Fondements de la statistique descriptive

Anicet E. T. Ebou, ediman.ebou@inphb.ci

Objectifs pédagogique

- Connaître les fondements de la statistique;
- Connaître les différents types de variable et connaître leurs propriétés principales;
- Connaître le type de graphique à utiliser en fonction de la variable étudiée et pouvoir le représenter avec un tableur ou R.

01

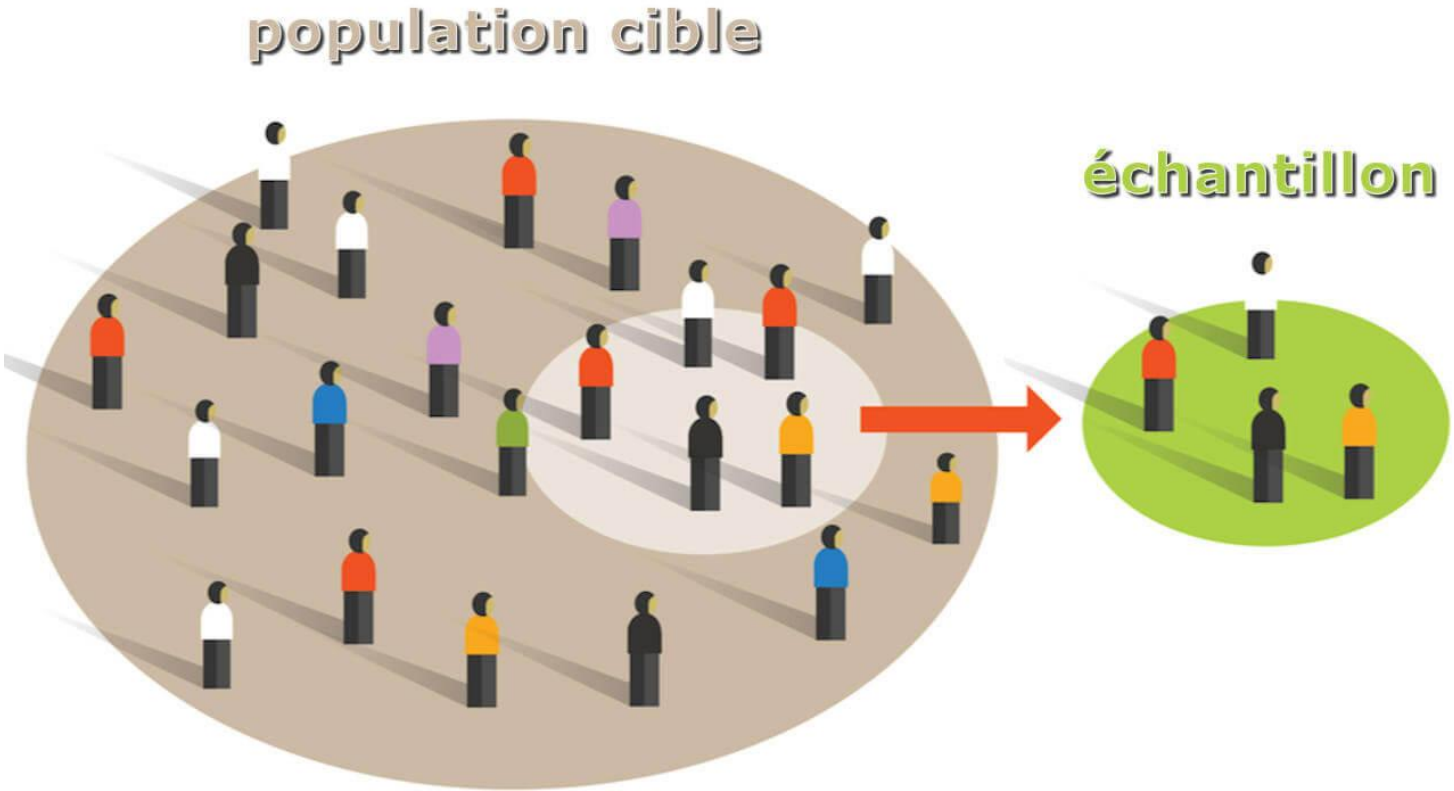
Définitions fondamentales

Définitions fondamentales

La statistique est l'ensemble de méthodes permettant de collecter, de décrire et d'analyser des observations (ou données).

