

Probabilités et Statistiques pour la Science des données

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

1ère séance, 2018/2019 - INPT

Introduction

- Les probabilités permettent d'étudier des phénomènes aléatoires.
- Les statistiques sont aujourd'hui utilisées dans presque tous les domaines.
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- L'objectif du cours est de développer chez les étudiants les compétences de base en statistiques comme porte d'entrée du monde de Big Data et Data Science.
- Le cours sera orienté vers des applications sur des logiciels tels que R et Python.

Plan

1 Partie 1.

- Rappels de probabilités.
- Statistiques descriptives.
- Echantillonnage, Estimation d'un paramètre, Estimation par intervalle de confiance.
- Tests paramétriques et non paramétriques.
- Test du χ^2 .
- Analyse de la variance à un facteur.
- Régression linéaire simple.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

2 Partie 2.

- Théorie de la décision, statistique fréquentiste et bayésienne
- Estimation, statistique exhaustive, information de Fisher et maximum de vraisemblance
- Éléments de statistique asymptotique, intervalles de confiance
- Théorie des tests

Expérience aléatoire, univers

Définition (Expérience aléatoire, univers)

- *On appelle expérience aléatoire une expérience sur un système dont le résultat n'est pas connu d'avance et peut varier si on répète cette expérience.*
- *On appelle univers l'ensemble des résultats possibles. Il est noté Ω .*

Exemple : [Jeter une pièce de monnaie, (P, F)], [lancer des dés, (1;6)], prélever des boules dans une urne...

Evènement, Terminologie

Définition (Evènement)

Un sous ensemble de Ω est appelé évènement. On note par A l'ensemble des évènements relatifs à l'expérience aléatoire.

Terminologie probabiliste:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Evènement certain	: Ω
Evènement impossible	: $\emptyset$
Evènement élémentaire	: $\{w\}$
Evènement contraire à A	: $\bar{A}$
A ou B	: $A \cup B$
A et B	: $A \cap B$
A implique B	: $A \subset B$
A et B incompatibles	: $A \cap B = \emptyset$
w réalise A	: $w \in A$

Notion de probabilité

Définition (Probabilité)

A chaque évènement A un poids $P(A)$ indiquant sa chance d'être réalisé si l'on effectue l'expérience aléatoire.

Définition (Probabilité (Formulation mathématique))

Étant donné un espace probabilisable $(\Omega, \mathcal{A})$, on appelle probabilité sur $(\Omega; \mathcal{A})$ toute application $P : \mathcal{A} \rightarrow \mathbb{R}$ satisfaisant aux trois propriétés suivantes:

- $\forall A \in \mathcal{A}, P(A) \geq 0$
- $P(\Omega) = 1$
- Pour toute suite (A_n) d'éléments deux à deux disjoints,

$$P(\cup_n A_n) = \sum_{n=0}^{+\infty} P(A_n)$$

Notion de probabilité

- Cas d'équiprobabilité: (Ω fini de cardinal n .)

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\forall A \in \mathcal{A}, P(A) = \frac{\text{Card}(A)}{n} \quad (1)$$

- La relation (1) est fausse dans le cas général.

Notion de probabilité

Proposition

Toute probabilité P possède les propriétés suivantes:

- 1 $P(\emptyset) = 0$
- 2 $\forall (A, B) \in \mathcal{A}^2, A \subset B \Rightarrow P(A) \leq P(B)$
- 3 *Formule de crible de Poincaré:* $\forall (A_1, A_2, \dots, A_n) \in \mathcal{A}^n$,

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\begin{aligned}
 P(\cup_{i=1}^n A_i) &= \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i_1 \leq i_2 \leq n} P(A_{i_1} \cap A_{i_2}) \\
 &+ \dots \\
 &+ (-1)^{k-1} \sum_{1 \leq i_1 \leq i_2 \leq \dots \leq i_k \leq n} P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) \\
 &+ (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)
 \end{aligned}$$

Probabilité conditionnelle

Définition

Soit B un évènement de probabilité non nulle. On appelle probabilité conditionnelle à B , ou probabilité sachant B , associée à P l'application:

$$P_B : \begin{cases} \mathcal{A} \rightarrow \mathbb{R} \\ A \rightarrow \frac{P(A \cap B)}{P(B)} \end{cases}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

On la note aussi $P(A/B)$.

⇒ La probabilité conditionnelle P_B est une probabilité satisfaisant aux trois propriétés déjà mentionnée.

⇒ Cas d'équiprobabilité: est ce que la probabilité P_B est uniforme?

⇒ Formule de probabilité enchainée:

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_n) &= P(A_1/A_2 \cap \dots \cap A_n)P(A_2/A_3 \cap \dots \cap A_n) \\ &\dots P(A_{n-1}/A_n)P(A_n) \end{aligned}$$

Formule de probabilité totale, Formule de Bayes

Définition

Soit $(B_k)_{1 \leq k \leq N}$ une partition de Ω telle que $\forall k, P(B_k) \geq 0$. Alors on a la formule de probabilité totale suivante:

$$\forall A, P(A) = \sum_{1 \leq k \leq N} P(A \cap B_k) = \sum_{1 \leq k \leq N} P(A/B_k)P(B_k)$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

⇒ Système complet $\{A_1; \bar{A}_1\}$, la formule de probabilité totale:

$$P(A) = P(A/A_1)P(A_1) + P(A/\bar{A}_1)P(\bar{A}_1)$$

⇒ Formule de Bayes:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Exercice

- **Exemple 1:** On considère une urne U_1 contenant deux boules blanches et une boule noire, et une urne U_2 contenant une boule blanche et une boule noire. On choisit une urne au hasard puis on prélève une boule dans cette urne. Les boules sont indiscernables au toucher.
Document not to be used for teaching. All rights reserved by the author. Dr. Mohamed ASSELLAOUI
 - 1 Quelle est la probabilité de tirer une boule blanche?
 - 2 Si la boule tirée est blanche, quelle est la probabilité que la boule soit extraite de l'urne U_1 ?

Indépendance des évènements

Définition

A est indépendant de B si $P(A/B) = P(A)$ ou si $P(B) = 0$.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

⇒ A est indépendant de B si et seulement si $P(A \cap B) = P(B)P(A)$.

⇒ Cas général: la famille $(A_k)_{k \in I}$ est dite famille d'évènements mutuellement indépendants si pour tout $J \subset I$, on a:

$$P(\cup_{j \in J} A_j) = \prod_{j \in J} P(A_j)$$

Fonction de répartition

Définition

- Soit $(\Omega; \mathcal{A}; P)$ un espace probabilisable. On appelle variable aléatoire une application X de l'univers Ω dans $\mathbb{R}$ dans la valeur $X(\omega)$ dépend du résultat obtenu de l'expérience aléatoire.
- La loi de probabilité sur les sous ensembles de $X(\Omega)$ est appelée loi de probabilité de la v.a X .
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- On appelle la fonction de répartition de X l'application:

$$F_X : \begin{cases} \mathbb{R} \rightarrow [0, 1] \\ t \rightarrow P(X \leq t) \end{cases}$$

$\Rightarrow F_X$ est croissante.

$\Rightarrow \lim_{t \rightarrow -\infty} F_X(t) = 0, \quad \lim_{t \rightarrow +\infty} F_X(t) = 1$

$\Rightarrow F_X$ est continue à droite en tout point de $\mathbb{R}$.

Types de v.a

Définition

- La v.a X est dite discrète si $\sum_{t \in \mathcal{X}(\Omega)} P(X = t) = 1$ (autrement dit si $X(\Omega)$ est fini ou dénombrable).
- La v.a X est dite continue si sa loi de probabilités est définie par une densité notée f_X telle que $\int_{-\infty}^{+\infty} f_X(t) dt = 1$ et $F_X(x) = \int_{-\infty}^x f_X(t) dt$.

Espérance et variance

- Le comportement moyen d'une v.a est donnée par l'espérance ou :
 - variable aléatoire discrète: $\mathbb{E}(X) = \sum_{x \in X(\Omega)} xP(X = x)$
 - variable aléatoire continue: $\mathbb{E}(X) = \int_{-\infty}^{+\infty} tf_X(t)dt$
- la dispersion de la variable autour de l'espérance est caractérisée par la variance donnée par:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

- Quelques propriétés de Var et $\mathbb{E}$:

$$\begin{aligned} Var(X) &\geq 0 \\ Var(aX + b) &= Var(aX) = a^2 Var(X) \\ Z &= \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}} \text{ est une variable centrée réduite.} \end{aligned}$$

Lois marginales, indépendance

Définition (Loi du couple)

Soient X et Y deux variables aléatoires. Dans le cas discret, la loi du couple (X, Y) est la loi de probabilité qui permet de lister l'ensemble des valeurs $\mathbb{P}(X = x, Y = y)$ pour tous les couple (x, y) . Dans le cas continu, la loi du couple (X, Y) revient à définir $f_{X,Y}(x, y)$ pour calculer $\mathbb{P}(X \in I, Y \in J)$ pour tout couple d'intervalles (I, J) . Les lois marginales de X et Y sont les lois de X et de Y .

