

Licence-S5 IDDL Année 2023/2024

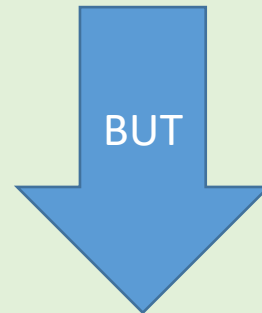
Traitement statistique des données en utilisant Logiciel R

Pr A. SOUFI

OBJECTIF

Statistique (Logiciel R)

- ☐ Utiliser des outils et des techniques pour analyser des données



- ☐ Prendre des décisions éclairées.

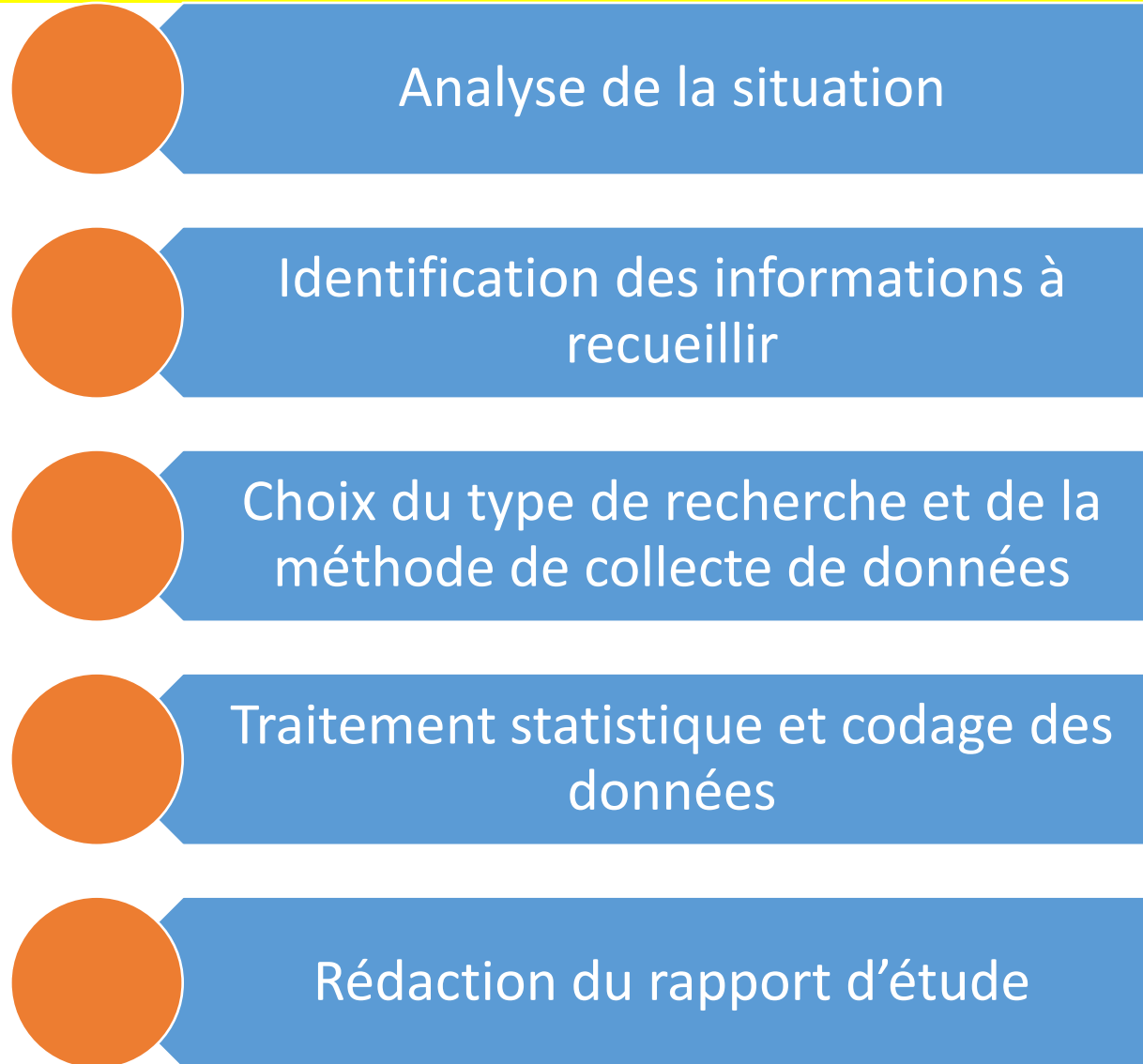
LA STATISTIQUE

La **statistique** (C'est à la fois **une science**, **une méthode** et **un ensemble de techniques**) est l'étude d'un phénomène pour :

- ✓ Collecter de données,
- ✓ Leur traitement,
- ✓ Leur analyse,
- ✓ Leur présentation
- ✓ L'interprétation des résultats

Afin de rendre les données compréhensibles

ÉTAPES D'UNE ÉTUDE STATISTIQUE



STATISTIQUE DESCRIPTIVE ET STATISTIQUE INFÉRENTIELLE

Analyse des données peut comporter plusieurs aspect qui sont regroupés sous deux thèmes soit :

☐ La statistique descriptive

➤ Dont le but est la description d un ensemble de données.


☐ La statistique inférentielle

➤ Dont le but est d'effectuer des estimations et des prévisions à partir d'un sous-ensemble de données (**Échantillon**)

OBJECTIF DE CETTE PARTIE DE FORMATION

- ❑ Organiser, présenter et décrire des données.
- ❑ Généraliser à une population, des caractéristiques observées sur des échantillons.

Plan

- I. **S**tatistiques descriptives, variable aléatoire et loi de probabilité
 - II. **É**chantillonnage
 - III. **E**stimation
 - IV. **T**ests d'hypothèse
- 
- Statistique inférentielle**

QUELQUES CONCEPTS FONDAMENTAUX

SÉRIE STATISTIQUE

Est simplement une liste de mesures obtenues généralement lors d'une étude ou de relevés de mesures.

Exemples de séries statistiques :

- Une liste avec l'âge de chaque garçon forme une série statistique : 4, 12, 7, 8
- La liste des couleurs favorites des garçons, forme aussi une série statistique: rouge, rouge, vert, noire
- Avec la nationalité des garçons, on obtient une autre série statistique: française, marocaine, allemande, française

LA POPULATION D'UNE SÉRIE STATISTIQUE

La **population** est l'ensemble des personnes ou des animaux ou des choses qui sont étudiées. (Donc, pas forcément des êtres humains !).

On se pose la question : « **Sur qui porte l'étude ?** Quelle personne, quel animal ou quelle chose ? ».

Exemples de population :

- Si on note l'âge des garçons de la classe, la population c'est les garçons.
- Si on relève leur couleur favorite, la population c'est aussi les garçons.
- Si j'étudie le sommet des montagnes, la population c'est les montagnes
- Si je note la vitesse des voitures, la population c'est les voitures.