La population correspond à l'ensemble des individus sur lequel porte l'étude et **l'échantillon** représente la fraction de cette population qui est réellement observée ou étudiée.

Définitions fondamentales



Définitions fondamentales

- **Une population statistique** est l'ensemble des éléments effectivement représentés par l'échantillonnage;
- **Un individu** désigne un élément d'un échantillon ou d'une population;
- **Un paramètre** résume des aspects de la population entière alors qu'**une statistique** est calculée à partir d'un échantillon.

02

Mesures et variables

Mesures et variables

- On s'intéresse à des **unités statistiques ou unités d'observation:** par exemple à des parcelles de terres.
- Sur ces unités, on mesure un caractère ou **une variable**, l'âge de l'arbre ou son diamètre, l'abondance de la bactérie.
- On suppose que la variable prend toujours une seule valeur sur chaque unité.
- Les variables sont désignées par simplicité par une lettre (X, Y, Z).

Mesures et variables

- Les **valeurs** possibles de la variable, sont appelées **modalités**.
- L'ensemble des valeurs possibles ou des modalités est appelé le **domaine** de la variable.

03

Typologie des variables

Typologie des variables

Variable qualitative: Une variable est dite qualitative quand les modalités sont des catégories. De façon plus spécifique, on a:

- **Variable qualitative nominale**: La variable est dite qualitative nominale quand les modalités ne peuvent pas être ordonnées.
- **Variable qualitative ordinale**: La variable est dite qualitative ordinale quand les modalités peuvent être ordonnées.

Typologie des variables

Le fait de pouvoir ou non ordonner les modalités est parfois discutable.

Par exemple: dans les catégories socioprofessionnelles, on admet d'ordonner les modalités: “ouvriers”, “employés”, “cadres”. Si on ajoute les modalités “sans profession”, “enseignant”, “artisan”, l'ordre devient beaucoup plus discutable.

Typologie des variables

Variable quantitative: Une variable est dite quantitative si toutes ses valeurs possibles sont numériques. On a alors:

- **Variable quantitative discrète:** Une variable est dite discrète, si l'ensemble des valeurs possibles est dénombrable.
- **Variable quantitative continue:** Une variable est dite continue, si l'ensemble des valeurs possibles est continu.

Typologie des variables

Ces définitions sont à relativiser, l'âge est théoriquement une variable quantitative continue, mais en pratique, l'âge est mesuré dans le meilleur des cas au jour près. Toute mesure est limitée en précision !

04

Série statistique

Série statistique

On appelle **série statistique** la suite des valeurs prises par une variable *X* sur les unités d'observation. Le nombre d'unités d'observation est alors souvent noté *n*. Les valeurs de la variable *X* sont notées x_1, x_2, x_3, …, x_n.

Série statistique

Exemple: On s'intéresse à la variable 'espece-arbre' notée *X* et à la série statistique des valeurs prises par *X* sur 20 arbres. La codification est :

- *Afzelia africana:* AA*;*
- *Ceiba pentandra:* CP*;*
- *Tectona grandis:* TG*;*
- *Terminalia ivorensis:* TI

Série statistique

Le domaine de la variable X est {AA, CP, TG, TI}.

Considérons la série statistique suivante:

TG, AA, CP, TG, TI, AA, CP, AA, TG, CP, AA, CP, CP, AA, TI, TI, TI, AA, AA, TI.

