Chapitre 3
Statistiques descriptives

Jean-Jil Duchamps*

ISIFC 2ème année, Statistiques pour l'ingénieur, 2020-2021

La **statistique descriptive**, statistique exploratoire ou analyse des données, a pour but de **résumer l'information** contenue dans les données de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes).

Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

1 Échantillons unidimensionnels : représentations graphiques et indicateurs

1.1 Terminologie

Les données dont nous disposons sont des mesures faites sur des **individus** (ou unités statistiques) issus d'une **population**. On s'intéresse à une ou plusieurs particularités des individus appelées **variables** ou **caractères**. L'ensemble des individus constitue l'**échantillon** étudié.

Exemple 1. Si l'échantillon est un groupe de TD à l'ISIFC,

- un individu est un étudiant
- la population peut être l'ensemble des étudiants de l'ISIFC, des élèves ingénieurs de France, des habitants de Besançon, etc...
- les variables étudiées peuvent être la taille, la filière choisie, la moyenne d'année, la couleur des yeux, la catégorie socio-professionnelle des parents,...

Si l'échantillon est constitué de tous les individus de la population, on dit que l'on fait un **recensement**. Il est extrêmement rare que l'on se trouve dans cette situation, essentiellement pour des raisons de coût. Quand l'échantillon n'est qu'une partie de la population, on parle de **sondage**. Le principe des sondages est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon. Pour que cela ait un sens, il faut que l'échantillon soit représentatif de la population. Il existe des méthodes pour y parvenir, dont nous ne parlerons pas ici.

Remarque 2. le mot "variable" désigne à la fois la grandeur que l'on veut étudier (variable statistique) et l'objet mathématique qui la représente (variable aléatoire).

*Laboratoire de mathématiques de Besançon, jean-jil.duchamps@univ-fcomte.fr.

Une variable statistique peut être **discrète** ou **continue**, **qualitative** ou **quantitative**. Les méthodes de représentation des données diffèrent suivant la nature des variables étudiées.

Quand on ne s'intéresse qu'au cas où on ne mesure qu'une seule variable sur les individus, on dit alors que l'on fait de la **statistique unidimensionnelle**. Dans ce cas, les données sont sous la forme de la série des valeurs prises par la variable pour les n individus, notées $x_1, \dots, x_n$. On supposera que ces données sont les réalisations de n variables aléatoires $X_1, \dots, X_n$ indépendantes et de même loi. On notera X une variable aléatoire de cette loi. Le terme d'**échantillon** désignera à la fois les séries $x_1, \dots, x_n$ et $X_1, \dots, X_n$.

Quand on mesure plusieurs variables sur les mêmes individus, on dit que l'on fait de la **statistique multidimensionnelle**.

L'objectif premier de la statistique descriptive est un objectif de représentation des données, et pas d'estimation. On peut cependant utiliser les outils de statistique descriptive dans un but d'estimation. Notamment, on s'intéressera au choix d'un modèle probabiliste pertinent, ce qui reviendra à estimer la fonction de répartition F ou la densité f de la variable aléatoire X sous-jacente, quand celle-ci est quantitative.

1.2 Représentations graphiques

1.2.a) Variables discrètes

Une variable discrète est une variable à valeurs dans un ensemble fini ou dénombrable. Mais l'ensemble des valeurs prises par cette variable dans un échantillon de taille n est forcément fini. Les variables qui s'expriment par des nombres réels sont appelées **variables quantitatives** ou **numériques** (ex : longueur, durée, coût,...). Les variables qui s'expriment par l'appartenance à une catégorie sont appelées **variables qualitatives** ou **catégorielles** (ex : couleur, catégorie socio-professionnelle, ...).

i) Variables discrètes qualitatives

Si la variable est qualitative, on appelle **modalités** les valeurs possibles de cette variable. L'ensemble des modalités est noté $E = \{e_1, \dots, e_k\}$. Par exemple, si la variable est le groupe sanguin d'un individu, l'ensemble des modalités est $E = \{\text{O, A, B, AB}\}$. Si on interroge $n = 200$ personnes, les données brutes se présenteront sous la forme d'une suite du type : A, B, B, A, O, ... Cette suite n'est pas lisible. La meilleure manière de représenter ces données est d'utiliser les fréquences absolues et relatives.

Définition 3. On appelle **fréquence absolue** de la modalités e_j le nombre total n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j : $n_j = \sum_{i=1}^{n} \mathbb{1}_{\{e_j\}}(x_i)$.

On appelle **fréquence relative** de la modalité e_j le pourcentage n_j/n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j.

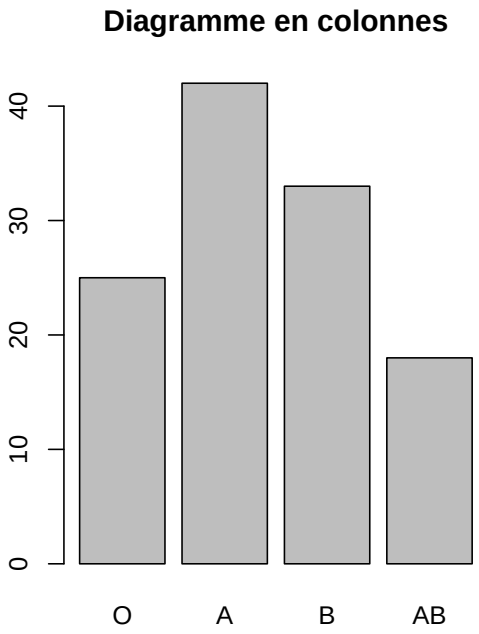
Dans l'exemple, on obtient le tableau ci-dessous :

