue** peut prendre une infinité de valeurs sous forme d'intervalle. La taille, le poids, la surface cultivée, la température moyenne sont des variables quantitatives continues. On obtient des valeurs à la précision de l'instrument de mesure près. Je ne mesure pas exactement 1m80 mais m'étant limité à mesurer ma taille au centimètre près, je sais seulement qu'elle est située entre 1m795 et 1m805.
Exemple 1 : l'unité statistique est la plante d'une parcelle de maïs.
   — les variables *nombre d'insectes foreurs sur la plante, nombre de noeuds, nombre de trous percés par les insectes foreurs* sont discrètes.
   — les variables *surface foliaire, hauteur de la plante, poids de l'épi* sont continues.
Exemple 2 : l'unité statistique est l'élève de CM2.
   — les variables *taille de fratrie, nombre de sports pratiqués* sont discrètes.
   — les variables *poids, hauteur, âge, tour de taille, tour de hanche* sont continues.
2) **qualitative** : ses valeurs sont des **modalités**, ou catégories, exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens.
On distingue des variables qualitatives **ordinales** ou **nominales**, selon que les modalités peuvent être naturellement ordonnées ou pas.
Une variable est **dichotomique** si elle n'a que deux modalités.
Exemple 1 : l'unité statistique est une parcelle de canne à sucre.
   — Les variables *type de sol, type de culture d'une exploitation, département d'origine, variété cultivée* sont nominales.
   — La variable *présence-absence du virus de la feuille jaune* est dichotomique.
   — La variable *degré d'infestation* (en notation visuelle) est ordinale.
Exemple 2 : l'unité statistique est un exploitant agricole.

— Les variables *taille vestimentaire, préférence plus ou moins marquée pour un engrais* sont ordinales.
— La variable *rendement à l'hectare* est quantitative mais peut être transformée en variable qualitative ordinale à 3 modalités : *faible*, *moyen*, *élevé*.

Exemple 3 : l'unité statistique est un établissement scolaire.
— Les variables *type d'établissement, département d'implantation* sont qualitatives nominales.
— La variable *présence-absence d'une cuisine scolaire* est dichotomique.
— Les variables *nombre d'élèves, effectif en personnel* sont quantitatives discrètes.
— Les variables *budget annuel de fonctionnement, taux de réussite à un examen de référence* sont quantitatives continues. La variable *taux de réussite* peut être transformée en variable qualitative ordinale à 5 modalités : *très faible*, *faible*, *moyen*, *élevé*, *très élevé*.

Exemple 4 : L'unité statistique est une sortie pédagogique d'un collège.
— Les variables *lieu visité, thème de la sortie* sont qualitatives nominales.
— La variable *présence-absence d'une personne ressource* est dichotomique.
— Les variables *nombre d'élèves, nombre de personnel encadrant* sont quantitatives discrètes.
— Les variables *prix de la sortie, durée de la sortie, distance parcourue* sont quantitatives continues. La variable *prix de la sortie* peut être transformée en variable qualitative ordinale à 4 modalités : *pas cher*, *moyen*, *cher*, *très cher*.

La statistique descriptive a pour objectif de synthétiser l'information contenue dans les jeux de données au moyen de tableaux, figures ou résumés numériques. Les variables statistiques sont analysées différemment selon leur nature (quantitative, qualitative).

2 Série univariée

2.1 Représentation d'une série univariée

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable. Rappelons que les observations effectuées pour une variable qualitative sont appelées **modalités** de la variable, plutôt que **valeurs**, ce dernier terme étant de préférence utilisé pour une variable quantitative.

Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes. Dans un document scientifique ou académique, il convient de les numéroter et de les légender. Cela facilite la lecture du document et permet de les référencer dans le texte.

Un tableau comprend 3 parties : le titre, le corps et la source d'information. Le titre permet de préciser le lieu, la période et les variables auxquels correspondent les données. La source d'information indique clairement s'il s'agit de données personnelles (recueillies par exemple par enquête ou par planification expérimentale) ou de données obtenues auprès d'un quelconque organisme ou média. Le corps du tableau dépend, lui, de la nature de la variable statistique étudiée.

2.1.1 Variable qualitative

A partir de l'observation d'une variable qualitative sur n individus statistiques, on peut construire un tableau dont le corps est :

Modalités	Effectifs	Fréquences
Modalité 1	n_1	f_1
Modalité 2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
Modalité i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
Modalité k	n_k	f_k
Totaux	n	1

TABLE 1 – Corps de tableau pour une variable qualitative.

où

n_i est l'effectif associé à la modalité i c'est-à-dire le nombre d'individus dans l'échantillon ayant cette modalité ;

n est la taille de l'échantillon (nombre total d'individus dans cet échantillon) ;

$f_i = n_i/n$ est la fréquence associée à la modalité i c'est-à-dire la proportion d'individus dans l'échantillon ayant cette modalité ;

k est le nombre de modalités distinctes observées dans l'échantillon.

Si la variable est ordinale, les modalités sont écrites dans l'ordre :

modalité $1 <$ modalité $2 < \cdots <$ modalité k.

Deux diagrammes permettent de représenter une variable qualitative : le **diagramme à secteurs angulaires (dit camembert)** et le **diagramme en bandes (dit tuyaux d'orgue)**.

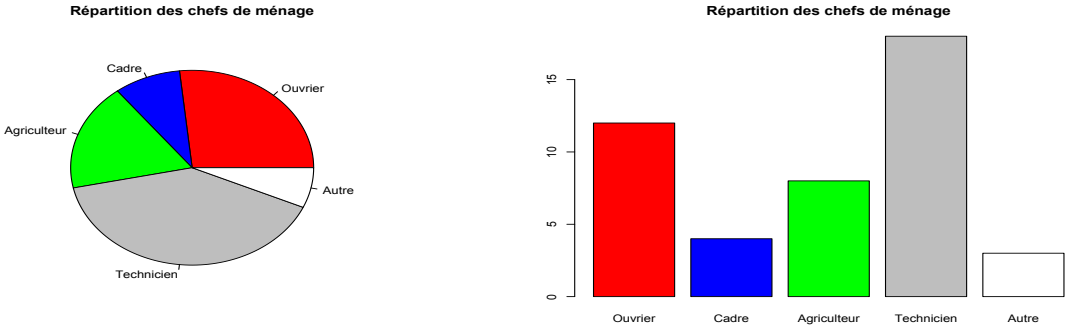


FIGURE 1 – Représentations d'une variable qualitative.

Le camembert est un disque partagé en secteurs, chaque secteur représentant une modalité et ayant une surface proportionnelle à la fréquence de cette modalité dans la série statistique.

Le diagrammes en bandes est un ensemble de rectangles de même largeur, séparés par un espace, chaque rectangle représentant une modalité et ayant une hauteur proportionnelle à la fréquence de cette modalité dans la série statistique.

2.1.2 Variable quantitative discrète

A partir de l'observation d'une variable quantitative discrète sur n individus statistiques, on peut construire un tableau dont le corps est donné par Table 2 :

Valeurs	Effectifs	Fréquences	Fréquences cumulées
x_1	n_1	f_1	F_1
x_2	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	F_k
Totaux	n	1	-

TABLE 2 – Corps de tableau pour une variable quantitative discrète.

où

n_i est l'effectif associé à la valeur x_i c'est-à-dire le nombre d'individus ayant cette valeur dans l'échantillon ;

n est la taille de l'échantillon (nombre total d'individus dans cet échantillon) ;

$f_i = n_i/n$ est la fréquence associée à la valeur x_i c'est-à-dire la proportion d'individus dans l'échantillon ayant cette valeur.

F_i est la fréquence cumulée en x_i c'est-à-dire la proportion d'individus dans l'échantillon ayant une valeur inférieure ou égale à x_i. Le calcul des F_i peut se faire façon récurrente de la manière suivante :

$$F_1 = f_1 \qquad \text{et} \quad F_i = F_{i-1} + f_i \quad \text{pour } i \in \{2, \cdots k\}.$$

k est le nombre de valeurs distinctes observées dans l'échantillon.

Les valeurs distinctes sont par ordre croissant dans le tableau :

$x_1 < x_2 < \cdots < x_k$.

Deux diagrammes permettent de représenter une variable quantitative discrète : le **diagramme en bâtons** et le **diagramme cumulatif**.

Le diagramme en bâtons associe à chaque valeur de la variable un segment vertical de hauteur proportionnelle à la fréquence de cette valeur dans la série statistique.

Le diagramme cumulatif est une courbe en escalier représentant les fréquences cumulées relatives.

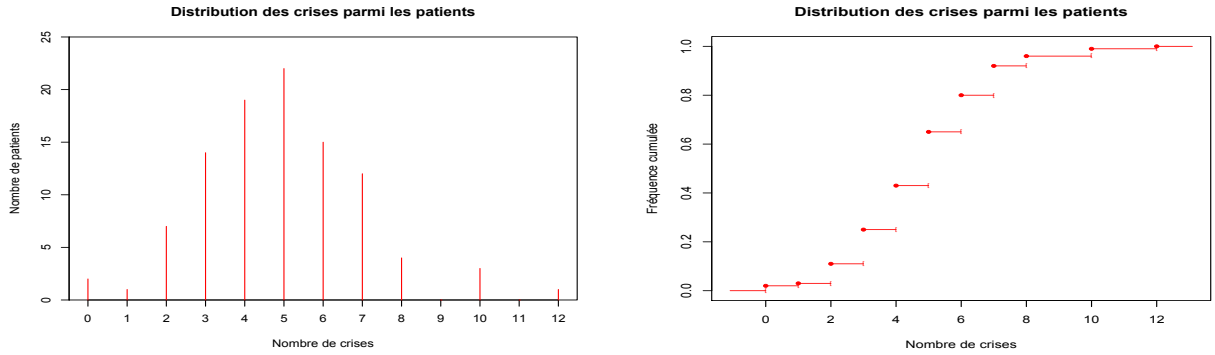


FIGURE 2 – Représentations d'une variable quantitative discrète.

2.1.3 Variable quantitative continue

A partir de l'observation d'une variable quantitative continue sur n individus statistiques (avec n suffisamment grand), on peut déterminer k classes statistiques et construire un tableau dont le corps est :

Classes statistiques	Effectifs	Fréquences	Fréquences cumulées
$]a_0, a_1]$	n_1	f_1	$F(a_1)$
$]a_1, a_2]$	n_2	f_2	$F(a_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$]a_{i-1}, a_i]$	n_i	f_i	$F(a_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$]a_{k-1}, a_k]$	n_k	f_k	$F(a_k)$
Totaux	n	1	-

TABLE 3 – Corps de tableau pour une variable quantitative continue.

où n_i est l'effectif associé à la classe $]a_{i-1}, a_i]$ c'est-à-dire le nombre d'individus ayant une valeur comprise entre a_{i-1} (exclus) et a_i dans l'échantillon ; n est la taille de l'échantillon (nombre total d'individus dans cet échantillon) ;

$f_i = n_i/n$ est la fréquence associée à la classe $]a_{i-1}, a_i]$ c'est-à-dire la proportion d'individus ayant une valeur comprise entre a_{i-1} (exclus) et a_i dans l'échantillon ;

$F(a_i)$ est la fréquence cumulée en a_i c'est-à-dire la proportion d'individus dans l'échantillon ayant une valeur inférieure ou égale à a_i. Le calcul des $F(a_i)$ peut se faire façon récurrente de la manière suivante :

$$F(a_1) = f_1 \qquad \text{et} \quad F(a_i) = F(a_{i-1}) + f_i \quad \text{pour } i \in \{2, \cdots k\}.$$

k est le nombre de valeurs distinctes observées dans l'échantillon.

Les bornes de classe vérifient bien évidemment : $a_0 < a_1 < a_2 < \cdots < a_k$.

Deux diagrammes permettent de représenter une variable quantitative continue : l'**histogramme** et la **courbe cumulative**.

L'histogramme est une juxtaposition de rectangles, chaque rectangle étant associé à une classe statistique et étant de surface (et non pas de hauteur) proportionnelle à la fréquence de cette classe.

La $i$ème classe statistique $]a_{i-1}, a_i]$ d'effectif n_i est associée à un rectangle de largeur $a_i - a_{i-1}$ et de hauteur $h_i = n_i/(a_i - a_{i-1})$. Notons que h_i est l'effectif par unité d'amplitude. Notons que l'on peut également, sans changer l'allure de l'histogramme, poser $h_i = f_i/(a_i - a_{i-1})$.

La courbe cumulative est une succession de segments de droite reliant le point $(a_{i-1}, F(a_{i-1}))$ au point $(a_i, F(a_i))$.

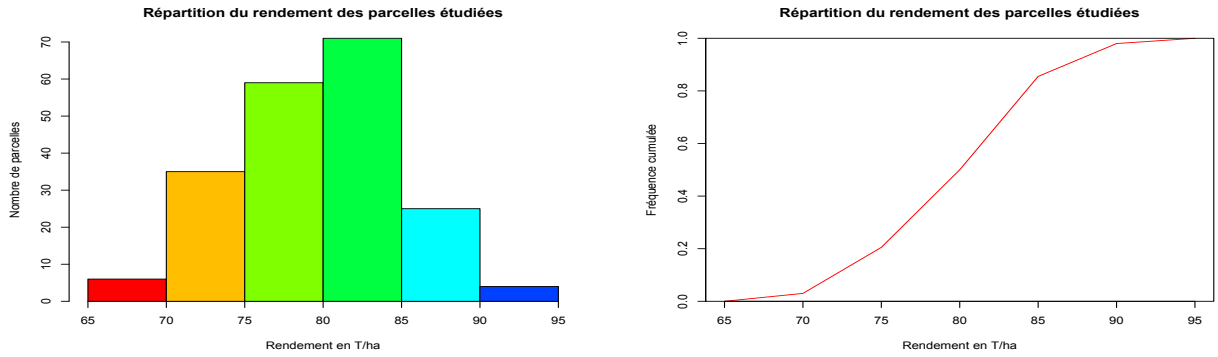


FIGURE 3 – Représentations d'une variable quantitative continue.

2.2 Résumés numériques d'une série univariée

Il est tout d'abord important de souligner que les **opérations arithmétiques n'ont aucun sens pour une variable qualitative codée numériquement ! !**.

Un exemple pour s'en convaincre. La variable *Numéro de département* indiquant le lieu de stage pour 150 individus dans Table 4 a un total de 22925 et une moyenne de 152,8. Ces valeurs n'apportent pas d'information sur la répartition des lieux de stage.

Département	Numéro	Effectif	Fréquence
Bouches du Rhône	13	27	0,18
Guadeloupe	971	5	0,03
Guyane	973	4	0,03
Hautes-Alpes	5	12	0,08
Martinique	972	7	0,05
Rhône	69	27	0,18
Seine	75	68	0,45
Totaux	-	150	1

TABLE 4 – Exemple de variable nominale codée numériquement. Lieu de stage à l'issue d'une formation.

Par conséquent, le présent paragraphe ne concerne que les variables quantitatives ! Des indicateurs statistiques de tendance centrale (résumés numériques donnant l'ordre de grandeur de la série statistique) et de dispersion (fournissant une idée de la variabilité dans la série statistique) sont présentés.

2.2.1 Indicateurs statistiques de tendance centrale

Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane, les quartiles et le mode.

La moyenne $\bar{x}$: La moyenne d'une série statistique vérifie :

$$\boxed{\text{Moyenne} = \frac{\text{Somme des valeurs de la série}}{\text{Nombre de valeurs dans la série}}}.$$

Exemple : La série statistique suivante représente les valeurs observées pour la variable *Nombre d'absences* au cours de l'année 2013-2014 pour les 25 élèves d'une classe de 6ème :

2 0 4 2 3 1 2 2 4 5 4 3 4 7 2 7 5 7 3 9 6 5 7 4 6 .

La moyenne est donc égale à

$$\frac{2+0+4+2+3+1+2+2+4+5+4+3+4+7+2+7+5+7+3+9+6+5+7+4+6}{25}$$

ce qui nous conduit à une moyenne d'absences par élève qui vaut $\frac{104}{25} = 4,16$.