Ici $n = 20$ et x_1 = TG, x_2 = AA, x_3 = CP, …, x_{20} = TI.

05

Etudes des variables

5.1

Variable qualitative nominale

Calcul des effectifs et fréquences

Une variable qualitative nominale a des valeurs distinctes qui ne peuvent pas être ordonnées. On note J le nombre de valeurs distinctes ou modalités. Les valeurs distinctes sont notées $x_1, \ldots, x_j, \ldots, x_J$.

On appelle **effectif** d'une modalité ou d'une valeur distincte, le nombre de fois que cette modalité (ou valeur distincte) apparaît.

On note n_j l'effectif de la modalité x_j. La **fréquence** d'une modalité est l'effectif divisé par le nombre d'unités d'observation $f_j = \frac{n_j}{n}, j = 1, \ldots, J.$

Exercice

A partir de l'exemple précédent, remplir le tableau statistique suivant:

X_j (Espèce d'arbre)	N_j (Effectif)	F_j (Fréquence)
n = 20		1

Diagramme en secteurs et diagramme en barres

Le tableau statistique d'une variable qualitative nominale peut être représenté par deux types de graphiques. Les effectifs sont généralement représentés par un **diagramme en barres** et les fréquences par un **diagramme en secteurs** (ou camembert ou *piechart* en anglais).

Diagramme en secteur en langage R

```
> pie(table1, radius = 1)
```

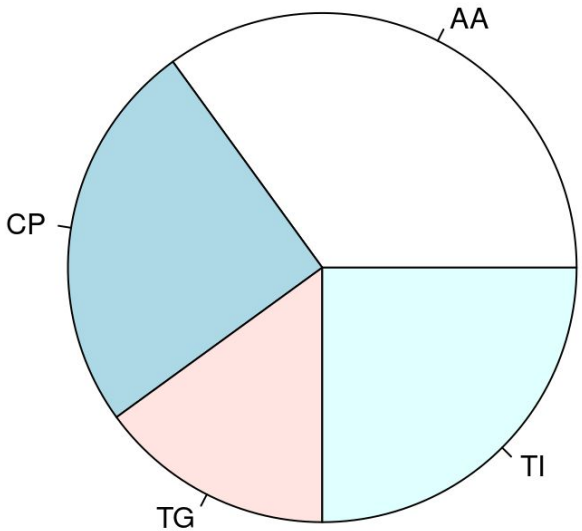


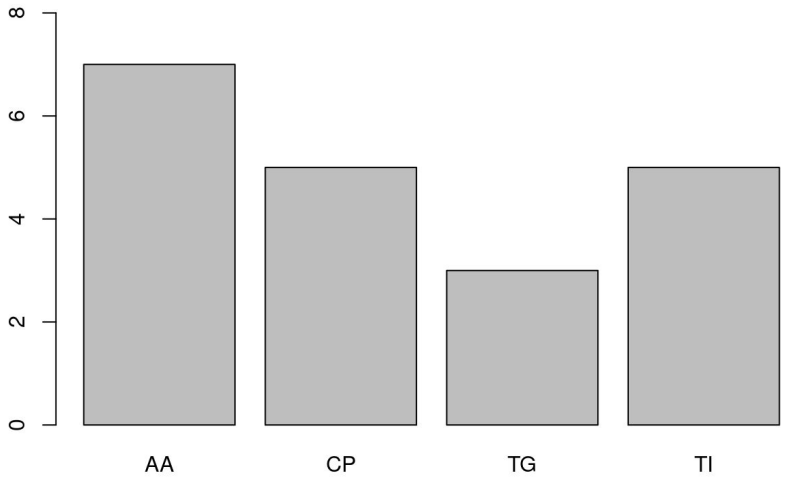
Diagramme en secteur en langage R

Il peut être souvent difficile de lire un diagramme en secteurs pour des fréquences proches qui vont alors résulter en un diagramme presque équitablement subdivisé. Il est alors préférable d'utiliser un diagramme en barre par exemple.