LE CARACTÈRE D'UNE SÉRIE

Le caractère d'une série statistique c'est ce qui est mesuré ou étudié pour la population.

On se pose la question «**Sur quoi porte l'étude ?**»

Exemples :

- Si je note l'âge des garçons, le caractère c'est l'âge.
- Si je relève leur couleur favorite, le caractère c'est leur couleur préférée.
- Si je mesure la hauteur des sommets de montagnes, le caractère c'est la hauteur.
- Si je mesure la vitesse des voitures, le caractère c'est la vitesse.



Une **série statistique** = une série de **valeurs**.

La **population** = quelles personnes, quelles choses, quels animaux on étudie ?

Le **caractère** = qu'est-ce qu'on étudie ?

MODALITÉS D'UN CARACTÈRE OU D'UNE VARIABLE STATISTIQUE

On désigne aussi par **modalités** les différentes valeurs présentées par les individus, d'une population relativement à une variable statistique

Exemple:

Variable « sexe » des Employer :

- ✓ La première modalité (Femme)
- ✓ La deuxième modalité (Homme)

VARIABLE QUANTITATIVE – VARIABLE QUALITATIVE

Les résultats de l'observation d'une variable pourront s'exprimer d'une manière qualitative ou quantitative selon qu'il sont mesurable ou non.

- **Variable quantitative** : si elle peut être soit mesurée, soit repérer par un nombre. (discrète ou continue)
- **Variable qualitative** : c est une valeur peut être définie par un code (nominale ou ordinale)

Exemple :

- ❑ le niveau de formation de l'employé (A, B et C)
- ❑ le sexe de l'employé (F et M)

EFFECTIFS

- Effectif n (fréquence absolue) d'une population est le nombre d'individus qui composent cette population.
- Effectif n_i relatif à la modalité i du caractère X est le nombre d'individus de la population qui présentent la modalité i .

TABLEAU DE DISTRIBUTION ($x_i; n_i$)

Pour construire un tel tableau, il suffit de calculer l'effectif de chaque modalité.



Pour calculer l'effectif de chaque modalité en utilisant le logiciel R à l'aide de la fonction **table()**

EXEMPLE PRATIQUE :

```
couleur <- c("rouge", "vert", "bleu", "rouge", "bleu", "vert")  
table(couleur)
```

FRÉQUENCES

La fréquence relative f_i d'une modalité i du caractère X est la proportion d'individus de la population qui présentent la modalité i , ainsi:

$f_i = \frac{n_i}{n}$ est la fréquence relative de la modalité i de X .

La somme des fréquences relatives est égale à 1 ou 100%.

EFFECTIFS ET FRÉQUENCES CUMULÉS

On utilise la *fréquence cumulée* pour déterminer le nombre d'observations qui se situent au-dessus (ou au-dessous) d'une valeur particulière dans un ensemble de données

- **Effectifs cumulés croissants(ascendants)**: l'effectif cumulé croissant jusqu'à une valeur **X** est le nombre d'observations strictement inférieures à x . (La **somme des effectifs des valeurs inférieures.**)
- **Effectifs cumulés décroissants(descendants)**: l'effectif cumulé décroissant jusqu'à une valeur **X** est le nombre d'observations strictement supérieur à x (La **somme des effectifs des valeurs supérieures.**)

EXEMPLE

Le tableau représentant les valeurs et les effectifs est la série statistique suivante :

Valeur	0	1	2	3	4	5
Effectif	3	3	12	6	3	3

- Si l'on s'intéresse à la question « Combien d'élèves ont moins de 3 frères et sœurs ? », on calcule ce que l'on appelle les **effectifs cumulés croissants**.

On les obtient **en additionnant les effectifs des valeurs inférieures**.

Par exemple, le nombre d'élèves ayant moins de 3 frères et sœurs est : $3 + 3 + 12 = 18$

DISTRIBUTIONS ET TABLEAUX STATISTIQUES

Si on décrit les individus selon un seul caractère, les tableaux statistiques sont à une dimension

Modalités du caractère	effectifs
C1	n1
C2	n2
C3	n3
C4	n4
C5	n5
C6	n6
total	n

CARACTÈRE QUANTITATIF CONTINU

On remplace C_i par l'intervalle $[e_{i-1}, e_i[$ (la classe)

$e_i - e_{i-1}$ est l'amplitude de la classe



Pour calculer la fréquence relative de chaque modalité d'une variable qualitative en utilisant le logiciel R, il suffit d'utiliser la fonction `prop.table()`. Cette fonction prend en entrée un tableau de effectifs , et renvoie un tableau de fréquences relatives

EXEMPLE PRATIQUE :

```
sondage <- data.frame(  
  yeux    = c("brun",    "brun",    "bleu",    "brun",    "vert",    "brun",    "bleu"    )  
  cheveux = c("brun",    "noir",    "blond",    "brun",    "brun",    "blond",    "brun"    )  
  genre    = c("féminin", "masculin", "féminin", "féminin", "masculin", "féminin", "masculin")  
)  
sondage
```

```
table(sondage$cheveux)
```

```
table(yeux = sondage$yeux, cheveux = sondage$cheveux)
```

```
prop.table(table(sondage$cheveux))
```

```
prop.table(table(yeux = sondage$yeux, cheveux = sondage$cheveux))
```


IMPORTATION DES DONNÉES DANS R

-----Depuis Excel

Utiliser la fonction `read_excel()` du package `readxl`

Exemple :

```
>install.packages("readxl")
```

```
>library("readxl")
```

```
>data <- read_excel("data.xlsx")
```



`file.choose()`

-----Depuis fichier text ou CSV

utilisant la fonction `read.table()` ou `read.csv()`

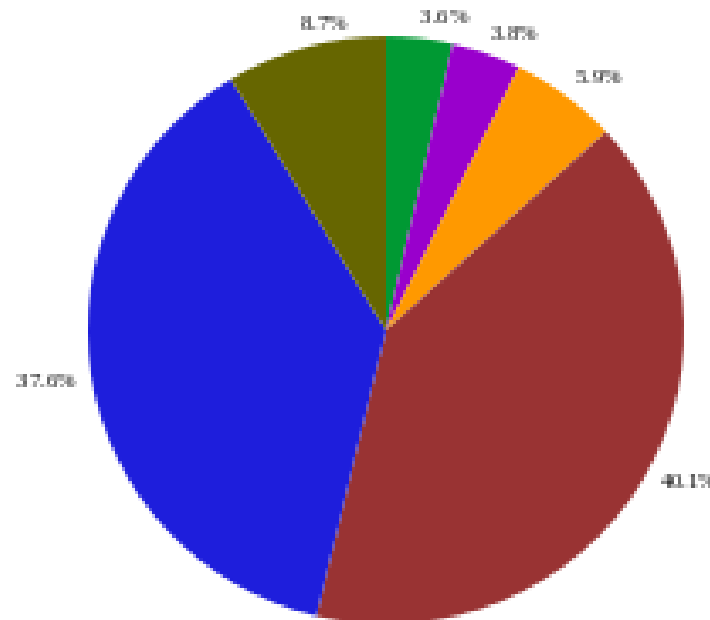
Exemple :

```
>data <- read.csv("data.csv")
```

