Université Claude Bernard Lyon-1

Stage PAF

Statistiques appliquées avec Geogebra

Anne Perrut

Janvier 2014

Table des matières

Introduction

Les exemples pour lesquels nous avons besoin d'informations objectives sont légion et les données chiffrées abondent dans la vie quotidienne, comme dans la recherche scientifique. Des organismes nous donnent chaque mois l'indice des prix, le nombre de chômeurs, le pourcentage de la population favorable à l'action du gouvernement... Par ailleurs, les ingénieurs tentent de prouver que leur production industrielle est plus fiable que celle de leur concurrent ; les chercheurs planifient des expériences pour montrer l'avantage de telle semence sur les semences habituelles. Dans tous les cas, des données sont collectées. Mais ces données sont souvent loin d'être des données exhaustives et donc on cherche à déduire des affirmations plus générales (la cote du ministre X augmente, les prix augmentent...), à partir de ces observations partielles. C'est ce qu'on appelle **la statistique inférentielle**.

Avant d'en arriver là dans le dernier chapitre, nous allons commencer par la phase préliminaire indispensable : **la statistique descriptive**, qui revient juste à décrire les données collectées. Pour obtenir des informations précises, ces données doivent être collectées soigneusement : pour limiter les dérives, il faut que les différentes observations soient faites dans des conditions identiques. De plus, il faut signaler qu'on ne peut que rarement rassembler des données exhaustives, car la tâche est immense, comme dans le cas de la cote du ministre X (il faudrait interroger la population toute entière). Parfois, l'observation implique la destruction, comme dans l'étude de la durée de vie d'une batterie. Ainsi, rassembler des données exhaustives revient à user la totalité de la production de batteries. En ce qui concerne les expériences scientifiques, on pourrait en faire toujours plus, jusqu'à quand ? Il faut donc se contenter d'observations partielles des phénomènes étudiés.

Ces observations partielles constituent ce qu'on appelle **un échantillon**, qui est un sous-ensemble de toutes les observations possibles. Cet ensemble de toutes les observations possibles est appelé **la population**. Chaque entité sur laquelle on procède aux mesures est appelée **un individu**. Attention, le mot *population* s'entend quelquefois comme l'ensemble des individus et quelquefois comme l'ensemble des mesures possibles.

Exemples :

- Considérons la population des personnes inscrites sur les listes électorales en France. Un individu est un électeur. On peut s'intéresser aux variables : âge, sexe, sympathisant de gauche/de droite, a voté aux précédentes élections.
- Soit la population des studios à louer à Lyon et Villeurbanne. Un individu est un studio. On peut mesurer les variables : loyer, superficie, présence d'une connection internet.

Geogebra permet depuis la version 4.2 de mener à bien de nombreuses études statistiques. Pour cela, deux possibilités :

- vous vous placez sur le tableur de geogebra et une icône “statistiques” (elle représente un histogramme) apparaît en haut à gauche. Il suffit ensuite de se laisser guider, au rythme des clics de la souris.
- vous vous placez sur le tableur de geogebra et tout une gamme d'instructions vous permet de calculer les indicateurs statistiques ou de tracer les principaux graphiques : histogrammes, boxplots...

Pour commencer, il faut donc afficher le tableur, qui s'utilise grosso modo comme celui de libre office ou excel.

`Affichage - Tableur`

Pour que ce tableur ressemble tout à fait à celui de libre office, il faut afficher le champ de saisie

`Affichage - Aspect - Tableur - Afficher le champ de saisie - Auto-complétion`

Vous êtes maitenant prêt à vous lancer dans les statistiques.

Chapitre 1

Statistique descriptive à une variable

On constate d'ores et déjà, à partir de ces exemples, que les variables/caractères sont de natures différentes. Une variable peut être **qualitative** (oui/non, homme/femme...) ou **quantitative** (à valeurs dans $\mathbb{R}$). Nous allons voir dans la suite de ce chapitre comment décrire simplement des données en trois étapes :

- **présentation des données**,
- **représentations graphiques**,
- calcul de **résumés numériques**.

suivant que la variable étudiée est une :

- **variable quantitative discrète** (elle ne prend qu'un nombre fini ou dénombrable de valeurs ; en général il s'agit d'entiers)
- **variable quantitative continue** (variable quantitative qui n'est pas discrète)
- **variable qualitative** (variable qui n'est pas quantitative).

Enfin, nous préciserons comment décrire la relation entre deux variables quantitatives.

1.1 Variables quantitatives discrètes

Le plus souvent, une variable quantitative discrète ne prend qu'un nombre fini et même petit (moins de 20) de valeurs. Par exemple, on peut s'intéresser au nombre d'enfants par femme, au nombre d'années d'étude des étudiants après le bac...

Exemple 1 *"The World Almanac and Book of Facts" (1975) a publié le nombre des grandes inventions mises au point chaque année entre 1860 et 1959, soit*

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

(source : base de données du logiciel R)

Notons $y = (y_1, ..., y_n)$ la suite des observations rangées par ordre croissant, n étant le nombre d'observations. Des données de ce type sont à présenter dans un tableau statistique

dont la première colonne est l'ensemble des r observations distinctes (ou **modalités**), classées par ordre croissant et notées $x_1, ..., x_r$. Puis on leur fait correspondre dans une seconde colonne leurs **effectifs**, c'est-à-dire leurs nombres d'occurrence, notés $n_1, ..., n_r$. Alors $\sum_i n_i = n$. On indique aussi les **fréquences** $f_i = n_i/n$ et les **fréquences cumulées** $F_i = \sum_{j=1}^{i} f_j$ pour tout $1 \leq i \leq r$. La plupart du temps, on donnera les fréquences sous forme de pourcentage. Pour notre exemple 1, on obtient

x_i	n_i	f_i	F_i
0	9	0.09	0.09
1	12	0.12	0.21
2	26	0.26	0.47
3	20	0.20	0.67
4	12	0.12	0.79
5	7	0.07	0.86
6	6	0.06	0.92
7	4	0.04	0.96
8	1	0.01	0.97
9	1	0.01	0.98
10	1	0.01	0.99
12	1	0.01	1

Venons-en maintenant aux représentations graphiques de base : le diagramme en bâtons et le diagramme cumulatif.

Le **diagramme en bâtons** se construit avec les modalités en abscisse et les effectifs en ordonnée. Quant au **diagramme cumulatif**, il s'obtient à partir des fréquences cumulées et c'est le graphe d'une fonction appelée **fonction de répartition empirique** et définie ainsi :

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x_i \leq x < x_{i+1} \qquad (i = 1, ..., r-1) \\ 1 & \text{si } x \geq x_r \end{cases}$$

Un certain nombre de grandeurs, qui forment **le résumé numérique** de l'échantillon, participent aussi à l'analyse des données. On peut les classer en deux catégories (pour commencer) : les paramètres de position et les paramètres de dispersion.