Diagramme en barre en langage R

```
> m <- max(v1)

> barplot(table1, ylim = c(0, m + 1))
```



5.2

Variable qualitative ordinale

Effectifs cumulés

Les valeurs distinctes d'une variable ordinale peuvent être ordonnées.

On peut alors calculer les effectifs cumulés à partir de la formule:

$$N_j = \sum_{k=1}^{j} n_k, j = 1, \ldots, J.$$

Avec $N_1 = n_1$ et $N_J = n$.

Fréquences cumulées

On peut de même calculer les fréquences cumulées:

$$F_j = \frac{N_j}{n} = \sum_{k=1}^{j} f_k, j = 1, \ldots, J.$$

Tableau statistique

On interroge 50 personnes sur leur dernier diplôme obtenu. La codification a été faite selon:

- Sans diplôme: Sd;
- Primaire: P;
- Secondaire: Se;
- Supérieur non-universitaire: Su;
- Universitaire: U.

Tableau statistique

Représenter le tableau statistique, à partir de la série suivant:

Sd Sd Sd Sd P P P P P P P P P P Se Se Se Se Se Se Se Se Se Se Se Se Se Se Su Su Su Su Su Su Su Su U U U U U U U U U U U U

x_j (dernier diplôme)	**n_j (effectif)**	***N_j (effectif cumulé)***	***f_j (fréquence)***	**F_j (fréquence cumulé)**

Diagramme en secteurs

Les fréquences d'une variable qualitative ordinale sont représentées au moyen d'un diagramme en secteur.

Diagramme en secteurs en langage R

```
> pie(table2, radius = 1)
```

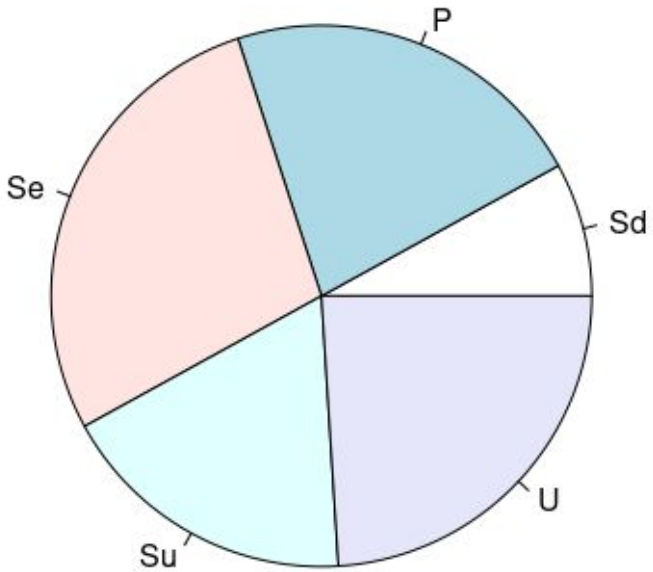


Diagramme en barre des effectifs

Les effectifs d’une variable qualitative ordinale sont représentés au moyen d’un diagramme en barres.