La formule de la moyenne, avec les notations de la table 2, est :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i.$$

La distribution d'effectifs correspondant aux nombres d'absences est :

Valeur distincte x_i	0	1	2	3	4	5	6	7	9
Effectif n_i	1	1	5	3	5	3	2	4	1

TABLE 5 – Nombre d'absences dans l'année pour une classe de 25 élèves.

La moyenne est alors calculée de la manière suivante :

$$\frac{1}{25}(1\times 0 + 1\times 1 + 5\times 2 + 3\times 3 + 5\times 4 + 3\times 5 + 2\times 6 + 4\times 7 + 1\times 9).$$

ce qui nous donne bien-sûr le même résultat 104/25=4,16.

Si les données sont sous forme de distribution d'effectifs en classes statistiques (table 3), n'ayant pas les valeurs exactes, on peut malgré cela calculer une moyenne approchée :

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

où c_i est le centre de la classe $]a_{i-1}, a_i]$ donc vaut $(a_{i-1} + a_i)/2$.

Exemple : La variable *Revenu mensuel du foyer* des 500 élèves d'un collège est étudiée. On obtient la *distribution en classes statistiques* indiquée dans Table 6.

La moyenne approchée du revenu mensuel est donc :

$$\frac{1}{500}(152\times 500 + 178\times 1500 + 90\times 2500 + 64\times 4000 + 16\times 7500).$$

ce qui nous donne 944000/500=1888.

Classe statistique $]a_{i-1}, a_i]$	Effectif n_i	Effectif cumulé N_i	Fréquence cumulée $F(a_i)$	Centre de classe c_i	Effectif par amplitude h_i
$]0, 1000]$	152	152	0,304	500	0,152
$]1000, 2000]$	178	330	0,660	1500	0,178
$]2000, 3000]$	90	420	0,840	2500	0,090
$]3000, 5000]$	64	484	0,968	4000	0,032
$]5000, 10000]$	16	500	1	7500	0,0032

TABLE 6 – Revenu mensuel du foyer pour les 500 élèves d'un collège.

Remarque : N'ayant pas les valeurs exactes des 500 revenus, nous avons pu fournir une valeur approchée de la moyenne en assimilant chaque valeur de classe au centre de classe.

La médiane Me **:** La médiane d'une série statistique est une valeur qui sépare la série en deux partie d'égale fréquence de telle sorte qu'il y a autant de valeurs inférieures à la médiane que de valeurs supérieures à la médiane.

On calcule la médiane en réordonnant la série statistique par ordre croissant.

Revenons à la série représentant la variable *Nombre d'absences* au cours de l'année 2013-2014 pour les 25 élèves d'une classe de 6ème. La série réordonnée est :

0 1 2 2 2 2 2 3 3 3 4 4 4 4 4 5 5 5 6 6 7 7 7 7 9.

La taille de la série est $n = 25$ donc impaire. Le rang de la médiane est $(n+1)/2 = 13$. La 13ième valeur de la série réordonnée est 4 donc $Me = 4$.

Dans le cas d'une distribution d'effectifs en classes statistiques (table 3), la médiane est calculée à partir de la formule dite d'interpolation linéraire suivante :

$$Me = a_{i^*-1} + \frac{0,5n - N_{i^*-1}}{n_{i^*}}(a_{i^*} - a_{i^*-1})$$

où la classe $]a_{i^*-1}, a_{i^*}]$ est celle vérifiant $F(a_{i^*-1}) < 0,5 \leq F(a_{i^*})$ et N_{i^*-1} l'effectif cumulé en a_{i^*-1}.

Considérons la distribution en classes statistiques fournie dans Table 6. Les deux valeurs de fréquence cumulée encadrant $0,5$ sont $F(1000) = 0,304$ et

$F(2000) = 0,660$. Par conséquent, la classe $]a_{i^*-1}, a_{i^*}]$ est $]1000; 2000]$ d'où

$$Me = 1000 + \frac{0,5 \times 500 - 152}{178}(2000 - 1000) \approx 1550.$$

Le revenu médian vaut à peu près 1550 ce qui signifie qu'il y a autant de revenus au dessus de 1550 que de revenus au dessous de cette valeur.

Le mode *Mo* **:** C'est la valeur la plus fréquente dans la série. Le mode n'est pas forcément unique. Si c'est le cas on parle de distribution unimodale. Sinon, on parle de distribution multimodale. Notons que le mode est calculable pour une variable qualitative.

Exemple : Pour la distribution représentée dans Table 5, les valeurs les plus fréquentes sont le 2 et le 4 puisque 5 élèves ont eu 2 absences au cours de l'année, et 5 autres ont eu 4 absences. Les autres effectifs sont inférieurs strictement à 5, on a donc deux modes : 2 et 4. La distribution est bimodale.

Dans le cas d'une distribution d'effectifs en classes statistiques, on parle de **classe modale** pour désigner la classe ayant la plus forte *fréquence par unité d'amplitude.*

Exemple : Pour la distribution représentée dans Table 6, les effectifs par unité d'amplitude h_i pour les 5 classes statistiques sont :

$$h_1 = 0,152; h_2 = 0,178; h_3 = 0,090; h_4 = 0,032; h_5 = 0,0032.$$

La classe modale correspond donc au plus grand des h_i, et est donc $]1000; 2000]$.

Le premier quartile Q_1 **:** c'est une valeur telle qu'il y a 25% de valeurs qui lui sont inférieures dans la série statistique et 75% qui lui sont supérieures.

Le troisième quartile Q_3 **:** c'est une valeur telle qu'il y a 75% de valeurs qui lui sont inférieures dans la série statistique et 25% qui lui sont supérieures.

Notons que la médiane est le second quartile. Les 3 quartiles s'obtiennent en séparant la série réordonnée en quatre parties d'égale fréquence.

Pour la série représentée dans Table 6, Q_1 est la 7ème valeur de la série réordonnée donc $Q_1 = 2$. D'autre part, Q_3 est la 18ème valeur de la série réordonnée donc $Q_3 = 5$.

Dans le cas d'une distribution d'effectifs en classes statistiques (table 3), le premier quartile est calculée à partir de la formule

$$Q_1 = a_{i^*-1} + \frac{0,25n - N_{i^*-1}}{n_{i^*}}(a_{i^*} - a_{i^*-1})$$

où la classe $]a_{i^*-1}, a_{i^*}]$ est celle vérifiant $F(a_{i^*-1}) < 0,25 \leq F(a_{i^*})$ et N_{i^*-1} l'effectif cumulé en a_{i^*-1}.

Le troisième quartile est calculée à partir de la formule

$$Q_3 = a_{i^*-1} + \frac{0,75n - N_{i^*-1}}{n_{i^*}}(a_{i^*} - a_{i^*-1})$$

où la classe $]a_{i^*-1}, a_{i^*}]$ est celle vérifiant $F(a_{i^*-1}) < 0,75 \leq F(a_{i^*})$ et N_{i^*-1} l'effectif cumulé en a_{i^*-1}.

Revenons à la distribution en classes statistiques fournie dans Table 6. Les deux valeurs de fréquence cumulée encadrant $0,25$ sont $F(0) = 0$ et $F(1000) = 0,304$. Par conséquent, la classe $]a_{i^*-1}, a_{i^*}]$ est $]0; 1000]$ d'où

$$Q_1 = 0 + \frac{0,25 \times 500 - 0}{152}(1000 - 0) \approx 822.$$

Les deux valeurs de fréquence cumulée encadrant $0,75$ sont $F(2000) = 0,660$ et $F(3000) = 0,840$. Par conséquent, la classe $]a_{i^*-1}, a_{i^*}]$ est $]2000; 3000]$ d'où

$$Q_3 = 2000 + \frac{0,75 \times 500 - 330}{90}(3000 - 2000) = 2500.$$

2.2.2 Indicateurs statistiques de dispersion

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance, l'écart-type et l'écart interquartile.

L'étendue : C'est l'écart entre la valeur maximale et la valeur minimale observées dans la série statistique.

Ainsi, pour la distribution statistique de Table 5, l'étendue vaut 9-0=9. Pour une distribution en classes statistiques, les valeurs minimale et maximale ne sont pas forcément connues. Avec les notations de Table 3, on pose que l'étendue vaut $a_k - a_0$.

Pour Table 6, l'étendue vaut $10000 - 0 = 10000$.

La variance v : La variance d'une série statistique vérifie :

$$\boxed{\text{Variance} = \frac{\text{Somme des carrés d'écart à la moyenne de la série}}{\text{Nombre de valeurs dans la série}}}.$$

La formule de la variance, avec les notations de la table 2, est :

$$v = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 \quad \text{ou de façon équivalente } v = \Big(\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2\Big) - \bar{x}^2.$$

Par exemple, pour la distribution statistique de Table 5, la variance vaut

$$\Big(\frac{0^2 + 1^2 + 5\times 2^2 + 3\times 3^2 + 5\times 4^2 + 3\times 5^2 + 2\times 6^2 + 4\times 7^2 + 9^2}{25}\Big) - 4,16^2$$

$= 4,77.$

Si les données sont fournies sous forme de distribution en classes statistiques (table 3), les valeurs exactes sur les différents individus statistiques ne sont pas disponibles. On peut cependant calculer une variance approchée en utilisant la formule :

$$v \approx \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \bar{x}^*)^2 \quad \text{ou aussi } v \approx \Big(\frac{1}{n}\sum_{i=1}^{k} n_i c_i^2\Big) - (\bar{x}^*)^2$$

où c_i est le centre de la classe $]a_{i-1}, a_i]$ donc vaut $(a_{i-1} + a_i)/2$

et $\bar{x}^*$ est la moyenne approchée.

Exemple : La variance approchée de la distribution en classes statistiques de Table 6 est

$$\Big(\frac{152\times 500^2 + 178\times 1500^2 + 90\times 2500^2 + 64\times 4000^2 + 16\times 7500^2}{500}\Big) - 1888^2$$

$= 2285456.$

L'écart type s **:** c'est la racine carrée de la variance. On a donc $\boxed{s = \sqrt{v}}$. Il s'exprime dans la même unité de mesure que la variable quantitative, ce qui n'est pas le cas de la variance.

On démontre par une inégalité dite de Bienaymé-Tchebichev que pour un réel positif a donné, la proportion de valeurs de la série statistique qui sont dans l'intervalle $[\bar{x} - a\times s; \bar{x} + a\times s]$ est supérieure à $1 - \dfrac{1}{a^2}$.

Ainsi plus des trois quarts des valeurs d'une série statistique sont dans l'intervalle $[\bar{x} - 2s; \bar{x} + 2s]$.

L'écart type de la distribution statistique dans Table 5 est $\sqrt{4,77} = 2,18$. Plus de 75% des valeurs sont dans l'intervalle $[4,16-2\times2,18; 4,16+2\times2,18]$ c'est-à-dire $[-0,20; 8,52]$. L'écart type approché de la distribution dans Table 6 est $\sqrt{2285456} = 1512$. Plus de 75% des valeurs sont dans l'intervalle $[1888 - 2 \times 1512; 1888 + 2 \times 1512]$ c'est-à-dire $[-1136; 4912]$.

L'écart interquartile I : C'est l'écart entre le premier et le troisième quartiles. $I = Q_3 - Q_1$. L'intervalle interquartile $[Q_1; Q_3]$ contient donc 50% des valeurs de la série. Il s'agit des valeurs les moins en queue de distribution.

L'écart interquartile de la distribution statistique de Table 5 est $5 - 2 = 3$.

L'écart interquartile de la distribution de Table 6 est $2500 - 822 = 1678$.

Les quartiles permettent de construire un diagramme représentant la distribution d'une variable quantitative : la **boîte à moustaches** aussi appelée **box-plot** (Figure 4). Ce diagramme est constitué d'une boîte dont la première arête est positionné en Q_1 et la seconde en Q_3 et de deux moustaches de longueur au plus égale à $1,5 \times (Q_3 - Q_1)$. La boîte symbolise 50% des valeurs (les valeurs en centre de distribution). La position de la médiane Me, séparant la boîte en deux, permet de visualiser l'éventuel étalement ou dissymétrie dans la répartition des valeurs. Les valeurs inférieures à $Q_1 - 1,5 \times I$ ou supérieures à $Q_3 + 1,5 \times I$ sont dites atypiques car trop éloignées des valeurs centrales. En l'absence de valeurs atypiques à gauche, l'extrémité de la moustache gauche est positionnée en la valeur minimale x_{min} qui vérifie alors $x_{min} > Q_1 - 1,5 \times I$. En l'absence de valeurs atypiques à droite, l'extrémité de la moustache droite est positionnée en la valeur maximale x_{max} qui vérifie alors $x_{max} < Q_3 + 1,5 \times I$. On verra que ce diagramme est très intéressant dans le cas d'une série bivariée quand une variable est quantitative et l'autre qualitative. En représentant simultanément les boîtes à moustaches de la variable quantitative pour chaque modalité de la variable qualitative, on peut avoir une idée du lien éventuel entre ces deux variables et comparer les étalements de la variable quantitative en fonction des modalités de la variable qualitative (Figure 7).

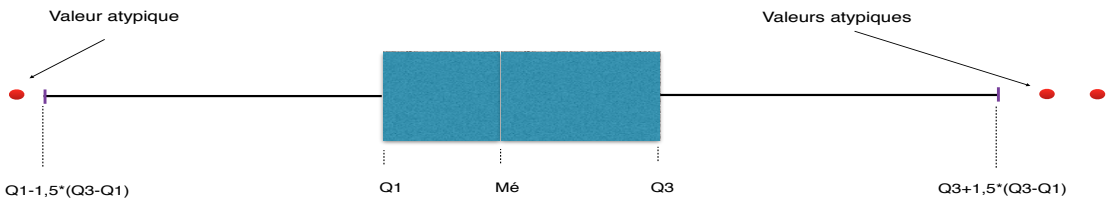


FIGURE 4 – Boîte à moustaches.

3 Série bivariée

3.1 Représentation d'une série bivariée

On va distinguer les méthodes de représentation d'une série bivariée selon la nature des deux variables statistiques concernées. Cependant, pour ce qui concerne les tableaux, leur corps a la même aspect décrit ci-dessous (table 7).

3.1.1 Tableaux de contingence

En effet, considérons deux variables statistiques A et B sur n individus statistiques. On note A_i selon le cas la ième modalité ou valeur ou classe statistique pour A. On note B_j selon le cas la jème modalité ou valeur ou classe statistique pour B. Le tri croisé de A et B nous donne le tableau de distribution d'effectifs (aussi dit tableau de contingence) suivant :

	B_1	$\cdots$	B_j	$\cdots$	B_J	Totaux
A_1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,J}$	$n_{1,.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,J}$	$n_{i,.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	$n_{I,1}$	$\cdots$	$n_{I,j}$	$\cdots$	$n_{I,J}$	$n_{I,.}$
Totaux	$n_{.,1}$	$\cdots$	$n_{.,j}$	$\cdots$	$n_{.,J}$	n

TABLE 7 – Corps de tableau pour le tri croisé de deux variables.

$n_{i,j}$ est l'effectif associé au couple (A_i, B_j),

$n_{i,.}$ est l'effectif associé à A_i,

$n_{.,j}$ est l'effectif associé à B_j,

I est le nombre de modalités (ou valeurs ou classes) distinctes observées dans l'échantillon pour A.

J est le nombre de modalités (ou valeurs ou classes) distinctes observées dans l'échantillon pour B.

Les variables considérées peuvent être de nature différente comme le montre l'exemple dans Table 8.

	Athlétisme	Basketball	Football	Totaux
$]140;150]$	14	3	20	37
$]150;160]$	25	10	32	67
$]160;170]$	41	27	59	127
$]170;180]$	30	19	45	94
$]180;190]$	18	35	29	82
Totaux	128	94	185	407

TABLE 8 – Tri croisé des variables *Sport préféré* et *Taille* pour les élèves d'un lycée.

3.1.2 Diagrammes pour deux variables qualitatives

On effectue pour chacune des modalités d'une des variables un diagramme représentant l'autre variable. Les figures ci-dessous nous fournissent deux exemples. Dans le cas où l'une des variables est ordinale, il convient de tenir compte de l'ordre des modalités dans la représentation.