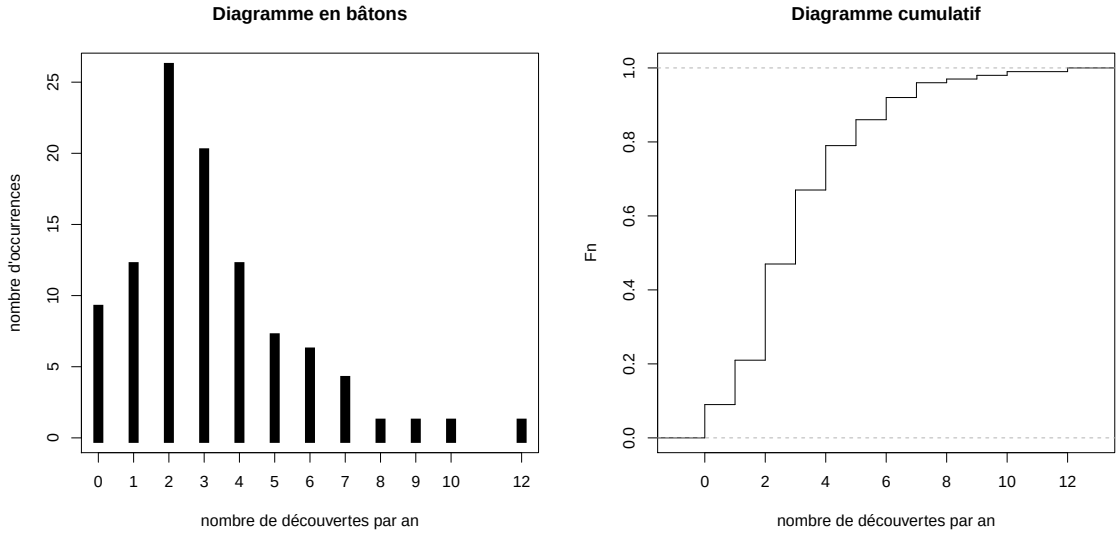
Commençons par les **paramètres de position** que sont la moyenne et la médiane. Ces grandeurs donnent un "milieu", une position centrale autour desquels les données sont réparties.

Définition 2 (paramètres de position) *La* **moyenne empirique** *est la moyenne arithmétique des observations :*

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{r} n_i x_i$$

La **médiane** *partage l'échantillon ordonné en deux parties sensiblement de même effectif : la moitié au moins des observations lui sont inférieures ou égales et la moitié au moins lui sont supérieures ou égales. Quand les observations sont triées, si le nombre d'observations*

FIGURE 1.1 – Grandes découvertes par an entre 1860 et 1959



n est impair, la médiane est $y_{(n+1)/2}$. Si n est pair, n'importe quel nombre de l'intervalle $[y_{n/2}, y_{n/2+1}]$ vérifie la définition. On convient la plupart du temps de prendre le milieu de cet intervalle pour médiane. La médiane peut aussi être définie tout simplement par $F_n^{-1}(1/2)$, où F_n est la fonction de répartition empirique.

La moyenne de l'échantillon de l'exemple 1 est 3,1. Sa médiane est 3. Ici la médiane et la moyenne sont proches, mais ce n'est pas toujours le cas (ex : 0,100,101).

Comme **paramètres de dispersion**, on peut citer la variance, l'écart interquartile et l'amplitude. Ils servent à mesurer la variabilité des données autour de la position centrale.

Définition 3 (paramètres de dispersion) *La* **variance empirique** *des observations est la moyenne du carré des écarts à la moyenne :*

$$s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{r} n_i(x_i - \bar{y})^2$$

On utilisera aussi par la suite le variance empirique modifiée :

$$s_{n-1}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{r} n_i(x_i - \bar{y})^2$$

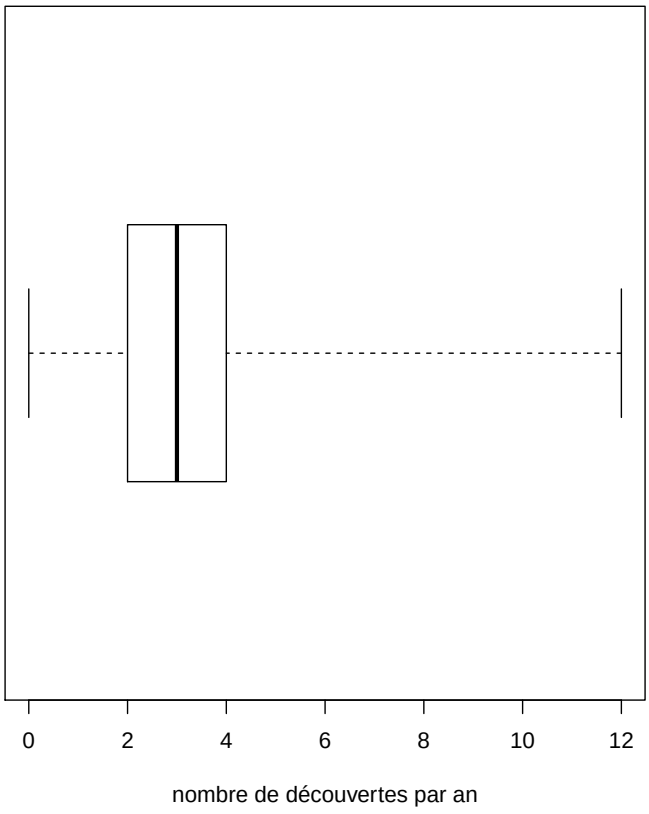
*L'***écart-type** *est la racine de la variance et on le note souvent s_n ou s_{n-1}, suivant qu'il est associé à la variance empirique ou à la variance empirique modifiée.*
Les **quartiles** *sont à rapprocher de la médiane : ils divisent l'échantillon en quatre sous-ensembles de même effectif. Un quart (au moins) des observations sont inférieures ou égales au premier quartile et trois quarts (au moins) des observations lui sont supérieures. On remarque que le premier quartile est $F_n^{-1}(1/4)$. Le deuxième quartile est la médiane. Le troisième quartile est supérieur à trois quarts des observations et est inférieur à un quart. C'est aussi $F_n^{-1}(3/4)$. La différence entre le troisième quartile et le premier quartile est l'***écart interquartile**.
*L'***amplitude** *(ou l'étendue) est la différence $y_n - y_1$.*

Pour les données de l'exemple 1, on obtient : $s^2_{n-1}(y) = 5,081$, $s_{n-1} = 2,254$, le premier quartile est 2, le troisième 4 et l'espace interquartile vaut donc 2. L'amplitude est 12.

Remarque : la fonction de répartition empirique F_n n'est pas bijective. Ainsi, l'image réciproque de 1/4 par cette fonction, $F_n^{-1}(1/4)$, est parfois un intervalle. On peut choisir n'importe quel point de cet intervalle pour faire office de premier quartile. Les programmes proposent de choisir la borne à gauche de l'intervalle, mais certains logiciels ou les calculatrices font un calcul légèrement différent. L'interprétation des résultats numériques reste la même quel que soit le calcul.

Le **diagramme en boîte** (parfois appelé diagramme en boîte à moustaches ou **boxplot** ou box-and-whisker plot) rassemble les informations liées aux quartiles, à la médiane et aux valeurs extrêmes. En pratique, on trace une boîte de largeur arbitraire à la hauteur des quartiles. On la coupe d'un trait au niveau de la médiane. Et tire des moustaches jusqu'aux minimum et maximum.

FIGURE 1.2 – Boxplot de l'exemple 1



Attention ! Souvent, on définit les valeurs extrêmes comme s'écartant du quartile le plus proche d'au moins 1,5 fois l'espace interquartile. On ajuste les moustaches aux valeurs observées et on note les valeurs extrêmes.

Geogebra vous permet de mettre en œuvre tous ces calculs et graphiques. Il est possible d'utiliser le menu **statistiques à une variable** avec les données brutes. Les fonctions suivantes détaillent les calculs :
`Unique[ <Liste Données> ]`
`Effectifs[<Booléen Cumul>, <Liste Données> ]`
`TableauEffectifs[ <Liste Données L> ]`
`Moyenne[<Liste Nombres>]`
`Médiane[ <Liste Nombres> ]`
`Q1[ <Liste Nombres> ]`
`Q3[ <Liste Nombres> ]`
`Barres[ <Série brute, <Largeur Barres> ]`
`DiagrammeBâtons[ <Liste d'abscisses>, <Liste d'ordonnées> ]`
`BoiteMoustaches[ <Ordonnée>, <Demi hauteur>, < Série brute>, <Booléen Aberrantes> ]`
Pour chacune de ces fonctions, on peut entrer soit les données brutes, soit les modalités et les effectifs.