Définition (Indépendance des v.a)

Les deux v.a X et Y sont dites indépendantes si la loi du couple est le produit des loi marginales, i.e,

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

dans le cas continue, on a $f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$

Covariance et corrélation

Définition (Covariance et corrélation)

Le covariance et la corrélation des v.a X et Y sont données par les quantités suivantes:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOUI.

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

- Cas particulier de X et Y indépendantes, on a $\text{Cov}(X, Y) = \text{Cor}(X, Y) = 0$;
- $\text{Cov}(X, X) = \text{Var}(X)$, et $\text{Cor}(X, X) = 1$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Loi normale CR et loi normale

Définition (Loi normale centrée réduite)

On dit qu'une v.a X suit la loi normale centrée réduite ($X \sim \mathcal{N}(0, 1)$) s'il a une densité f_X définie par:

$$\forall t \in \mathbb{R}, f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\triangleright \mathbb{E}(X) = 0, \quad \text{Var}(X) = 1.$$

Définition (Loi normale)

On dit qu'une v.a X suit la loi normale ($X \sim \mathcal{N}(\mu, \sigma^2)$) s'il a une densité f_X définie par:

$$\forall t \in \mathbb{R}, f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

Loi normale CR et loi normale

1 $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1).$

2 Si $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sont indépendantes, alors
 $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$

3 La gaussienne modélise plusieurs situation réelles.

Document not to be used for teaching. All rights reserved by the author (Dr. Mohamed ASSELJAOUI)

Loi des grands nombres et TCL

Théorème (Loi des grands nombres)

Soit $S_n = \sum_{i \leq n} X_i$ où (X_n) est une suite de v.a indépendantes de même loi d'espérance m et de variance σ^2 , alors presque surement, on a

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\lim_{n \rightarrow +\infty} \frac{S_n}{n} = m$$

Théorème (Théorème Central Limite)

Le loi de $\frac{S_n - nm}{\sqrt{n\sigma}}$ tend vers une v.a $\bar{X} \sim \mathcal{N}(0, 1)$.

Loi de Bernoulli

- ▷ X est suit la loi de Bernoulli si $X(\Omega) = \{0, 1\}$ et $P(X = 1) = p$ et $P(X = 0) = 1 - p$.
- ▷ X peut être interprétée comme l'indicatrice du succès.
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- ▷ On note $X \sim \mathcal{B}(p)$.
- ▷ $\mathbb{E}(X) = p$ et $\text{Var}(X) = p(1 - p)$.
- ▷ Le calcul des moments de X sont facilités par le fait que $X^n = X, \forall n \geq 1$

Loi Binomiale

- ▷ X suit la loi Binomiale si $X(\Omega) = [0; n]$ et

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \forall k \in [0; n] .$$
- ▷ X peut être interprétée une répétition de n fois d'une v qui suit la loi de Bernoulli.
 Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- ▷ On note $X \sim \mathcal{B}(n; p)$.
- ▷ $\mathbb{E}(X) = np$ et $\text{Var}(X) = np(1 - p)$.
- ▷ Soit (X_n) une suite de v.a indépendantes qui suivent toutes une loi de Bernoulli de paramètre p . Alors leur somme $Z = \sum_{i=1}^n X_i$ suit la loi binomiale $\mathcal{B}(n; p)$.

Loi Binomiale

- Exercice 2: Soit une urne contenant 8 boules noires et 5 boules rouges indiscernables au toucher. On procède à 10 extractions successives d'une boule, avec remise. Quelle est la loi du nombre de boules rouges obtenues ?

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSALLAOUI.

Loi de Poisson

- ▷ X suit la loi de Poisson de paramètre λ si $X(\Omega) = \mathbb{N}$ et
 $\forall k \in \mathbb{N}, P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$
- ▷ On note $X \sim \mathcal{P}(\lambda)$ Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- ▷ La loi de Poisson peut être interprétée comme la loi limite d'une loi binomiale $\mathcal{B}(n; \frac{\lambda}{n})$
- ▷ $\mathbb{E}(X) = \lambda, \text{Var}(X) = \lambda$

Loi de Poisson

- ▷ Exercice: La variable aléatoire X suit la loi de Poisson de paramètre λ . On pose

$$Y = \frac{1}{(X+1)(X+2)}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Calculer l'espérance de Y .

- ▷ Solution: On a $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ et on a :

$$\mathbb{E}(Y) = \sum_{k=0}^{+\infty} \frac{1}{(k+1)(k+2)} e^{-\lambda} \frac{\lambda^k}{k!} = \frac{e^{-\lambda}}{\lambda^2} (e^{\lambda} - 1 - \lambda)$$

Loi géométrique

- ▷ X suit la loi de Poisson de paramètre p si $X(\Omega) = \mathbb{N}^*$ et $P(X = k) = p(1 - p)^{k-1}$
- ▷ On note $X \sim \mathcal{G}(p)$ Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- ▷ X désigne le nombre de répétitions d'une expérience de Bernoulli nécessaires pour obtenir un succès.
- ▷ $\mathbb{E}(X) = \frac{1}{p}$, $\text{Var}(X) = \frac{1-p}{p^2}$.

Loi uniforme discrète

- ▷ X suit la loi uniforme sur $\{1; \dots; n\}$ si $\forall i \in \{1; \dots; n\}, P(X = i) = \frac{1}{n}$
- ▷ On note $X \sim \mathcal{U}(\{1; \dots; n\})$.
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.
- ▷ Il s'agit d'une loi dont tous les poids de probabilité sont identiques.
- ▷ $\mathbb{E}(X) = \frac{n+1}{2}, \text{Var}(X) = \frac{n^2-1}{12}$.

Loi uniforme discrète

- ▷ Exercice: Une personne souhaitant rentrer chez elle a un trousseau de n clefs. Déterminer le nombre moyen d'essais nécessaires dans chacun des deux cas suivants:

1 Cas 1: La personne élimine après chaque essai la clef qui n'a pas convenu.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

2 Cas 2: La personne remet dans le trousseau après chaque essai la clef qui n'a pas convenu.

■ Solution:

1 Loi uniforme: $E(X) = \frac{n+1}{2}$

2 Loi géométrique: $E(X) = n$

Loi hypergéométrique

- ▷ X suit la loi hypergéométrique de paramètres $(N; n; p)$ si

$X(\Omega) = [\max(0, n - N(1 - p)); \min(n, Np)]$ si

$$\forall i \in [1; n], P(X = i) = \frac{\binom{Np}{n} \binom{N(1-p)}{n-k}}{\binom{N}{n}}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOUI.

- ▷ On note $X \sim \mathcal{H}(N; n; p)$
- ▷ Il s'agit d'une loi dont tous les poids de probabilité sont identiques.
- ▷ $\mathbb{E}(X) = np, \text{ Var}(X) = \frac{N-n}{N-1} np(1-p).$

Loi uniforme continue

- ▷ X suit la loi uniforme continue sur $[a, b]$ si elle a pour densité f définie par

$$f(x) = \begin{cases} 0 & \text{si } x \notin [a; b] \\ \frac{1}{b-a} & \text{si } x \in [a; b] \end{cases}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- ▷ On note $X \sim \mathcal{U}_{[a,b]}$
- ▷ $\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$

Loi uniforme continue

- ▷ Exercice: Soient (X_n) une suite de v.a indépendantes qui suivent toutes la loi uniforme sur $[0; 1]$. Déterminer la loi de $Y = \max_{1 \leq i \leq n} \{X_i\}$.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELAOUI.