REPRÉSENTATIONS GRAPHIQUES

DIAGRAMME CIRCULAIRE

- Est un moyen de représenter une série statistique dont le caractère est qualitatif. Il est obtenu en découpant un disque en secteurs dont les mesures d'angle sont proportionnelles à l'effectif. ($\text{angle} = (n_i \cdot 360) / N$)





Pour créer un diagramme circulaire en R, on utilise la fonction `geom_bar()` du package `ggplot2`

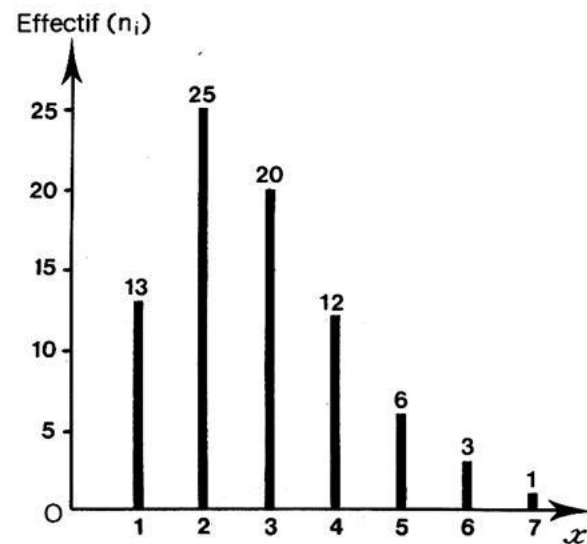
Exemple :

```
#data <- data.frame(catégorie = c("Chocolat", "Vanille",  
  "Fraise"),valeur = c(40, 30, 30))  
  
#ggplot(data, aes(x = "", y = valeur, fill = catégorie))  
  +geom_bar(width = 1, stat = "identity") +coord_polar(theta =  
  "y") +scale_fill_manual(values = c("red", "blue", "green"))  
  +labs(fill = "Catégorie",title = "Diagramme circulaire des  
  préférences")
```

DIAGRAMME À BÂTONS (x_i, n_i)

Est un moyen de représenter une série statistique dont le caractère est quantitatif discret.

Si x_1, \dots, x_p sont les valeurs possibles prises par le caractère et si les effectifs correspondants sont n_1, \dots, n_p , il est constitué par les segments qui relient le point $(x_k, 0)$ au point (x_k, n_k) .



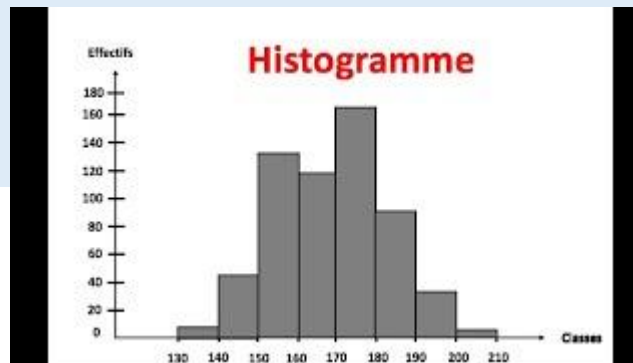


```
#library(ggplot2)
#data <- data.frame( genre = c("Homme", "Femme"), nombre =
  c(10, 20) )
# ggplot(data, aes(x = genre, y = nombre)) + geom_bar(stat =
  "identity")
```

HISTOGRAMME

Est un moyen de représenter une série statistique dont le caractère est quantitatif Continu.

Si la série statistique est donnée par les classes $([a_i, a_{i+1}[$), il est constitué par des rectangles dont la base est le segment $[a_i, a_{i+1}[$ (sur l'axe des réels) et l'aire est proportionnelle à l'effectif de la classe.





Pour tracer un histogramme en R, nous pouvons utiliser la fonction **hist()**.

Cette fonction prend en entrée un vecteur de données et crée un histogramme à partir de ce vecteur.

----Créer un vecteur de données

```
#scores <- rnorm(100, mean = 50, sd = 10)
```

----Tracer l'histogramme

```
#hist(scores)
```

----Changer la largeur des barres de l'histogramme

```
#hist(scores, binwidth = 5)
```

----Changer la couleur des barres de l'histogramme

```
#hist(scores, col = "blue")
```


CARACTÉRISTIQUES DE TENDANCE CENTRALE ET DE DISPERSION

Mesures descriptives les plus couramment utilisées sont résumées dans le tableau ci-après:

Tendance centrale

- Moyenne arithmétique
- Médiane
- Mode

Dispersion

- Étendue
- Écart-type
- Coefficient de variation

Position

- Quantiles

MESURE DE TENDANCE CENTRALE

- ❑ **Moyenne Arithmétique** : la somme des données de la série divisée par le nombre de données. (**sous-R: mean()**)
- ❑ **Médian** : Mesure de tendance centrale d'une série ordonnée qui partage en deux les données de la série de sorte qu'au plus 50% des données lui sont inférieures et au plus 50% lui sont supérieures.
(**sous-R : median()**)
- ❑ **Mode** : Donnée de la série qui revient le plus fréquemment.
(**sous-R: Pour calculer la mode en R, il faut d'abord calculer l'effectif de chaque valeur de la variable. L'effectif est le nombre de fois qu'une valeur apparaît dans la série de données. Une fois que les effectifs sont calculés, il faut trouver la valeur qui a l'effectif le plus élevé en utilisant la fonction **which.max()** . Cette valeur est le mode.**)

MESURE DE DISPERSION

- ❑ Donnent une indication de *la variation* des données. Elles résument comment les modalités sont *homogènes* ou *hétérogènes*
- ❑ Décrire la dispersion consiste à mesurer la divergence des modalités par *rapport la moyenne*

MESURE DE DISPERSION

- **La variance** : Mesure la ***divergence*** des données par rapport à la moyenne. De façons techniques, c'est la moyenne des écarts au carré des modalités par rapport à la moyenne.