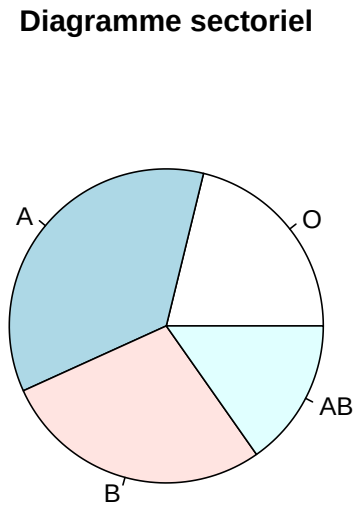
Les représentations graphiques correspondantes sont de deux types :

- **diagrammes en colonnes** ou en **bâtons** : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de cette modalité
- **diagrammes sectoriels** ou **camemberts** : à chaque modalité correspond un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative de cette modalité

Groupe sanguin	O	A	B	AB
Fréq. absolue	25	42	33	18
Fréq. relative	21%	36%	28%	15%

TABLE 1 – Groupes sanguins d'un groupe de 118 personnes.





ii) Variables discrètes quantitatives

Quand la variable est quantitative, on utilise les mêmes représentations à l'aide des fréquences absolues et relatives. La différence fondamentale entre les représentations pour des variables qualitatives et quantitatives tient au fait qu'il existe un ordre naturel sur les modalités (qui sont des nombres réels) pour les variables quantitatives, alors qu'aucun ordre n'est prédéfini pour les variables qualitatives. C'est pourquoi les diagrammes en bâtons sont toujours utilisés, mais pas les diagrammes sectoriels. Par exemple, on a effectué une enquête auprès de 1000 couples en leur demandant notamment leur nombre d'enfants. Le tableau ci-dessous donne les fréquences.

Nombre d'enfants	0	1	2	3	4	5	6	>6
Fréq. absolue	235	183	285	139	88	67	3	0
Fréq. relative	23.5%	18.3%	28.5%	13.9%	8.8%	6.7%	0.3%	0%

TABLE 2 – Nombres d'enfants de 1000 couples.

Exercice 4. Tracer le diagramme en bâtons pour cet exemple.

1.2.b) Variables continues

Quand la variable étudiée est continue, les représentations du type diagramme en bâtons sont sans intérêt, car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1. Il existe deux types de représentations graphiques :

- l'histogramme et le polygone des fréquences qui lui est associé.
- la fonction de répartition empirique, qui permet notamment de construire des graphes de probabilités.

Dans notre cours, on considérera uniquement l'histogramme.

Ce type de représentation nécessite d'ordonner les données. Si l'échantillon initial est noté $x_1, \dots, x_n$, l'échantillon ordonné sera noté $x_1^*, \dots, x_n^*$.

Exemple : Une entreprise qui fabrique des dispositifs électroniques de mesure de la tension artérielle décide de faire des tests de fiabilité d'un nouveau modèle développé. Un échantillon de 10 appareils est prélevé. On note le temps (en journées) au bout duquel on observe un dysfonctionnement sur chaque appareil. On obtient les valeurs suivantes :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

L'échantillon ordonné est :

5.4 9.5 24.3 35.7 57.1 67.3 91.6 118.4 170.9 251.3

Le principe de cette représentation est de regrouper les observations "proches" en classes. Pour cela, on se fixe une borne inférieure de l'échantillon a $0 < x_1^*$ et une borne supérieure a $k > x_n^*$. On partitionne l'intervalle $]a_0; a_k]$, contenant toutes les observations, en k intervalles $]a_{j-1}; a_j]$ appelés **classes**. La largeur de la classe j est $l_j = a_j - a_{j-1}$. Si toutes les classes sont de même largeur $l = (a_k - a_0)/k$, on dit que l'on fait un **histogramme à pas fixe**. Si les l_j ne sont pas tous égaux, on dit que l'on fait un **histogramme à pas variable**. On appelle **effectif** de la classe j le nombre d'observations appartenant à cette classe :

$$n_j = \sum_{i=1}^{n} \mathbb{1}_{]a_{j-1};a_j]}(x_i).$$

La **fréquence** (ou **fréquence relative**) de la classe j est n_j/n.

Définition 5 (Histogramme)**.** L'**histogramme** est la figure constituée des rectangles dont les **bases** sont les classes (largeur l_j) et dont les **aires** sont égales aux fréquences de ces classes ($A_j = n_j/n$). Autrement dit, la hauteur du j-ème rectangle est $h_j = A_j/l_j = n_j/nl_j$.

On voit que la représentation proposée par l'histogramme dépend de plusieurs paramètres (les bornes inférieure et supérieure a_0 et a_k , le nombre et la largeur des classes). Cela fait que plusieurs histogrammes peuvent être dessinés à partir des mêmes données et avoir des allures assez différentes, pouvant donner lieu à des **interprétations trompeuses**.