`Histogramme[true,{1,2,...,14},{5,3,0,2,...},true,1/100]` trace un histogramme, avec les effectifs cumulés, en calculant les densités des classes et non pas juste les effectifs, avec un facteur d'échelle 1/100. On obtient ainsi le diagramme des fréquences cumulées. Une autre méthode pour obtenir ce diagramme est le `DiagrammeEscaliers[ <Liste d'abscisses>, <Liste d'ordonnées> ]` .

1.2 Variables quantitatives continues

On étudie ici les variables aléatoires continues et les variables discrètes avec un grand nombre de modalités.

Soit $y = (y_1, ..., y_n)$ un échantillon de n mesures d'une variable continue. Il n'est pas pertinent de calculer les fréquences empiriques car elles sont le plus souvent égales à $1/n$. Par contre, le résumé numérique et le diagramme cumulatif sont les mêmes que dans le cas discret. Ainsi, la moyenne empirique est donnée par

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

La variance empirique, la variance empirique modifiée et les écarts-types associés sont

$$s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s_n(y) = \sqrt{s_n^2(y)}$$

$$s_{n-1}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s_{n-1}(y) = \sqrt{s_{n-1}^2(y)}$$

À la place du diagramme en bâtons, on trace un **histogramme**. Pour ce faire, notons a et b respectivement la valeur observée minimale et la valeur observée maximale. On découpe l'intervalle $[a, b]$ en k petits intervalles appelés classes, avec k compris entre 6 et 12 (en

gros $1 + \ln n / \ln 2$). Notons ces classes $[a_0, a_1[,...,[a_{k-1}, a_k]$, avec $a = a_0$ et $b = a_k$. On peut présenter les données dans un tableau en indiquant : les classes rangées par ordre croissant, les effectifs n_i des classes (nombres d'observations dans chaque classe), les fréquences n_i/n, les amplitudes L_i des classes ($L_i = a_i - a_{i-1}$), les **densités des observations** dans chaque classe, $h_i = \frac{f_i}{L_i}$. L'histogramme est composé de rectangles dont
- les bases sont les classes
- les hauteurs sont les densités des classes (ou les hauteurs sont proportionnelles aux densités, ce qui donne le même graphique).

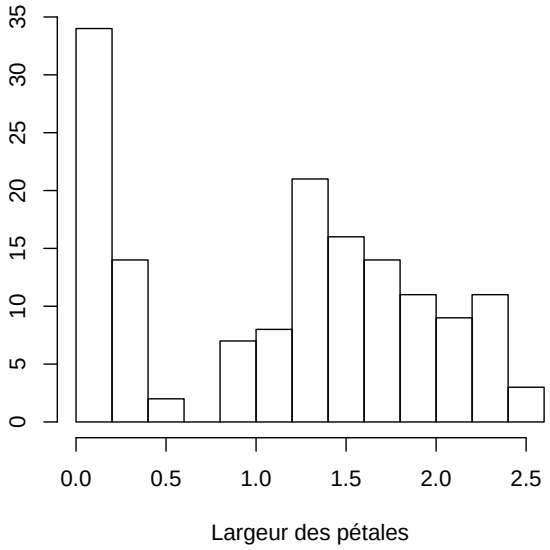
On peut choisir des classes de même largeur, comme c'est le cas par défaut dans de nombreux logiciels. Pour déterminer le nombre de classes non vides, plusieurs calculs ont été proposés par d'éminents statisticiens : je cite par exemple la règle de Sturges, de Scott, de Freedman-Diaconis.

$$\begin{aligned} \text{Sturges} \qquad & k = \lceil 1 + \log(n)/\log(2) \rceil \\ \text{Scott} \qquad & k = \lceil 3.5 s_{n-1} n^{-1/3} \rceil \\ \text{Freedman-Diaconis} \qquad & k = \lceil 2(Q3 - Q1) n^{-1/3} \rceil \end{aligned}$$

où $\lceil x \rceil$ représente l'entier immédiatement supérieur à x.

Exemple 4 *On a mesuré en centimètres la largeur des pétales de 150 iris (données de Edgar Anderson, 1935). Voir l'histogramme à la figure 1.3.*

FIGURE 1.3 – Histogramme de l'exemple 4



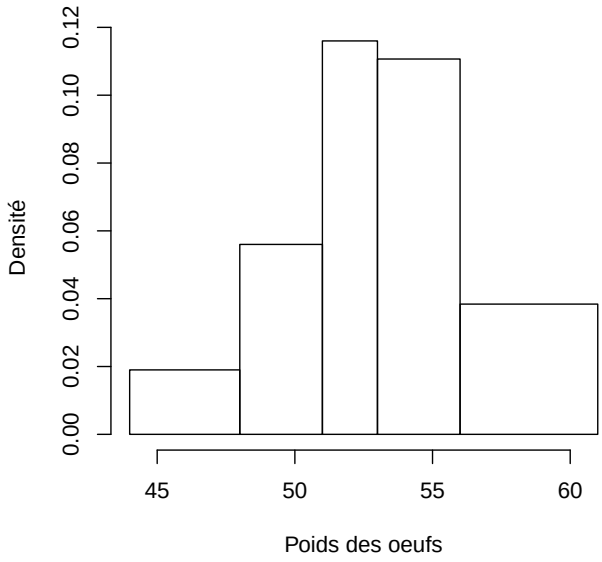
Interprétation : l'échantillon n'est pas homogène. On est peut-être en présence de 3 espèces d'iris, ces trois espèces ayant une largeur de pétales petite (autour de 0.1 cm), moyenne (autour de 1.3 cm) ou grande (proche de 2 cm).

Mais on peut aussi choisir des classes d'effectif différent.

Exemple 5 *On obtient, sur un échantillon de 250 œufs qu'on a pesé,*

Masse des œufs en gramme	*Effectif*	*Fréquence*	*Amplitude*	*Densité*
44-48	*19*	*0,076*	*4*	*0,019*
48-51	*42*	*0,168*	*3*	*0,056*
51-53	*58*	*0,232*	*2*	*0,116*
53-56	*83*	*0,332*	*3*	*0,1107*
56-61	*48*	*0,192*	*5*	*0,0384*

FIGURE 1.4 – Histogramme de l'exemple 5



Pour tracer un histogramme,
Les fonctions suivantes détaillent les calculs :

```
Moyenne[<Liste Nombres>]
Médiane[ <Liste Nombres> ]
Q1[ <Liste Nombres> ]
Q3[ <Liste Nombres> ]
Histogramme[ <Liste Bornes Classes>, <Liste Données>, <Densité True|False> , <Echelle> (optionnel) ]
BoiteMoustaches[ <Ordonnée>, <Demi hauteur>, < Série brute>, <Booléen Aberrantes> ]
DiagrammeEscaliers[ <Liste d'abscisses>, <Liste d'ordonnées>, <Booléen Reliés>, <Style des points> ]
```

Remarque : Les diagrammes en bâtons, histogrammes et boxplots sont très importants dans l'étude des échantillons. Ils permettent, nous le verrons plus tard, d'avoir une idée précise de la distribution du caractère étudié dans le population totale. On peut ainsi observer si cette distribution est symétrique, concentrée sur les petites valeurs, concentrée

sur les grandes valeurs... Les graphes permettent enfin de nettoyer les données en repérant les données aberrantes, en particulier grâce au boxplot. Ces valeurs aberrantes peuvent fausser les conclusions des tests statistiques et il est donc nécessaire de les identifier pour pouvoir les ôter de l'échantillon.