Loi exponentielle

- ▷ X suit la loi exponentielle de paramètre λ si elle a pour densité:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ \lambda e^{-\lambda x} & \text{si } x \geq 0 \end{cases}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- ▷ On note $X \sim \mathcal{E}(\lambda)$.
- ▷ $\mathbb{E}(X) = \frac{1}{\lambda}$, $\text{Var}(X) = \frac{1}{\lambda^2}$.
- ▷ Montrer que X suit une loi exponentielle si et seulement si

$$\forall s \in \mathbb{R}, \forall t \geq 0, P(X > s + t | X > t) = P(X > s)$$

Plan

1 Partie 1.

- Rappels de probabilités.
- Statistiques descriptives.
- Echantillonnage, Estimation d'un paramètre, Estimation par intervalle de confiance.
- Tests paramétriques et non paramétriques.
- Test du χ^2 .
- Analyse de la variance à un facteur.
- Régression linéaire simple.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

2 Partie 2.

- Théorie de la décision, statistique fréquentiste et bayésienne
- Estimation, statistique exhaustive, information de Fisher et maximum de vraisemblance
- Éléments de statistique asymptotique, intervalles de confiance
- Théorie des tests

Statistique descriptive

Statistique descriptive :

1 Statistique descriptive univariée (une dimension)

- Vocabulaire de la statistique
- Représentations d'une série statistique

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

2 Statistique descriptive bivariée (deux dimensions)

- Présentation des données
- Distributions conjointes, marginales et conditionnelle
- Indépendance statistique

Vocabulaire de la statistique

- ★ **Les statistiques** désignent des grandeurs que l'on calcule, ou que l'on est capable de calculer, sur un ensemble de données observée.
- ★ **La statistique** désigne à la fois un ensemble de données observées et les méthodes de recueil, de traitement et d'analyse de celles-ci.
- ★ **Une population** désigne tout objet statistique étudié. Elle est composé d'**individus** appelés **unités statistiques**. Lorsque la population est trop importante pour être connue entièrement, on prélève un **échantillon représentatif**.
- ★ **La statistique** étudie les caractéristiques des individus dans le but de décrire la population.
- ★ **Les caractères** qui caractérise des individus d'une population objet de l'étude statistique sont aussi appelés **variables**.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSFI, AOU.

Catégories des variables statistiques

- ✓ Les variables peuvent être **qualitatives** ou **quantitatives**;
- ✓ Les variables ont différentes situations possibles appelées **modalités**;
- ✓ L'individu ne peut avoir qu'une seule modalité d'une variable;
- ✓ Les variables qualitatives se présentent sous deux catégories:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

1 **nominales** lorsque leurs modalités ne peuvent pas être ordonnées.

Exemple: Catégorie socioprofessionnelle, situation matrimoniale...;

2 **ordinales** lorsque leurs modalités peuvent être ordonnées.

Exemple: niveau d'anglais, appréciation d'un étudiant...

Catégories des variables statistiques

✓ Les variables quantitatives sont des variables numériques. Elles se présentent à leur tour en deux catégories:

- 1 Les variables quantitatives dont le nombre de modalités est fini ou dénombrable appelées **variables discrètes**.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Exemple: Nombre de salles de cours, Nombre d'enfants ...;

- 2 Les variables dont les modalités sont infinies appelées **variables continues**.

Exemple: Poids, taille, le revenu mensuel ...;

Exemples de variables

Table: Exemple de caractère nominal

P	Population	Effectif total=36
i	unités statistiques	Chaque étudiant $i \in \{1, \dots, n\}$
X	Caractère	Le sexe
X_f, X_m	Modalités	Féminin ou masculin
n_f, n_m	Effectif associé à chaque modalité	16 hommes, 20 femmes

Table: Exemple de caractère quantitatif

E	Echantillon	Effectif total=36
i	unités statistiques	Chaque étudiant $i \in \{1, \dots, n\}$
X	Caractère	La note à l'examen
$\{x_f, \dots, x_N\}$	Valeurs	6, 9, 10, 11, 12, 14, 16, 17.5, 19
$\{n_f, \dots, n_N\}$	Effectif associé à chaque valeur	1, 1, 3, 4, 10, 5, 5, 4, 3

Représentation des données, Notations

- Pour chaque valeur ou modalité x_i de la variable, on note n_i le nombre d'effectif de x_i avec $\sum_i n_i = n$. La fréquence correspondante $f_i = \frac{n_i}{n}$ et le pourcentage $100f_i$.
- Pour les variables qualitatives nominales, on définit le tableau statistique suivant:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

x_i	n_i	f_i

- Pour les variables discrètes ou les variables qualitatives ordinales, on définit également les effectifs cumulés $N_i = \sum_{j=1}^i n_j$ et les fréquences cumulées $F_i = \sum_{j=1}^i f_j$ (interprétée comme la proportion des individus pour lesquels $X < n_{i+1}$). Le tableau statistique correspondant:

x_i	n_i	N_i	f_i	F_i

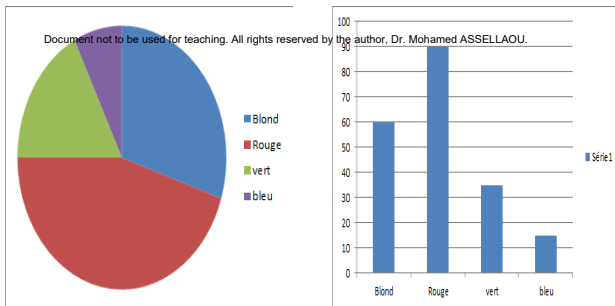
Représentation des données, Notations

- Les variables continues sont généralement regroupés en k classes notées $[a_i, a_{i+1}]$.
- Pour chaque classe, on note c_i le centre de la classe, d_i l'amplitude de l'intervalle, n_i l'effectif, f_i la fréquence ainsi que les fréquences cumulées $F_i = \sum_{j=1}^i f_j$ (interprétée comme la proportion des individus pour lesquels $X < a_{i+1}$). Le tableau statistique correspondant:

$[a_i, a_{i+1}[$	$c_i = \frac{a_{i-1} + a_i}{2}$	$d_i = a_{i+1} - a_i$	n_i	N_i	f_i	F_i

Représentation graphique

- Les variables qualitatives nominales se présentent graphiquement sous forme de :
 - 1 Camembert (diagramme en secteur).
 - 2 Diagramme en barre des effectifs. **Exemple:** couleurs des individus: Blond= 60, Rouge= 90, Vert= 35, Bleu= 15:



- Les variables qualitatives ordinales et quantitatives discrètes se présentent graphiquement sous forme de:
 - 1 Camembert (diagramme en secteur).
 - 2 Diagramme en barre des effectifs, des fréquences ou des fréquences cumulées.

Exemple: Le niveau d'anglais d'une population de 40 personnes se présente comme suit: Effectif de $A_1 = 10$, $A_2 = 27$, $B_1 = 10$, $B_2 = 20$, $C_1 = 15$, $C_2 = 13$.

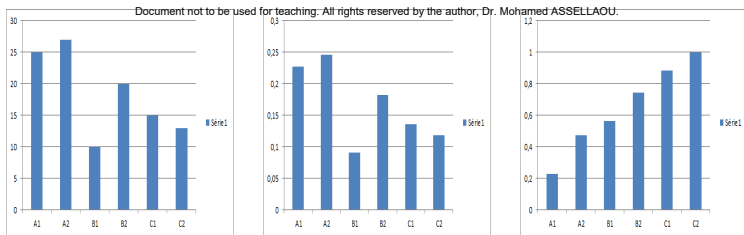


Figure: Diagrammes en barre des effectifs, des fréquences et des fréquences cumulées.

Représentation graphique: Variable continue

- Les variables quantitatives continues se présentent graphiquement en:

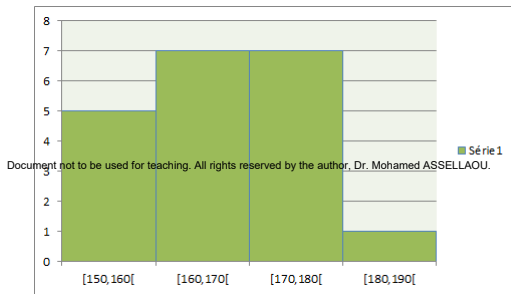
- 1 Histogramme en effectif ou en fréquence;
- 1 Diagramme des fréquences cumulées;
- 2 Boîte à moustache.

Exemple : la taille des individus en cm. 157 - 158 - 158 - 158 - 159 - 159 - 161 - 161 - 162 - 164 - 167 - 168 - 169 - 171 - 171 - 171 - 174 - 175 - 178 - 179 - 186 . Le tableau statistique correspond est:

$[a_i, a_{i+1}[$	c_i	d_i	n_i	N_i	f_i	F_i
$[150, 160[$	155	10	5	5	0.25	0.25
$[160, 170[$	165	10	7	12	0.35	0.60
$[170, 180[$	175	10	7	19	0.35	0.95
$[180, 190[$	185	10	1	20	0.05	1.00

Représentation graphique: Variable continue

■ Histogramme en effectif:



- La boîte à moustache: la partie centrale de la boîte est constituée d'un rectangle dont la longueur est la **distance interquartile** $|Q_3 - Q_1|$. Les moustaches sont des segments qui s'étendent de part et d'autre de la boîte jusqu'au **premier décile** D_1 pour la moustache inférieure, jusqu'au **dernier décile** D_9 pour la moustache supérieure.