En pratique, il est conseillé de suivre les **règles** suivantes :

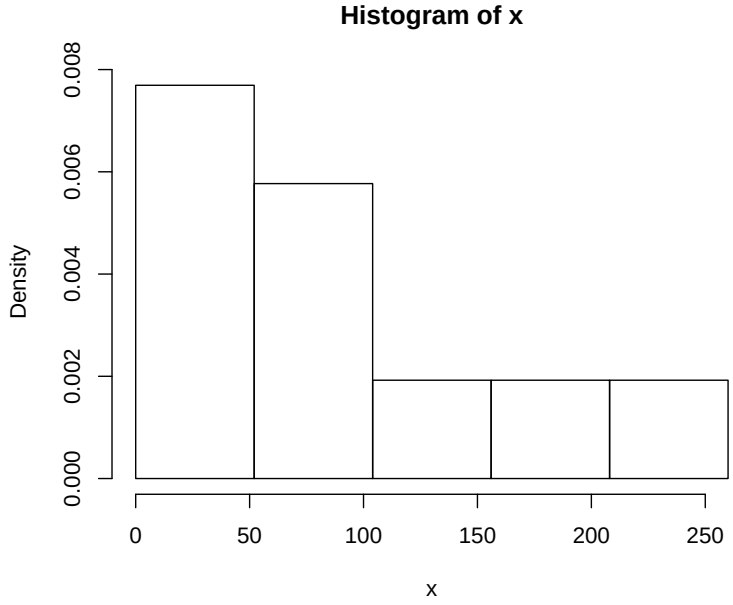
- Il est recommandé d'avoir entre 5 et 20 classes. La *règle de Sturges* préconise de choisir un nombre de classes égal à $k \approx 1 + log_2(n) = 1 + ln(n)/ln(2)$. Cela donne par exemple $k = 5$ pour $n \leqslant 22$, $k = 6$ pour $23 \leqslant n \leqslant 45$, etc...

- Le choix des bornes a_0 et a_k doit être fait de façon à respecter une certaine homogénéité des largeurs de classes. Un choix fréquent est $a_0 = x_1^* - 0.025(x_n^* - x_1^*)$ et $a_k = x_n^* + 0.025(x_n^* - x_1^*)$.

Reprenons l'exemple des dispositifs électroniques. On a $n = 10$ données, donc la règle de Sturges dit de choisir $k = 5$ classes. Comme $x_1^* = 5.4$ et $x_n^* = 251.3$, la règle énoncée donne $a_0 = -0.747$ et $a_5 = 257.4$, qu'on peut arrondir à $a_0 = 0$ et $a_5 = 260$. Si on veut un histogramme à 5 classes de même largeur, cette largeur sera donc $h = 260/5 = 52$. On obtient alors le tableau suivant :

classes $]a_{j-1}, a_j]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_j	4	3	1	1	1
fréquences n_j/n	40%	30%	10%	10%	10%
hauteurs n_j/nh	0.0077	0.0058	0.0019	0.0019	0.0019

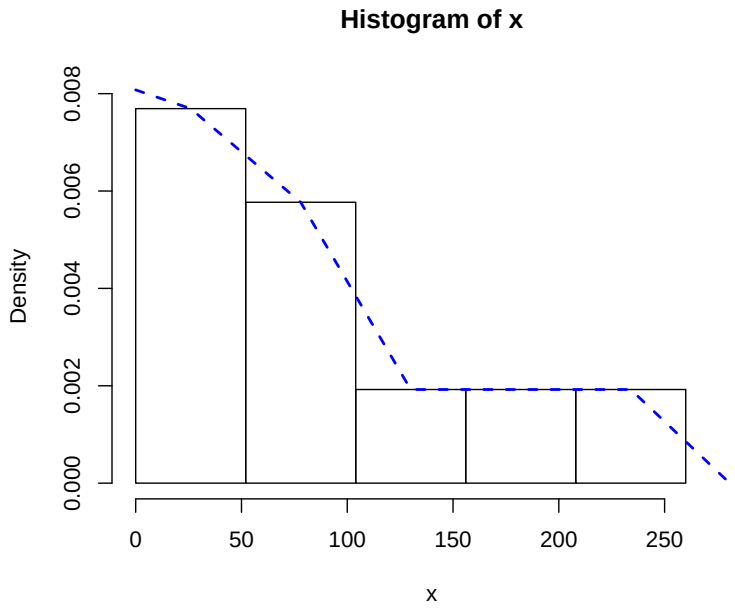
et l'histogramme correspondant est donné par la figure suivante :



L'histogramme fournit bien une visualisation de la répartition des données. Ici, le phénomène marquant est la concentration des observations sur les petites valeurs et le fait que, plus la durée de vie grandit, moins il y a d'observations. Autrement dit, la densité de la variable aléatoire représentant la durée de vie d'un dispositif électronique est une fonction décroissante.

L'histogramme n'est pas une approximation satisfaisante de la densité dans la mesure où c'est une fonction en escalier, alors que la densité est en général une fonction continue. Une meilleure approximation est le **polygone des fréquences**, c'est à dire la ligne brisée reliant les milieux des sommets des rectangles, et prolongée de part et d'autre des bornes de l'histogramme de sorte que l'aire sous le polygone soit égale à 1 (comme une densité).