(sous-R : var(plage))

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

- **L'écart-type** : L'écart-type est la ***racine carrée de la variance***

(sous-R : sd())

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

COEFFICIENT DE VARIATION

Coefficient de variation permet d'évaluer *l'importance relative* de la *dispersion* d'une distribution. Il permet ainsi de *comparer* la dispersion de *2 distributions* qui n'ont pas la même *unité de mesure*. Il est donné par :

Sous-R : `cv()`

$$CV(x) = \frac{s_x}{\bar{x}} * 100$$

- Indique le degré d'*homogénéité* d'une distribution.
- Un coefficient de variation moins que **15%** , une indication d'une bonne homogénéité de la distribution des données

QUARTILES

- Le **quartile inférieur** est la valeur du milieu du premier ensemble, dans lequel 25 % des valeurs sont inférieures à Q_1 et 75 % lui sont supérieures. le premier quartile prend la notation Q_1 .
- Le **quartile supérieur** est la valeur du milieu du deuxième ensemble, dans lequel 75 % des valeurs sont inférieures à Q_3 et 25 % lui sont supérieures. Le troisième quartile prend donc la notation Q_3

(Sous-R: `quantile(data, probs=c(0.25,0.50,0.75))`)

LA FONCTION SUMMARY()

Cette fonction renvoie un résumé des statistiques descriptives d'un vecteur de données, y compris les quartiles.

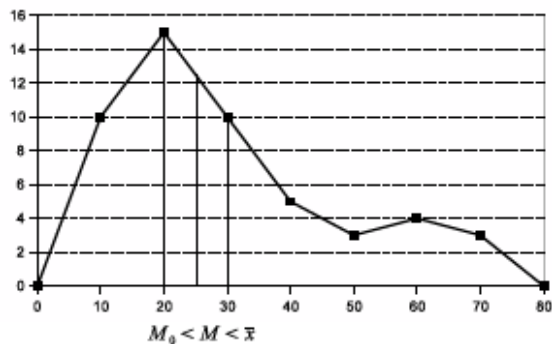
Exemple :

```
summary(data)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	2.50	5.00	5.50	7.50	10.00

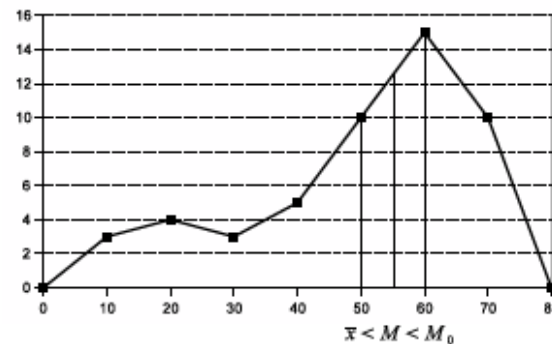
MESURES DE FORME: COEFFICIENTS DE DISSYMMÉTRIE

La distribution des valeurs est **dissymétrique à droite** si la portion du polygone des fréquences située à droite du sommet est plus longue que l'autre.



Dans ce cas $M_0 < M < \bar{x}$

La distribution des valeurs est **dissymétrique à gauche** si la portion du polygone des fréquences située à gauche du sommet est plus longue que l'autre.



Dans ce cas $\bar{x} < M < M_0$

$$\beta_1 = \frac{3(\bar{x} - M)}{\sigma}$$

Sa valeur est généralement comprise entre -1 et +1:

$\beta_1 < 0$ distribution dissymétrique à gauche

$\beta_1 = 0$ distribution symétrique

$\beta_1 > 0$ distribution dissymétrique à droite

Sous R : La fonction `skewness()` de la librairie `e1071`

CORRÉLATION LINÉAIRE

- ❑ **Objectif** : Etudier la relation bivariée, entre deux variables quantitatives X et Y, mesurées sur les mêmes individus.
- ❑ **Coefficient** de corrélation linéaire (r) : c est un indice qui permet d'estimer la force du lien entre les variable

NB : Plus la corrélation est forte, plus la valeur de r s'approche de 1(ou -1)

(sous-R : `cor(plage_Y,plage_X)`)

RAPPEL SUR LES NOTIONS DE PROBABILITÉ UTILISÉES EN STATISTIQUES

PLAN

- ✓ **Variable Aléatoire**
- ✓ **Loi normale**

DÉFINITION

Une variable aléatoire X est une fonction de l'ensemble fondamental Ω à des valeurs dans \mathbb{R} , $X : \Omega \rightarrow \mathbb{R}$

❖ Lorsque la variable X ne prend que des valeurs discrètes, on parle de **variable aléatoire discrète**

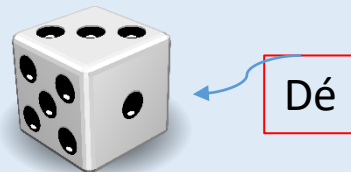
EXEMPLE

On lance deux Dés distincts et on s'intéresse à la somme des points. On note X une variable aléatoire, elle est définie par :

$X: \Omega \rightarrow \mathbb{R}$ avec $\Omega = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$

$(\omega_1, \omega_2) \rightarrow \omega_1 + \omega_2$

L'ensemble des valeurs possibles de X est $\{2, 3, \dots, 12\}$



FONCTION DE RÉPARTITION

En théorie des probabilités, la fonction de répartition d'une variable aléatoire réelle X est la fonction qui, à tout réel x , donne la probabilité que X soit inférieur ou égal à x

$$F(x) = P(X \leq x)$$

EXEMPLE

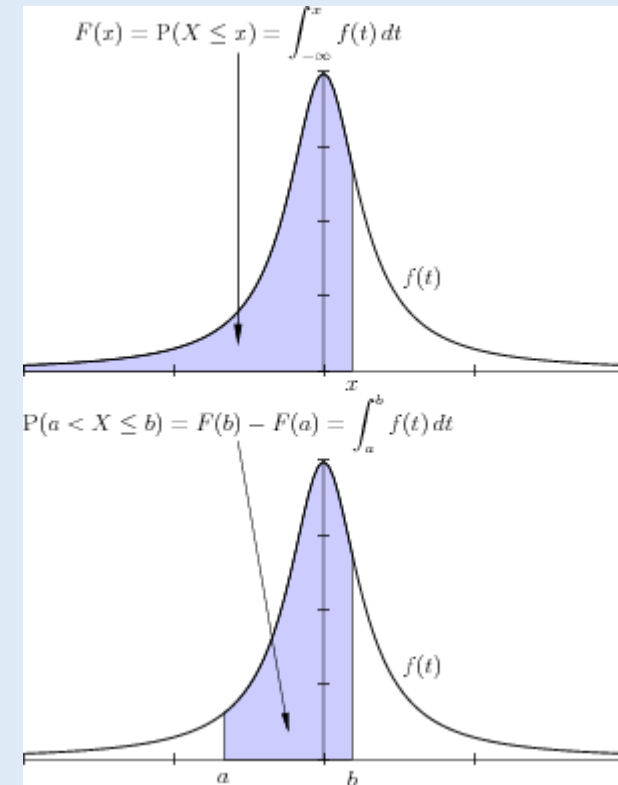
Si « X est la variable aléatoire qui représente la température à un instant donné »

Alors « $F(0)$ est la probabilité que la température soit inférieure ou égale à 0 et $F(10)$ est la probabilité que la température soit inférieure ou égale à 10, et ainsi de suite. »

DENSITÉ DE PROBABILITÉ D'UNE VARIABLE CONTINUE

La **densité de probabilité** f d'une variable aléatoire continue est une fonction qui décrit la probabilité qu'une variable aléatoire prenne une valeur dans un intervalle donné. Elle est définie comme suit :

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx.$$



ESPÉRANCES MATHÉMATIQUES

L'espérance mathématique $E[X]$ d'une variable aléatoire se définit comme la moyenne des valeurs prises par cette variable, pondérées par leurs probabilités :

✓ Dans le cas d'une variable discrète :
$$E[X] = \sum_{k=-\infty}^{+\infty} k P_X(k)$$

✓ Pour une variable continue :
$$E[X] = \int_{-\infty}^{+\infty} x p_X(x) dx$$

NB : $E[X] \Leftrightarrow$ la moyenne notée \bar{X}

EXEMPLE

Si « l'on sait que 10 % des ampoules ont une durée de vie de 100 heures, 20 % une durée de vie de 200 heures, 30 % une durée de vie de 300 heures, etc., »

Alors « l'espérance mathématique de X est :

$$\begin{aligned} E(X) &= (0,1) * 100 + (0,2) * 200 + (0,3) * 300 + \dots \\ &= 250 \end{aligned}$$

EXEMPLE SOUS-R