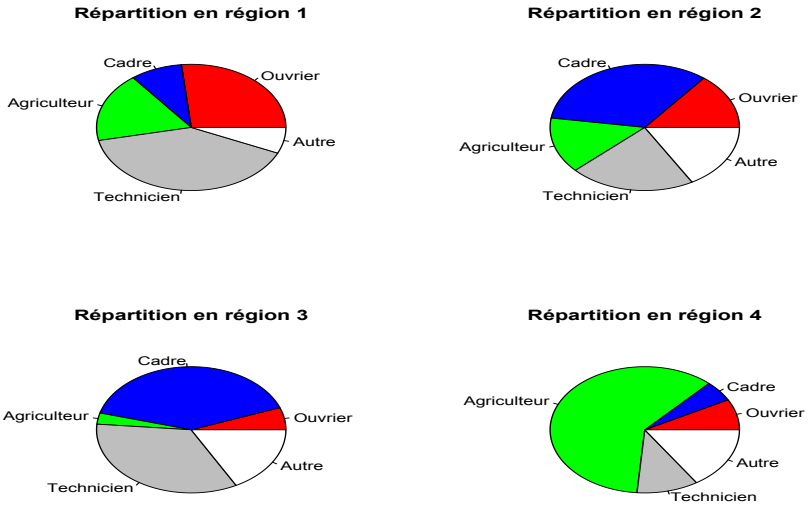


FIGURE 5 – Exemple de représentation de 2 variables qualitatives.

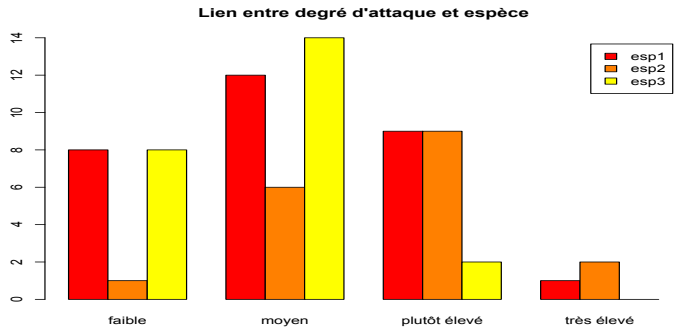


FIGURE 6 – Exemple de 2 variables qualitatives, une nominale et l'autre ordinale.

3.1.3 Diagrammes pour cas mixte

Quand une variable est qualitative et l'autre est quantitative, le diagramme le plus approprié est souvent le box-plot (aussi appelé diagramme boîte à moustaches), à raison d'un box-plot par modalité de la variable qualitative. En effet, si le nombre de modalités de la variable qualitative n'est pas trop élevé, chaque box-plot permet de visualiser, pour la variable quantitative, la position des quartiles, l'éventuelle dissymétrie de la distribution et de vérifier l'existence de valeurs atypiques (Figure 4).

On peut également effectuer un autre type de diagramme pour chaque modalité de la variable qualitative (histogramme, diagramme en bâtons,...) selon que la variable quantitative est discrète ou continue.

Dans Figure 7, la variable *Revenu du foyer* est représentée par rapport à la variable *Région de résidence*. On remarquera que les histogrammes et les boîte à moustaches donnent des informations similaires mais différentes sur la variable quantitative. Un autre exemple est donné dans Figure 9 où l'on présente la répartition des joueurs de volley-ball d'un lycée en fonction de la variable quantitatifve *Hauteur* et la variable dichotomique *Sexe*.

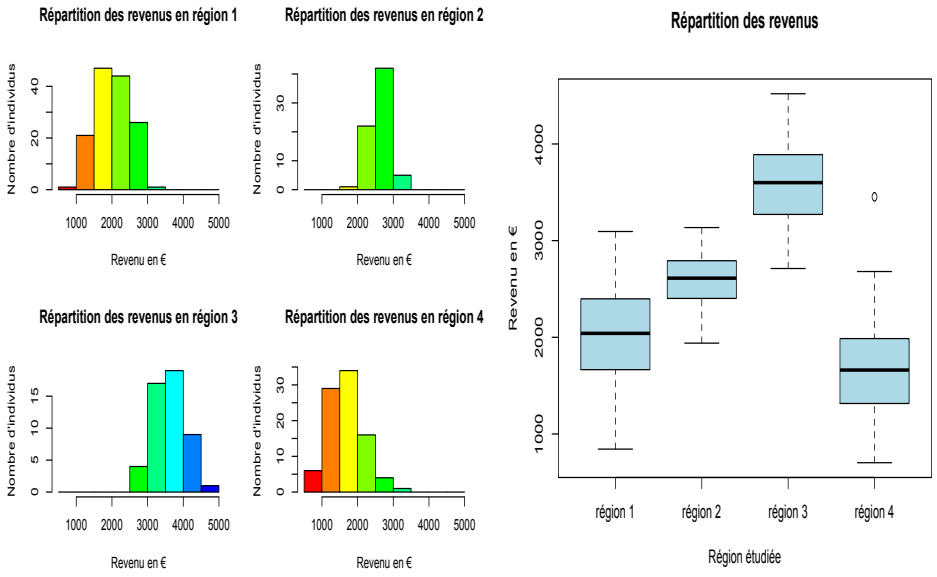


FIGURE 7 – Exemple de représentation d'une variable qualitative et d'une variable quantitative continue.

3.1.4 Diagrammes pour deux variables quantitatives

Quand les deux variables sont quantitatives, on utilise la représentation en nuage de points : chaque individu statistique est représenté dans le plan par un point de coordonnées égales aux valeurs observées sur cet individu. La proximité de deux points dans le plan correspond à la similarité des couples de valeurs associées aux deux variables statistiques. Le nuage peut avoir une allure particulière qui nous renseigne sur le lien éventuel entre les deux variables (allure rectiligne, exponentielle, parabolique, etc, ...). En général, un nuage sphérique indique une absence de corrélation entre les variables mais ceci est à confirmer par un test statistique (programme de statistique inférentielle). Un exemple d'une telle représentation est fournie par Figure 8 pour les variables *Hauteur* et *Poids* mesurées sur des 23 élèves d'une classe.

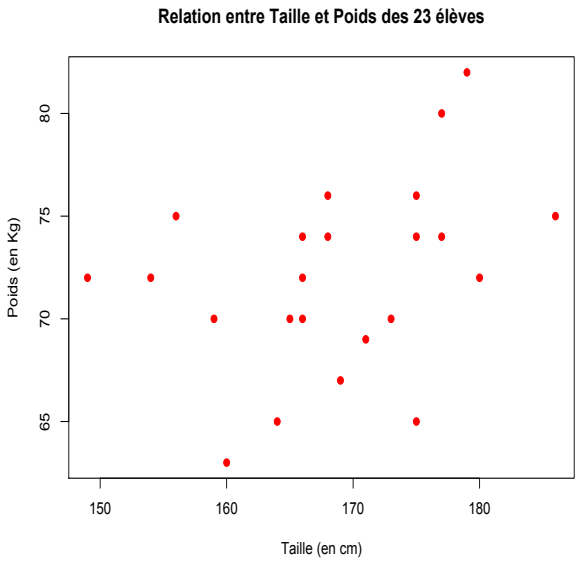


FIGURE 8 – Nuage de points.

3.2 Résumés numériques d'une série bivariée

Quand notre série est bivariée, on peut pour chacune des deux variables statistiques mener séparément une étude univariée. Aux résumés numériques vus dans le cas d'une série univariée viennent se rajouter des résumés concernant les liaisons éventuelles entre les deux variables. Il s'agit de la covariance, le coefficient de corrélation, le critère du khi-deux d'indépendance

3.2.1 Cas de deux variables qualitatives

A partir d'un tableau de contingence telle que Table 7, on peut calculer une valeur χ^2 appelée statistique du khi-deux d'indépendance, mesurant l'écart entre les effectifs observés $n_{i,j}$ et les effectifs *attendus* $\nu_{i,j}$ sous l'hypothèse d'indépendance entre les deux variables :

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{i,j} - \nu_{i,j})^2}{\nu_{i,j}}$$

avec $\nu_{i,j} = \dfrac{n_{i,.} \times n_{.,j}}{n}$.

Cette statistique est nulle dans le cas d'indépendance parfaite (profils lignes identiques et profils colonnes identiques). Plus les profils diffèrent, plus elle est élevée.

Le tableau des $\nu_{i,j}$ associé au tableau des $n_{i,j}$ est appelé **tableau d'indépendance parfaite**. Il correspond à un tableau ayant les mêmes totaux en lignes et mêmes totaux en colonnes que notre tableau de contingence observé, et ayant ses profils lignes identiques, et ses profils colonnes identiques aussi.

Exemple : Le tri croisé de la performance d'élève de sixième et le collège d'appartenance pour une commune de Guadeloupe nous fournit le tableau présenté dans Table 9.

	Faible	moyenne	Elevée	Très élevée	Totaux
Collège 1	2	15	7	1	25
Collège 2	5	10	12	3	30
Collège 3	3	7	9	6	25
Totaux	10	32	28	10	80

TABLE 9 – Tri croisé des variables *Performance* et *Collège* pour les élèves de 3 collèges.

Le tableau d'indépendance parfaite associé à celui de Table 9 est :

	Faible	moyenne	Elevée	Très élevée	Totaux
Collège 1	3,1	10	8,8	3,1	25
Collège 2	3,8	12	10,4	3,8	30
Collège 3	3,1	10	8,8	3,1	25
Totaux	10	32	28	10	80

TABLE 10 – Tableau d'indépendance parfaite associé à Table 9.

On remarquera que les $\nu_{i,j}$ (qui sont appelés **effectifs attendus sous l'hypothèse d'indépendance parfaite**) ne sont pas forcément entiers.

Le calcul de la statistique du khi-deux d'indépendance nous donne la valeur $\chi^2 = 9,42$. Pour déclarer qu'il y a indépendance ou pas entre ces deux variables, un test statistique est nécessaire. Il s'agit du test dit d'indépendance du khi-deux (programme de statistique inférentielle).

Dans le cas de deux variables ordinales, le coefficient de corrélation des rangs de Spearman r_S (voir paragraphe suivant) permet de mesurer la concordance (r_S proche de 1) ou la discordance (r_S proche de -1) entre les classements des individus statistiques basés sur ces variables.

3.2.2 Cas de deux variables quantitatives

Soient X et Y deux variables quantitatives dont le tri croisé donne le tableau de contingence Table 7 où A_i représente la valeur x_i de X, et B_j la valeur y_j de Y. Un premier indicateur de liaison entre X et Y est la covariance.

La covariance : La covariance entre deux variables statistiques X et Y observées sur les mêmes individus est le nombre :

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j}(x_i - \bar{x})(y_j - \bar{y})$$

où $\bar{x}$, $\bar{y}$ sont les moyennes respectives de X et Y.

Si $s_{xy} > 0$, cela indique que les variables ont tendance à varier dans le même sens.

Si $s_{xy} < 0$, cela indique qu'elles ont tendance à varier en sens contraire.

Le coefficient de corrélation (linéaire) : C'est le rapport r de la covariance entre X et Y sur le produit des écart-types : $r = \dfrac{s_{xy}}{s_x s_y}$

où s_{xy} est la covariance entre X et Y, et s_x, s_y les écarts-types respectifs de X et Y.

Ce coefficient est toujours compris entre -1 et + 1 et ne dépend pas des unités de mesure utilisées pour X et Y.

S'il est proche de + 1 ou - 1 , X et Y sont très corrélées linéairement : le nuage de points est presque aligné le long d'une droite (croissante si $r = +1$, décroissante si $r = -1$). S'il n'y a aucun lien (linéaire) entre X et Y, ce coefficient est proche de zéro.

Le coefficient de Spearman (ou coefficient de corrélation des rangs) : C'est, dans le cas de deux variables quantitatives X et Y mesurées sur les mêmes individus, le coefficient de corrélation entre le rang des individus pour X et le rang des individus pour Y, noté r_S.

Exemple : Le nuage de points présenté en Figure 8 provient des valeurs obtenues pour 23 élèves d'une classe de lycée. En respectant l'ordre de recueil sur les 23 élèves pour chacune des variables, les données se présentent comme indiqué ci-dessous :

Taille en cm

154 168 165 166 180 177 171 173 175 186 159 175 169 166 160 164 166 168 156 177 149 175 179

Poids en Kg :

72 74 70 70 72 74 69 70 76 75 70 74 67 74 63 65 72 76 75 80 72 65 82

Ainsi le premier élève a une taille de 154 cm et un poids de 72 Kg, le deuxième a une taille de 168 cm et un poids de 74 Kg, etc...

On peut calculer d'abord des résumés numériques pour chacune des deux séries univariées :

Indicateur statistique	*Taille*	*Poids*
Minimum	149,0	63,0
Premier quartile	164,5	70,0
Médiane	168,0	72,0
Moyenne	168,6	72,0
Troisième quartile	175,0	74,5
Maximum	186,0	82,0
Etendue	37,0	19,0
Ecart type	8,9	4,5
Ecart interquartile	10,5	4,5

TABLE 11 – Indicateurs statistiques univariés des données présentée en Figure 8.

Indicateur statistique	*(Taille, Poids)*
Covariance	14,109
Corrélation linéaire	0,341
Corrélation de Spearman	0,386

TABLE 12 – Indicateurs statistiques bivariés des données présentée en Figure 8.

Des tests de nullité du coefficient de corrélation (en statistique inférentielle) permet de vérifier le degré de liaison entre ces variables quantitatives.

3.2.3 Cas mixte

Une manière de quantifier le lien éventuel entre une variable qualitative X et une variable quantitative Y est de calculer le coefficient dit de détermination entre ces deux variables.

Le coefficient de détermination R^2 : c'est un indicateur dont la valeur comprise entre 0 et 1 est la proportion de variabilité de Y expliquée par X.

Comment la variabilité de Y est-elle quantifiée ? Si X possède I modalités, on considère que chacune de ces modalités correspond à un groupe d'individus statistiques. Le $i$ème groupe est composé des individus sur lesquels la $i$ème modalité de X a été observée.

Notons n_i l'effectif, $\bar{y}_i$ la moyenne et s_i^2 la variance du $i$ème groupe.

La variabilité totale de Y est égale à ns_y^2

(= effectif global $\times$ variance globale).

La variabilité intra-groupe est égale à $\sum_{i=1}^{I} n_i s_i^2$

(= somme de effectif de groupe $\times$ variance de groupe).

La variabilité inter-groupe est égale à $\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2$

(= somme des carrés d'écart entre moyenne de groupe et moyenne globale).

On a la formule dite d'analyse de variance :

$$\boxed{\text{Variabilité totale de } Y = \text{variabilité intra-groupe} + \text{variabilité inter-groupe.}}$$

et en fait $\boxed{R^2\text{=variance intra-groupe/variabilité totale.}}$

R^2 est d'autant plus proche de un que la variabilité intra-groupe est proche de zéro (très fort lien entre X et Y). R^2 est d'autant plus proche de zéro que la variabilité inter-groupe est proche de zéro (absence de lien entre X et Y).

La variabilité inter-groupe est donc appelée variabilité de Y expliquée par X.
La variabilité intra-groupe est appelée variabilité résiduelle.

Prenons un exemple simple : la figure 9 représente la distribution de la

variable *Hauteur* en fonction de la variable *Sexe* pour l'ensemble des 32 joueurs de volley-ball du lycée A. Quelle est la proportion de variabilité de *Hauteur* expliquée par *Sexe* ? Cette dernière variable possède deux modalités donc on peut faire un groupe pour chacune d'elles : le groupe 1 des garçons avec un effectif $n_1 = 14$, une hauteur moyenne $\bar{y}_1 = 1,76$ et une variance $s_1^2 = 0,0028$; le groupe 2 des filles avec un effectif $n_2 = 18$, une hauteur moyenne $\bar{y}_2 = 1,69$ et une variance $s_1^2 = 0,0008$. D'autre part, la variance des hauteurs est $s_y^2 = 0,0030$.

La variabilité totale de *Hauteur* est donc $ns_y^2 = 32 \times 0,0030 = 0,096$. La variabilité intra-groupe est $n_1 s_1^2 + n_2 s_2^2 = 14 \times 0,0028 + 18 \times 0,0008 = 0,0536$. La proportion de variabilité de *Hauteur* expliquée par *Sexe* est donc

$$R^2 = 0,0536/0,096 = 0,558 = 55,8\%.$$

Des tests de nullité du coefficient de détermination (en statistique inférentielle) permet de vérifier le degré de liaison entre la variable qualitative et la variable quantitative.