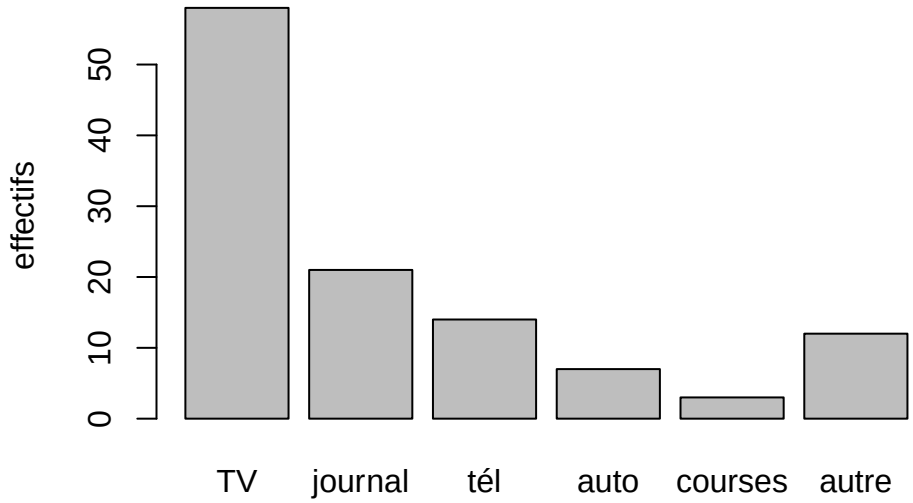
1.3 Variables qualitatives

Les modalités d'une variable qualitative ne sont pas numériques. De fait, on ne peut pas calculer les grandeurs statistiques telles que la moyenne, la variance... On peut faire un tableau pour présenter les données en indiquant pour chaque modalité l'effectif et la fréquence. Les trois représentations graphiques sont les diagrammes en colonne, les diagrammes en barre et le diagramme en secteur ("camembert").

Exemple 6 *On a interrogé des étudiants qui se rongent les ongles. Voici les circonstances pendant lesquelles ils pratiquent cette mauvaise habitude.*

activité	*effectif*
regarder la télévision	*58*
lire un journal	*21*
téléphoner	*14*
conduire une auto	*7*
faire ses courses	*3*
autre	*12*

FIGURE 1.5 – Diagramme des ongles rongés (exemple 6)



Chapitre 2

Statistique des données bivariées

Nous avons vu précédemment comment décrire une variable mesurée sur un échantillon. La plupart du temps, les données sont multivariées, c'est-à-dire qu'on dispose de plusieurs variables mesurées sur chaque individu. Et on cherche à comprendre la relation entre ces variables. Nous nous limiterons ici à deux caractères (variables) observés sur les individus d'un échantillon : par exemple, le poids et la taille, le taux d'ozone et le taux de monoxyde de carbone dans l'air...

Soit un échantillon de n individus sur lesquels on a mesuré deux variables quantitatives : les données sont les couples

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Tout d'abord, on étudie $(x_1, ..., x_n)$ et $(y_1, ..., y_n)$: représentations graphiques, calcul de la moyenne, de l'écart-type, détection de valeurs aberrantes... Puis on veut voir si les variables sont liées, de quelle nature et de quelle force est leur relation, si on peut prédire une des variables à partir de l'autre.

On peut se faire une idée précise de cette relation par une première représentation graphique : le tracé du caractère y en fonction du caractère x.

Exemple 7 *À Lyon, un agent immobilier a établi un tableau comprenant les prix en milliers d'euros et les surfaces en m^2 des 24 appartements vendus dans l'année.*

Y (k€)	75	140	400	134	395	250	160	125	189	125	175	150
X (m^2)	28	50	196	55	190	110	60	48	90	35	86	65
Y	77	122	100	163	42	39	187,5	100	135	157	42	251
X	32	52	40	70	28	30	105	52	80	60	20	100

(voir graphe ci-dessous)

2.1 Le coefficient de corrélation

Le diagramme d'un caractère en fonction de l'autre fournit une indication visuelle de la relation qui lie les caractères. Si la relation semble proche d'une relation linéaire, on

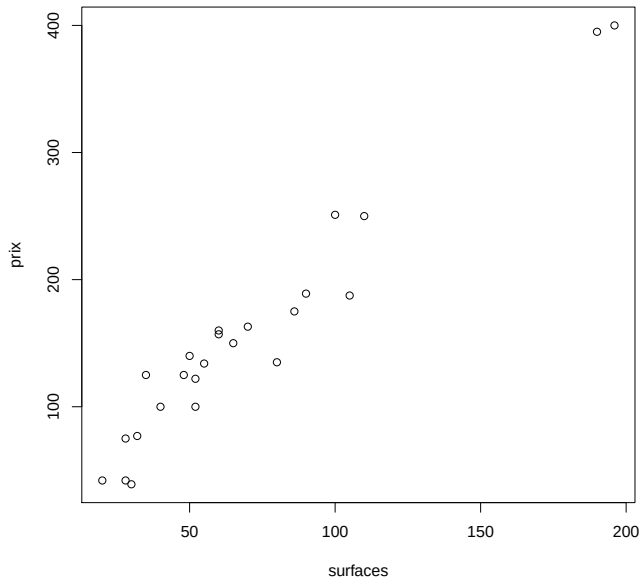


FIGURE 2.1 – diagramme de l'exemple 7

peut la quantifier par le calcul du **coefficient de corrélation** r :

$$r = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

où $\bar{x}$ et $\bar{y}$ représentent les moyennes empiriques respectivement des observations $(x_1, ..., x_n)$ et $(y_1, ..., y_n)$, et σ_x et σ_y sont les écarts-type (non modifiés) des mêmes séries d'observations.

Il est fondamental de garder en mémoire les propriétés de ce coefficient :
- r est toujours compris entre -1 et 1
- la valeur absolue de r indique la force de la relation linéaire et son signe indique sa direction.

Le coefficient de corrélation de l'exemple 7 vaut 0,9717.

Précaution : un fort coefficient de corrélation n'indique pas nécessairement une relation de cause à effet. Par exemple, relevons dans plusieurs villes de tailles différentes le nombre d'homicides perpétrés par mois et le nombre d'écoles élémentaires. À n'en pas douter, ces variables seront fortement corrélées. De là à conclure que les homicides surviennent à cause des écoliers ou de leurs professeurs, et qu'il suffit de fermer les écoles pour régler le problème de la violence...

Geogebra : dans l'outil **Statistiques à deux variables**, geogebra donne aussi le coefficient de Spearman qui calcule la corrélation sur les rangs des observations.

2.2 La droite de régression linéaire

On étudie souvent la relation entre deux variables sur un échantillon pour déduire l'une des variables connaissant l'autre :
- le bûcheron veut prédire le volume de bois coupé d'un arbre en mesurant le diamètre du tronc à deux mètres du sol,
- le médecin veut prédire la quantité d'alcool dans le sang à partir des mesures d'un alcootest,
- l'agent immobilier veut donner un prix à partir de la superficie d'un logement.

Ainsi, une des deux variables est explicative (mettons x) et l'autre à prédire (mettons y). Si on observe graphiquement une relation linéaire entre les deux variables mesurées sur un échantillon, impression confirmée par le calcul du coefficient de corrélation, on peut prédire y à partir de x grâce à une droite idéale située au plus près des points observés. Cette droite est appelée **droite des moindres carrés** ou **droite de régression** et a pour équation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

où la pente est

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

et l'ordonnée à l'origine est

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Remarque : cette droite minimise la quantité $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ où $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Dans le cas de l'exemple 7, la droite des moindres carrés a pour équation $\hat{y} = 16.436 + 1.985x$ (voir tracé ci-dessous).