Résumés numériques (indicateurs)

- Plusieurs indicateurs typiques permettent de résumer une série statistique.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- 1 Certains indicateurs caractérisent **la tendance centrale** comme **la moyenne**, **la médiane**, **le mode** et
- 2 D'autres indicateurs caractérisent **la dispersion** comme **la variance**, **l'écart type**, **l'étendue**, **1^{er} quartile**, **3^{ème} quartile** ...

Indicateurs de tendance centrale

- La moyenne d'un échantillon de n observations d'une population noté $\{x_i\}_{i \leq n}$ est donnée par la formule suivante:

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

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Le mode qui est égal à une donnée de l'échantillon désigne le point milieu de la classe ayant le plus grand effectif.
- La médiane de l'échantillon précédent est donnée par la formule suivante:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{si } n \text{ est impair.} \\ \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{si } n \text{ est pair.} \end{cases}$$

Calcul de la moyenne arithmétique

- **Exemple 1:** Soit la série de chiffres suivants (8, 5, 9, 13, 25). La moyenne arithmétique correspondante est:

$$\bar{x} = \frac{8 + 5 + 9 + 13 + 25}{5} = 12$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- **Exemple 2:** Maintenant on considère que certains chiffres de la série ci-dessus ont un effectif > 1 :

x_i	5	8	9	13	25
n_i	2	2	2	3	1

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n n_i x_i = \frac{5 * 2 + 8 * 2 + 9 * 2 + 13 * 3 + 25 * 1}{10} = 10.8$$

Calcul d'autres types de moyenne

- **Moyenne quadratique:** La moyenne quadratique d'une série statistique $\{x_i\}$ associé aux effectifs $\{n_i\}$ avec $i \in \{1, ..n\}$:

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n (n_i x_i^2)}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- **Moyenne géométrique:** Elle correspond à la valeur:

$$Q = \left[\prod_{i=1}^n (x_i^{n_i}) \right]^{\frac{1}{n}}$$

- **Moyenne harmonique:** Elle correspond à la valeur:

$$H = \frac{n}{\sum_{i=1}^n \frac{n_i}{x_i}}$$

Calcul de la médiane

- **Exemple 1:** Quelle est la médiane des séries ci-dessous:

1 (3, 4, 7, 8, 9)?

2 (3, 4, 7, 8, 8.5, 9)?

- **Exemple 2:** Effectifs groupés par valeurs: Quelle est la médiane de la série ci-dessous:

x_i	n_i	N_i	f_i	F_i
2	2	2	0.066	0.066
8	3	5	0.1	0.166
9	4	9	0.133	0.3
10	4	13	0.133	0.433
11	5	18	0.167	0.6
12	3	21	0.1	0.7
13	6	27	0.2	0.9
15	1	28	0.033	0.933
18	2	30	0.066	1

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSEL LAOU.

Calcul de la médiane

- **Exemple 3:** Effectifs groupés par classe: Quelle est la médiane de la série ci-dessous:

$[a_i, a_{i+1}[$	n_i	N_i
$[0, 5[$	2	2
$[5, 10[$	7	9
$[10, 15[$	18	27
$[15, 20[$	3	30

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$M_e = a_i^* + d_i \left[\frac{\frac{n}{2} - N_{i-1}}{n_i} \right]$$

Calcul du mode

■ **Exemple 1:** Quelle est le mode des séries suivantes:

1 (8, 7, 4, 5, 6)

2 (8, 5, 5, 5, 5, 4, 4, 4, 7, 7, 6, 6, 6, 6, 6)

Document not to be used for teaching. All rights reserved by the author Dr. Mohamed ASSELLAOU.

$[a_i, a_{i+1}[$	n_i	N_i
$[0, 5[$	2	2
$[5, 10[$	7	9
$[10, 15[$	18	27
$[15, 20[$	3	30

$$\text{Mode} = a_i^* + d_i \frac{n_i - n_{i-1}}{2n_i - n_{i+1} - n_{i-1}}$$

Calcul du mode

■ Exemple 1: Amplitudes inégales:

x_j	n_j	N_j	$\frac{n_j}{d_j}$
$[0, 10[$	9	9	0.9
$[10, 12[$	9	18	4.5
$[12, 20[$	12	30	1.5

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\text{Mode} = a_j^* + d_j \frac{\frac{n_j}{d_j} - \frac{n_{j-1}}{d_{j-1}}}{2 \frac{n_j}{d_j} - \frac{n_{j+1}}{d_{j+1}} - \frac{n_{j-1}}{d_{j-1}}}$$

Formes d'une distribution à l'aide des indicateurs de tendance centrale

- Distribution parfaitement symétrique: Moyenne=Médiane= Mode.

Exemple:

x_j	1	2	3	4	5
n_j	2	4	5	4	2

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOUI.

- Distribution étalée à droite: Moyenne>Médiane > Mode. Exemple:

x_j	1	2	3	4	5
n_j	10	8	6	4	2

- Distribution étalée à gauche: Moyenne<Médiane< Mode. Exemple:

x_j	1	2	3	4	5
n_j	2	4	6	8	10

Indicateurs de dispersion

- L'étendue de l'échantillon est égale à la différence entre la valeur maximale et la valeur minimale des données x_i
- L'écart interquartile est:

$$IQT = Q_3 - Q_1$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

où le quartile Q_1 et le quartile Q_3 représentent respectivement les médianes des parties inférieure et supérieure obtenues par la médiane de l'échantillon.

- Les quantiles permettant de diviser la population en $p \geq 2$ sous-populations d'effectifs égaux. Par exemple, les déciles d'une série statistique partagent la série en dix parties de même effectif. En pratique, seuls les premier et dernier déciles, respectivement notés D_1 et D_9 , sont utilisés.

Calcul de l'étendue et des quartiles

- **Exemple:** Trouver l'étendue des deux séries correspondant au x notes de deux élèves A et B et en déduire le rapport de dispersion des notes des deux élèves.

A	8	9	10	11	12
B	2	4	10	16	18

- **Exemple:** Calculer les quartiles des trois séries ci-dessous:

x_j	1	3	4	6	7	9	11	12	14	15	16	17	19	20
-------	---	---	---	---	---	---	----	----	----	----	----	----	----	----

x_j	1	3	4	6	7	9	11	12	13	14	15	16	17	20
-------	---	---	---	---	---	---	----	----	----	----	----	----	----	----

n_j	3	1	2	3	3	2	3	1	2	1	2	2	1	1
-------	---	---	---	---	---	---	---	---	---	---	---	---	---	---

$[a_i, a_{i+1}[$	$[0, 4[$	$[4, 8[$	$[8, 12[$	$[12, 16[$	$[16, 20[$
n_j	4	8	5	6	4

Exemple

- **Exemple** le tableau suivant regroupant les données relatives aux salaires annuels nets dans l'industrie par catégorie socio-professionnelle.

	Cadres	Professions intermédiaires	Employés	Ouvriers
D_1	28 900	19 370	13 380	14 650
Q_1	35 560	22 810	15 400	16 710
$\tilde{x}$	43 910	27 180	18 960	19 620
Q_3	56 970	32 710	23 360	23 420
D_9	76 870	39 210	28 540	27 920
$\bar{x}$	50 600	28 670	20 310	20 780

Exemple: Boite à moustache

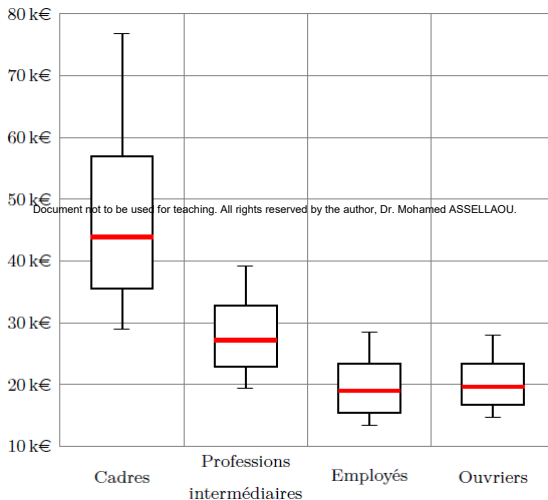


Figure: Boite à moustache pour les données relatives aux salaires annuels nets.