```
library(stats)
mu <- 250
sigma <- 50
x <- seq(0, 500, 1)
# F fonction de répartition de X
F <- 1 - pnorm(x - mu, mean = mu, sd = sigma)
# Graphique de la fonction de répartition de X :
plot(x, F, type = "l", xlab = "x", ylab = "F(x)")
# La densité de probabilité de X
f <- dnorm(x - mu, mean = mu, sd = sigma) / sigma
# graphique de la densité de probabilité de X
plot(x, f, type = "l", xlab = "x", ylab = "f(x)")
E <- mu
```

MÉDIANE

- On appelle médiane d'une variable aléatoire X , un réel m tel que :

$$P(X \leq m) \geq 1/2 \leq P(X \geq m)$$

LOIS DE PROBABILITÉ

La loi de probabilité est une fonction qui décrit la probabilité qu'une variable aléatoire prenne une valeur donnée ou un ensemble donné de valeurs :

Il existe deux principaux types de lois de probabilité :

1. Les lois de probabilité discrètes ne peuvent prendre qu'un nombre fini ou dénombrable de valeurs. Par exemple, le résultat d'un lancer de dé est une variable aléatoire discrète, car il peut prendre les valeurs $\{1,2,3,4,5,6\}$
2. Les lois de probabilité continues peuvent prendre n'importe quelle valeur dans un intervalle donné. Par exemple, la température à un instant donné est une variable aléatoire continue, car elle peut prendre n'importe quelle valeur entre 100 C° et 300 C° .

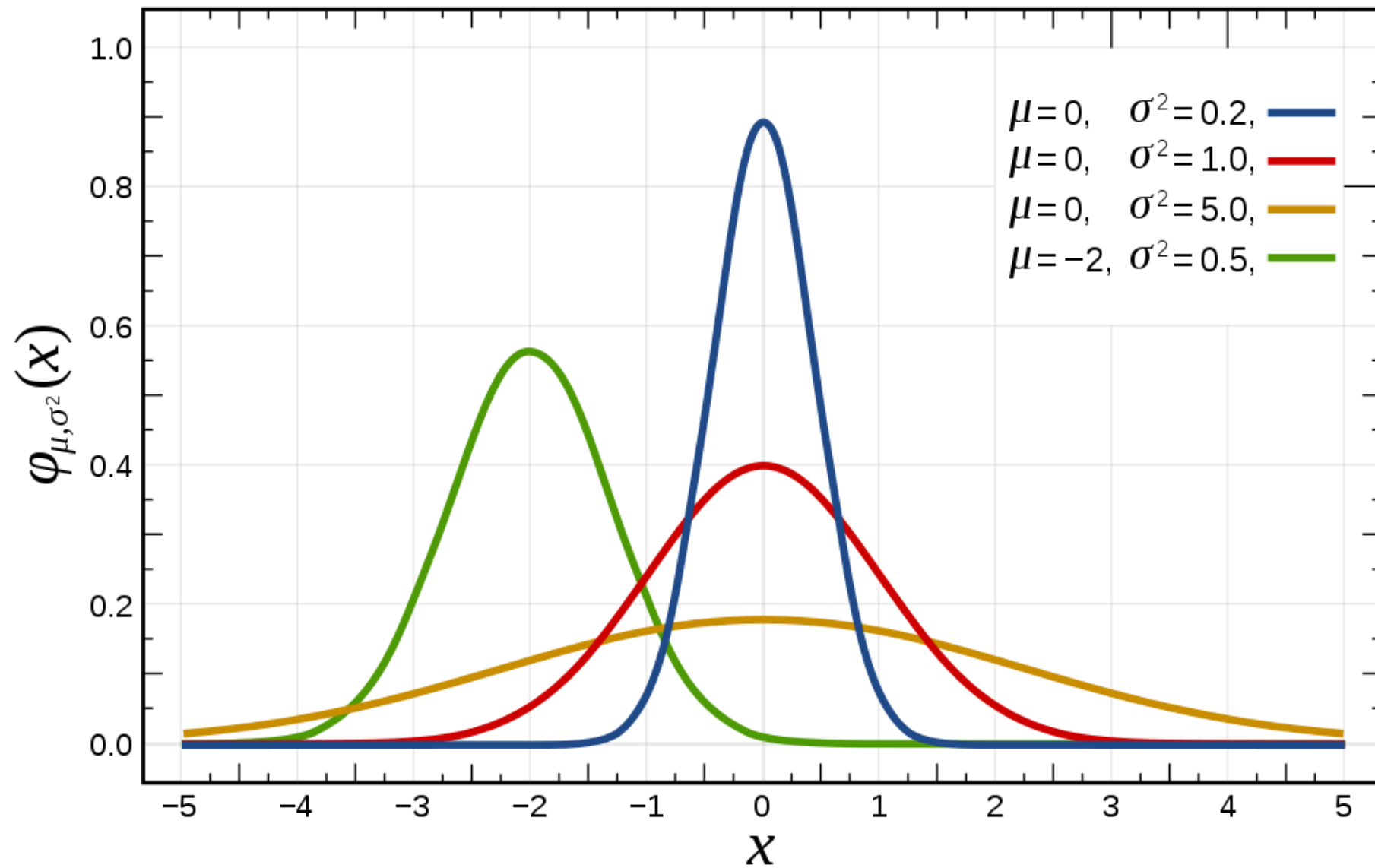
LA LOI NORMALE

- Est une loi de probabilité continue qui décrit de nombreux phénomènes aléatoires, tels que la taille des personnes, la durée de vie des composants électroniques, ou le prix des actions.
- La loi normale est caractérisée par deux paramètres :
 1. La moyenne μ , qui correspond à la valeur centrale de la distribution
 2. L'écart-type σ , qui correspond à la dispersion des valeurs autour de la moyenne.

La densité de probabilité d'une loi normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Les courbes représentent la fonction
densité de probabilité de la loi normale



Notation

Lorsqu'une variable aléatoire X suit une loi normale, elle est dite *gaussienne* ou *normale* et il est habituel d'utiliser la notation avec la variance σ^2 :

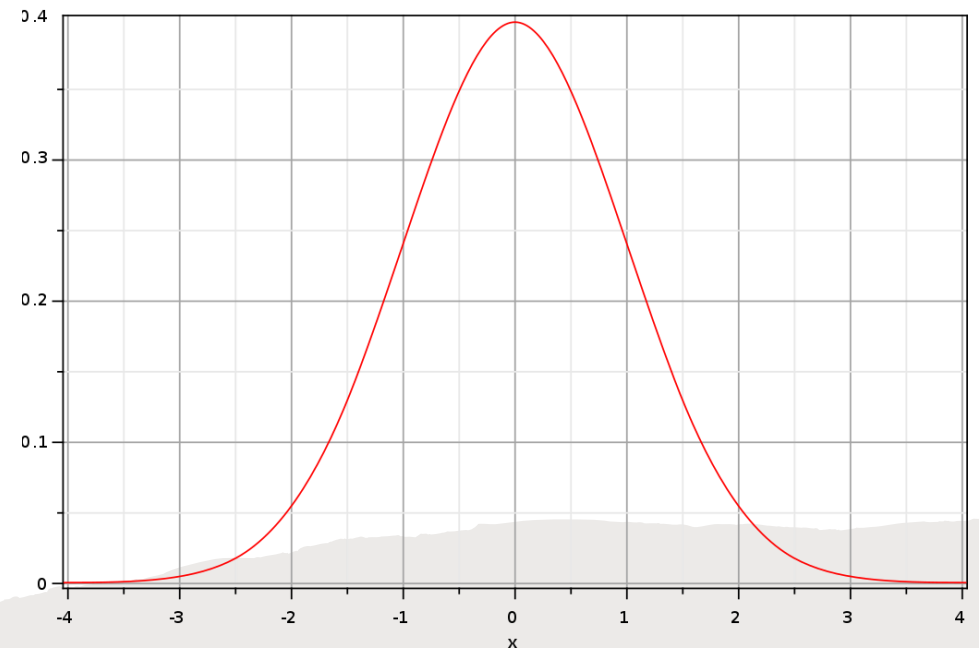
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Loi normale centrée réduite

La loi normale centrée réduite est une loi normale dont la moyenne est égale à 0 et l'écart-type est égal à 1. Elle est souvent notée $N(0,1)$.

La densité de probabilité d'une loi normale centrée réduite est donnée par la formule suivante :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$



PASSER D'UNE LOI NORMALE À LA LOI NORMALE CENTRÉE RÉDUITE

Il suffit de soustraire la moyenne de la distribution à chaque valeur de la variable aléatoire, puis de diviser le résultat par l'écart-type.

La formule suivante résume cette transformation :

$$Z = (X - \mu) / \sigma$$

EXEMPLE

On suppose que la taille des personnes suive une loi normale avec une moyenne de 170 cm et un écart-type de 10 cm.

Donc SI < La variable aléatoire X qui représente la taille d'une personne suit donc une loi normale $N(170,10)$.>

Alors < Pour passer à la loi normale centrée réduite, on calcule la variable aléatoire Z suivante :

$$Z = (X - \mu) / \sigma = (\text{Taille} - 170) / 10 >$$

SIMULATION DE LA LOI NORMALE EN R

- Pour simuler une loi normale en R, nous pouvons utiliser la fonction `rnorm()`.
- Cette fonction prend deux arguments : la moyenne et l'écart-type de la loi normale.

Exemple :

pour simuler 100 valeurs d'une loi normale centrée réduite (moyenne = 0, écart-type = 1), nous pouvons utiliser le code suivant :