Diagramme en barre des effectifs en langage R

```
> barplot(table2)
```

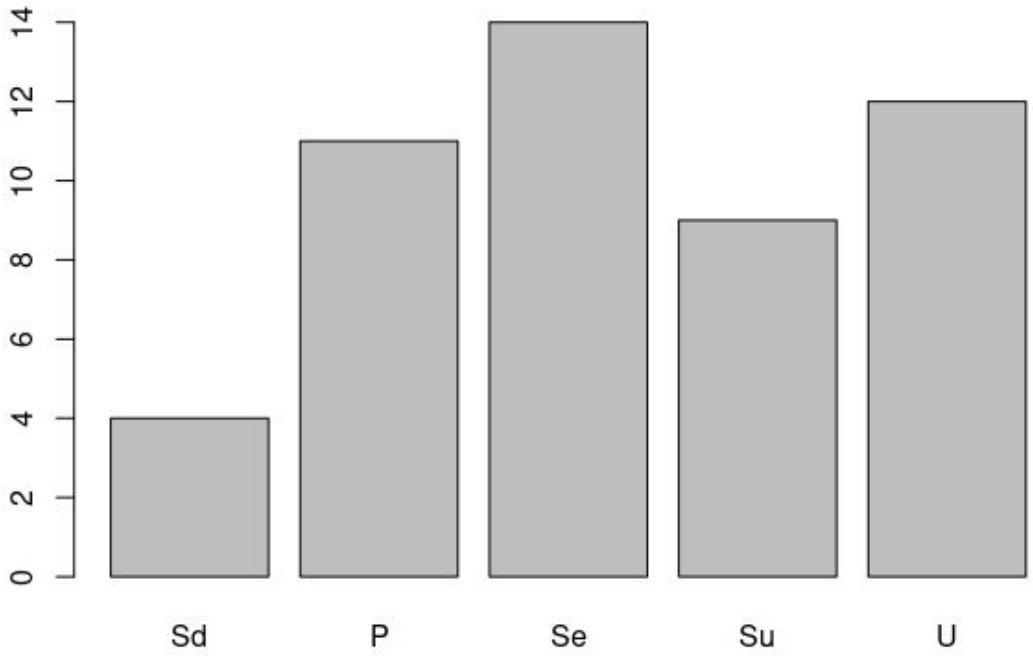


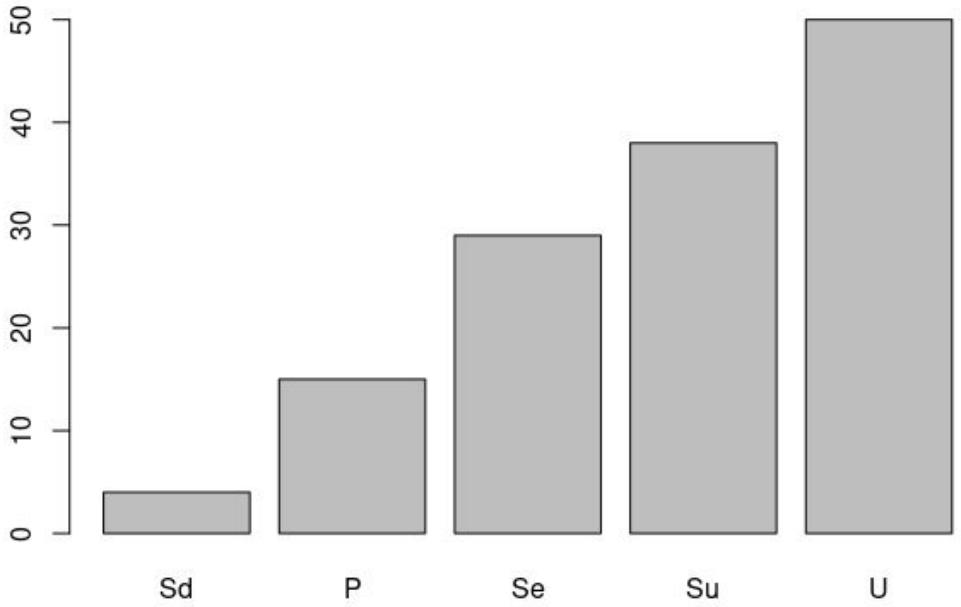
Diagramme en barres des effectifs cumulés

Les effectifs cumulés d’une variable qualitative ordinale sont représentés au moyen d’un diagramme en barres.

Diagramme en barres des effectifs cumulés en R

```
> table3 <- cumsum(table2)

> barplot(table3)
```



5.3

Variable quantitative discrète

Tableau statistique

La variable étant quantitative, on peut calculer les mêmes paramètres que précédemment, à savoir, l'effectif, l'effectif cumulé, la fréquence et la fréquence cumulée de chaque modalité.