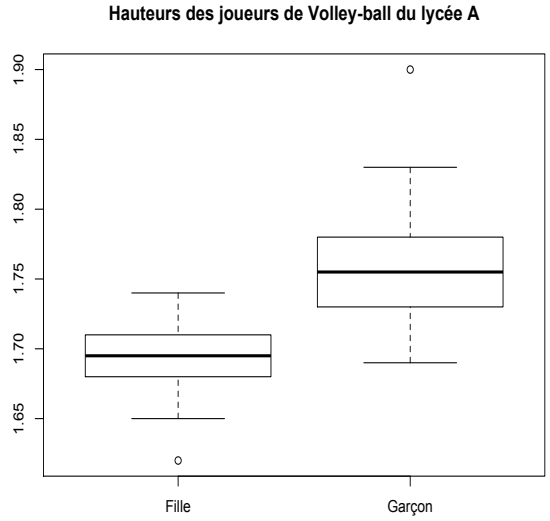


FIGURE 9 – Lien entre variable qualitative et variable quantitative continue.

4 Mini-Glossaire de Statistique Descriptive

Amplitude d'une classe statistique : En statistique univariée, une classe statistique est un intervalle. L'amplitude de la classe $]a_{i-1}\dagger; a_i]$ est $a_i - a_{i-1}$. Exemple : la classe $]16\,;43]$ est d'amplitude $43 - 16 = 27$ (unités de mesure). L'amplitude est donc la longueur de l'intervalle représentant la classe statistique.

Caractère qualitatif : Un caractère statistique est qualitatif si ses valeurs, ou modalités, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme,$\cdots$, n'ont pas de sens. Exemples : *Espèce, Statut marital, Sexe, Profession, Numéro de département de naissance ; Etat du temps constaté à une station expérimentale ; Variété de plante, Etat sanitaire, numéro de parcelle, Région.*

Un caractère qualitatif est dit **ordinal** s'il existe une hiérarchie dans ses modalités.

Un caractère qualitatif est dit **nominal** s'il n'y a pas de hiérarchie dans ses modalités.

Caractère quantitatif : Un caractère statistique est quantitatif si ses valeurs sont des nombres sur lesquels des opérations arithmétiques telles que somme, moyenne,$\cdots$, ont un sens. Exemples : *Taille, Poids, Salaire, Rendement, Note à un examen, PNB/habitant, Espérance de vie, Nombre d'habitants, Taux d'infestation.*

Un caractère quantitatif est dit **discret** si les valeurs possibles sont des nombres isolés sur l'axe réel. Dans la pratique, il s'agit souvent de données de comptage. Par exemple, si l'individu statistique est une plante, les caractères *nombre d'attaques d'un parasite, nombre de feuilles* sont quantitatifs discrets.

Un caractère quantitatif est dit **continu** s'il peut prendre toutes les valeurs dans un intervalle réel. Par exemple, si l'individu statistique est une plante, les caractères *hauteur, surface foliaire, biomasse* sont quantitatifs continus.

Caractère statistique (ou variable statistique) : propriété (ou valeur) disctinctive observée ou mesurée sur l'individu statistique. Il peut être qualitatif ou quantitatif.

Classe modale : C'est la classe ayant le plus grand effectif par unité d'amplitude. Dans le cas d'une classe modale unique, on parle de distribution continue **unimodale**.

Classe statistique : Intervalle correspondant à des valeurs observées pour un caractère quantitatif continu. Par exemple, dans le cas de *hauteur de plante* mesurée en cm, on peut établir les classes :

$$]0; 25],]25; 50],]50; 75],]75; 100],]75; 100],]100; 150].$$

Notons qu'on peut représenter la distribution d'une variable quantitative discrète en classes statistiques si elle prend de très nombreuses valeurs. Par exemple, si l'on étudie la pullulation d'insectes ravageurs sur des plantes, on peut avoir les classes statistiques suivantes

pour la variable *Nombre d'insectes sur la plante* :

$$[0;10],]10;100],]100;1000],]1000;5000].$$

Remarque : Les classes statistiques sont exclusives c'est-à-dire une valeur observée appartient à une classe et une seule.

Coefficient de corrélation (linéaire) : Le coefficient de corrélation entre deux variables statistiques X et Y observées sur les mêmes individus est le nombre r vérifiant : $r = \dfrac{s_{xy}}{s_x s_y}$

où s_{xy} est la covariance entre X et Y, et s_x, s_y les écarts-types respectifs de X et Y.

Ce coefficient est toujours compris entre -1 et + 1.

S'il est proche de + 1 ou - 1 , X et Y sont très corrélées linéairement : le nuage de points est presque aligné le long d'une droite (croissante si $r = +1$, décroissante si $r = -1$). S'il n'y a aucun lien entre X et Y, ce coefficient est nul, ou presque nul.

Coefficient de Spearman (ou coefficient de corrélation des rangs) : C'est, dans le cas de deux variables quantitatives X et Y mesurées sur les mêmes individus, le coefficient de corrélation entre le rang des individus pour X et le rang des individus pour Y.

Coefficient de variation : C'est le rapport écart-type sur la moyenne. Il est calculé pour des variables quantitatives positives : taille, durée, poids. C'est un nombre sans dimension (c'est-à-dire qu'il est indépendant du choix des unités de mesure). Il permet de comparer la dispersion autour de la moyenne de variables statistiques ayant des échelles ou des unités de mesure différentes.

Courbe cumulative : On l'utilise quand la variable quantitative est continue. Pour la tracer, on relie par des segments de droite les points $(a_i, F(a_i))$ pour $i = 0, \cdots, k$, les a_i étant les limites des k classes statistiques concernées et $F(a_i)$ la fréquence cumulée en a_i.

Diagramme circulaire (ou à secteurs angulaires ou camembert) : Il s'agit d'un disque divisé en sections angulaires. Chaque section correspond à une modalité de la variable qualitative et a un angle proportionnel à la fréquence de cette modalité.

Diagramme cumulatif : C'est le tracé de la fonction qui à tout x associe $F(x) =$ proportion d'observations $\leq x$. On l'utilise dans le cas d'une variable quantitative discrète et l'on obtient une courbe dite en escalier.

Diagramme en bandes : Chaque valeur distincte de la variable qualitative est représentée par une bande verticale de longueur l'effectif ou la fréquence associée à cette valeur.

Diagramme en bâtons : Chaque valeur distincte de la variable quantitative discrète est représentée par un bâton vertical de longueur l'effectif ou la fréquence associée à cette valeur.

Diagramme en étoiles : Si on a plusieurs variables quantitatives, on peut représenter chaque individu statistique par un polygone. Les valeurs pour un individu sont représentées par des points reliés entre eux par des segments de manière à former un polygone. Il y a donc autant d'arêtes dans le polynome associé que de variables étudiées. Ainsi, si on étudie 5 variables, on a un pentagone, 6 un hexagone, etc...

Diagramme figuratif : Chaque modalité de la variable qualitative est représentée par une image (ordinateur, maison, plante, avion,...) rappelant la variable (ou la population) statistique étudiée, et de taille proportionnelle à la fréquence de cette modalité.

Distribution statistique : Ensemble des modalités, valeurs, ou classes d'une variable, avec les effectifs observés correspondants. Une distribution d'effectifs univariée est la donnée de $(x_1, n_1), \cdots, (x_k, n_k)$, où les x_i sont les valeurs distinctes du caractère statistique et n_i l'effectif associé x_i

Ecart interquartile : C'est la différence I entre le 1er et le 3ème quartile : $I = Q_3 - Q_1$.

Ecart-type : pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, l'écart-type noté s_x est donné par la formule :

$$s_x = \sqrt{\frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)}$$

où $\bar{x}$ est la moyenne de la série.

Etendue : C'est l'écart entre la plus petite et la plus grande valeur dans la série statistique.

Fractiles (ou quantiles) : On appelle fractiles des valeurs divisant une série en plusieurs parties. Pour une valeur α comprise entre 0 et 1, le fractile d'ordre α noté q_α est, par définition, tel que la proportion de valeurs inférieures à q_α vaut α. On a donc $F(q_\alpha) = \alpha$. Les fractiles divisant la série en k parties d'effectifs égaux ont parfois une dénomination commune : Les 3 quartiles divisent la série en 4 parties d'effectifs égaux, les 9 déciles en 10, les 99 centiles en 100. Les 3 quartiles sont notés Q_1, Q_2, Q_3 (Q_2 étant la médiane).

Fréquence (ou fréquence relative) : C'est la proportion (ou le pourcentage) d'individus pour lesquels une variable statistique a pris une valeur donnée. Si, sur 150 familles, 50 ont 2 enfants, on dira que la fréquence f_i correspondant à la valeur $x_i = 2$ de la variable *nombre d'enfants*, est : 0,33 ou 1/3 ou 33,33%.

Fréquence cumulée : Résultat de l'addition, de proche en proche, des fréquences d'une distribution observée, soit en commençant par le 1er :

$F_1 = f_1$, $F_2 = f_1 + f_2$, $\cdots$, $F_i = f_1 + f_2 + \cdots + f_i$ (fréquences cumulées croissantes),

soit en commençant par le dernier (en notant k le nombre total de valeurs distinctes) :

$F_k^* = f_k, F_{k-1}^* = f_k + f_{k-1}, \cdots, F_i^* = f_k + f_{k-1} + \cdots + f_i$ (fréquences cumulées décroissantes).

Histogramme : Graphique permettant de représenter une distribution continue regroupée en classes : rectangles juxtaposés dont les bases sont les classes, et les surfaces sont proportionnelles aux effectifs (ou fréquences) associés. Donc les hauteurs de rectangle sont proportionnelles aux **effectifs par unité d'amplitude** : pour la classe $]a_{i-1}, a_i]$ d'effectif n_i, la hauteur du rectangle associée est $h_i = n_i/(a_i - a_{i-1})$.

Indépendance : Deux variables statistiques X et Y sont dites indépendantes si la distribution de Y conditionnelle à $X = x$, pour tout x, ne dépend pas de x. Cela signifie que les profils des lignes du tableau de contingence sont identiques, ou de faÁon équivalente que les profils des colonnes du tableau de contingence sont identiques, et donc que la distribution de fréquences conditionnelle est égale à la distribution de fréquences marginale.

Indicateur statistique (ou résumé numérique) : C'est un nombre permettant de résumer numériquement les traits principaux d'une distribution statistique. On parle aussi de résumé numérique. On distingue principalement deux types d'indicateurs :

- les indicateurs de position (ou de tendence centrale) qui donne une idée de l'ordre de grandeur de la série : moyenne, médiane, mode, quartile,...
- les indicateurs de dispersion qui donnent une idée de la variabilité dans la série : étendue, variance, écart-type, écart interquartile,...

Inégalité de (Bienaymé)-Tchébichev : Pour toute série statistique $x_1, \cdots, x_n$ de moyenne $\bar{x}$ et d'écart-type s_x , la proportion de valeurs dans l'intervalle $[\bar{x} - k \times s_x; \bar{x} + k \times s_x]$ est supérieure à $1 - \dfrac{1}{k^2}$, pour tout nombre $k \geq 1$. Par exemple (pour $k = 2$), plus de 75% des valeurs sont dans : $[\bar{x} - 2s_x; \bar{x} + 2s_x]$, c'est-à-dire s'écartent de la moyenne de moins de 2 écart-types.

Intervalle interquartile : C'est l'intervalle dont les bornes sont le 1er et le 3ème quartile : $[Q_1, Q_3]$. Il contient 50% des observations ; rappelons que 25% des valeurs de la série statistique sont inférieures à Q_1 et 25% sont supérieures à Q_3.

Intervalle médian : C'est l'intervalle dont toutes les valeurs vérifient la propriété de la médiane pour la série statistique étudiée.

Médiane : C'est le fractile d'ordre 0.5. La médiane est notée M_e et vérifie $F(M_e) = 0.5$. Il y a autant de valeurs inférieures à M_e que supérieures à M_e dans la série statistique.

Mode : C'est la valeur la plus fréquente dans la série statistique. Le mode n'est pas forcément unique. Quand il existe plusieurs modes, la distribution statistique est dite multimodale.

Moyenne : C'est la somme des valeurs divisée par le nombre de valeurs. Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la moyenne notée

$\bar{x}$ est donné par la formule : $\bar{x} = \frac{1}{n}(n_1x_1 + \cdots + n_kx_k)$.

Nuage de points : Ensemble de points isolés représentés dans un graphique cartésien. Une séries à deux caractères quantitatifs $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$ peut être représentée par les n points $M_1, M_2, \cdots, M_n$ de coordonnées $(x_1, y_1); (x_2, y_2); \cdots; (x_n, y_n)$.

Population statistique : Une population statistique est un ensemble d'éléments sur lesquels porte une étude. Exemples : ensemble des électeurs d'une région ; ensemble des accidents de la route dans une zone, pendant une période ; ensemble de parcelles cultivées sur lesquelles on peut mesurer un rendement ; ensemble de pays pour lesquels on dispose de données géographiques ou économiques, ...

Profil : C'est une distribution conditionnelle de fréquences (et non d'effectifs). Dans un tableau de contingence à I lignes et J colonnes, le profil de la ligne i est obtenu en divisant les effectifs $n_{i1}, n_{i2}, \cdots, n_{iJ}$ de cette ligne par la somme $n_{i.}$ de ces effectifs. On obtient : $\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \cdots, \frac{n_{iJ}}{n_{i.}}$. De même, le profil de la colonne j est : $\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \cdots, \frac{n_{Ij}}{n_{.j}}$. où $n_{.j}$ est la somme des effectifs de cette colonne.

Quantiles : Voir fractiles.

Quartiles : Ce sont les 3 fractiles d'ordre 0,25, 0,5 et 0,75 notés respectivement Q_1, Q_2, Q_3. Ils divisent la distribution statistique en quatre parties d'égale fréquence. Q_1 est le premier quartile, Q_3 le troisième. Q_2 est la médiane.

Résumé numérique : Voir indicateur statistique.

Série statistique (ou distribution observée) : Séquence des modalités, ou valeurs d'une variable statistique. L'ordre correspond souvent à l'ordre chronologique de recueil des observations.

Statistique Descriptive : Ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer des données nombreuses et variées.

Statistique Descriptive univariée : La Statistique Descriptive univariée étudie un seul caractère statistique, et ne s'intéresse donc pas aux liens éventuels entre plusieurs caractères.

Statistique Descriptive bivariée : La Statistique Descriptive bivariée concerne l'extraction d'information sur deux caractères statistiques, et leurs liens éventuels.

Statistique Descriptive multivariée : La Statistique Descriptive multivariée analyse un nombre k (> 2) de variables mesurées ou observées simultanément sur les mêmes individus. Elle permet de mettre en évidence le type de lien existant éventuellement entre ces variables.

Statistique Inférentielle : La Statistique Inférentielle utilise la théorie des probabilités pour extrapoler à toute la population statistique, des résultats observés sur des échantillons. Elle inclut *l'Estimation Statistique* et la *Théorie des Tests d'hypothèses*.

Tableau de contingence : C'est le tableau d'effectifs obtenu par tri croisé d'une série bivariée (ou multivariée).

Tri à plat d'une série statistique brute : C'est l'inventaire des modalités ou valeurs rencontrées dans la série, avec les effectifs correspondants.

Tri croisé d'une série bivariée : C'est l'inventaire des modalités ou valeurs rencontrées conjointement dans une série comportant deux variables mesurées pour chaque individu statistique, avec les effectifs correspondants.

Variable statistique (ou caractère statistique) : propriété (ou valeur) disctinctive observée ou mesurée sur l'individu statistique. Elle peut être qualitative ou quantitative.

Variance : Pour une distribution d'effectifs $(x_1, n_1), \cdots, (x_k, n_k)$, où x_i a pour effectif associé n_i, la variance notée s_x^2 est donnée par la formule :

$s_x^2 = \frac{1}{n}(n_1(x_1 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2)$. La variance est le carré de l'écart-type.