Indicateurs de dispersion

- La variance de la population dénotée σ^2 est définie par:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 \right) - \bar{x}^2$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Le coefficient de variation mesure la dispersion relative des données autour de la moyenne:

$$CV = \frac{\sigma}{\bar{x}}$$

Indicateurs de dispersion

- **Exemple 1:** Calculez la variance et le coefficient de variation des séries suivantes:

x_i	2	5	7	1	9	13	6	15	8	16
-------	---	---	---	---	---	----	---	----	---	----

x_i	2	6	9	11	15
n_i	5	9	4	3	5

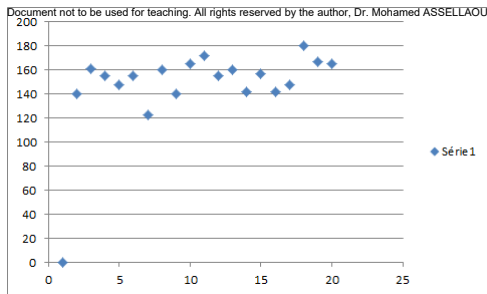
- **Exercice :** On connaît les salaires mensuels nets bruts des 200 employés de la même entreprise, à 10 ans d'intervalle. Les données sont regroupées par classe. Le nombre d'employés est passé de 200 à 1994 à 280 en 2004. Est ce que la dispersion des salaires a augmenté?

Salaires	Effectifs 1994	Effectifs 2004
1000-2000	40	56
2000-3000	70	118
3000-4000	80	92
4000-5000	5	10
5000-10000	5	4

Tableaux et graphiques

- Séries quantitatives connues individuellement: **Exemple:** Est ce qu'il ya une corrélation entre les données de la série suivantes:

$\{(140, 38.2), (161, 44.3), (155, 46.1), (148, 38.2), (155, 50.5)$
 $, (123, 22.4), (160, 40.4), (140, 34.7), (165, 50.5), (172, 50.5), (155, 38.1)$
 $, (160, 57.3), (142, 39.3), (157, 46.1), (142, 37.1), (148, 45.9), (180, 66.3)$
 $, (167, 60), (165, 50.5)\}$



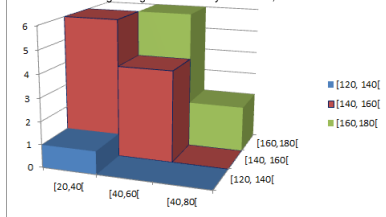
⇒ Droite de "tendance".

Tableaux et graphiques

■ Séries quantitatives groupées

x, y	$[20,40[$	$[40,60[$	$[40,80[$
$[120, 140[$	1	0	0
$[140, 160[$	6	4	0
$[160, 180[$	0	6	2

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.



Graphique et graphiques

■ Séries qualitatives: Exemple:

Sexe, Statut	Actifs occupés	Chomeurs	Inactifs
Masculin	5	3	1
Féminin	4	3	4

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

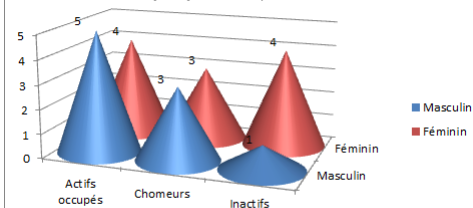


Tableau de contingence, distribution marginales

- Le tableau de contingence de deux séries statistiques se présente sous la forme:

x, y	y_1	y_2	$\dots$	y_j	$\dots$	y_q	$n_{i.} = \sum_j n_{ij}$
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2q}	$n_{2.}$
$\dots$	Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.						
x_p	n_{p1}	n_{p2}	$\dots$	n_{pj}	$\dots$	n_{pq}	$n_{p.}$
$n_{.j} = \sum_i n_{ij}$	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.j}$	$\dots$	$n_{.q}$	$n_{..} = \sum_{ij} n_{ij}$

- ✓ n_{ij} représente l'effectif des individus ayant la modalité i de x et la modalité j de y .
- ✓ $n_{i.}$ représente tout l'effectif des individus ayant la modalité i de x
- ✓ $n_{.i}$ représente tout l'effectif des individus ayant la modalité i de y
- ✓ Les fréquences marginales se calculent en divisant les effectifs marginaux sur la somme des effectifs.

Moyenne et variance marginales

- Les moyennes marginales de x et de y sont données par les formules:

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^p n_{i.} x_i$$

$$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^q n_{.j} y_j$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Les variances marginales de x et de y sont données par les formules:

$$\sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^p n_{i.} x_i^2 - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^q n_{.j} y_j^2 - \bar{y}^2$$

Moyennes et variances conditionnelles

- Les distributions conditionnelles s'obtiennent par la fixation d'une valeur des deux variables.
- Pour chaque colonne i (valeur i de y) on fait ressortir la moyenne conditionnelle correspondante $\bar{x}_i$ et de même pour la moyenne conditionnelle $\bar{y}_j$ obtenue en fixant la valeur de la ligne j de x .
- Les variances sont également correspondante à $\bar{x}_i$ et $\bar{y}_j$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Moyenne et variance marginales

- **Exemple:** Soit le tableau de contingence suivant:

x, y	1	4	$n_{i.}$
2	3	5	8
8	4	12	16
$n_{.j}$	7	17	24

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- ✓ Calculer les moyennes et variances marginales et conditionnelles de x et de y ?

Indépendance

- On définit les fréquences relatives de la modalité (i, j) :

$$f_{ij} = \frac{n_{ij}}{n}$$

- L'indépendance entre les variables se traduit de manière simple par l'égalité concernant les fréquences relatives:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$f_{ij} = f_{i.} \times f_{.j} \quad (3)$$

- l'équation (3) se traduit en termes des effectifs comme suit:

$$n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

Mesure d^2

- On définit la mesure de liaison d^2 entre deux variables nominales comme suit :

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$d^2 := \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

- Si les deux variables nominales sont indépendantes alors d^2 est nul;

Deux variables ayant deux modalités

- Soient X et Y deux variables de modalités respectives: x_1, x_2, y_1, y_2 et d'effectifs:

$X \setminus Y$	y_1	y_2
x_1	a	b
x_2	c	d

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Le coefficient d^2 se définit par l'expression suivante:

$$d^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Rapport de corrélation empirique

- Le rapport de corrélation empirique entre la variable qualitative X ayant k modalités d'effectifs $n_1, n_2, \dots, n_k$ et la variable quantitative Y se définit par la relation suivante:

$$e^2 = \frac{\left(\frac{1}{n} \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \right)}{\sigma_Y^2}$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

où $\bar{y}_i$ est la valeur moyenne de y pour la modalité i de X

$$\sigma_Y^2 = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_i^j - \bar{y})^2$$

- $e^2 \in [0, 1]$
- $e^2 = 0$ signifie qu'il ya pas de dépendance en moyenne
- $e^2 = 1$, pour une modalité i de X , tous les individus ont la même valeur et ceci pour toutes les valeurs de l'indice.

Corrélation entre deux variables quantitatives

- Le coefficient de corrélation linéaire r entre deux variables quantitatives mesure le caractère linéaire entre ces deux variables:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\left(\frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y} \right)}{\sigma_X \sigma_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

- Si $|r| = 1$, il y a une relation linéaire parfaite entre les deux variables
- Si les variables sont indépendantes, r est nul.
- Si r est nul, les deux variables sont indépendantes **linéairement**.

Construction de la droite de régression linéaire

- La droite de régression passe par le point moyen (point de coordonnées $(\bar{x}, \bar{y})$) et qui permet de minimiser la somme des carrés des écarts des observations.

- La droite de la régression linéaire a la forme suivante:

$$y = ax + b, \quad \text{où } a = r \frac{\sigma_Y}{\sigma_X}, \quad b = \bar{y} - a\bar{x}$$

- La droite de régression linéaire permet de faire des prévisions.

Théorie d'échantillonnage

- **Prélèvement d'un échantillon représentatif de la population ou échantillon aléatoire** par des techniques appropriées. Cela relève de la théorie de l'échantillonnage.
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOUI.
- **Etude des caractéristiques de cet échantillon**, issu d'une population dont on connaît la loi de probabilité.

Théorie d'échantillonnage

Définition (Echantillon)

Un échantillon aléatoire est un n -uplet $(X_1, X_2, \dots, X_n)$ de n variables aléatoires indépendantes suivant la même loi qu'une variable X appelée variable aléatoire parente. Une réalisation de l'échantillon sera notée $(x_1, x_2, \dots, x_n)$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Définition (Statistique)

Soit X une variable aléatoire et $(X_1, X_2, \dots, X_n)$ un échantillon de X . Une statistique T est une variable aléatoire fonction mesurable de $(X_1, X_2, \dots, X_n)$, i.e,

$$T(X) = T(X_1, X_2, \dots, X_n)$$

Distribution d'échantillonnage de moyenne

- La statistique $\bar{X}$ (moyenne empirique de l'échantillon) est défini par:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Soient m et σ^2 la moyenne et la variance de la variable parente X , alors:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\begin{aligned} \mathbb{E}(\bar{X}) &= m \\ \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \end{aligned}$$

- D'après la loi des grands nombres, on a :

$$\bar{X} \rightarrow m \quad \text{quand} \quad n \rightarrow +\infty$$

Distribution d'échantillonnage de moyenne

- Rappel (Théorème Central Limite): Soient $(X_1, \dots, X_n)$ n variables aléatoires de même loi d'espérance m et de variance σ^2 alors:

$$\frac{X_1 + \dots + X_n - nm}{\sigma\sqrt{n}} = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \quad \text{converge en loi vers } Z \sim \mathcal{N}(0, 1)$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Pour une taille d'échantillon n suffisamment grande, on peut considérer que $\bar{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$
- **Application:** Soit X la v.a représentant le succès pendant une suite de n répétitions indépendantes dont la probabilité est égale à p . Soit $F = \frac{X}{n}$ la fréquence empirique du succès pendant n répétitions. Si n est suffisamment grand, déterminer la loi de F ?