Remarque : plus r est grand en valeur absolue, plus la droite est proche des points observés sur l'échantillon, et plus sera précise la prédiction. On dit alors que l'ajustement est bon.

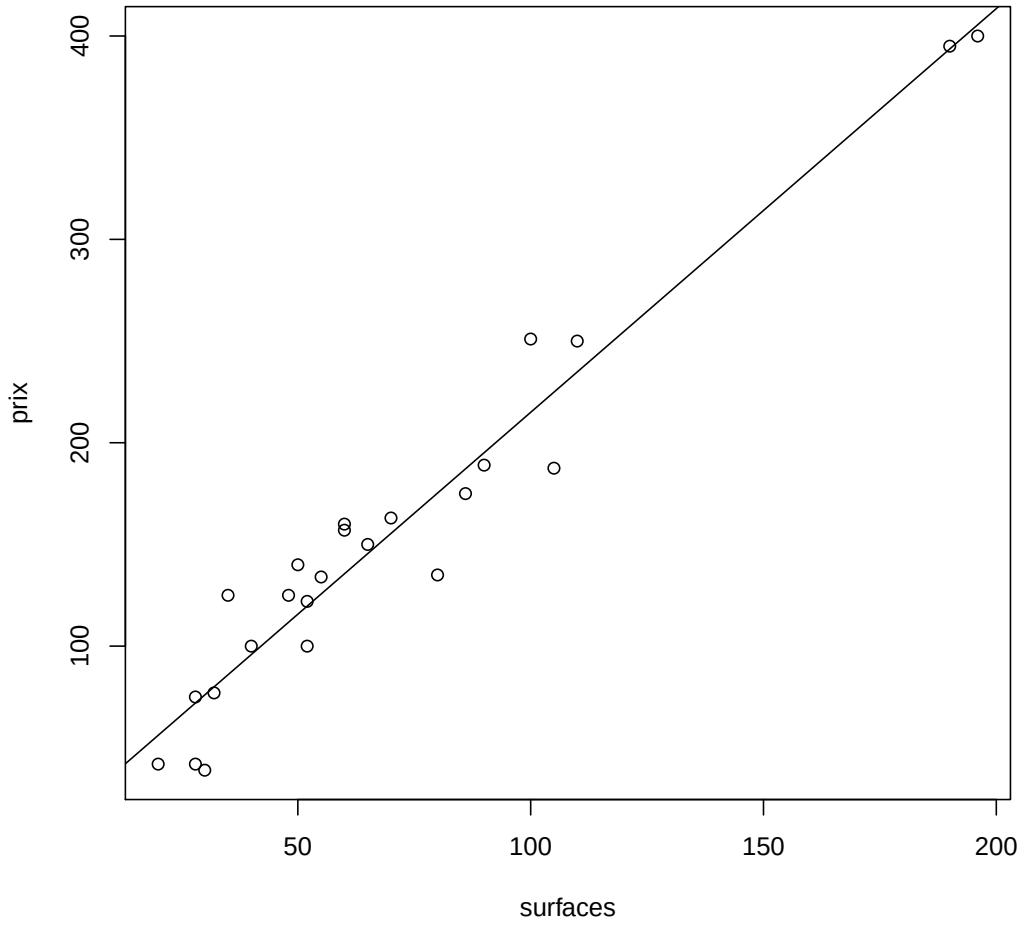


FIGURE 2.2 – droite des moindres carrés de l'exemple 7

Chapitre 3

Simulation de variables aléatoires

Savoir simuler une variable aléatoire est un exercice intéressant qui utilise la loi uniforme sur $[0, 1]$. Cela permet aussi de mieux appréhender les comportements des variables aléatoires, ainsi que les théorèmes limites. On peut esquisser du calcul de Monte Carlo, qui utilise des simulations quand le calcul analytique ne suffit pas. On peut aussi s'amuser à fabriquer un générateur congruentiel de nombres aléatoires, histoire de faire un peu d'arithmétique.

3.1 Simulation : Aléa

Au commencement, on trouve toujours des nombres de type `rand` ou `AléaUniforme[ 0,1 ]` : on part de nombres uniformément répartis entre 0 et 1. On les transforme pour obtenir des réalisations de variables aléatoires de lois diverses.

3.2 Simulation de variables discrètes

Pour une loi quelconque, on utilise **le découpage en intervalles**. Soit une v.a. discrète X à valeurs dans $\{1, ..., K\}$, dont la loi est donnée par les $p_k = P[X = k]$. Soit U une v.a. de loi uniforme sur $[0, 1]$. On pose $V = i$, où i est l'unique élément de $\{1, ..., K\}$ vérifiant $\sum_{j=1}^{i-1} p_j \leq U < \sum_{j=1}^{i} p_j$. Montrer que V a la même loi que X.

En déduire une manière de simuler la v.a. X de loi de Bernoulli $\mathscr{B}(p)$. Puis simuler la v.a. X de loi : $P[X = 1] = 1/2$, $P[X = 2] = 1/6$, $P[X = 3] = 1/3$.

Pour des lois particulières, on peut imaginer d'autres manières de faire. Par exemple, on veut piocher un élément dans l'ensemble $\{1, ..., n\}$, pour simuler un lancer de dé. Si on ne dispose que d'une fonction `AléaUniforme[0,1]` qui fournit un nombre u entre 0 et 1, il suffit de calculer la partie entière de $1 + n \times$ `AléaUniforme`$[0, 1]$.

3.3 Simulation de variables continues

Une technique universelle passe par l'inverse de la fonction de répartition. Soit une loi continue que nous voulons simuler. Notons F sa fonction de répartition. On suppose, pour plus de simplicité, que F est bijective sur son support. Soit U une v.a. de loi uniforme sur

$[0,1]$. On pose $X = F^{-1}(U)$. Quelle est la loi de X ? Calculons sa fonction de répartition.

$$P[X \leq x] = P[F^{-1}(U) \leq x] = P[U \leq F(x)] = F(x)$$

car la fonction de répartition de la loi uniforme est l'identité sur $[0,1]$. Et donc X suit la loi voulue.

Les inverses des fonctions de répartition sont souvent disponibles dans les tableurs.

Mise en œuvre pour la loi exponentielle.

Pour la loi normale, on utilise souvent la méthode de Box-Muller : soient U et V deux v.a. uniformes sur $[0,1]$. On pose

$$X = \sqrt{-2\log V}\sin(2\pi V)$$
$$Y = \sqrt{-2\log V}\cos(2\pi V)$$

Alors X et Y sont des v.a. normales centrées réduites, indépendantes.

Sous Geogebra, voici les commandes utiles pour la simulation :
`AléaBinomiale, AléaEntreBornes, AléaNormale, AléaPoisson, AléaUniforme`
`InverseBinomiale, InverseCauchy, InverseExponentielle, InverseFDistribution, InverseGamma, InverseHyperGéométrique, InverseKhiCarré, InverseNormale, InversePascal, InversePoisson, InverseTDistribution, InverseWeibull`
Complément : `ElémentAuHasard, Echantillon[ <Liste>, <Taille>, <booléen> ]`

Chapitre 4

Estimation d'une proportion

On considère une caractéristique que possède une proportion p d'individus dans la population (par exemple, la proportion de boursiers dans l'enseignement supérieur, le pourcentage des abstentionnistes aux prochaines élections municipales...). Soit un échantillon de taille n pris dans la population et X le nombre d'individus dans cet échantillon qui possède la caractéristique étudiée. Alors X suit une loi binomiale $\mathscr{B}(n,p)$ et cette v.a. peut s'écrire comme une somme de n v.a. de loi de Bernoulli $\mathscr{B}(p)$. On peut appliquer le théorème de de Moivre-Laplace (version préliminaire du théorème central limite) à ces variables de Bernoulli. Posons $F_n = X/n$, la proportion d'individus ayant la caractéristique étudiée dans notre échantillon. Sa loi peut être approchée par une loi normale :

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{F_n - p}{\sqrt{p(1-p)/n}}$$

suit approximativement une loi normale centrée réduite.