Quelques exemples de diagrammes représentant une distribution statistique

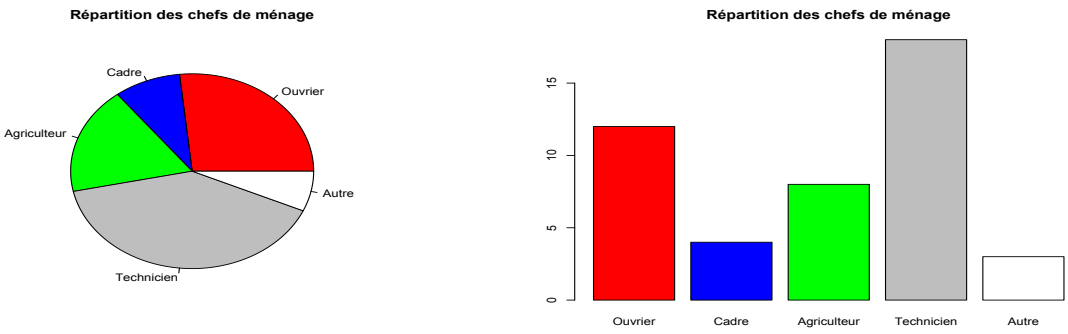


FIGURE 10 – Exemples de représentation d'une variable qualitative.

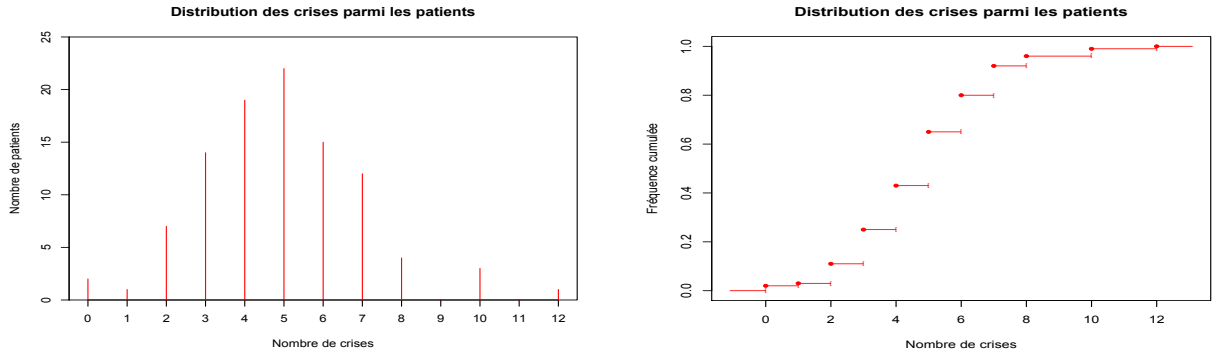


FIGURE 11 – Exemple de représentation d'une variable quantitative discrète.

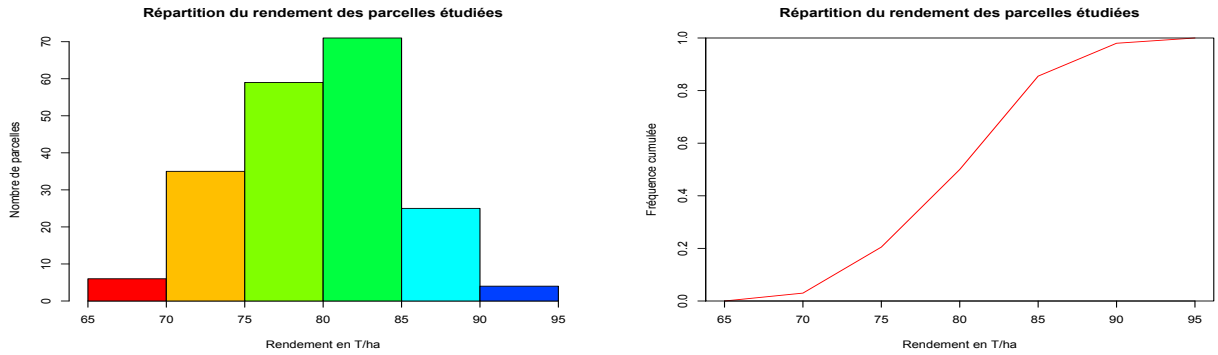


FIGURE 12 – Exemples de représentation d'une variable quantitative continue.

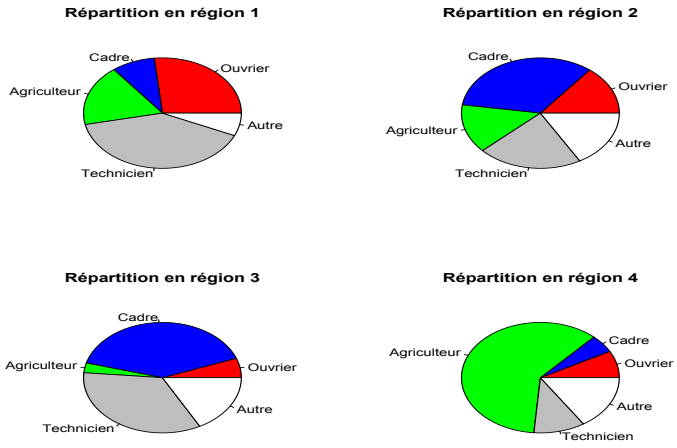


FIGURE 13 – Exemple de représentation de 2 variables qualitatives.

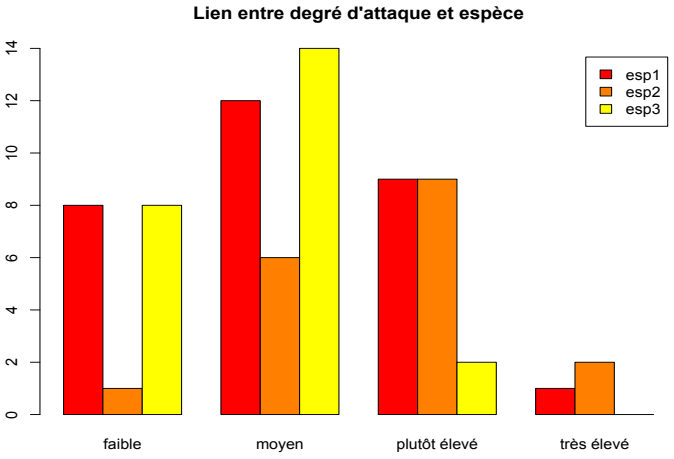


FIGURE 14 – Exemple de 2 variables qualitatives,une nominale et l'autre ordinale.

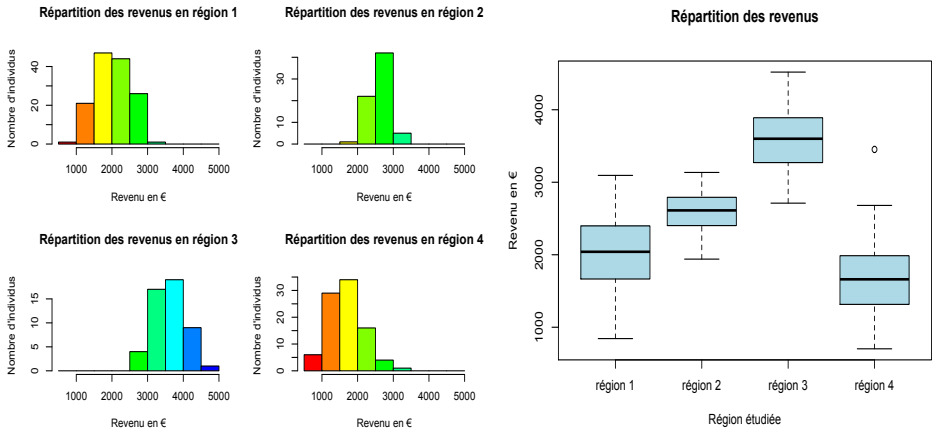


FIGURE 15 – Exemple de représentation d'une variable qualitative et d'une variable quantitative continue.

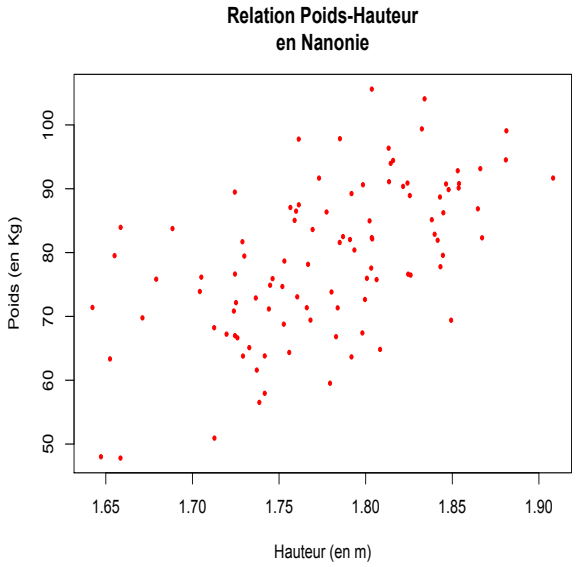
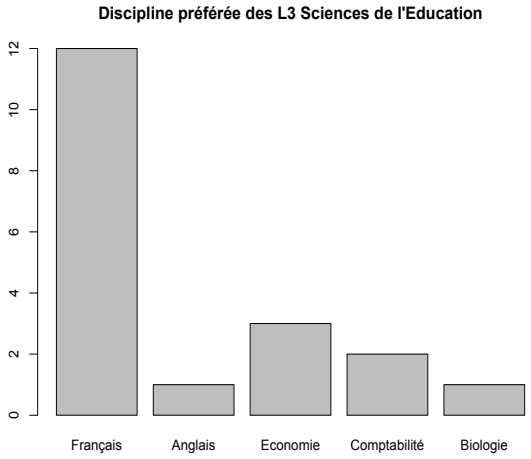
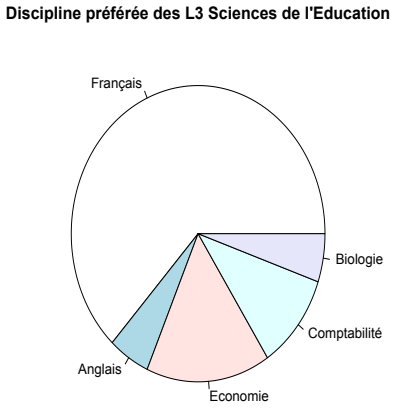


FIGURE 16 – Exemple de représentation de deux variables quantitatives.

5 Exercices

Exercice 1 *Préciser les diagrammes ci-dessous ? Quel est le caractère représenté ?*





Exercice 2 *Classer les variables ci-dessous selon leur type :*

Langue maternelle, Taille, Pays d'origine, Profession, Sexe, Nationalité, Poids, Pointure, Race, Couleur des yeux, Dextérité, Nombre d'enfants, Revenu mensuel, Taux d'endettement.

Exercice 3 *Proposer des exemples de variable quantitative transformée en variable qualitative. Préciser les modalités de cette dernière.*

Exercice 4 *A quels types de variable correspondent ces propriétés ?*

1. *Ses valeurs ne possèdent pas d'ordre. Elles sont uniquement définies par des noms.*
2. *Elle s'exprime toujours à l'aide d'une unité de mesure.*
3. *Ses valeurs sont des noms mais correspondent à une hiérarchisation (c'est-à-dire possèdent un certain ordre)*
4. *Ses valeurs peuvent être n'importe quel nombre sur un intervalle.*
5. *Ses valeurs sont des nombres particuliers. Par conséquent, elle ne prend pas toutes les valeurs sur un intervalle.*

Exercice 5 *Parmi ces assertions, préciser celles qui sont* **vraies**, *celles qui sont* **fausses**.

1. *On appelle variable, une caractéristique que l'on étudie*
2. *La tâche de la Statistique Descriptive est de recueillir des données*
3. *La tâche de la Statistique Descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques*
4. *Les valeurs pouvant être mesurées pour une variable quantitative sont appelées valeurs possibles de la variable quantitative*
5. *Une variable est quantitative si ses valeurs sont des nombres, sinon c'est une variable qualitative*
6. *En Statistique, on classe les variables selon différents types*
7. *Les valeurs des variables qualitatives sont aussi appelées modalités*
8. *La variable Sexe est dichotomique*
9. *Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une et une seule modalité*
10. *Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative*
11. *La variable quantitative poids d'automobile peut être reclassée en compacte, intermédiaire et grosse*
12. *En pratique, lorsqu'une variable quantitative discrète prend un grand nombre de valeurs distinctes, on la traite comme continue*

Exercice 6 *Soit la liste suivante des prénoms d'un groupe d'étudiants suivis entre parenthèses d'une indication du nombre de livres lus dans l'année (A = peu, B = moyen, C = beaucoup, D = exceptionnel) :*

Pierre (C), Paul (C), Jacques (A), Ralph (B), Abdel (A), Sidonie (B), Henri (C), Paulette (B), Farida (B), Laure (C), Kevin (D), Carole (B), Marie-Claire (A), Jeanine (C), Julie (C), Ernest (C), Cindy (C), Vanessa (D), José (C), Aurélien (C).

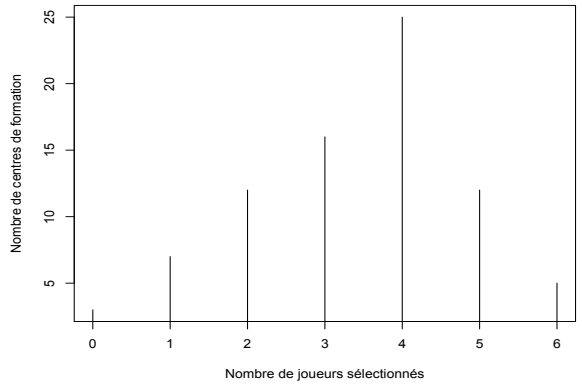
1. *Quelle est la nature de la variable* appétit de lecture *?*
2. *Construire le tableau représentatif de cette distribution.*
3. *Représenter cette distribution à l'aide d'un diagramme en tuyaux d'orgue.*

Exercice 7 *Calculer les fréquence relatives et tracer le diagramme le plus adapté à la répartition du groupe sanguin d'un groupe d'élèves :*

O : 140 ; A : 147 ; AB : 63.