Etude de la statistique S^2

- On définit la statistique S^2 comme suit:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Propriétés de S^2 :

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$S^2 = \frac{1}{n} \left(\sum_{i=1}^n X_i^2 \right) - \bar{X}^2$$

$$S^2 = \frac{1}{n} \left(\sum_{i=1}^n (X_i - m)^2 \right) - (\bar{X} - m)^2$$

- S^2 converge presque sûrement vers σ^2

Etude de la statistique S^2

- L'espérance et la variance de S^2 données par les formules suivantes:

$$\mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$\text{Var}(S^2) = \frac{n-1}{n^3} [(n-1)\mu_4 - (n-3)\sigma^4]$$

où μ_4 est le moment centré d'ordre 4 de X .

- La variable $\frac{S^2 - \frac{n-1}{n} \sigma^2}{\sqrt{\text{Var}(S^2)}} \rightarrow Z \sim \mathcal{N}(0, 1)$

Loi de Chi-deux et de Student

Loi de χ^2 (Chi-deux)

Soient $X_1, \dots, X_p \sim \mathcal{N}(0, 1)$ p v.a indépendantes, alors la loi de Chi-deux à p degré de liberté χ_p^2 la loi de la variable $Z = \sum_{i=1}^p X_i^2$

□ $\mathbb{E}(Z) = p$ et $Var(Z) = 2p$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Loi de Student

Soient $U \sim \mathcal{N}(0, 1)$ et $X \sim \chi_n^2$ deux v.a indépendantes, alors $T_n = \frac{U}{\sqrt{\frac{X}{n}}}$ est une variable de Student à n degrés de liberté.

□ $\mathbb{E}(T_n) = 0$ si $n > 1$ et $Var(T_n) = \frac{n}{n-2}$ si $n > 2$.

Echantillon gaussien

- Ici $X \sim \mathcal{N}(m, \sigma^2)$.
- Etude de la moyenne $\bar{X}$: $\bar{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$ loi exacte (même pour n petit).

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Etude de S^2 : On a:

$$n \frac{S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - m}{\sigma} \right)^2 - \left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \right)^2$$

Echantillon gaussien

- La loi de $\frac{nS^2}{\sigma^2}$ est une loi de chi deux à $(n - 1)$ degrés de liberté, i.e, $\chi^2(n - 1)$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Les statistiques $\bar{X}$ et S^2 sont indépendantes.

- La variable $T = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \sqrt{\frac{n-1}{\frac{nS^2}{\sigma^2}}} = \frac{\bar{X} - m}{S} \sqrt{n-1}$ est une variable de Student à $n - 1$ degré de liberté.

Estimation à un paramètre

- L'estimation ponctuelle consiste à chercher des estimateurs pour les paramètres d'une population $(m, \sigma, \dots)$ à l'aide d'un échantillon de n observations issues de cette population.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Les lois des grands nombres permettent de prendre $\bar{X}$ et S^2 pour estimer m et σ . La fréquence empirique f d'un événement est une estimation de la probabilité p . Les variables $\bar{X}$, S^2 et f sont appelées estimateurs de m , σ et p .

Qualités d'un estimateur

- On note θ le paramètre à estimer et T un estimateur de θ .
- T est dit convergent si T converge vers θ lorsque la taille de l'échantillon tend vers l'infini.
- $\bar{X}$, S^2 et f sont des estimateurs convergents.
- T est une v.a dont on suppose connue la loi de probabilité pour une valeur de θ fixée.

Estimateur sans biais

□ L'estimateur T est dit sans biais si $\mathbb{E}(T) = \theta$

□ L'estimateur T est dit biaisé si $\mathbb{E}(T) \neq \theta$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

□ L'estimateur T est dit asymptotiquement sans biais si $\mathbb{E}(T) \rightarrow \theta$ lorsque $n \rightarrow \infty$.

□ **Exemple:** $\bar{X}$ est un estimateur sans biais de m .

S^2 est un estimateur biaisé de σ^2 et asymptotiquement sans biais.

$S^{*2} = \frac{n}{n-1} S^2$ est un estimateur sans biais de S^2

Estimateur sans biais

- Erreur quadratique moyenne donne:

$$\mathbb{E}[(T - \theta)^2] = \text{Var}(T) + (\mathbb{E}(T) - \theta)^2$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Entre deux estimateurs sans biais, le plus précis au sens de l'erreur quadratique moyenne est celui de variance minimale.
- S'il existe un estimateur de θ sans biais, de variance minimale alors il est unique presque sûrement.

Estimation par intervalles

- L'estimation par intervalle de confiance d'un paramètre θ consiste à associer à un échantillon à n observations, un intervalle aléatoire I :

$$P(\theta \in I) = 1 - \alpha$$

où $1 - \alpha$ est appelé seuil de confiance et α est appelé risque.

- Soient X la variable parente dont la loi de probabilité dépend de θ , $(X_1, \dots, X_n)$ un échantillon de X et T un estimateur de θ fonction de l'échantillon, alors la loi de probabilité de T permet de déterminer à α fixé, des valeurs t_1 et t_2 telles que:

$$P(t_1 \leq \theta - T \leq t_2) = 1 - \alpha$$

on a alors

$$P(T + t_1 \leq \theta \leq T + t_2) = 1 - \alpha$$

Propriétés d'un intervalle de confiance

- t_1 et t_2 doivent vérifier $P(t_1 + T \leq \theta \leq T + t_2) = 1 - \alpha$ donc:

$$P(\theta - T < t_1) = \alpha_1, \quad P(\theta - T > t_2) = \alpha_2 \quad \text{avec} \quad \alpha = \alpha_1 + \alpha_2$$

- Un intervalle est symétrique si $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Un intervalle de confiance à droite ou à gauche respectivement si $\alpha_1 = 0$ et $\alpha_2 = \alpha$ ou $\alpha_1 = \alpha$ et $\alpha_2 = 0$
- La largeur de l'intervalle augmente lorsque α diminue
- La largeur de l'intervalle diminue quand la taille de l'échantillon augmente.

Construction d'un intervalle de confiance

- **Application:** La résistance à l'éclatement des citernes à essence dans une usine est une variable normale de moyenne inconnue et d'écart-type égal à 13, i.e $\sim \mathcal{N}(m, 13^2)$. Des essais sur 16 réservoirs conduisent à une résistance moyenne à l'éclatement de 1215.

On veut construire un intervalle de confiance à risques symétriques ayant une probabilité égale à 0,95 de contenir la résistance moyenne à l'éclatement des citernes produites par ce constructeur.

NB: Chercher dans la table de la loi normale centrée réduite la valeur correspondante à $0.975 = (1 + 0.95)/2$ de sorte que

$$P\left(\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq \text{la valeur dans la table}\right) = 0.975.$$

- **Correction:** $m \in \left[1215 - 1.96 \frac{13}{4}, 1215 + 1.96 \frac{13}{4}\right]$

Moyenne d'une loi normale de moyenne m et d'écart type σ

- **Ecart type σ connu:** $\bar{X}$ est un estimateur de m et on sait que $\bar{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$. En utilisant la table de loi normale centrée réduite pour n et α donnés, on trouve a tel que:

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

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Si $\bar{x}$ est la moyenne de l'échantillon de taille n , alors on a:

$$m \in \left[\bar{x} - a \frac{\sigma}{\sqrt{n}}, \bar{x} + a \frac{\sigma}{\sqrt{n}} \right]$$

Exemple d'application:

- **Application** On suppose que le poids d'un nouveau née est une variable aléatoire de loi inconnue de moyenne inconnue et d'écart-type égal à 0.5 kg. Le poids moyen des 49 enfants nés pendant le même mois a été de 3,6 kg.

1 Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau née pendant ce mois.