Ce théorème nous permet de trouver l'intervalle de fluctuation asymptotique pour la proportion empirique F_n :

$$F_n \in \left[p - 1.96\sqrt{\frac{p(1-p)}{n}}, p + 1.96\sqrt{\frac{p(1-p)}{n}}\right]$$

avec probabilité 0.95. On peut l'illustrer dans un premier temps à n fixé, puis dans un second temps, en faisant varier n (pour les plus intrépides).

- Pour $n = 50$ et $p = 0.4$, simuler 100 réalisations de F_n. Les tracer successivement, puis ajouter l'intervalle de fluctuation sur le graphique. On peut refaire cette simulation pour $n = 200$.
- Pour n variant de 1 à 100, calculer la suite des valeurs de F_n et tracer la ligne brisée. Ajouter l'intervalle de fluctuation.

La prise de décision est très simple. On observe la réalisation f de F_n. On calcule l'intervalle de fluctuation de F_n sous l'hypothèse que $p = p_0$. Si f est dans l'intervalle, notre observation ne contredit pas notre hypothèse : on dit que le test n'est pas significatif. Par contre, si f n'est pas dans l'intervalle, alors notre observation n'est pas en accord avec notre hypothèse et on affirme que le test est significatif : la vraie proportion (dans notre population toute entière) n'est pas p_0. Bien sûr, on a la probabilité 0.05 de se tromper, dans ce cas.

On peut aussi estimer p à partir de F, toujours en vertu du théorème de De Moivre-Laplace. Pour cela, il faut estimer l'écart-type de F qui est $\sqrt{p(1-p)/n}$, par $\sqrt{F(1-F)/n}$. On obtient alors un intervalle de confiance pour la proportion p inconnue, de niveau de confiance 95%

$$\left[f - 1.96\sqrt{\frac{f(1-f)}{n}}, f + 1.96\sqrt{\frac{f(1-f)}{n}}\right]$$

Attention : cet intervalle n'est valide que lorsque $n \geq 20$.

Dans Geogebra, on trouve un outil tout prêt qui donne les intervalles de confiance : il suffit de se promener dans l'outil *calcul des probabilités* puis dans l'onglet *statistiques*. La prise de décision se fait grâce à une variable aléatoire qu'on appelle statistique de test :

$$Z = \frac{F - p_0}{\sqrt{p_0(1-p_0)/n}}$$

On rejettera l'hypothèse si Z ne se situe pas entre -1.96 et 1.96. Geogebra calcule la probabilité : p-valeur $= P_0[|Z| > 1.96]$ où P_0 s'entend comme la probabilité quand $p = p_0$. Si la p-valeur est inférieur à 0.05, alors le test est significatif.

Chapitre 5

Échantillons et lois d'échantillonnage

Nous nous intéressons à une population, c'est-à-dire un ensemble d'individus. Les individus peuvent être des patients dans un hôpital, des succursales d'une chaîne de magasins, des automobiles à la sortie de l'usine... Nous ne pouvons pas étudier la population tout entière, c'est pourquoi nous allons piocher un échantillon de n individus dans la population. Pour plus de simplicité dans les calculs, nous allons procéder à un échantillonnage simple avec remise (les individus sont piochés uniformément, indépendamment).

Nous nous bornerons à étudier un caractère de nos individus (date de la dernière vaccination, chiffre d'affaires du mois précédent, présence d'un défaut dans le pare-brise...). Ce caractère, cette variable est supposée quantitative.

À partir de l'étude de nos n mesures, nous devons en déduire de l'information sur le caractère dans la population tout entière.

5.1 Échantillons

Commençons par quelques simulations. On considère une classe de 30 élèves. On leur donne une note pour le DS n° 1, en piochant un nombre uniformément entre 0 et 20. À l'aide d'un tableur, calculons la moyenne et la variance de la classe. Recommençons pour le DS n° 2,..., DS n° 5. Puis calculons la moyenne de chaque élève. Enfin, calculons la moyenne des moyennes individuelles, ainsi que la variance. On observe que les moyennes des élèves sont beaucoup plus stables que leurs notes aux DS.

Choisissez maintenant une loi, discrète ou continue. Simulez un échantillon de taille $n = 100$, noté $(x_1, ..., x_n)$. Calculez la suite des moyennes empiriques

$$\overline{x}_1 = x_1, \overline{x}_2 = (x_1 + x_2)/2, ..., \overline{x}_n = (x_1 + \cdots + x_n)/n$$

Tracez cette suite.

Dans la suite, nous considérerons un échantillon, c'est-à-dire une suite de v.a. $(X_i)_{1 \leq i \leq n}$ indépendantes et identiquement distribuées, d'espérance m et de variance σ^2. Nous note-

rons toujours n la taille de cet échantillon. On définit la moyenne empirique de l'échantillon :

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Proposition 8

$$E\Big[\overline{X}_n\Big] = m, \qquad \mathrm{Var}\Big(\overline{X}_n\Big) = \sigma^2/n$$

Ainsi, quand n tend vers l'infini, $\overline{X}_n$ a sa variance qui tend vers 0 : elle converge vers une constante, qui est forcément son espérance m. On dit que $\overline{X}_n$ converge en probabilité. Les statisticiens disent que $\overline{X}_n$ est un estimateur convergent ou consistant de m. Et donc, pour n assez grand, en calculant la moyenne empirique de vos observations, vous pouvez avoir une idée assez précise de la moyenne théorique m. Reste à répondre aux questions : que veut dire "n assez grand", "assez précis" ?

Pour résumer, on utilisera la moyenne de l'échantillon pour estimer la moyenne sur la population toute entière. Il existe une théorie de l'estimation : un estimateur est une v.a., une fonction de notre échantillon. On connaît plusieurs manières de construire des estimateurs, les estimateurs ont diverses propriétés. Parmi les propriétés des estimateurs, on aime qu'ils soient sans biais (leur espérance est le paramètre à estimer), leur variance doit être faible pour qu'ils soient peu sensibles aux fluctuations d'échantillonnage, on aime connaître leur loi ou au moins une loi approchée. Nous avons déjà démontré que la moyenne empirique est sans biais. On peut aussi montrer que la moyenne empirique est l'estimateur de variance minimale dans beaucoup de situations usuelles. Autrement dit, c'est le meilleur estimateur (dans un certain sens). Quant à sa loi, nous y venons...

5.2 Intervalles de confiance

On s'intéresse à la loi de $\overline{X}_n$. Quand on ne connaît pas la loi d'une v.a., on peut avoir une idée assez précise à partir de la loi d'un échantillon. Par exemple, choisissons une loi (discrète ou continue). Simulons un échantillon de grande taille, $(x_1, ..., x_N)$. Traçons l'histogramme. Simulons à nouveau un échantillon de la même taille, $(y_1, ..., y_N)$. Traçons l'histogramme des $(x_i + y_i)/2$. On a simulé un échantillon de $\overline{X}_2$. On continue le procédé pour simuler un échantillon de $\overline{X}_3$, $\overline{X}_5$, $\overline{X}_{10}$.