Exercice 8

1. *Commenter le diagramme ci-dessous.*
2. *Quelle est la variable représentée ? Quel est l'individu statistique ?*
3. *Retrouver le tableau d'effectifs associé à ce diagramme.*
4. *Calculer la moyenne de joueurs sélectionnés par centre de formation*
5. *Calculer l'écart type du nombre de joueurs sélectionnés.*



Exercice 9 *Pour un collège de Baie-Mahault, la distribution d'effectifs du lieu de résidence d'élève est :*

Code du lieu de résidence	*Effectif*
97122	*148*
97170	*122*
97139	*25*
97129	*59*

1. *Représenter cette distribution à l'aide d'un diagramme.*
2. *Le calcul de la moyenne de cette distribution a t'il un sens ?*

Exercice 10 *La manière de choisir un échantillon est-elle un facteur important pour pouvoir tirer des conclusions fiables à partir d'un échantillon ? La taille d'un échantillon influence t-elle les conclusions tirées de cet échantillon ?*

Exercice 11 *Le service de statistiques d'un rectorat a enregistré les actes de violence au sein des collèges au cours de l'année scolaire 2013-2014 ayant conduit à un procès verbal. Un total de 75 actes ont ainsi été présentés dans le tableau suivant :*

Incident	*Violence verbale entre élèves*	*Violence physique entre élèves*	*Violence verbale et physique entre élèves*	*Violence envers un adulte*
Nombre	*23*	*17*	*28*	*? ?*

1. *Compléter le tableau puis préciser la population statistique , le caractère étudié et son type.*
2. *Représenter cette distribution à l'aide d'un graphique.*

Exercice 12 *Soit la série statistique correspondant aux revenus mensuels du foyer de 28 élèves en milliers d'euros.*

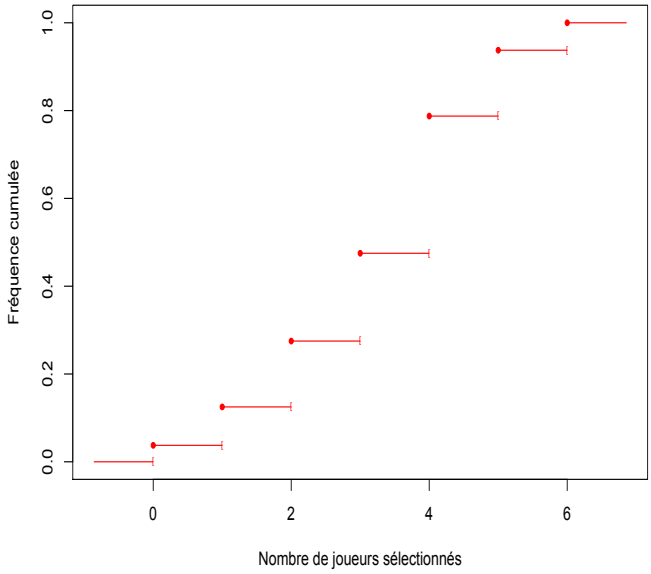
5,2 8,4 1,8 3,1 13,7 12,1 19,5 2,4 1,6 2,7 19,3 10,4 19,8 2,5 1,5 2,1 7,4 2,5 3,0 13,5 7,1 8,2 1,4 3,2 1,3 1,2 1,9 1,1

1. *Présenter cette série sous forme de distribution d'effectifs à 4 classes statistiques d'égales amplitudes.*
2. *Calculer les fréquences et les fréquences cumulées.*
3. *Construire l'histogramme associé à cette distribution d'effectifs.*
4. *Construire la courbe cumulative associée à cette série.*

Exercice 13 *A quelles formes de présentation de données correspondent ces propriétés ?*

1. *Il donne une bonne idée des données, mais on lui préfère en général les graphiques.*
2. *Il n'est pas nécessaire de lire des nombres. D'un simple coup d'oeil, on a une vision d'ensemble des données.*
3. *Il représente les fréquences ou les effectifs par des barres dont les hauteurs égalent les fréquences.*
4. *Plus le nombre de données est grand, plus cette présentation est inefficace. Elle ne nous donne pas une bonne vue d'ensemble.*

Exercice 14 *Quel est le type du diagramme ci-dessous ? Quelle est la variable statistique considérée ?*



Exercice 15 *On fait une étude sur la population de Guadeloupe. On veut savoir s'il y a un lien entre*

- *la langue maternelle et le niveau de scolarité*
- *le niveau de scolarité et le revenu*
- *le quotient intellectuel et le revenu*
- *le quotient intellectuel et le sexe*

Préciser les variables statistiques à considérer. Pour chacune d'elles, préciser leur type, les modalités ou valeurs qu'elles peuvent prendre.

Exercice 16 *Parmi ces assertions, préciser celles qui sont* **vraies**, *celles qui sont* **fausses**.

1. *La moyenne d'une série de valeurs distinctes peut être supérieure à la valeur maximale.*
2. *La moyenne d'une série de valeurs distinctes peut être inférieure à la valeur minimale.*
3. *La variance peut être strictement négative.*
4. *L'écart type n'est jamais strictement inférieur à zéro.*

Exercice 17 *On a demandé aux enfants d'une classe : Combien y a-t-il d'enfants dans votre famille ? La collecte des données nous fournit les données brutes :*

1, 2, 1, 3, 1, 4, 2, 1, 3, 1, 2, 5, 2, 1, 1, 3, 2, 1, 2, 3, 1, 1, 1, 2, 4, 2, 1, 3.

1. *Présenter le tableau d'effectifs associé à cette série.*
2. *Calculer la moyenne, la médiane et le mode de cette série statistique.*
3. *Quels sont l'étendue et l'écart type de cette distribution.*

Exercice 18

1. *Comment appelle t-on l'ensemble A de tous les objets que l'on étudie ?*
2. *Comment appelle t-on un sous-ensemble choisi dans A ?*
3. *Comment appelle t-on un élément de A ?*
4. *Comment appelle t-on le nombre d'objets composant une population ou un échantillon ?*
5. *Lorsque l'on veut connaître certaines caractéristiques d'une population, on dit qu'on enquête sur la population : Vrai* $\square$ *Faux* $\square$
6. *Une enquête peut être réalisée auprès de toute la population ou sur un échantillon : Vrai* $\square$ *Faux* $\square$
7. *Une corrélation est une enquête réalisée auprès de toute la population : Vrai* $\square$ *Faux* $\square$
8. *Les tableaux et graphiques sont utilisés pour donner une meilleure vue d'ensemble des données : Vrai* $\square$ *Faux* $\square$

Exercice 19 *La répartition des moyennes annuelles des 400 élèves de sixième d'un établissement vous est donnée sous la forme d'un tableau :*

Moyenne annuelle	*Nombre d'élèves*
]0 ; 4]	*85*
]4 ; 8]	*112*
]8 ; 10]	*98*
]10 ; 12]	*67*
]12 ; 15]	*23*
]15 ; 20]	*15*

1. *Quelle est la population statistique ? Quelle est la variable étudiée ?*
2. *Quelle est la valeur approchée de la moyenne des moyennes annuelles ?*

3. *Quelle est l'écart-type approché des moyennes annuelles ?*
4. *Quelle est la médiane des moyennes annuelles ?*
5. *Quelle est l'étendue, l'écart interquartile de cette distribution ?*

Exercice 20 *Vingt étudiants ont choisi leur module de langue de la façon suivante : ESPAGNOL, ESPAGNOL, ANGLAIS, PORTUGAIS, ANGLAIS, ESPAGNOL, PORTUGAIS, PORTUGAIS, ALLEMAND, ANGLAIS, ESPAGNOL, ANGLAIS, ANGLAIS, ESPAGNOL, ESPAGNOL, ESPAGNOL, ANGLAIS, PORTUGAIS, ALLEMAND, ANGLAIS.*

1. *Déterminer la distribution de fréquences de cette série statistique.*
2. *Préciser le type de variable étudiée puis donner son mode.*

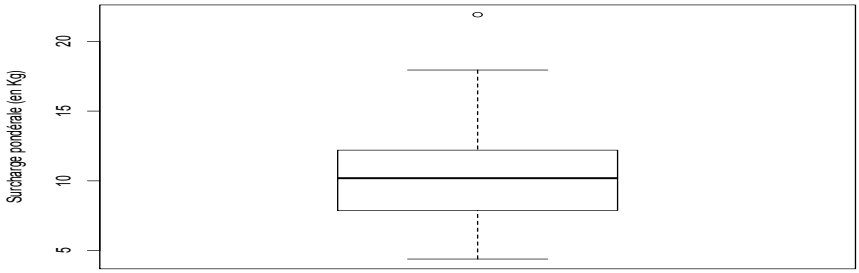
Exercice 21 *Cinquante Ficus soumis à des conditions expérimentales identiques sont mesurés. La série statistique obtenue est la suivante : 24 37 41 25 29 41 32 21 24 27 28 34 12 23 32 31 27 26 38 54 42 35 48 27 20 34 28 29 37 56 31 33 24 26 18 26 54 32 48 13 43 53 45 26 35 40 56 61 28 31.*

1. *Classer les données en classes d'amplitude 5.*
2. *Représenter la distribution d'effectifs sous forme de tableau puis d'histogramme.*
3. *Tracer la boîte à moustaches de cette distribution.*

Exercice 22 *Un quartier résidentiel comprend 99 unités d'habitation ayant une valeur locative moyenne de 1000 EUR et une valeur locative médiane de 900 EUR. Deux nouvelles unités d'habitation sont construites dans le quartier : l'une a une valeur locative de 700 EUR et l'autre, une villa luxueuse, a une valeur locative de 11400 EUR.*

1. *Quelles sont les nouvelles moyenne et médiane de valeur locative pour le quartier ?*
2. *Pouvait-on s'attendre à de tels résultats ?*

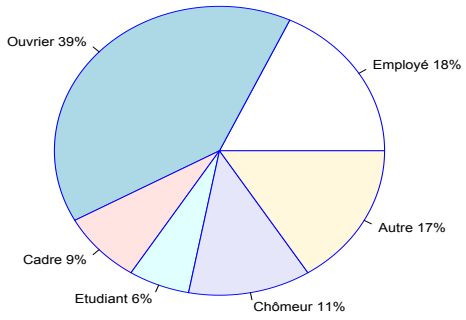
Exercice 23 *Une étude sur l'alimentation des enfants scolarisés fournit le diagramme suivant concernant leur surcharge pondérale.*



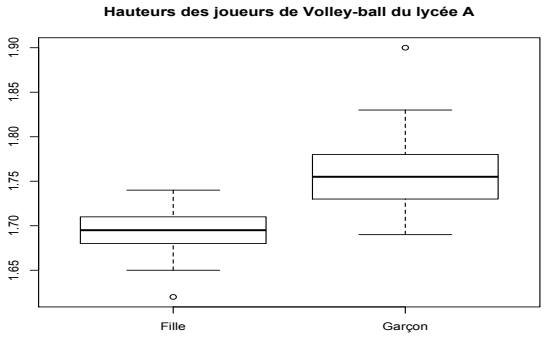
1. *Estimer les quartiles et interpréter les.*
2. *Estimer l'intervalle interquartile et interpréter.*
3. *Commenter la valeur extérieure.*

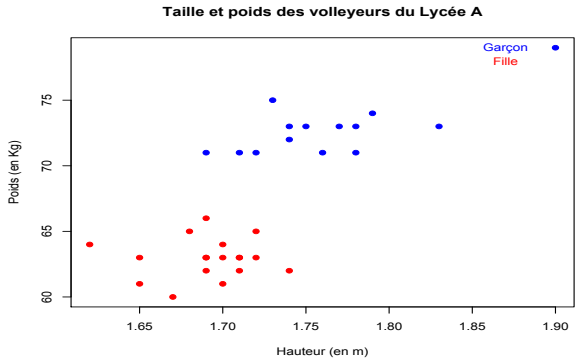
Exercice 24 *On considère la situation professionnelle du chef de famille pour les élèves d'un établissement scolaire. Vérifier que le tableau et le diagramme suivants correspondent à la même distribution.*

Situation professionnelle	*Employé*	*Ouvrier*	*Cadre*	*Etudiant*	*Chômeur*	*Autre*
Effectif	*27*	*60*	*12*	*9*	*18*	*24*



Exercice 25 *Commenter les graphiques ci-dessous, en essayant d'y extirper le maximum d'information.*





Exercice 26 *Le tableau ci-dessous donne, en pourcentage, la répartition de la population active selon le secteur d'activité, dans cinq pays. Proposer une représentation graphique de ces données.*

	primaire	*secondaire*	*tertiaire*
Allemagne	*24*	*44*	*32*
USA	*13*	*38*	*49*
France	*5*	*51*	*44*
Italie	*42*	*32*	*26*
Russie	*44*	*29*	*27*

Exercice 27

1. *Commenter les deux pages suivantes en essayant dêtre le plus exhaustif que possible (PCS =Profession ou Catégorie Socioprofessionnelle).*
2. *Montrer qu'il n'y a pas de valeurs extérieures supérieures pour le pourcentage de néobacheliers issus du département ou de départements limitrophes en 2011-2012 mais que l'université de Montpellier a une valeur extérieure inférieure pour ce pourcentage.*
3. *Montrer que l'université de la Réunion a une valeur extérieure supérieure pour le pourcentage de néobacheliers issus de PCS défavorisées en 2011-2012.*

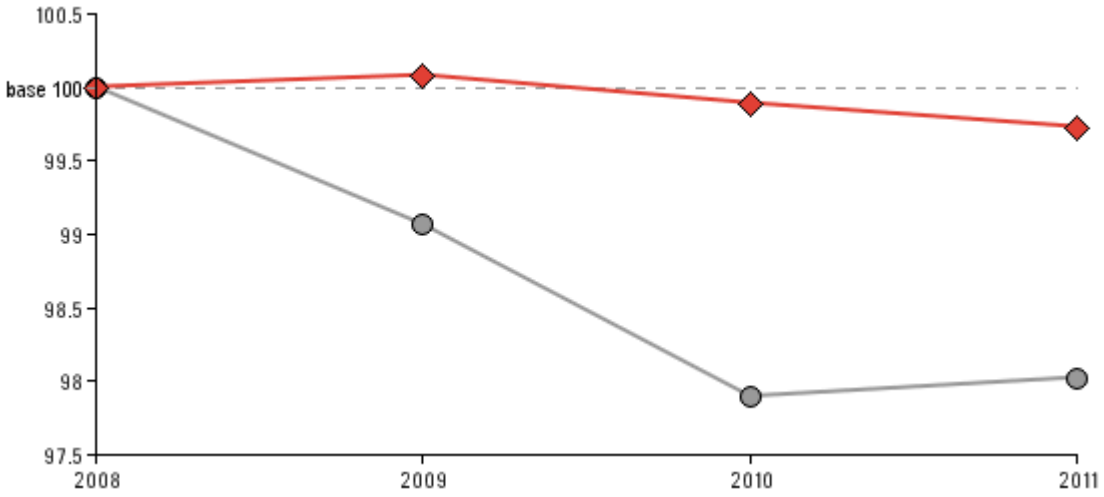
% des néobacheliers issus du département ou des départements limitrophes

Définition : Part des néobacheliers issus du département ou des départements limitrophes de leur unité d'inscription (département d'obtention du baccalauréat) parmi les néobacheliers de l'établissement (inscriptions principales).

Source : MESR-DGESIP/DGRI-SIES : SISE

	-	2008-09	2009-10	2010-11	2011-12
Antilles-Guyane	-	99,2	99,3	99,1	98,9
Les universités françaises	-	84,7	83,9	82,9	83,0

Évolutions historiques comparées (université et référence nationale) - indice base 100 en 2008-09

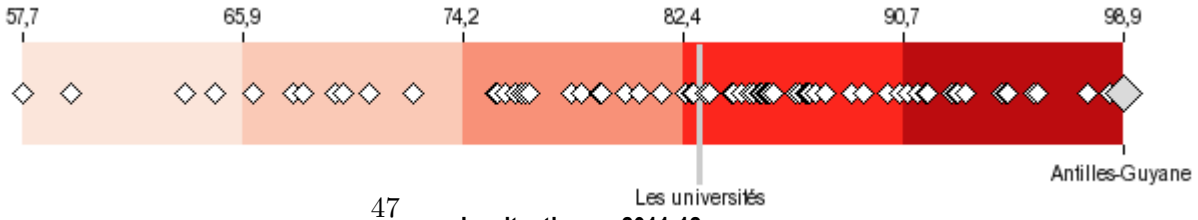


Comparaison inter-universitaire en 2011-12

premier décile	premier quartile	médiane	dernier quartile	dernier décile
69,4	78,2	85,3	91,2	94,4

Positionnements et valeurs				
77. Montpellier 3	...	4. Valenciennes	3. Polynésie	2. La Réunion	**1. Antilles-Guyane**
57,7		95,8	97,6	98,4	**98,9**

Répartition des universités par ordre croissant en 2011-12



La situation en 2011-12

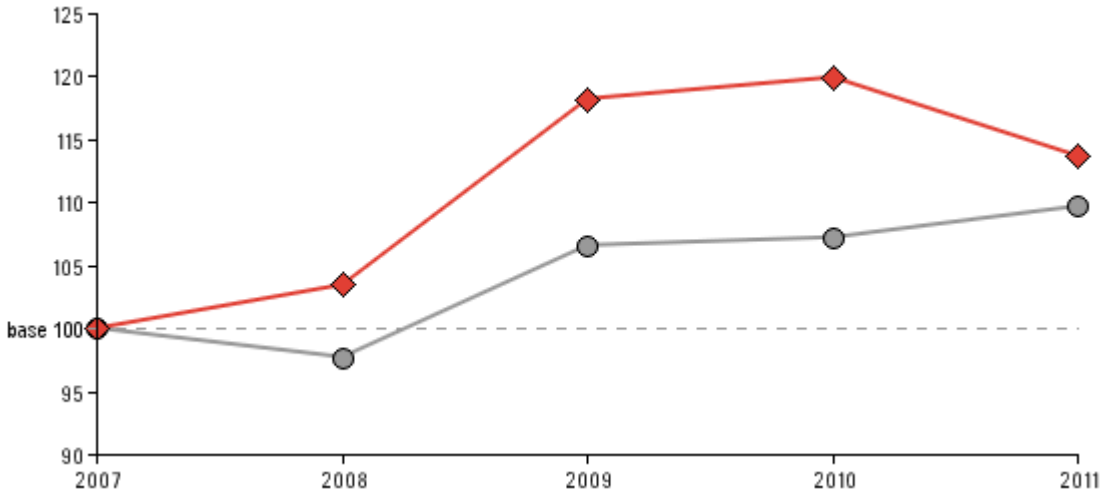
% de néobacheliers issus de PCS défavorisées

Définition : Part de néobacheliers issus de PCS défavorisées (ouvrier qualifié, ouvrier non qualifié, ouvrier agricole, retraité employé et ouvrier, chômeur n'ayant jamais travaillé, personne sans activité professionnelle) parmi les néobacheliers de l'établissement.