2 Quel serait le niveau de confiance d'un intervalle de longueur 0.1 kg centrée en 3.6 pour ce poids moyen ?

- **Correction:** 1)- $\left[3.6 - 1.96 \frac{0.5}{\sqrt{49}}, 3.6 + 1.96 \frac{0.5}{\sqrt{49}} \right]$, 2)- Chercher la valeur V_a correspondante à $2a = \frac{0.1 * \sqrt{49}}{0.5}$ dans la table de la loi normale centrée réduite. Le niveau de confiance étant égal à $2V_a - 1$

Moyenne d'une loi normale de moyenne m et d'écart type σ

- **Ecart type σ inconnu**: $\bar{X}$ est un estimateur de m et puisque σ étant inconnu, on a S un estimateur de σ et:

$$\frac{\bar{X} - m}{\frac{S}{\sqrt{n-1}}} \text{ suit une loi de Student à } (n-1) \text{ degrés de liberté.}$$

En utilisant la table de la loi de Student pour n et α fixés, on a trouve la valeur de $t_{\frac{\alpha}{2}}$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{X} - m}{\frac{S}{\sqrt{n-1}}} \leq t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Si $\bar{x}$ et s sont respectivement la moyenne et l'écart type d'un échantillon de taille n , alors:

$$m \in \left[\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} \right]$$

Si $n \geq 30$ la loi de Student approche $\mathcal{N}(0, 1)$

Exemple d'application

- **Application:** Un laboratoire est chargé de préciser la résistance à l'éclatement, en kg/cm^2 , des réservoirs à essence des camions-citernes d'un constructeur donné. On considère que la résistance à l'éclatement de ces citernes est **une variable de loi inconnue, de moyenne inconnue et d'écart-type inconnu**. Des essais sur **50** réservoirs conduisent à une résistance moyenne à l'éclatement de **1215 kg/cm^2** et un écart type **$s = 13 \text{ kg/cm}^2$** .

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOUI.

On veut construire un intervalle de confiance à risques symétriques ayant une probabilité égale à **0,95** de contenir la résistance moyenne à l'éclatement des citernes produites par ce constructeur.

- **Correction:** $1215 - 1.96 * \frac{13}{\sqrt{50}} \leq \bar{X} \leq 1215 + 1.96 * \frac{13}{\sqrt{50}}$

Variance d'une loi normale

□ **Moyenne m connue:** Le meilleur estimateur de σ^2 :

$$T = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$$

On a également $\frac{nT}{\sigma^2} \sim \chi_n^2$. Donc après lecture de la table de la loi Chi-deux à α et n fixés, on obtient les a et b tels que:

$$P(a \leq \frac{nT}{\sigma^2} \leq b) = 1 - \alpha$$

Alors on obtient $\sigma^2 \in \left[\frac{nT}{b}, \frac{nT}{a} \right]$

Si $n \geq 30$, $\frac{\sqrt{2T}}{\sigma} - \sqrt{2n-1} \sim \mathcal{N}(0, 1)$

Variance d'une loi normale

- **Moyenne m est inconnue**: On a $\frac{nS^2}{\sigma^2} \sim \chi_{n-1}^2$. Donc après lecture de la table de la loi Chi-deux à α et n fixés, on obtient les a et b tels que:

$$P\left(a \leq \frac{nS^2}{\sigma^2} \leq b\right) = 1 - \alpha$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Alors on obtient $\sigma \in \left[\sqrt{\frac{nS^2}{b}}, \sqrt{\frac{nS^2}{a}} \right]$

Si S^{2*} est l'estimateur de σ^2 alors on aura:

$$\sigma \in \left[\sqrt{\frac{(n-1)S^{*2}}{b}}, \sqrt{\frac{(n-1)S^{*2}}{a}} \right]$$

Si $n \geq 30$, $\frac{\sqrt{2}S}{\sigma} - \sqrt{2n-3} \sim \mathcal{N}(0, 1)$

Intervalle de confiance pour une proportion

- Soit n très grand taille d'une population dont la proportion p d'individus possède un certain caractère. L'estimation de p est f par un échantillon de taille n . La variable $nf \sim \mathcal{B}(n, p)$ mais comme n est très grand on aura:

$$nf \sim \mathcal{N}(np, np(1-p)) \quad \text{ou encore} \quad f \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

En utilisant la table de loi normale centrée réduite pour n et α donnés, on trouve a tel que:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$P\left(\frac{|p - f|}{\sqrt{\frac{p(1-p)}{n}}} \leq a\right) = 1 - \alpha$$

l'intervalle de confiance de la proportion est donné par:

$$(p - f)^2 \leq a^2 \frac{p(1-p)}{n}$$

On peut considérer la méthode où le $p = \frac{1}{2}$ dans le second membre (p minimise le second membre): $p \in \left[f - \frac{a}{2\sqrt{n}}, f + \frac{a}{2\sqrt{n}}\right]$

Exemple d'application

- **Ex 1:** On prélève 25 pièces dans une production industrielle. Une étude préalable a montré que le diamètre de ces pièces suivait une loi normale $\mathcal{N}(10, 4)$. Entre quelles valeurs a-t-on 90 chances sur 100 de trouver le diamètre moyen de ces 25 pièces et leur écart-type?
- **Ex 2:** On cherche à estimer une proportion p d'une population avec une précision de 0.01. On choisit un risque $\alpha = 0,05$, la valeur de la table de la loi normale nous donne $a = 1,96$. Quelle est la taille de l'échantillon à prévoir?
- **Ex 3:** Si $X \sim \mathcal{N}(3, 0.36)$,

 - 1 Calculer $P(X > 3.6)$
 - 2 Trouver a tel que $P(X > a) = 0.05$
- **Correction:** 1)- $9.34 \leq \bar{X} \leq 10.66$, $1.49 \leq S \leq 2.41$, 2)- $n \geq 98^2$

Exemple de Saporta: Niveau de pluies dans une région

- Le niveau naturel de pluies $\sim \mathcal{N}(600, 100^2)$.
- Des entrepreneurs prétendent que le procédé d'insémination des nuages au moyen d'iodure d'argent pourrait augmenter le niveau moyen de pluies de **50mm**. Le résultat de ce procédé pendant 10 ans donne:

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Année	1951	1952	1953	1954	1955	1956	1957	1958	1959
mm	510	614	780	512	501	534	603	788	650

- Il y a deux hypothèses à tester:
 - 1 H_0 : Le procédé est sans effet
 - 2 H_1 : Le procédé augmente réellement le niveau de pluies.

Exemple de Saporta: Niveau de pluies dans une région

- Les agriculteurs étaient décidés d'abandonner H_0 et d'accepter H_1 si le résultat obtenu par les mesures fait parti d'une éventualité improbable qui n'avait que 5% de se réaliser.
- Sous l'hypothèse H_0 , $\bar{X} \sim \mathcal{N}(600, \frac{100^2}{9})$. Le procédé est adopté (H_1) si $P(\bar{X} > t) = \alpha = 0.05$
- Comme $\frac{\bar{X}-600}{\frac{100}{3}} \sim \mathcal{N}(0, 1)$, on a $\frac{t-600}{\frac{100}{3}} = 1.64$ d'après la table de $\mathcal{N}(0, 1)$, on en déduit que $k = 655mm$
- On obtient la règle de décision:
 - $\{\bar{X} > 655\}$ est appelée région de rejet de H_0
 - $\{\bar{X} < 655\}$ est appelée région d'acceptation de H_0
- Comme $\bar{X} = 610, 2mm$ l'hypothèse H_0 sera adoptée.
- Sous l'hypothèse H_1 , $\bar{X} \sim \mathcal{N}(650, \frac{100^2}{9})$, l'hypothèse H_0 est adoptée si $\{\bar{X} < 655\}$, $P(\bar{X} < 655) = P(\frac{\bar{X}-650}{\frac{100}{3}} < 0.15) = \beta = 0.56$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Notions sur les tests

- Un test est un procédé qui permet de choisir entre deux hypothèses:

Décision \ Vérité	H_0	H_1
H_0	$1 - \alpha$	β
H_1	α	$1 - \beta$

Document not to be used for teaching. All rights reserved by the author, Dr Mohamed ASSELLAOUI.

- α s'appelle le risque de première espèce
- β s'appelle le risque de deuxième espèce

Etapes d'un test statistique

- Choix du risque α
- Choix des hypothèses H_0 et H_1
- Détermination de la variable de décision
- Détermination de la région critique $W : P(W/H_0) = \alpha$
- Calcul éventuel de la puissance du test : $P(W/H_1) = 1 - \beta$
- Calcul, sur l'échantillon, de la valeur expérimentale de la variable de décision
- Conclusion du test: rejet ou acceptation de H_0 .

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Catégories d'un test statistique

- **Un test paramétrique**: permet de vérifier si une caractéristique d'une population, que l'on notera θ , satisfait une hypothèse H_0 ou non. On distingue les hypothèses simples des hypothèses composites:

- une hypothèse simple : $H_0 : \theta = \theta_0$,
Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- une hypothèse composite est du type $H_0 : \theta \in A$ où A est une partie non réduite à un singleton.