Dans l'exemple précédent, le polygone des fréquences est :

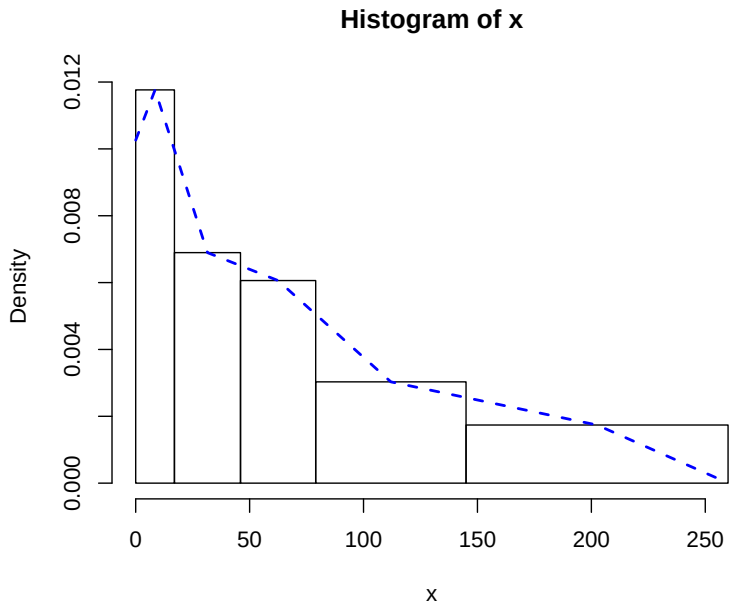


L'inconvénient d'un histogramme à pas fixe est que certaines classes peuvent être très chargées et d'autres pratiquement vides. Par exemple ici, la classe 1 contient plus d'observations à elle seule que les classes 3, 4 et 5 réunies. Pour connaître la répartition des observations dans les classes chargées, on a envie de scinder celles-ci. De même, on peut regrouper des classes trop peu chargées.

Une façon d'y parvenir est de faire en sorte que toutes les classes aient le même effectif. Dans ce cas, elles ne peuvent pas être de même largeur. Les bornes des classes sont cette fois aléatoires, puisqu'elles sont fonction des observations.

Dans l'exemple des dispositifs électroniques, on peut faire en sorte d'avoir 2 *observations par classe. On détermine par exemple les limites des classes en prenant le milieu de deux observations ordonnées successives. On obtient alors le tableau et l'histogramme suivants :*

classes $]a_{j-1}, a_j]$	$]0, 17]$	$]17, 46]$	$]46, 79]$	$]79, 145]$	$]145, 260]$
largeurs h_j	17	29	33	66	115
effectifs n_j	2	2	2	2	2
fréquences n_j/n	20%	20%	20%	20%	20%
hauteurs n_j/nh_j	0.0118	0.0069	0.0061	0.0030	0.0017

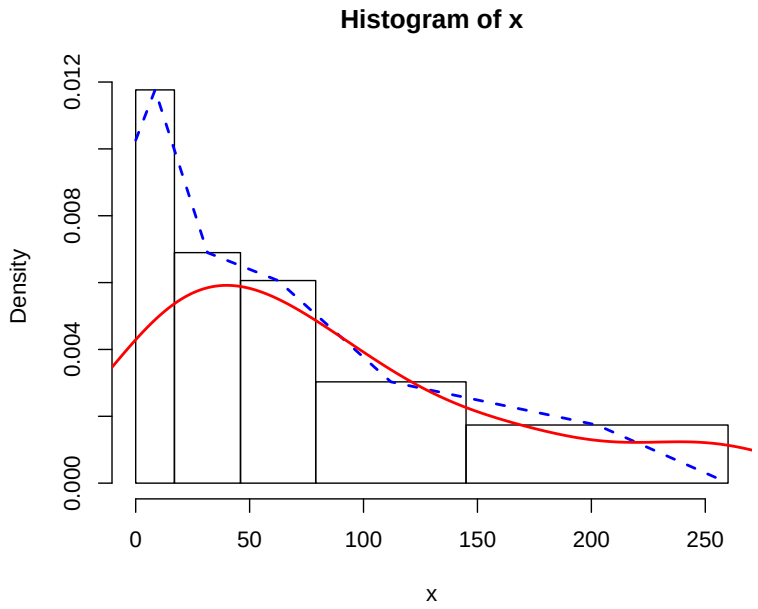


On constate que cet histogramme décrit plus finement la distribution que le précédent. C'est toujours le cas des histogrammes à classes de même effectif. Mais leur usage est moins répandu que celui des histogrammes à classes de même largeur, car ils sont moins faciles à tracer.

On voit que des histogrammes distincts sur les mêmes données peuvent être sensiblement différents. Donc il faudra se méfier des histogrammes si on veut estimer la densité des observations. On se contentera de dire que l'histogramme et, mieux encore, le polygone des fréquences, donnent une allure générale de cette densité.

Par exemple ici, il est clair que la forme des deux histogrammes et polygones n'est pas très éloignée de la densité d'une loi exponentielle ($f(x) = \lambda e^{-\lambda x}$). En revanche, ils ne ressemblent pas du tout à la densité d'une loi normale (en forme de cloche). On en conclura qu'il est très peu probable que la durée de vie d'une ampoule soit de loi normale, et qu'il est possible, voire vraisemblable, qu'elle soit de loi exponentielle. Ce jugement est pour l'instant purement visuel. Il faudra l'affiner par des techniques quantitatives plus précises.

Avec R, on peut estimer rapidement la densité de la loi d'un échantillon avec la fonction `density` (graphe en rouge).