Tableau statistique

Une forêt est composée de 50 parcelles, et la variable Z représente le nombre d'arbres par parcelle. Faire le tableau statistique contenant les effectifs cumulés et les fréquences cumulées à partir de la série statistique suivante:

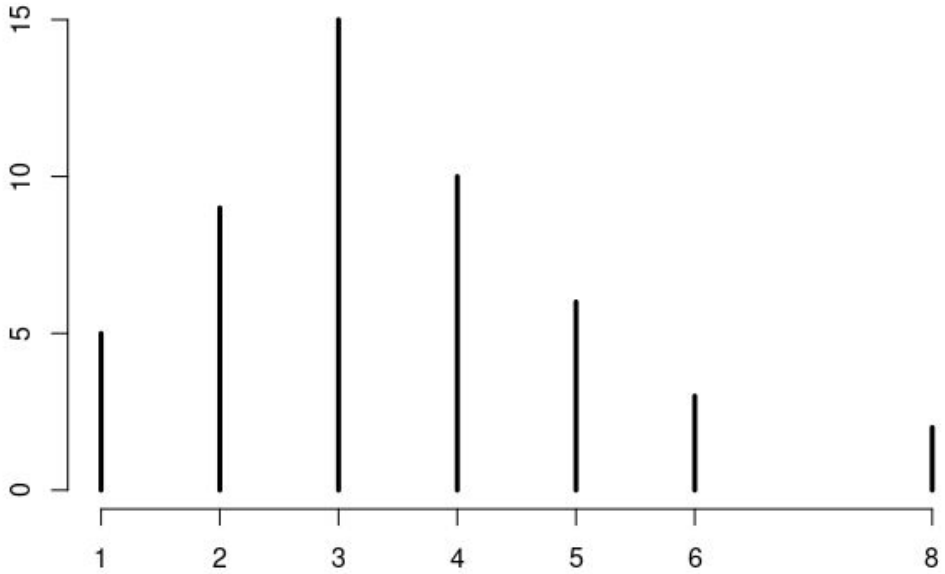
1	1	1	1	1	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	6	6	8	8

Diagramme en bâtonnets

Quand la variable est discrète, les effectifs sont représentés par des bâtonnets.

Diagramme en bâtonnets en langage R

```
> plot(table4,type = "h",xlab = "", ylab = "", main = "",frame = 0,lwd = 3)
```