```
n <- 100
```

```
x <- rnorm(n, mean = 0, sd = 1)
```

```
hist(x)
```

EXERCICE

On suppose que Z suit la $N(0,1)$ calculer :

- $P[Z \leq 1,26]$
- $P[Z \leq -0,94]$

On suppose que X suit la loi $N(18,4)$ calculer :

- $P[16,72 \leq X \leq 18,94]$

SOUS-R : SIMULATION DE LA LOI NORMALE

```
qnorm(0.975)      # quantile d'ordre 0.975 de la loi N(0,1)
dnorm(0)           # valeur de la fonction de densité en 0 de la loi N(0,1)
pnorm(1.96)        # valeur de la fonction de répartition en 1.96 de la loi N(0,1)
rnorm(20)          # génération de 20 réalisations indépendantes suivant la loi N(0, 1)
rnorm(10,mean=5,sd=0.5) # génération de 20 réalisations indépendantes suivant la loi N(5, 0.25)
```


ECHANTILLONNAGE

ECHANTILLONNAGE

- On a une population $P(\mu, \sigma, p)$ les paramètres connus .
- On prélève des échantillons ch_1, ch_2, \dots, ch_p
- le contexte des tirages d'échantillons :
 - **Échantillonnage aléatoire simple** : chaque unité de la population a la même probabilité d'être sélectionnée. C'est la méthode la plus simple et la plus équitable.
 - On peut effectuer un tirage avec remise ou un tirage sans remise.

DISTRIBUTION D'ÉCHANTILLONNAGE DES MOYENNES

- Pour chaque échantillon on calcule la moyenne \bar{x}_i
- On dit que on a une La distribution d'échantillonnage des moyennes
- Donc on peut calculer la moyenne des moyennes $\mu_{\bar{x}_i} = m$ (Moy pop)
- L'cart type $\sigma_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}}$
- NB : dans le cas d'une population fini de taille N

$$\sigma_{\bar{x}_i} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

DISTRIBUTION D'ÉCHANTILLONNAGE DES PROPORTIONS

- Pour chaque échantillon on calcule la proportion p_i (fréquence relative d'une modalité)
- On dit que on a une La distribution d'échantillonnage des proportions
- Donc on peut calculer la moyenne des proportions $\mu_{p_i} = p$
- L'écart type $\sigma_{p_i} = \sqrt{\frac{p(1-p)}{n}}$

DANS LE CAS DE LA TAILLE $n \geq 30$

- La variable aléatoire \bar{X} associé au \bar{x}_i suit la loi normal $N(m, \frac{\sigma}{\sqrt{n}})$ et F suit la loi normal $N(p, \sqrt{\frac{p(1-p)}{n}})$

DANS LE CAS D UNE POPULATION AVEC UNE DISTRIBUTION NORMAL

Quel que soit la taille n des échantillons ,

- La variable aléatoire \bar{X} associé au \bar{x}_i suit la loi normal $N(m, \frac{\sigma}{\sqrt{n}})$ et F suit la loi normal $N(p, \sqrt{\frac{p(1-p)}{n}})$

EXERCICE 1

Une machine découpe des rondelles de diamètre moyen 20 mm et d'écart type 2 mm. On prélève un échantillon de 100 pièces.

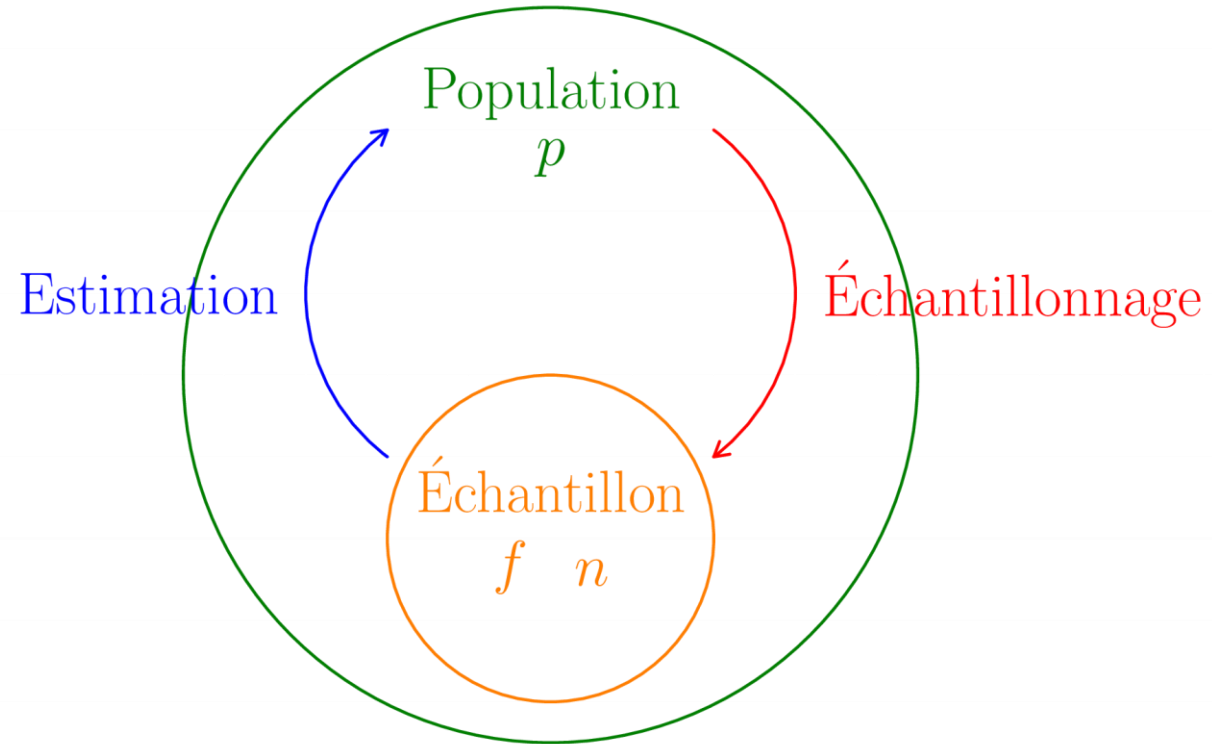
Déterminer la probabilité que la moyenne des diamètres de cet échantillon soit inférieure à 20,4 mm

EXERCICE 2

8% des rondelles sont défectueuses. On prélève un échantillon de 100 pièces. Déterminer la probabilité que la proportion de rondelles défectueuses dépasse 10%.

ESTIMATION

Estimation est le processus inverse de l'échantillonnage



ESTIMATION

❑ On a une population on veut estimer la moyenne m et/ou la proportion p

❑ Types d'estimateurs :

- Estimateurs ponctuels
- Estimateurs intervalles de confiance

ESTIMATION PONCTUELLE

✓ On a une Population (moyenne m , écart type s , proportion p) et on prélève un échantillon **CH** de taille n .

Si

On calcule Les paramètres de **CH** sont : \bar{x} moyenne , σ_e Ecart type et f la proportion des individus possédant un caractère donné.