On constate que l'estimation de densité obtenue ne ressemble pas à la densité d'une loi exponentielle. En fait, cette méthode n'est efficace que si l'on a beaucoup de données, ce qui est loin d'être le cas dans cet exemple.

1.3 Indicateurs statistiques

Les représentations graphiques présentées dans la section précédente ne permettent qu'une analyse visuelle de la répartition des données. Pour des variables quantitatives, il est intéressant de donner des indicateurs numériques permettant de caractériser au mieux ces données. On donne en général deux indicateurs : un indicateur de localisation et un indicateur de dispersion.

1.3.a) Indicateurs de localisation ou de tendance centrale

Le but est de donner un ordre de grandeur général des observations, un nombre unique qui résume au mieux les données. On pense immédiatement à la moyenne des observations.

i) La moyenne empirique

La **moyenne empirique** de l'échantillon est la moyenne arithmétique des observations, notée $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Son interprétation est évidente.

Pour l'exemple des dispositifs électroniques, $\overline{x}_{10} = 83.15$*, donc on dira que la durée de vie moyenne d'un de ces dispositif est de* $83.15h$*. Les représentations graphiques nous ont amenés à admettre que la durée de vie d'une ampoule était une variable aléatoire de loi exponentielle. On rappelle que l'espérance de la loi* $\exp(\lambda)$ *est* $1/\lambda$*. D'après la loi des grands nombres, la moyenne empirique converge presque sûrement vers l'espérance de la loi. Il est donc logique de considérer qu'une estimation de* λ *est* $1/\overline{x}_{10} = 0.012$*. Cette valeur est cohérente avec la valeur trouvée à l'aide du graphe de probabilités,* 0.013*. On retrouvera ce principe d'estimation plus tard, sous le nom de méthode des moments.*

ii) Les extrema

La plus petite valeur $x_1^* = \min\{x_1, \ldots, x_n\}$ et la plus grande valeur $x_n^* = \max\{x_1, \ldots, x_n\}$ d'un échantillon sont évidemment des indications intéressantes. Leur moyenne $\frac{x_1^* + x_n^*}{2}$ est un indicateur de localisation.

Dans notre exemple, $\frac{x_1^* + x_n^*}{2} = 128.35$.

Problème : Les deux indicateurs que l'on vient de définir sont très sensibles aux valeurs extrêmes. En particulier, il arrive parfois qu'une série statistique présente des valeurs aberrantes, c'est à dire des valeurs exagérément grandes ou petites par rapport aux autres valeurs de l'échantillon. Par exemple, ce serait le cas si une durée de vie était égale à 0.01 ou 10000. En général, la présence d'une valeur aberrante est due à une erreur de saisie ou une erreur dans l'expérience ayant abouti à cette observation. Il faut alors l'éliminer avant d'effectuer l'analyse statistique. Il existe des méthodes de détection des valeurs aberrantes, mais il est souvent difficile de décider si une valeur est aberrante ou pas. Aussi est-il important de disposer d'indicateurs qui ne soient pas trop sensibles aux valeurs aberrantes. Or la moyenne est très sensible : si une des observations est extrêmement grande, elle va tirer la moyenne vers le haut.

La médiane empirique est un indicateur de localisation construit pour être insensible aux valeurs aberrantes.

iii) La médiane empirique

La médiane empirique de l'échantillon, notée $\tilde{x}_n$ ou $\tilde{x}_{1/2}$, est un réel qui partage l'échantillon ordonné en deux parties de même effectif. La moitié des observations sont inférieures à $\tilde{x}_n$ et l'autre moitié lui sont supérieures. Il y a donc une chance sur deux pour qu'une observation soit inférieure à la médiane, et évidemment une chance sur deux pour qu'une observation soit supérieure à la médiane.

Si n est impair, la médiane empirique est la valeur située au centre de l'échantillon ordonné : $\tilde{x}_n = x^*_{\frac{n+1}{2}}$.

Si n est pair, n'importe quel nombre compris entre $x^*_{\frac{n}{2}}$ et $x^*_{\frac{n}{2}+1}$ vérifie la définition de la médiane. Par convention, on prend en général le milieu de cet intervalle : $\tilde{x}_n = \left(x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}\right)/2$.

L'expression de la médiane montre bien que c'est un indicateur qui n'est pas sensible aux valeurs aberrantes. Pour l'illustrer, considérons les deux échantillons suivants :

1 3 5 8 10 1 3 5 8 10000

Dans les deux cas, la médiane empirique vaut 5 alors que dans le premier cas, la moyenne empirique vaut 5.4 et dans le deuxième cas, la moyenne empirique vaut 20003.4. La moyenne est fortement influencée par la valeur aberrante 10000 du deuxième échantillon, alors que la médiane ne l'est pas du tout.

Dans l'exemple des dispositifs électroniques, $\tilde{x}_{10} = (57.1 + 67.3)/2 = 62.2$. On constate que la médiane est ici nettement inférieure à la moyenne : la durée de vie moyenne est de $83.1h$, et pourtant un dispositif sur deux tombera en panne avant $62.2h$ de fonctionnement. C'est ce qu'on avait déjà observé sur l'histogramme, et qui peut se remarquer directement sur les données.

iv) Caractérisation des indicateurs de localisation

Un indicateur de localisation c est fait pour résumer au mieux à lui seul l'ensemble des observations. L'erreur commise en résumant l'observation x i par c peut être quantifiée par une distance ou un écart entre ces deux valeurs : $d(x_i, c)$. L'erreur moyenne commise sur tout l'échantillon est $e = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c)$. Un bon indicateur de localisation doit minimiser cette erreur globale. L'indicateur c optimal est obtenu en annulant la dérivée de e par rapport à c.