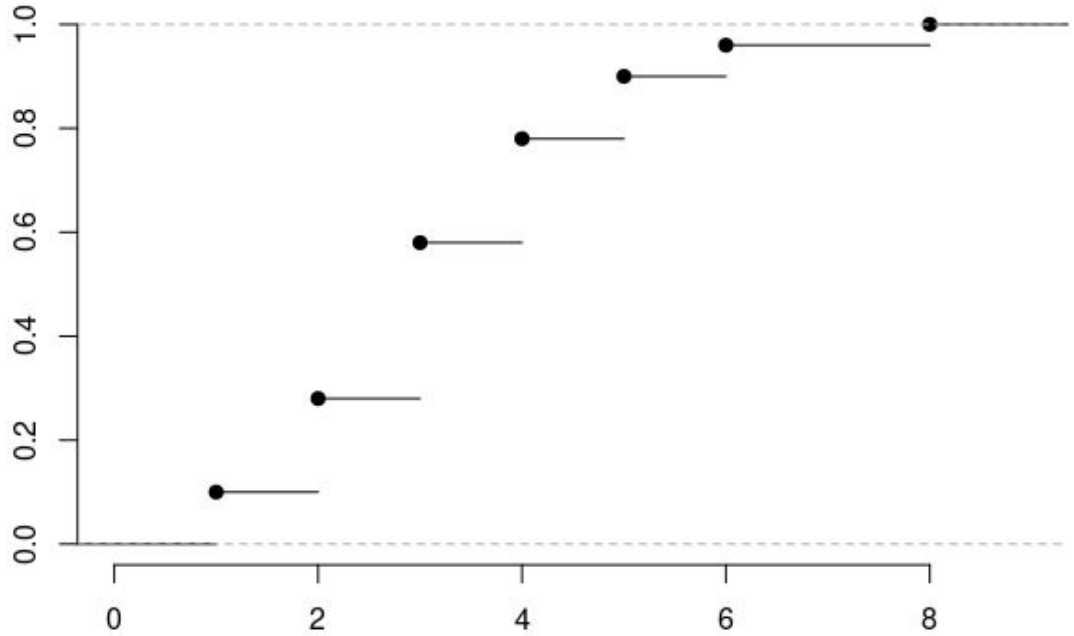
Fonction de répartition

Les fréquences cumulées sont représentées au moyen de la fonction de répartition. Cette fonction est définie dans $\mathbb{R}$ sur [0, 1], et vaut:

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_j & x_j \leq x < x_{j+1} \\ 1 & x_J \leq x. \end{cases}$$

Fonction de répartition en R

```
> plot(ecdf(z),xlab = "", ylab = "", main = "",frame = 0)
```



5.4

Variable quantitative continue

Regroupement en classe

Une variable quantitative continue peut prendre une infinité de valeurs possibles. Le domaine de la variable est alors $\mathbb{R}$ ou un intervalle de $\mathbb{R}$.

Pour faire des représentations graphiques et construire le tableau statistique, il faut procéder à des regroupements en classe.

Le tableau regroupé en classe est souvent appelé **distribution groupée**.

Regroupement en classe

Si $[c^-_j; c^+_j]$ désigne la classe j, on note, de manière générale:

- c^-_j la borne inférieure de la classe j;
- c^+_j la borne supérieure de la classe j;
- $c_j = (c^+_j + c^-_j)/2$ le centre de la classe j;
- $a_j = c^+_j - c^-_j$ l'amplitude de la classe j.

Regroupement en classe

- n_j l'effectif de la classe j;
- N_j l'effectif cumulé de la classe j;
- f_j la fréquence de la classe j;
- F_j la fréquence cumulée de la classe j.

Regroupement en classe

La répartition en classes des données nécessite de définir *a priori* le nombre de classes *J* et donc l'amplitude de chaque classe. En règle générale, on choisit au moins cinq classes de même amplitude. Cependant, il existe des formules qui nous permettent d'établir le nombre de classes et l'intervalle de classe (l'amplitude) pour une série statistique de *n* observations.

Règle de Sturges

Sturges a proposé une valeur approximative pour le nombre *J* en fonction de la taille *n* de l'échantillon: $J = 1 + log_2(n)$

Cependant, vu que le logarithme de 2 n'est pas facilement retrouvé sur les calculatrices, cette formule peut-être modifiée pour qu'on passe en base 10.

$$J = 1 + \frac{10}{3} log_{10}(n)$$

Règle de Yule

La construction de la règle de Sturges se base sur une distribution symétrique, de distribution binomiale ou gaussienne. Dès que les données ont une distribution asymétrique, ou présentent des valeurs largement étalées, le nombre de classes n'est pas optimal. La règle de Yule permet de contourner cette règle: $J = 2.5\sqrt[4]{n}.$

Regroupement en classe

L'intervalle de classe est obtenue ensuite de la manière suivante:

longueur de l'intervalle = $(x_{max} - x_{min})/J$, où x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

Tableau statistique

Il faut arrondir le nombre de classe *J* à l'entier le plus proche. Par commodité, on peut aussi arrondir la valeur obtenue de l'intervalle de classe.

A partir de la plus petite valeur observée, on obtient les bornes des classes en additionnant successivement l'intervalle de classe (l'amplitude).