Source : MESR-DGESIP/DGRI-SIES : SISE

	2007-08	2008-09	2009-10	2010-11	2011-12
Antilles-Guyane	27,5	28,5	32,5	33,0	31,3
Les universités françaises	20,1	19,7	21,5	21,6	22,1

Évolutions historiques comparées (université et référence nationale) - indice base 100 en 2007-08

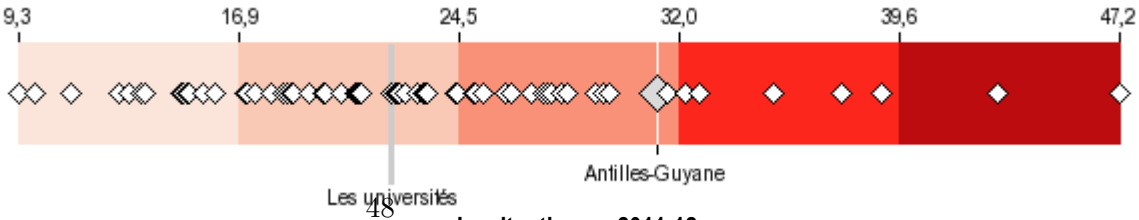


Comparaison inter-universitaire en 2011-12

premier décile	premier quartile	médiane	dernier quartile	dernier décile
14,9	18,0	21,6	26,1	31,3

Positionnements et valeurs						
77. Paris 6	...	10. Littoral	**9. Antilles-Guyane**	8. Lille 3	...	1. La Réunion
9,3		29,6	**31,3**	31,6		47,2

Répartition des universités par ordre croissant en 2011-12



La situation en 2011-12

6 Corrigés des exercices

Corrigé de l'exercice 1 *Il s'agit, pour le graphique de gauche, du diagramme en bandes (ou en tuyaux d'orgue) et, pour le graphique de droite, du diagramme à secteurs angulaires (ou camembert). Le caractère (ou variable) représenté est* la discipline préférée des L3 sciences de l'éducation.

Corrigé de l'exercice 2 *Les types possibles sont : qualitatif nominal (QN), qualitatif ordinal (QO), quantitatif discret (QD) et quantitatif continu (QC).*

Langue maternelle (QN), Taille (QC), Pays d'origine (QN), Profession (QN), Sexe (QN), Nationalité (QN), Poids (QC), Pointure (QD), Race (QN), Couleur des yeux (QN), Dextérité (QO), Nombre d'enfants (QD), Revenu mensuel (QC), Taux d'endettement (QC).

Corrigé de l'exercice 3 *Les variables quantitatives dans le tableau c-dessous peuvent être transformées en variable qualitative ordinale. Les modalités de cette dernière sont précisées dans la seconde colonne.*

Variable quantitative	*Modalités enviseageables*
Hauteur	*Petit, Moyen, Grand*
Poids	*Très léger, Léger, Moyen, Lourd, Très lourd*
Rendement	*Faible, Moyen, Elevé*
Chiffre d'affaire	*Modéré, Moyen, Important, Très important*
Cylindrée	*Petite, Moyenne, Grosse*

Corrigé de l'exercice 4 *Les types possibles sont : qualitatif nominal (QN), qualitatif ordinal (QO), quantitatif discret (QD) et quantitatif continu (QC).*

1. *QN*
2. *QD ou QC*
3. *QO*
4. *QC*
5. *QD mais attention QN est aussi possible (penser au code postal ou au numéro de département)*

Corrigé de l'exercice 5

1. *VRAI*
2. *FAUX*

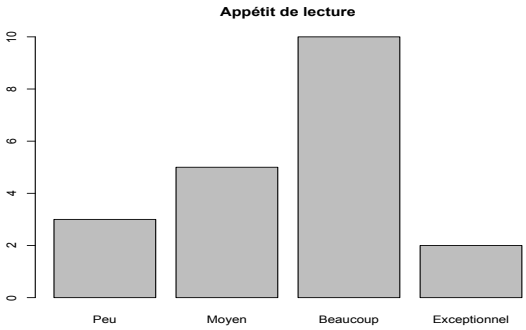
3. *VRAI*
4. *VRAI*
5. *FAUX*
6. *VRAI*
7. *VRAI*
8. *VRAI*
9. *VRAI*
10. *VRAI*
11. *VRAI*
12. *VRAI*

Corrigé de l'exercice 6

1. *L'*appétit de lecture *est une variable qualitative ordinale*
2.

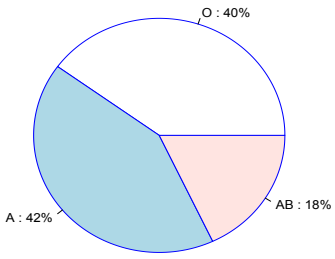
Modalités	*Effectifs*	*Fréquences*
Peu	*3*	*0,15*
Moyen	*5*	*0,25*
Beaucoup	*10*	*0,50*
Exceptionnel	*2*	*0,10*
Total	*20*	*1*

3.



Corrigé de l'exercice 7 *Les fréquence relatives sont*

pour O : 0,40 ; pour A : 0,42 ; pour AB : 0,18.



Corrigé de l'exercice 8

1. *Il s'agit d'un diagramme en bâtons représentant la variable* Nombre de joueurs sélectionnés. *Les valeurs de cette variable vont de 0 à 6. La valeur la plus fréquente est le 4.*
2. *La variable représentée est* le nombre de joueurs sélectionnés. *L'individu statistique est le centre de formation.*
3.

Valeurs	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs	*2*	*7*	*12*	*16*	*25*	*12*	*4*

4. *La moyenne de joueurs sélectionnés par centre de formation est*

$$\frac{2\times 0+7\times 1+12\times 2+16\times 3+25\times 4+12\times 5+4\times 6}{2+7+12+16+25+12+4}=\frac{263}{78}=3{,}37.$$

5. *Pour calculer l'écart type du nombre de joueurs sélectionnés, on calcule d'abord la variance. Une formule pour la variance est*

   Variance= la moyenne des carrés moins le carré de la moyenne.

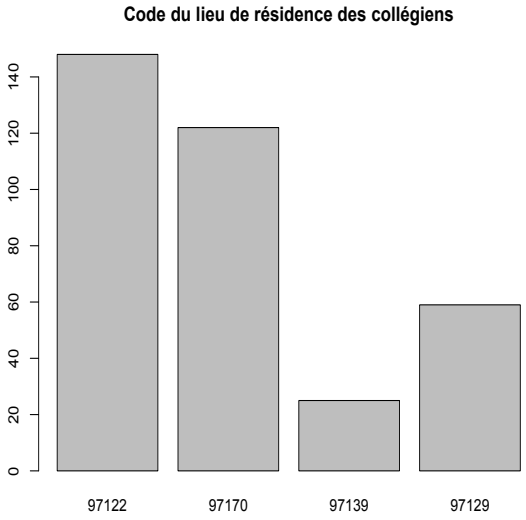
   La moyenne des carrés est

$$\frac{2\times 0^2+7\times 1^2+12\times 2^2+16\times 3^2+25\times 4^2+12\times 5^2+4\times 6^2}{2+7+12+16+25+12+4}=\frac{1043}{78}=13{,}37.$$

   donc la variance vaut $13,37-(3,37^2)=2,01$ *et l'écart type vaut* $\sqrt{2,01}=1,42$.

Corrigé de l'exercice 9

1. *Cette distribution peut être représentée par un diagramme en bandes :*



2. *La variable est numérique (car codée) mais son type est qualitatif nominal. La moyenne n'a pas de sens sur une telle variable.*

Corrigé de l'exercice 10 *La taille et la manière de choisir un échantillon est un facteur important pour pouvoir tirer des conclusions fiables. Il faut un nombre suffisamment élevé d'individus pour éviter des biais. L'échantillon doit être représentatif.*

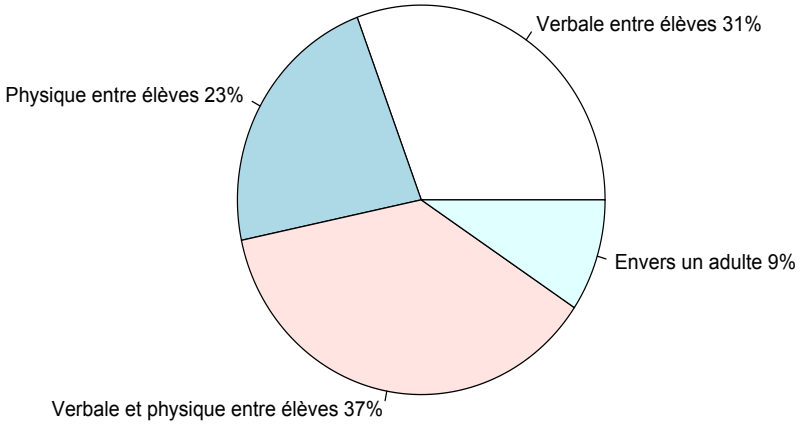
Corrigé de l'exercice 11

1. *On complète le tableau de sorte que le total fasse 75. Le nombre d'incidents avec violence envers un adulte est donc 7.*

Incident	*Violence verbale entre élèves*	*Violence physique entre élèves*	*Violence verbale et physique entre élèves*	*Violence envers un adulte*
Nombre	*23*	*17*	*28*	*7*

La population statistique est l'ensemble des incidents ayant conduit à un procès verbal. Le caractère étudié est le type de violence. Il s'agit d'un caractère qualitatif nominal.

2.



Corrigé de l'exercice 12

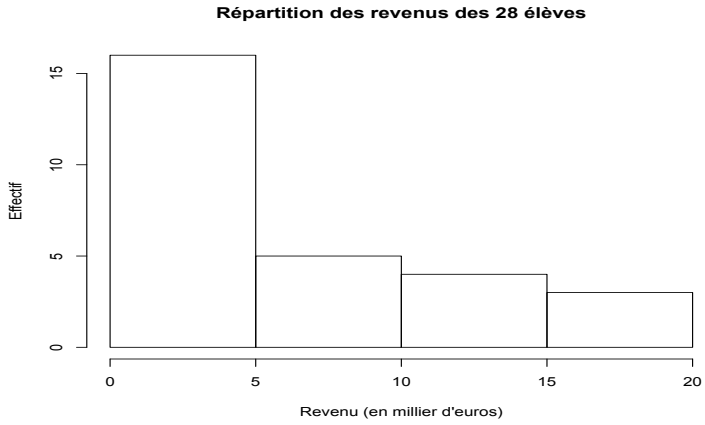
1.

Classes de revenu	*Effectifs*
]0 ; 5]	*16*
]5 ; 10]	*5*
]10 ; 15]	*4*
]15 ; 20]	*3*

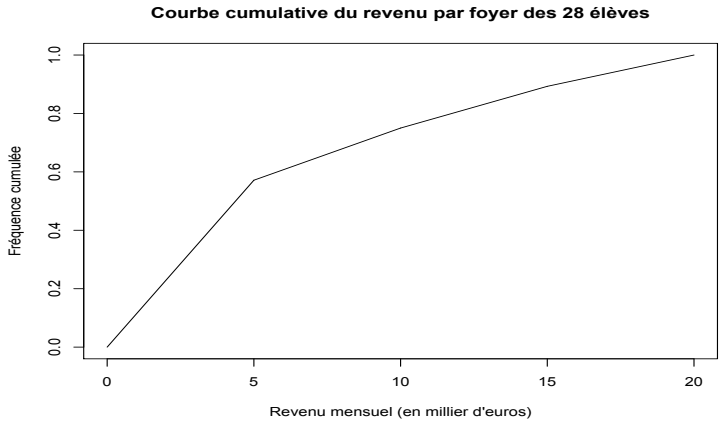
2.

Classes de revenu	*Effectifs*	*Fréquences*	*Fréquences cumulées*
]0 ; 5]	*16*	*0,57*	*0,57*
]5 ; 10]	*5*	*0,18*	*0,75*
]10 ; 15]	*4*	*0,14*	*0,89*
]15 ; 20]	*3*	*0,11*	*1*

3.



4.



Corrigé de l'exercice 13

1. *Le tableau*
2. *Le diagramme*
3. *Le diagramme en bandes (dit aussi en tuyaux d'orgue)*
4. *Le tableau*

Corrigé de l'exercice 14 *Le graphique est le diagramme cumulatif de la variable quantitative discrète* Nombre de joueurs sélectionnés.

Corrigé de l'exercice 15 *Les variables statistiques à considérer sont :*

la langue maternelle et le sexe (variables qualitatives nominales)

le niveau de scolarité (variable qualitative ordinale)

le quotient intellectuel (variable quantitative discrète)

et le revenu (variable quantitative continue)

Les modalités possibles sont

langue maternelle : Français, Anglais, Créole, Espagnol,...

sexe : Garçon, Fille

niveau de scolarité : CP, Sixième, Troisième, Seconde, Terminale,...

Les valeurs possibles sont

quotient intellectuel : 80, 84, 100, 110, 13, 160,...

revenu : toute valeur entre 0 et 10000 par exemple.

Corrigé de l'exercice 16

1. *FAUX*
2. *FAUX*
3. *FAUX*
4. *VRAI*

Corrigé de l'exercice 17

1. *Le tableau d'effectifs associé à cette série est :*

Valeurs	*1*	*2*	*3*	*4*	*5*
Effectifs	*12*	*8*	*5*	*2*	*1*

2. *La moyenne de cette série est :*

$$\frac{12\times 1+8\times 2+5\times 3+2\times 4+1\times 5}{12+8+5+2+1}=\frac{56}{28}=2.$$

$n = 28$ donc $n/2 = 14$ est pair. La médiane est donc la moyenne des 14ième et 15ième valeurs. La médiane est donc égale à (2+2)/2=2. Le mode est la valeur la plus fréquente donc 1.

3. *L'étendue est la différence entre la plus grande valeur et la plus petite donc vaut 5-1=4.*
4. *Pour calculer l'écart type, on calcule d'abord la variance. Comme la moyenne est un nombre entier, on peut utiliser la formule suivante pour la variance :*

   $\boxed{\text{Variance= la moyenne des carrés d'écart à la moyenne.}}$

   soit

$$\frac{12 \times (1-2)^2 + 8 \times (2-2)^2 + 5 \times (3-2)^2 + 2 \times (4-2)^2 + 1 \times (5-2)^2}{12+8+5+2+1} = \frac{34}{28} = 1,21.$$

   donc l'écart type vaut $\sqrt{1,21} = 1,10$.

Corrigé de l'exercice 18

1. *L'ensemble A de tous les objets que l'on étudie est la population statistique.*
2. *Un sous-ensemble choisi dans A est un échantillon.*
3. *Un élément de A est un individu statistique.*
4. *Le nombre d'objets composant une population est la taille de la population. Le nombre d'objets composant un échantillon est la taille d'échantillon.*
5. *Lorsque l'on veut connaître certaines caractéristiques d'une population, on dit qu'on enquête sur la population : Vrai $\boxed{x}$ Faux $\square$*
6. *Une enquête peut être réalisée auprès de toute la population ou sur un échantillon : Vrai $\boxed{x}$ Faux $\square$*
7. *Une corrélation est une enquête réalisée auprès de toute la population : Vrai $\square$ Faux $\boxed{x}$*
8. *Les tableaux et graphiques sont utilisés pour donner une meilleure vue d'ensemble des données : Vrai $\boxed{x}$ Faux $\square$*

Corrigé de l'exercice 19 *La répartition des moyennes annuelles des 400 élèves de sixième d'un établissement vous est donnée sous la forme d'un tableau :*

Classes statistiques	*Centres de Classe*	*Effectifs*	*Effectifs cumulés*
]0 ; 4]	*2*	*85*	*85*
]4 ; 8]	*6*	*112*	*197*
]8 ; 10]	*9*	*98*	*295*
]10 ; 12]	*11*	*67*	*362*
]12 ; 15]	*13,5*	*23*	*385*
]15 ; 20]	*17,5*	*15*	*400*

1. *La population statistique est formée des 400 élèves de sixième. La variable étudiée est la moyenne annuelle de l'élève.*
2. *La valeur approchée de la moyenne de cette distribution statistique est calculée en utilisant les centres de classe. On obtient :*

$$\frac{85\times 2+112\times 6+98\times 9+67\times 11+23\times 13,5+15\times 17,5}{85+112+98+67+23+15}=\frac{3034}{400}=7,585.$$

3. *Pour calculer l'écart-type approché des moyennes annuelles, on calcule d'abord la variance approchée. On peut utiliser ici la formule*

   Variance= la moyenne des carrés moins le carré de la moyenne.