- Si on choisit l'écart quadratique, $e = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$. La valeur de c qui minimise cette erreur est obtenue en annulant la dérivée de e par rapport à c :

$$\frac{\partial e}{\partial c} = -\frac{2}{n}\sum_{i=1}^{n}(x_i - c) = -2(\bar{x}_n - c)$$

  qui vaut 0 pour $c = \bar{x}_n$. La moyenne empirique est donc la valeur qui résume le mieux l'échantillon au sens dit "des moindres carrés".
- Si on choisit $e = \dfrac{1}{n}\sum_{i=1}^{n}|x_i - c|$, on obtient $c = \tilde{x}_n$.
- Si on choisit $e = \dfrac{1}{n}\sup_{i=1}^{n}|x_i - c|$, on obtient $c = (x_1^* + x_n^*)/2$.

Il est donc justifié d'utiliser ces trois quantités comme indicateurs de localisation.

1.3.b) Indicateurs de dispersion ou de variabilité

Pour exprimer les caractéristiques d'un échantillon, il est nécessaire de compléter les indicateurs de localisation par des indicateurs de dispersion, qui mesureront la variabilité des données.

i) Variance et écart-type empiriques

Si on choisit la distance euclidienne, on a vu que $c = \bar{x}_n$. L'indicateur de dispersion correspondant est donc $s_n^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. Il est appelé **variance empirique** de l'échantillon, et mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

Proposition 6. $s_n^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$.

L'**écart-type empirique** de l'échantillon est la racine carrée de la variance empirique : $s_n = \sqrt{s_n^2}$. Il s'exprime dans la même unité que les données, ce qui rend son interprétation plus facile que celle de la variance.

Cependant, la variabilité doit toujours se comparer à la valeur moyenne. Par exemple, si on s'intéresse aux températures, une variabilité de 10° n'a pas le même sens si la température moyenne de référence est 12° ou 10000° . Des données présentent une forte variabilité si l'écart-type est grand par rapport à la moyenne.

Aussi on définit le **coefficient de variation empirique** de l'échantillon par

$$cv_n = \frac{s_n}{\bar{x}_n}.$$

L'intérêt de cet indicateur est qu'il est sans dimension.

Si $cv_n > 0.15$, l'échantillon possède une variabilité significative.

Si $cv_n \leqslant 0.15$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de tout l'échantillon.

ii) L'étendue

L'étendue d'un échantillon est $e_n = x_n^* - x_1^*$. Cet indicateur est moins riche que la variance empirique et est évidemment très sensible aux valeurs aberrantes. Il est employé couramment en contrôle de qualité, notamment pour détecter ces valeurs aberrantes.

iii) Quantiles empirique

Les **quantiles empiriques** sont des valeurs qui partagent l'échantillon ordonné en un certain nombre de parties de même effectif.

- s'il y a 2 parties, on retrouve la médiane empirique $\tilde{x}_n$.
- s'il y a 4 parties, on parle de quartiles, notés $\tilde{q}_{n,1/4}$, $\tilde{q}_{n,1/2}$, $\tilde{q}_{n,3/4}$. On a $\tilde{q}_{n,1/2} = \tilde{x}_n$
- s'il y a 10 parties, on parle de déciles, notés $\tilde{q}_{n,1/10}, \ldots, \tilde{q}_{n,9/10}$
- s'il y a 100 parties, on parle de centiles, notés $\tilde{q}_{n,1/100}, \ldots, \tilde{q}_{n,99/100}$
- etc.

Plus généralement, les **quantiles empiriques** de l'échantillon $x_1, \ldots, x_n$ sont définis par :

$$\forall p \in]0;1[, \quad \tilde{q}_{n,p} = \begin{cases} \frac{1}{2}(x_{np}^* + x_{np+1}^*) & \text{si } np \text{ est entier,} \\ x_{\lfloor np \rfloor + 1}^* & \text{sinon.} \end{cases}$$

où $\lfloor \cdot \rfloor$ désigne la partie entière. Pour $p = 1/2$, on retrouve bien l'expression de la médiane empirique $\tilde{x}_n$.

Dans l'exemple des dispositifs électroniques, on n'a que 10 données, donc seuls les quartiles ont un sens. On connaît déjà la médiane empirique $\tilde{q}_{n,1/2} = \tilde{x}_n = 62.2$. On obtient $\tilde{q}_{n,1/4} = x_3^* = 24.3$ et $\tilde{q}_{n,3/4} = x_8^* = 118.4$.

Par ailleurs, $[\tilde{q}_{n,1/4}; \tilde{q}_{n,3/4}]$ est un intervalle qui contient la moitié la plus centrale des observations. Sa largeur $\tilde{q}_{n,3/4} - \tilde{q}_{n,1/4}$ est un indicateur de dispersion, appelé **distance inter-quartiles**, qui est insensible aux valeurs aberrantes. Dans l'exemple des dispositifs électronique, elle vaut $94.1h$. On définit de la même manière des distances inter-déciles, inter-centiles,...