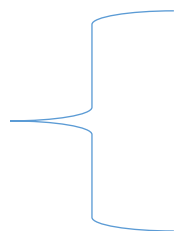
Alors

$$m = \bar{x}$$

$$s = \sigma_e \sqrt{\frac{n}{n-1}}$$

$$p = f$$

ESTIMATION PAR INTERVALLE DE CONFIANCE

- Estimation de m 
 - s connu
 - s inconnu
- Estimation de p

ESTIMATION DE LA MOYENNE m PAR INTERVALLE DE CONFIANCE

LE CAS 1 : ECART TYPE σ CONNU

Rappel :

- On suppose que X suivre la loi normale $N(m, \sigma)$
- On veut trouver un intervalle I centré en m avec la probabilité de 95% que x soit dans I , c'est à dire que $P(m - h \leq X \leq m + h) = 0,95$
- On passe à la loi normale centrée réduite on aura : $T = \frac{X-m}{\sigma}$
- Donc $P(m - h \leq X \leq m + h) = 0,95 \leftrightarrow P\left(\frac{-h}{\sigma} \leq T \leq \frac{h}{\sigma}\right) = 0,95$
- $P\left(\frac{-h}{\sigma} \leq T \leq \frac{h}{\sigma}\right) = 0,95 \leftrightarrow 2P(T \leq \frac{h}{\sigma}) - 1 = 0,95 \leftrightarrow h = 1,96\sigma$

LE CAS 1 : ECART TYPE σ CONNU

À mémoriser:

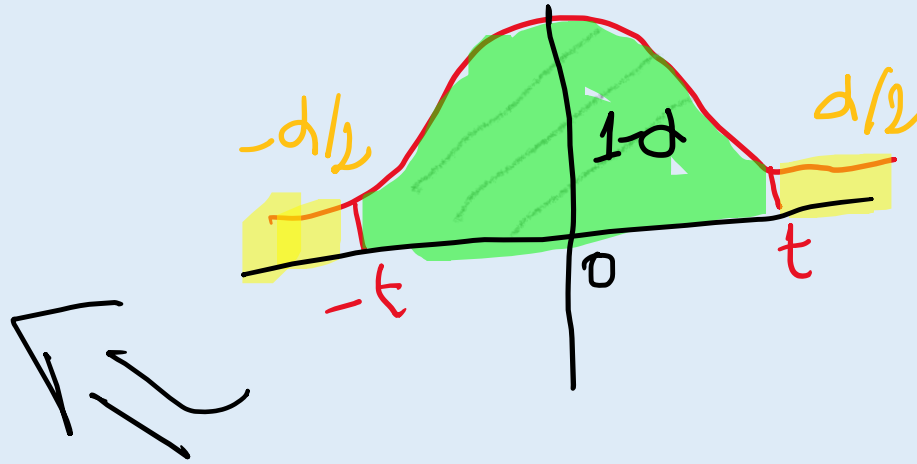


$T \rightarrow N(0,1)$ Si on cherche $P(-t \leq T \leq t) = 0,97 \Leftrightarrow$
 $2\pi(t) - 1 = 0,97 \rightarrow \pi(t) = \frac{1,97}{2} \rightarrow t = 2,17$

En général :

$$P(-t \leq T \leq t) = 1 - \alpha$$

si $\alpha = 5\%$ $t = 1,96$
si $\alpha = 3\%$ $t = 2,17$



LE CAS 1 : ECART TYPE σ CONNU

D'après la loi de l'échantillonnage on $\bar{X} \rightarrow N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

On sait que : $P\left(m - t \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq m + t \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

Question : on veut estimer la moyenne m par un intervalle de confiance avec un coefficient de confiance $2\pi(t) - 1 = 1 - \alpha$.

\bar{x} est déjà calculé c est la moyenne de l'échantillon donc

$$P\left(\bar{x} - t \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + t \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

LE CAS 1 : ECART TYPE σ CONNU

L'intervalle de confiance de m avec un coefficient de confiance

$$\text{est : } I_{\alpha} = \left[\bar{x} - t \frac{\sigma}{\sqrt{n}}; \bar{x} + t \frac{\sigma}{\sqrt{n}} \right]$$

tel que : \bar{x} : est la moyenne calculée depuis l'échantillon

σ : Ecart type connu de la population

n : est la taille de l'échantillon prélevé

t : est calculé à partir de $2\pi(t) - 1 = 1 - \alpha$

LE CAS 1 : ECART TYPE σ INCONNU

- On calcule l'estimation ponctuelle de $\sigma = \sigma_e \sqrt{\frac{n}{n-1}}$

$$I_\alpha = \left[\bar{x} - t \frac{\sigma_e}{\sqrt{n-1}} ; \bar{x} + t \frac{\sigma_e}{\sqrt{n-1}} \right]$$

ESTIMATION DE LA PROPORTION p PAR INTERVALLE DE CONFIANCE

D'après la loi d'échantillonnage $F \rightarrow N \left(p, \sqrt{\frac{p(1-p)}{n}} \right)$ donc

$$P \left(p - t \sqrt{\frac{p(1-p)}{n}} \leq F \leq p + t \sqrt{\frac{p(1-p)}{n}} \right) \text{ Alors}$$

$$P \left(p - t \sqrt{\frac{p(1-p)}{n}} \leq p \leq p + t \sqrt{\frac{p(1-p)}{n}} \right)$$

$$I_{\alpha} = \left[f - t \sqrt{\frac{f(1-f)}{n-1}}; f + t \sqrt{\frac{f(1-f)}{n-1}} \right]$$

ESTIMATION PAR IC EN R



EXEMPLE 1 : ESTIMATION DE L'INTERVALLE DE CONFIANCE D'UNE MOYENNE

Supposons que nous disposions d'un échantillon de taille 100 de valeurs de poids. La moyenne de l'échantillon est de 75 kg. Nous souhaitons estimer l'intervalle de confiance de cette moyenne à 95 %.