Tableau statistique: exercice

On mesure la taille en centimètres de 50 plants d'hévéa sur une parcelle. Construire le tableau statistique à partir des observations suivantes:

152	152	152	153	153
154	154	154	155	155
156	156	156	156	156
157	157	157	158	158
159	159	160	160	160
161	160	160	161	162
162	162	163	164	164
164	164	165	166	167
168	168	168	169	169
170	171	171	171	171

Histogramme

L'histogramme consiste à représenter les effectifs (resp. les fréquences) des classes par des rectangles contigüs dont la surface (et non la hauteur) représente l'effectif (resp. la fréquence).

Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe *j* est donc donnée par: $h_j = \frac{n_j}{a_j}$

Histogramme

- On appelle h_j la densité d'effectif;
- L'aire de l'histogramme est égale à l'effectif total n, puisque l'aire de chaque rectangle est égale à l'effectif de la classe j: $a_j \times h_j = n_j$.

Pour un histogramme des fréquences on a: $d_j = \frac{f_j}{a_j}$

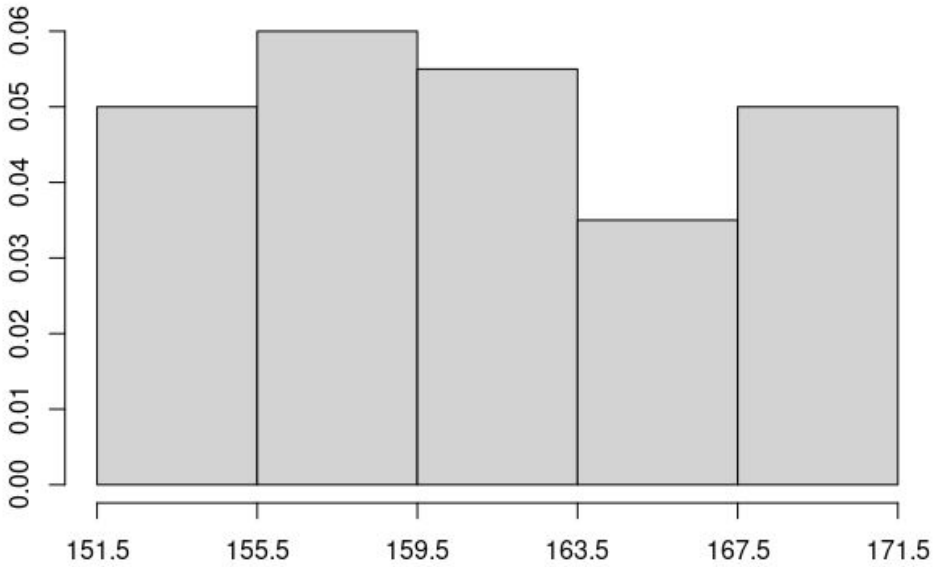
- On appelle d_j la densité de fréquence;
- L'aire de l'histogramme est égale à 1, puisque l'aire de chaque rectangle est égale à la fréquence de la classe j: $a_j \times d_j = f_j$.

Histogramme

Dans le cas de classes de même amplitude certains auteurs et logiciels représentent l'histogramme avec les effectifs (resp. les fréquences) reportés en ordonnée, l'aire de chaque rectangle étant proportionnelle à l'effectif (resp. la fréquence) de la classe.

Histogramme en langage R

```
> hist(z)
```



Fonction de répartition

La fonction de répartition *F(x)* est une fonction de $\mathbb{R}$ dans [0, 1], qui est définie par:

$$F(x) = \begin{cases} 0 & x < c_1^- \\ F_{j-1} + \frac{f_j}{c_j^+ - c_j^-}(x - c_j^-) & c_j^- \leq x < c_j^+ \\ 1 & c_J^+ \leq x \end{cases}$$

Fonction de répartition en langage R