Remarque 7. On considère les observations $x_1, \ldots, x_n$ comme des réalisations de variables aléatoires $X_1, \ldots, X_n$. Ainsi, toutes les quantités définies dans ce chapitre sont elles-mêmes des réalisations de variables aléatoires :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x}, \quad \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2,$$

$$\forall p \in]0;1[, \quad \tilde{Q}_{n,p} = \begin{cases} \frac{1}{2}(X^*_{np} + X^*_{np+1}) & \text{si } np \text{ est entier,} \\ X^*_{\lfloor np \rfloor + 1} & \text{sinon.} \end{cases}$$

2 Statistique bidimensionnelle : régression linéaire et moindres carrés

Le but de la régression linéaire est de rechercher une relation stochastique qui lie deux ou plusieurs variables. Les domaines dans lesquels cette technique est souvent employée sont la physique, chimie, astronomie, biologie, médecine, géographie, économie, etc.

2.1 Motivation et introduction

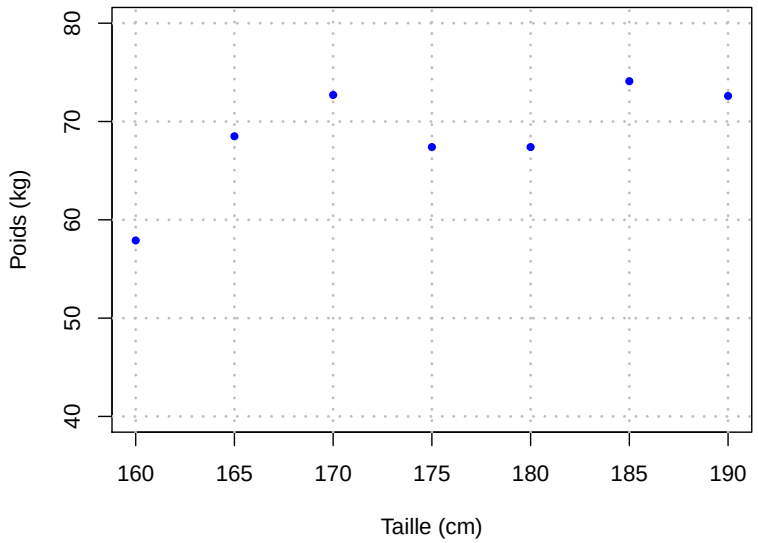
2.1.a) Relation entre deux variables

Considérons X et Y deux variables aléatoires. Par exemple, X représente la taille d'un individu et Y son poids.

But : savoir comment Y varie en fonction de X.

Dans la pratique, on s'intéresse à un échantillon de n individus, on relève le poids et la taille pour chaque individu i et on obtient un tableau d'observations ou données pairées.

Observations	1	2	3	4	5	6	7
Taille (cm)	160	165	170	175	180	185	190
Poids (kg)	57.9	68.5	72.7	67.4	67.4	74.1	72.6



On cherche à mettre en évidence une relation entre Y et X, c'est-à-dire, on cherche une fonction f qui satisfait

$$Y = f(X).$$

2.1.b) Relation déterministe

Dans certains cas, la relation est exacte. Par exemple, si X en euros, Y en dollars, ou bien si X distance ferroviaire, Y prix du billet, etc. Dans ces cas, on a

$$Y = f(X)$$

où f est une fonction déterminée (exemples pour f : fonctions linéaires, fonctions affines,etc.).

Remarque 8. On utilisera le terme de fonction "linéaire" pour désigner une fonction "affine", par exemple $f(X) = \beta_0 + \beta_1 X$, où β_0, β_1 sont des réels fixés. Par exemple, si X représente les degrés Celsius et Y les degrés Farenheit, on a $Y = 32 + 9/5X$. Ici, on a en identifiant $\beta_0 = 32$ et $\beta_1 = 9/5$. Souvent, on sait que la relation entre X et Y est linéaire mais les coefficients sont inconnus.

En pratique, pour mettre en évidence cette relation, on considère un échantillon de n individus et on vérifie que les données sont alignées.

Si c'est le cas, on a un **modèle linéaire déterministe**.

Si ce n'est pas vérifié, on va alors chercher la **droite qui ajuste le mieux l'échantillon**, c'est-à-dire on va chercher un **modèle non déterministe**. Les n observations vont permettre de vérifier si la droite candidate est adéquate.

2.1.c) Relation stochastique

La plupart des cas ne sont pas des modèles linéaires déterministes (la relation entre X et Y n'est pas exacte). Reprenons l'exemple de la taille X et du poids Y. Les données ne sont pas alignées.

Une hypothèse raisonnable est de supposer que X et Y sont liés (on parlera plus tard de corrélation, je vous conseille de regarder la vidéo "Chocolat, corrélation et moustache de chat" de la chaîne youtube La Statistique expliquée à mon chat). Dans l'exemple précédent, plus un individu est grand, plus il est lourd. On a donc, pour chaque individu i, une relation de ce genre

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

où ε_i est une variable aléatoire qui représente le comportement individuel.

Dans notre exemple, Y augmente quand X augmente, on dit qu'on a un **modèle linéaire stochastique envisageable**.