```
# Importation du package stats
```

```
library(stats)
```

```
# Définition des données
```

```
x <- c(70, 72, 73, ..., 82, 84)
```

```
# Estimation de l'intervalle de confiance
```

```
t.test(x, conf.level = 0.95)
```

La sortie de la fonction `t.test()` est la suivante :

One-sample t test

data: x

t = 2.308, df = 99, p-value = 0.026

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

69.93364 79.96636

Ce résultat indique que la moyenne de la population est comprise entre 69,9 et 79,9 kg avec une probabilité de 95 %.

EXEMPLE 2 : ESTIMATION DE L'INTERVALLE DE CONFIANCE D'UNE PROPORTION

Supposons que nous disposions d'un échantillon de taille 1000 de personnes, dont 500 sont favorables à une nouvelle loi. Nous souhaitons estimer l'intervalle de confiance de cette proportion à 95 %.

```
# Importation du package stats

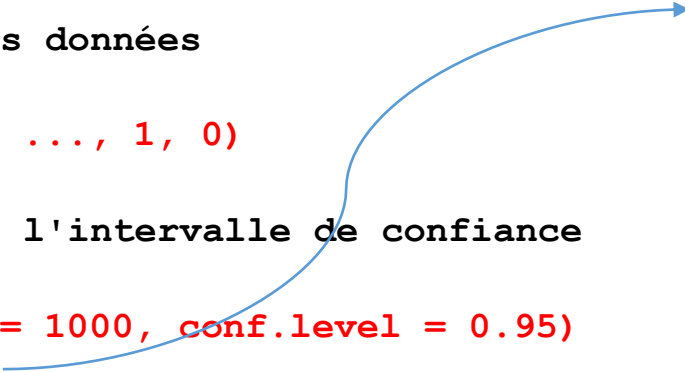
library(stats)

# Définition des données

x <- c(1, 1, 0, ..., 1, 0)

# Estimation de l'intervalle de confiance

prop.test(x, n = 1000, conf.level = 0.95)
```



La sortie de la fonction `prop.test()` est la suivante :

```
One-sample proportions test

data: x
number of successes: 500
number of trials: 1000
sample estimate: 0.5
confidence interval:
0.463344 0.536656
hypothesis test:
p-value = 0.0000000000000000444
```

Ce résultat indique que la proportion de personnes favorables à la nouvelle loi dans la population est comprise entre 46,3 % et 53,7 % avec une probabilité de 95 %.

EXERCICE

Une enquête concernant la santé, et portant sur 3000 adolescents d'un certain pays européen de 12 à 20 ans, a dénombré 570 adolescents ayant pris un psychotrope au cours des 12 mois précédant l'enquête. Parmi les 1 400 filles, 378 ont pris un psychotrope.

1. Donner une estimation ponctuelle de la fréquence de consommation de psychotropes chez les adolescents de ce pays
2. Estimer par intervalle de confiance à 95 % la fréquence de consommation de psychotropes chez les adolescents de ce pays.
3. Même questions chez les filles, puis les garçons.
4. Quelle devrait être la taille de l'échantillon pour que la marge d'erreur à 95 % dans l'estimation de la fréquence de consommation de psychotropes chez les adolescents de ce pays soit inférieure à 1 %, en supposant que la fréquence observée de consommation de psychotropes n'est pas modifiée.

SOLUTION

$$1) p = f = \frac{570}{3000}$$

$$2) \bar{I}_2 = \left[f - 1,96 \sqrt{\frac{f(1-f)}{n-1}} ; f + 1,96 \sqrt{\frac{f(1-f)}{n-1}} \right]$$

$$n = 3000$$

$$3) 2.22.22$$

TESTS D'HYPOTHÈSES

- Référence : [Tutoriel Tests d'hypothèses - Tout simplement expliqué - DATAtab](#)