   La moyenne des carrés est :

$$\frac{85\times 2^2+112\times 6^2+98\times 9^2+67\times 11^2+23\times 13,5^2+15\times 17,5^2}{85+112+98+67+23+15}=\frac{29202,5}{400}=73,01.$$

   La variance vaut donc $73,01-(7,585)^2=15,47$ *et l'écart type est égal à* $\sqrt{15,47}=3,93$.
4. *Pour calculer la médiane d'une distribution en classes statistiques, on procède ainsi : On calcule* $n/2$. *on obtient 200. Les effectifs cumulés encadrant* $n/2$ *sont donc 197 et 295 et correspondent aux limites de classe 8 et 10. La médiane appartient donc à la classe]8 ;10]. On applique alors la formule dite d'interpolation :*

$$Me=a_{i^*-1}+\frac{0,5n-N_{i^*-1}}{n_{i^*}}(a_{i^*}-a_{i^*-1})=8+\frac{200-197}{98}\times(10-8)=8+\frac{3}{98}\times 2=8,06.$$

5. *L'étendue pour une distribution en classes statistiques est l'écart entre la plus grand limite de classe et la plus petite. Elle vaut donc ici* $20 - 0 = 20$. *Pour calculer l'écart interquartile, il faut calculer les quartiles* Q_1 *et* Q_3. *Pour calculer* Q_1, *On calcule* $0,25 \times n$. *on obtient 100. Les effectifs cumulés encadrant* $0,25 \times n$ *sont donc 85 et 197 et correspondent aux limites de classe 4 et 8. Le premier quartile* Q_1 *appartient donc à la classe]4 ;8]. On applique alors la formule dite d'interpolation :*

$$Q_1 = a_{i^*-1} + \frac{0,25n - N_{i^*-1}}{n_{i^*}}(a_{i^*} - a_{i^*-1}) = 4 + \frac{100 - 85}{112} \times (8 - 4) = 4 + \frac{15}{112} \times 4 = 4,54.$$

Pour calculer Q_3, *On calcule* $0,75 \times n$. *on obtient 300. Les effectifs cumulés encadrant* $0,75 \times n$ *sont donc 295 et 362 et correspondent aux limites de classe 10 et 12. Le troisième quartile* Q_3 *appartient donc à la classe]10 ;12]. En appliquant la formule d'interpolation, on a :*

$$Q_3 = a_{i^*-1} + \frac{0,75n - N_{i^*-1}}{n_{i^*}}(a_{i^*} - a_{i^*-1}) = 10 + \frac{300 - 295}{67} \times (12 - 10) = 10 + \frac{5}{67} \times 2 = 10,15.$$

L'écart interquartile est $Q_3 - Q_1 = 10,15 - 4,54 = 5,56$.

Corrigé de l'exercice 20

1. *La distribution de fréquences de cette série statistique est :*

Langue	*Fréquence*
ESPAGNOL	*7/20=0,35*
ANGLAIS	*7/20=0,35*
PORTUGAIS	*4/20=0,20*
ALLEMAND	*2/20=0,10*
Total	*20*

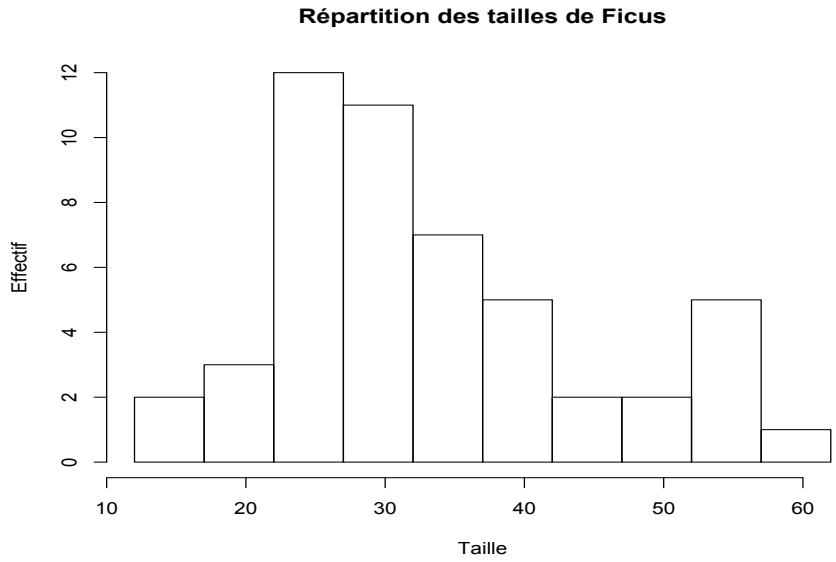
2. *La variable est qualitative nominale. Elle possède deux modes : ESPAGNOL et ANGLAIS.*

Corrigé de l'exercice 21

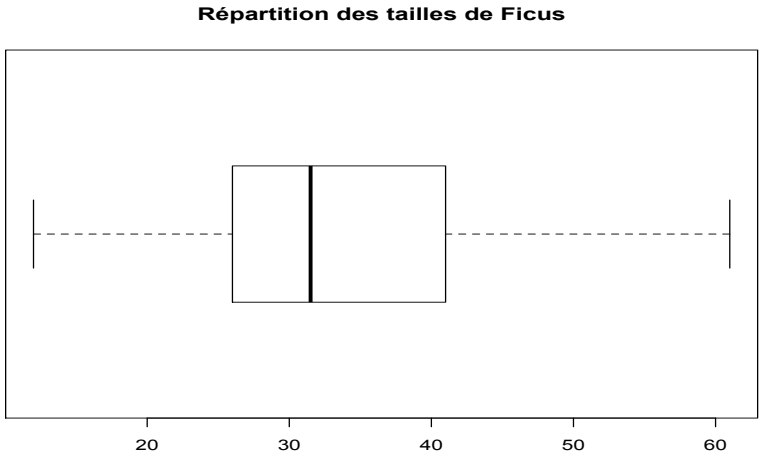
1. *La valeur minimale est 12 et la valeur maximale 61. On peut prendre les limites de classe suivantes : 12 17 22 27 32 37 42 47 52 57 62 qui fournissent des classes d'égale amplitude 5.*

2.

Classes statistiques	*]12 ;17]*	*]17 ;22]*	*]22 ;27]*	*]27 ; 32]*	*]32 ;37]*	*]37 ;42]*	*]42 ;47]*	*]47 ; 52]*	*]52 ;57]*	*]57 ; 62]*
Effectifs	*2*	*3*	*12*	*11*	*7*	*5*	*2*	*2*	*5*	*1*



3.



Corrigé de l'exercice 22

1. *Le nouveau total des mesures de valeur locative est*

$$(99 \times 1000) + 700 + 11400 = 111100.$$

Le nouveau total d'individus statistiques est 99+2=101. La nouvelle moyenne est donc $111100/101 = 1100$. *D'autre part,* 700 *est au dessous de la médiane* 900 *et 11400 est au dessus de cette médiane, donc la nouvelle médiane est égale à* 900.

2. *On pouvait s'attendre à une augmentation de la moyenne car l'une des deux nouvelles valeurs est très nettement au dessus de la moyenne initiale. Pour la médiane, on rajoute autant de valeurs de part et d'autre de la médiane initiale, donc elle reste inchangée.*

Corrigé de l'exercice 23

1. *Le premier quartile vaut à peu près 8, la médiane 10 et le troisième quartile 12. L'interprétation est qu'il y a autant d'enfants ayant une surcharge pondérale supérieure à 10 qu'inférieure à 10. D'autre part, 25% des enfants ont une surcharge inférieure à 8 et 25% une surcharge supérieur à 12.*
2. *L'intervalle interquartile vaut à peu près [8 ;10] : 50% des surcharges sont dans cette fourchette centrale.*
3. *Un enfant a une surcharge (à peu près égale à 21) qui est nettement supérieure aux autres valeurs observées dans l'étude.*

Corrigé de l'exercice 24

Le nombre total d'élèves dans l'étude est $27 + 60 + 12 + 9 + 18 + 24 = 150$.

Le tableau des fréquences est donc

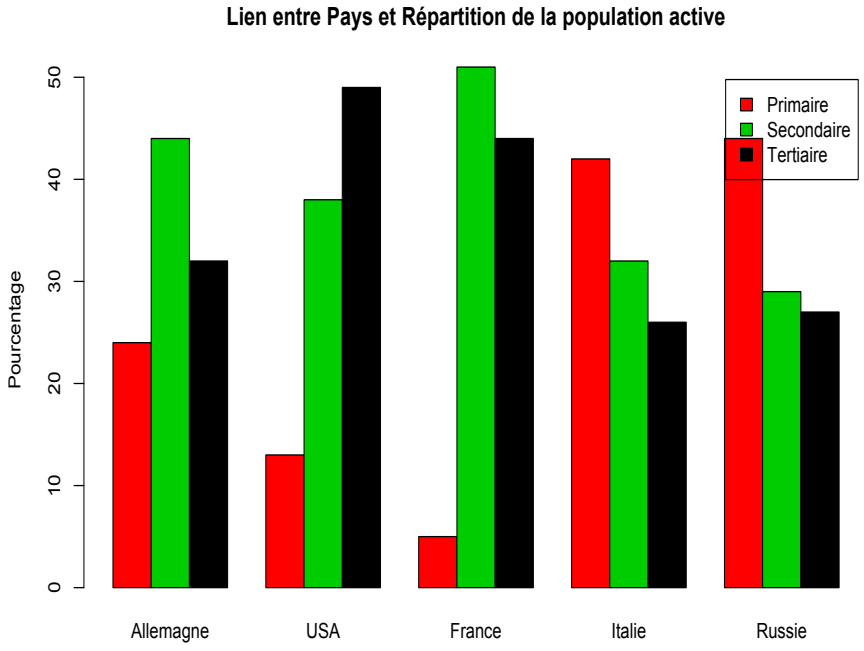
Situation professionnelle	*Employé*	*Ouvrier*	*Cadre*	*Etudiant*	*Chômeur*	*Autre*
Fréquence	$27/150 = 0,18$	$60/150 = 0,40$	$12/150 = 0,08$	$9/150 = 0,06$	$18/150 = 0,12$	$24/150 = 0,16$

Les fréquences pour les catégories {Cadre, Ouvrier, Chômeur, Autres} *ne correspondent pas à celles données par le diagramme. Le tableau d'effectifs et le diagramme ne correspondent donc pas à la même distribution.*

Corrigé de l'exercice 25 *Le premier graphique représente la répartition des tailles de joueurs de volley-ball d'un lycée selon leur sexe. Il s'agit donc d'une représentation croisée du couple de variables (Sexe, Hauteur) par boîtes à moustaches. La répartition des filles s'étale principalement de 1m65 à 1m74 avec une valeur extérieure inférieure (1m62). La médiane est proche de 1m69 et l'intervalle interquartile, qui représente la fourchette des 50% de valeurs centrales, est à peu près [1m68 ;1m72]. La taille chez les garçons va de 1m69 à 1m83 avec une valeur extérieure supérieure (1m90). La médiane vaut 1m75 et l'intervalle interquartile [1m74 ;1m78]. Les garçons sont donc globalement plus grands que les filles, mais certains filles sont plus grandes que certains garçons.*

Le deuxième graphique correspond à un nuage de points représentant les variables Hauteur et Poids des joueurs de volley-ball du lycée. Notons que la variable Sexe apparait également sous forme de couleur attribuée à chacune de ses modalités. On visualise donc sur ce graphique trois variables (deux quantitatives continues et une qualitative nominale). Les informations sur les variables Hauteur et Sexe sont donc plus détaillées ici. On observe que les poids s'étalent de 71Kg à 78Kg pour les garçons et de 60 à 66Kg pour les filles.

Corrigé de l'exercice 26 *La répartition de la population active selon le secteur d'activité dans cinq pays peut être représentée à l'aide d'un diagramme en bandes par pays.*



Corrigé de l'exercice 27

1. *Un exemple de rédaction de commentaires est :*
   Les deux figures présentent des résultats d'une étude sur les néobacheliers inscrits à l'université des Antilles et de la Guyane (UAG) au cours des années universitaires 2007-2008 à 2011-2012. Une comparaison est faite avec l'ensembles des universités hexagonales. La source d'information est le ministère de l'enseignement supérieur et de la recherche (MESR), direction générale de l'enseignement supérieur et de l'insertion professionnelle (DGESIP).
   La première page concerne le pourcentage de néobacheliers issus du département ou de départements limitrophes. Ce pourcentage est la part des néobacheliers issus du département ou des départements limitrophes de leur unité d'inscription (département d'obtention du baccalauréat) parmi les néobacheliers de l'établissement (inscriptions principales). Pour l'UAG, ce pourcentage est proche de 99% alors que pour l'ensemble des 77 universités françaises, le pourcentage moyen au cours des années étudiées varie de 82,9% à 84,7%.
   Pour l'année universitaire 2011-2012, 10% des universités ont ce pourcentage inférieur à 69,4% et il est inférieur à 85,3% pour la moitié d'entre elles. 10% des universités ont ce pourcentage supérieur à 94,4%. L'UAG se classe première en 2011-2012 avec un pourcentage de 98,9%.
   La seconde page concerne le pourcentage de néobacheliers issus de Profession ou Catégorie Socioprofessionnelle (PCS) défavorisées. Ce pourcentage est la part des néobacheliers issus de PCS défavorisées (ouvrier qualifié, ouvrier non qualifié, ouvrier agricole, retraité employé et ouvrier, chômeur n'ayant jamais travaillé, personne sans activité professionnelle) parmi les néobacheliers de l'établissement. Pour l'UAG, ce pourcentage varie entre 27,5% et 33,0% pour les années universitaires 2007-2008 à 2011-2012, alors que pour l'ensemble des 77 universités françaises, le pourcentage moyen au cours des années étudiées varie de 19,7% à 22,1%.
   Pour l'année universitaire 2011-2012, 10% des universités ont ce pourcentage inférieur à 14,9% et il est inférieur à 21,6% pour la moitié d'entre elles. 10% des universités ont ce pourcentage supérieur à 31,3%. L'UAG se classe en 9ième position en 2011-2012 avec un pourcentage de 31,3%.
2. *Pour le pourcentage de néobacheliers issus du département ou de départements limitrophes en 2011-2012, le premier quartile est 78,2 et le troisième 91,2. L'écart interquartile I est donc égal à 13,0. Par conséquent,* $1,5 \times I = 19{,}5$ *donc* $Q_1 - 1{,}5 \times I = 58{,}7$. *D'autre part,* $Q_3 + 1{,}5 \times I = 110{,}7$. *Il n'y a, bien-sûr, pas de valeurs supérieures à* $110{,}7$ *donc pas de valeurs extérieures supérieures. Par contre, pour*

l'université de Montpellier, la valeur est $57{,}7 < 58{,}7$ *donc est extérieure inférieure.*

3. *Pour le pourcentage de néobacheliers issus de PCS défavorisées en 2011-2012, le premier quartile est 18,0 et le troisième 26,1. L'écart interquartile* I *est donc égal à 8,1. Par conséquent,* $1,5 \times I = 12,15$ *et* $Q_3 + 1,5 \times I = 38,25$*. Comme la valeur pour l'université de la Réunion est* $47,2 > 38,25$*, il s'agit donc d'une valeur extérieure supérieure.*