Quelle que soit la loi de départ, $\overline{X}_n$ suit une loi normale pour n assez grand. C'est le théorème de la limite centrale.

Théorème 9 (Théorème de la limite centrale) *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de v.a. indépendantes, de même loi, de moyenne m et de variance σ^2. Alors, quand n est assez grand,*

$$Z = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \text{ suit approximativement une loi } \mathscr{N}(0,1)$$

Autrement dit, pour tous $a < b$, quand n tend vers l'infini,

$$P\Big[a \le \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \le b\Big] \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x$$

Ce théorème nous fournit des réponses quant au comportement de $\overline{X}_n$ quand la taille de l'échantillon est grande (en pratique, supérieure à 20). Dans le cas où la loi du caractère étudié X est normale, la loi de $\overline{X}_n$ est exactement une loi normale.

Théorème 10 *Soit $(X_1, ..., X_n)$ un échantillon de la loi $\mathcal{N}(m, \sigma^2)$. Alors*

$$Z = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \text{ suit exactement une loi } \mathcal{N}(0,1)$$

On connaît maintenant la nature des fluctuations de $\overline{X}_n$ autour de la moyenne théorique m. Ceci va nous permettre de donner la précision de notre estimation. Mais comme le support de la loi normale est $\mathbb{R}$, on ne peut pas borner de manière déterministe $|\overline{X}_n - m|$. Par contre, on a établi quelle est la loi de $\overline{X}_n - m$. Et cette loi, la loi normale, a des observations très concentrées. D'après les propriétés de la loi normale, quand n est grand (mettons supérieur à 20), on sait que

$$P[m - 2\sigma/\sqrt{n} \leq \overline{X}_n \leq m + 2\sigma/\sqrt{n}] = 0.954$$

ou, de manière équivalente,

$$P[\overline{X}_n - 2\sigma/\sqrt{n} \leq m \leq \overline{X}_n + 2\sigma/\sqrt{n}] = 0.954$$

Ce qui peut se traduire par : quand on estime m par $\overline{X}_n$, l'erreur faite est inférieure à $2\sigma/\sqrt{n}$, pour 95,4% des échantillons. Ou, avec une probabilité de 95,4%, la moyenne inconnue m est dans l'intervalle $[\overline{X}_n - 2\sigma/\sqrt{n}, \overline{X}_n + 2\sigma/\sqrt{n}]$. Généralisons ce raisonnement.

Définition 11 *On peut associer à chaque incertitude α ($0 < \alpha < 1$), un intervalle qui contient la vraie moyenne m avec une probabilité égale à $1-\alpha$. Un tel intervalle est appelé intervalle de confiance de niveau de confiance $1 - \alpha$.*

Définition 12 *Soit Z une v.a.. Le fractile supérieur d'ordre α de la loi de Z est le réel z qui vérifie*

$$P[Z \geq z] = \alpha$$

On notera z_α ce fractile supérieur d'ordre α. Le fractile inférieur d'ordre α de la loi de Z est le réel z qui vérifie

$$P[Z \leq z] = \alpha$$

Proposition 13 *Soit $z_{\alpha/2}$ le fractile supérieur d'ordre $\alpha/2$ de la loi normale $\mathcal{N}(0,1)$. L'intervalle*

$$[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

est un intervalle de confiance pour la moyenne, de niveau de confiance $1 - \alpha$.

(preuve importante)

Remarque : soit Z une v.a. de loi $\mathcal{N}(0,1)$, $z_{\alpha/2}$ vérifie

$$P[Z \leq -z_{\alpha/2}] = P[Z \geq z_{\alpha/2}] = \frac{\alpha}{2}, \qquad P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1 - \alpha$$

Seules quelques valeurs de α sont utilisées habituellement. Les trois valeurs communes sont :
- $\alpha = 0.01$, et $z_{0.005} = 2.58$,
- $\alpha = 0.05$, et $z_{0.025} = 1.96$,
- $\alpha = 0.1$, et $z_{0.05} = 1.645$.

On appelle aussi intervalle de confiance la réalisation de l'intervalle précédent

$$[\bar{x}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

Un problème subsiste : celui du calcul de σ, qui est la plupart du temps inconnu. Dans ce cas, on l'estime par

$$S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

Pour des échantillons grands, remplacer l'écart-type par l'écart-type empirique a un effet négligeable.

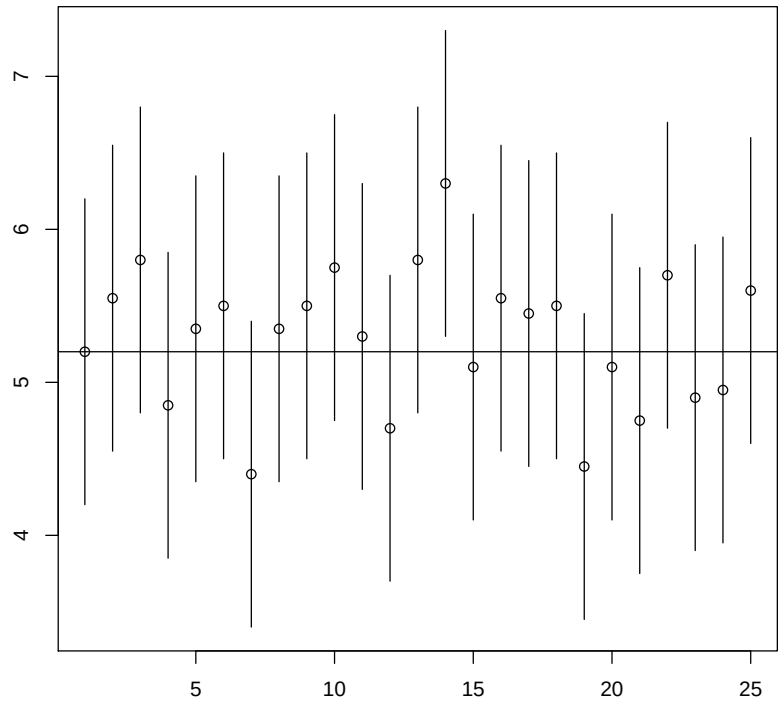
Voici un exemple théorique. Soit un caractère X qui suit une loi de Poisson de paramètre inconnu λ (ex : nombre de suicides ayant lieu dans le métro lyonnais chaque année). On rappelle que le paramètre d'une loi de Poisson est aussi sa moyenne. On cherche à estimer par un intervalle de confiance le paramètre λ. Supposons que λ vaut 5.2 et générons 25 échantillons de taille 20 de loi de Poisson $\mathscr{P}(5.2)$ pour visualiser des intervalles de confiance qui auraient pu être donnés suite à un échantillonnage sur 20 années. Avec geogebra, `TMoyenneEstimée[ <Liste Données Échantillon>, <Niveau>]`.

5.3 Tests statistiques

Le but des tests d'hypothèses est de contredire ou de confirmer une affirmation concernant la population étudiée, en se fondant sur l'observation d'un échantillon. On supposera dans cette section que les échantillons sont de grande taille (mettons supérieure à 20).

Par exemple, on change le traitement médical d'une population de patients. Leur durée de vie moyenne s'en trouve-t-elle allongée ? On modifie la composition du carburant d'un moteur. Le rendement moyen est-il significativement amélioré ? Le bruit moyen le long de la nationale 7 à Changy est-il supérieur à 95 dB ?