2.1.d) Cadre et contexte

Définition 9. On appelle **modèle de régression simple** le modèle suivant :

$$Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$$

avec :

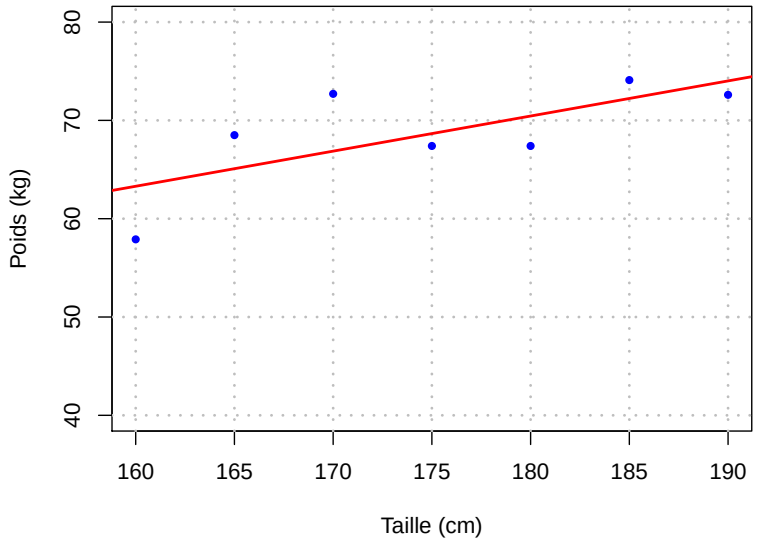
- Y_i v.a. dite variable à expliquer ;
- X_i v.a. dite variable explicative (prédicteur, régresseur, déterministe) ;
- ε_i v.a. dite erreur de précision (résidu).

Hypothèses : ε_i indépendants (mesures indépendantes), de même loi (mesure effectués), centrés et de variance σ^2.

Remarque 10. Si ε_i suit une $\mathcal{N}(0, \sigma^2)$, on dit que le modèle linéaire est gaussien.

2.2 La méthode des moindres carrés

On cherche à trouver une droite qui minimise la somme des distances au carré, de chaque point à la droite (verticalement). Dans notre exemple précédent, cette droite est représentée en rouge sur le graphique.



Définition 11. On appelle **erreur quadratique moyenne** :

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

avec $\hat{y}_i = \beta_1 x_i + \beta_0$.

Problème de moindres carrés : trouver (β_0, β_1) qui minimisent δ^2.

Solution : On cherche β_0, β_1 tels que :

$$0 = \frac{\partial \delta^2}{\partial \beta_0} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)$$
$$0 = \frac{\partial \delta^2}{\partial \beta_1} = -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i - \beta_1 x_i - \beta_0)$$

i.e.

$$0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i - \beta_0$$
$$0 = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \beta_0 \frac{1}{n}\sum_{i=1}^{n} x_i.$$

On pose

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\overline{x_n y_n} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i, \quad \overline{x_n^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

alors résoudre le système précédent revient à résoudre

$$A\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \begin{pmatrix}\bar{y}_n\\ \overline{x_n y_n}\end{pmatrix}$$

avec

$$A = \begin{pmatrix}1 & \bar{x}_n\\ \bar{x}_n & \overline{x_n^2}\end{pmatrix}.$$

On a

$$det(A) = \overline{x_n^2} - (\bar{x}_n)^2 = s_x^2,$$

et on a $det(A) = 0 \iff \forall i,\ x_i = \bar{x}_n$.

Proposition 12. *Les coefficients de la droite de régression qui minimisent δ^2 ont pour expression :*

$$\hat{\beta}_0 = \bar{y}_n - \frac{c_{xy}}{s_x^2}\bar{x}_n,$$

$$\hat{\beta}_1 = \frac{c_{xy}}{s_x^2},$$

avec $c_{xy} = \overline{x_n y_n} - \bar{x}_n\bar{y}_n$. La droite des moindres carrés a pour expression :

$$y = \frac{c_{xy}}{s_x^2}(x - \bar{x}_n) + \bar{y}_n.$$

2.3 Indicateurs statistiques bidimensionnels

Définition 13. On appelle **covariance empirique** entre les x_i et les y_i la quantité :

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n).$$

Proposition 14. $c_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}_n\bar{y}_n$.

Définition 15. Le **coefficient de corrélation linéaire empirique** entre les x_i et les y_j est donné par :

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

où $s_x = \sqrt{s_x^2} = \sqrt{\overline{x_n^2} - (\bar{x}_n)^2}$ et $s_y = \sqrt{s_y^2} = \sqrt{\overline{y_n^2} - (\bar{y}_n)^2}$.

Proposition 16. *1.* $|r_{xy}| \leqslant 1$*,*

2. si les (x_i, y_i) *sont alignés sur une même droite, alors* $|r_{xy}| = 1$*, et réciproquement.*

Remarque 17. Cette propriété sert à évaluer la pertinence de la régression : en général, on dira que le modèle linéaire est pertinent si $|r_{xy}| \geqslant 0.95$ (cette valeur de 0.95 est arbitraire, il faut simplement utilise un seuil qui fait consensus dans votre champ d'études).

2.4 L'erreur quadratique minimale

Définition 18. δ^2_{min} est l'erreur quadratique moyenne obtenue pour les valeurs optimales de β_0 et β_1 :

$$\delta^2_{min} = \delta^2(\hat{\beta}_0, \hat{\beta}_1).$$

Proposition 19. $\delta^2_{min} = s_y^2(1 - r_{xy}^2)$.

Remarque 20. $\delta^2_{min} = 0$ signifie que les points sont alignés. Cette erreur donne une indication de la qualité des données (en fait, δ^2_{min} estime σ^2).