L'hypothèse de travail (durée de vie allongée, rendement amélioré, bruit supérieur à 95 dB) s'appelle l'**hypothèse alternative** H_1. L'affirmation contraire est appelée **hypothèse nulle** H_0 et elle consiste souvent à supposer que la situation n'a pas évolué, ou que la situation est conforme à ce qui est communément acquis. Construire un test revient à établir une règle de décision pour le rejet ou non de H_0 en fonction des observations. Il faut garder à l'esprit que le rejet de H_0 doit être fondé sur des arguments forts, car il devra entraîner une modification des habitudes et donc une importante dépense (par exemple, mettre sur le marché un nouveau médicament, changer de fournisseur de carburant, acheter le nouveau carburant plus cher, aménager l'entrée de ville de Changy).



Exemple 14 (Temps de montage) *Un ouvrier spécialisé d'une chaîne de montage passe un temps variable sur chaque pièce. La loi de ce temps a pour moyenne 270 s et pour écart-type 24 s. Une modification technique du montage pourrait diminuer ce temps. Pour le tester, on chronomètre $n = 38$ montages avec la modification technique et on obtient une moyenne empirique $\overline{X}_n$.*
Notons μ le temps moyen que passe l'ouvrier sur chaque pièce en mettant en œuvre la nouvelle technique. La question est de tester l'hypothèse $H_0 : \mu = 270$ contre l'hypothèse $H_1 : \mu \neq 270$. On doit donc établir quelles sont les valeurs de $\overline{X}_n$ qui vont conduire à rejeter H_0.

Dans cet exemple, la moyenne observée sur l'échantillon devra être significativement différente de 270 pour que H_0 soit rejetée. Dans ce cas, il existera toujours une probabilité de se tromper (H_0 est vraie, mais on la rejette), car le fait de travailler à partir d'un échantillon entraine nécessairement une incertitude. On fixe cette probabilité, qui doit être petite, à α avec souvent $\alpha = 0.05$. On dit que **le test est de niveau** $\alpha = 5\%$. On pourrait bien sûr diminuer cette probabilité, mais alors la règle de décision nous conduirait à toujours accepter H_0 et ce n'est pas satisfaisant. Il existe par ailleurs une probabilité de se tromper en acceptant H_0 (c'est l'erreur de seconde espèce), mais nous ne l'étudierons pas.

Dans notre exemple, il est raisonnable de penser que les échantillons dont la moyenne empirique $\overline{X}_n$ s'écarte beaucoup de 270 vont conduire à rejeter H_0. Ainsi, la règle de décision revient à déterminer une région de rejet de H_0 de la forme $|\overline{X}_n - 270| > c$. La

probabilité α sera dans ce cas

$$P[|\overline{X}_n - 270| > c] = \alpha = 0.05 \qquad \text{quand } H_0 \text{ est vraie}$$

Une fois α fixée, on peut déterminer c. En effet, on sait que quand H_0 est vraie, $\overline{X}_n$ suit approximativement une loi normale $\mathcal{N}(270, 24/\sqrt{38})$. Ainsi,

$$\alpha = 0.05 = P[|\overline{X}_n - 270| > c] = 2P\left[\frac{\overline{X}_n - 270}{24/\sqrt{38}} > \frac{c}{24/\sqrt{38}}\right] = 2P\left[Z > \frac{c}{24/\sqrt{38}}\right]$$

où Z suit une loi normale centrée réduite. Or on sait que $P[Z \leq -1.96] = P[Z > 1.96] = 0.025$. Ainsi,

$$\frac{c}{24/\sqrt{38}} = 1.96$$

ou $c = 7.63$. La région de rejet de H_0 est donc : $|\overline{X}_n - 270| > 7.63$. Ce qui revient au même que dire que 270 n'est pas dans l'intervalle de confiance. Ou encore, que le temps moyen observé est dans l'intervalle de fluctuation.
On peut aussi décrire la région de rejet en fonction de Z par $|Z| > 1.96$ où $Z = \frac{\overline{X}_n - 270}{24/\sqrt{38}}$.
On appelle Z **statistique de test** (il s'agit d'une grandeur que l'on calcule à partir de l'échantillon et de H_0 et qui permet de prendre la décision).

Une fois la région de rejet définie, il est facile de mesurer la force avec laquelle on rejette H_0, si c'est pertinent. Reprenons l'exemple précédent. On a fixé un seuil 1.96 pour la statistique Z, mais plus la valeur de Z observée sera différente de 270, plus le rejet de H_0 se fera avec force. Ainsi, on calcule **la p-valeur** (p-value en anglais) qui est la probabilité sous H_0 pour que la statistique de test prenne une valeur plus extrême que celle qu'on observe. Une p-valeur extrêmement petite (inférieure à 0,001) signifie que **le test est extrêmement significatif** : on rejette H_0 avec une probabilité de se tromper très proche de 0. Une p-valeur comprise entre 0.001 et 0.01 correspond à **un test très significatif**. Une p-valeur entre 0.01 et 0.05 correspond à **un test significatif** et une p-valeur supérieure à 0.05 conduit à accepter H_0.

Exemple 14 Faisons quelques applications numériques. On veut réaliser un test de niveau $\alpha = 0.05$. Alors le seuil est 1.96.
- On observe une moyenne empirique calculée sur 38 observations égale à 267. Alors

$$z_{obs} = (267 - 270)/(24/\sqrt{38}) = -0.77 > -1.96$$

donc on accepte H_0.
- On observe une moyenne empirique égale à 260. Alors

$$z_{obs} = (260 - 270)/(24/\sqrt{38}) = -2.57 < -1.96$$

donc on rejette H_0. De plus, la p-valeur vaut $P[|Z| > |-2.57|] = 0.01$. Le test est donc significatif.

Dans la plupart des cas, l'écart-type est inconnu. On l'estime alors, comme pour les intervalles de confiance, par S_{n-1} l'écart-type empirique.

Mise en œuvre d'un test (1)
1- Choix des hypothèse H_0 et H_1.
2- Détermination de la statistique et de la forme de la région de rejet.
3- Choix de α et calcul de la région de rejet.
4- Décision, au vu de l'échantillon et calcul de la p-valeur si nécessaire.

On peut aussi plus simplement baser sa décision sur la p-valeur.
Mise en œuvre d'un test (2)
1- Choix des hypothèse H_0 et H_1.
2- Détermination de la statistique de test et de la forme de la région de rejet.
3- Choix de α et calcul de la région de la p-valeur grâce aux observations.
4- Décision.

Quand on effectue un test de la moyenne, on peut vouloir tester si la moyenne est différente de la valeur μ_0 contenue dans H_0, ou si elle est supérieure ou inférieure. Dans chacun de ces trois cas, la région de rejet aura une forme différente. Considérons la statistique

$$Z = \frac{\overline{X}_n - \mu_0}{S_{n-1}/\sqrt{n}}$$

où n est la taille de l'échantillon, $\overline{X}_n$ la moyenne empirique et S l'écart-type empirique. Notons z la valeur observée de Z. Voici un récapitulatif :

Hypothèses	Région de rejet	p-valeur
$H_0 \ : \ \mu = \mu_0$ $H_1 \ : \ \mu > \mu_0$	$Z \geq 1.645$	$p = P[Z > z_{obs}]$
$H_0 \ : \ \mu = \mu_0$ $H_1 \ : \ \mu < \mu_0$	$Z \leq -1.645$	$p = P[Z < z_{obs}]$
$H_0 \ : \ \mu = \mu_0$ $H_1 \ : \ \mu \neq \mu_0$	$\lvert Z\rvert \geq 1.96$	$p = P[\lvert Z\rvert > \lvert z_{obs}\rvert]$

Commandes geogebra : `TMoyenneEstimée, TTest`.
Il est aussi possible de mettre en œuvre tous les tests dans le menu **calcul de probabilités**.