- **Un test non paramétrique** permet de tester une propriété (sa loi de probabilité, indépendance, homogénéité)

Moyenne de $\mathcal{N}(m, \sigma^2)$: Test unilatéral

- **Ecart type σ connu**: Le test repose sur la variable de décision $\bar{X}$.

$$H_0 : m = m_0, \quad H_1 : m > m_0$$

On sait que $\bar{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$. Si H_0 est adoptée $\bar{X} \sim \mathcal{N}(m_0, \frac{\sigma^2}{n})$. La région critique est $\{\bar{X} > k\}$

Pour n et α donnés, on peut trouver la valeur de $z = \frac{k - m_0}{\frac{\sigma}{\sqrt{n}}}$ à partir de la table de la loi normale centrée réduite, telle que:

$$P(\bar{X} > k \mid H_0) = P\left(\frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} > \frac{k - m_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

La valeur z étant connue, vous pouvez en faire sortir k pour un échantillon donné de taille n donnée. Si la valeur de $\bar{x}$ obtenue par cet échantillon vérifie $\{\bar{x} > k\}$ alors H_0 est rejetée.

Moyenne de $\mathcal{N}(m, \sigma^2)$: Test bilatéral

- **Ecart type σ connu**: Le test repose sur la variable de décision $\bar{X}$.

$$H_0 : m = m_0, \quad H_1 : m \neq m_0$$

On sait que $\bar{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$. Si H_0 est adoptée $\bar{X} \sim \mathcal{N}(m_0, \frac{\sigma^2}{n})$. La région critique est $\{|\bar{X} - m_0| > k\}$

Pour n et α donnés, on peut trouver la valeur de $z = \frac{k}{\frac{\sigma}{\sqrt{n}}}$ à partir de la table de la loi normale centrée réduite, telle que:

$$P(|\bar{X} - m_0| > k \mid H_0) = P\left(\frac{|\bar{X} - m_0|}{\frac{\sigma}{\sqrt{n}}} > \frac{k}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

La valeur z étant connue, vous pouvez en faire sortir k pour un échantillon donné de taille n donnée. Si la valeur de $\bar{x}$ obtenue par cet échantillon vérifie $\{|\bar{X} - m_0| > k\}$ alors H_0 est rejetée.

Moyenne de $\mathcal{N}(m, \sigma^2)$

- **Ecart type σ inconnu**: Le test repose sur la variable de décision $\bar{X}$.
 $H_0 : m = m_0, H_1 : m \neq m_0$. On sait que :

$$T_{n-1}(m) = \frac{\bar{X} - m}{\frac{s}{\sqrt{n-1}}} \text{ suit une loi de Student à } (n-1) \text{ degrés de liberté.}$$

Si H_0 est adoptée, $T_{n-1}(m_0)$ suit la loi de Student à $n-1$ degrés de liberté. La région critique est $\{|T_{n-1}(m_0)| > k\}$ et on cherche k tel que:

$$P(|T_{n-1}(m_0)| > k \mid H_0) = \alpha$$

Si la valeur de $\bar{X}$ obtenue par l'échantillon vérifie $\{|T_{n-1}(m_0)| > k\}$ alors H_0 est rejetée.

- **Exemple**: $H_0 : m = 30, H_1 : m > 30$. Un échantillon de 15 a donné: $\bar{x} = 37.2, s = 6.2$, La valeur critique à $\alpha = 0.05$ donne 1.761 et on peut calculer que $t = \frac{37.2-30}{\frac{6.2}{\sqrt{14}}} = 4.35 \geq 1.761$ donc l'hypothèse H_0 est à rejeter.

Variance d'une loi normale

□ **Moyenne m connue:** Le meilleur estimateur de σ^2 :

$$T = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2$$

On a également $\frac{nT}{\sigma_0^2} \sim \chi_n^2$. Si on prend $H_0 : \sigma = \sigma_0$, $H_1 : \sigma > \sigma_0$, la région critique est $T > k$ avec

$$P\left(\frac{nT}{\sigma_0^2} > \frac{nk}{\sigma_0^2}\right) = \alpha$$

La valeur de $z = \frac{nk}{\sigma_0^2}$ est obtenue grâce à la table de Chi 2

Sur un échantillon de taille n , vous pouvez donc conclure la valeur de k . Soit T la variance de cet échantillon calculée à partir d'une moyenne connue m_0 . Si $T > k$, H_0 est rejetée.

Variance d'une loi normale

- **Moyenne m est inconnue**: On a $\frac{nS^2}{\sigma^2} \sim \chi_{n-1}^2$. On considère $H_0 : \sigma = \sigma_0$, $H_1 : \sigma > \sigma_0$.

Si H_0 est adoptée, $\frac{nS^2}{\sigma_0^2} \sim \chi_{n-1}^2$. La région critique est $S^2 > k$ avec

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$P\left(\frac{nS^2}{\sigma_0^2} > \frac{nk}{\sigma_0^2}\right) = \alpha$$

La valeur de $z = \frac{nk}{\sigma_0^2}$ est obtenue grâce à la table de Chi 2.

Sur un échantillon de taille n , vous pouvez donc conclure la valeur de k . Soit S^2 la variance de cet échantillon. Si $S^2 > k$, H_0 est rejetée.

Intervalle de confiance pour une proportion

- Soit n très grand taille d'une population dont la proportion p d'individus possède un certain caractère. L'estimation de p est f par un échantillon de taille n . On a

$$f \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

On considère le test suivant sur la proportion: $H_0: p = p_0$, $H_1: p \neq p_0$, la région critique est déterminée par:

$$\alpha = P\left(|f - p_0| > k | H_0\right) = P\left(\frac{|f - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > \frac{k}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$$

La valeur de k est obtenue grâce à la table de $\mathcal{N}(0, 1)$

- **Exemple** Un sondage sur un échantillon de 625 cadres révèle qu'il y a 48% d'entre eux qui utilise Internet. Or une hypothèse a été émise comme quoi la moitié des cadres utilise Internet. Le sondage est-il contradictoire avec cette hypothèse aux niveaux de confiance suivants: 95%, 90%, 99% ?

Comparaison de deux échantillons gaussiens

$$X_1 \sim \mathcal{N}(m_1, \sigma_1), \text{ et } X_2 \sim \mathcal{N}(m_2, \sigma_2)$$

- Les hypothèses sont :

$$H_0 : m_1 = m_2 \text{ et } \sigma_1 = \sigma_2, \quad H_1 : m_1 \neq m_2 \text{ et } \sigma_1 \neq \sigma_2$$

- Test des variances: On a

$$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2, \quad \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2$$

Document not to be used for teaching / All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Sous l'hypothèse $H_0 : \sigma_1 = \sigma_2$, on a le rapport des deux estimateurs de σ_1 et σ_2 , avec $\frac{n_1 S_1^2}{n_1 - 1} \geq \frac{n_2 S_2^2}{n_2 - 1}$:

$$F_{n_1-1; n_2-1} = \frac{\frac{n_1 S_1^2}{n_1 - 1}}{\frac{n_2 S_2^2}{n_2 - 1}}$$

La région critique ici est représentée par $P(F_{n_1-1; n_2-1} > k) = \alpha$. La valeur k est obtenue grâce à la table de la loi de Fisher.

Comparaison de deux échantillons gaussiens

$$X_1 \sim \mathcal{N}(m_1, \sigma_1), \text{ et } X_2 \sim \mathcal{N}(m_2, \sigma_2)$$

- Test des moyennes: supposons que $\sigma_1 = \sigma_2 = \sigma$, on a:

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi_{n_1+n_2-2}^2, \quad \bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(m_1 - m_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

Sous l'hypothèse $H_0 : m_1 = m_2$, la variable suivante est une variable de Student:

$$T_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(n_1 S_1^2 + n_2 S_2^2\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sqrt{n_1 + n_2 - 2}$$

La région critique est représentée par $P(|T_{n_1+n_2-1}| > k) = \alpha$.

Comparaison de deux pourcentages

- f_1 et f_2 présentent les pourcentages d'individus ayant un certain caractère et p_1 et p_2 les probabilités correspondantes.
- $H_0 : p_1 = p_2 = p, H_1 : p_1 \neq p_2$
- Soient les deux v.a F_1 , et F_2 dont les deux réalisations respectives f_1 et f_2 . Sous H_0 , on a

$$F_1 \sim \mathcal{N}\left(p, \frac{p(1-p)}{n_1}\right), F_2 \sim \mathcal{N}\left(p, \frac{p(1-p)}{n_2}\right)$$

On a :

$$F_1 - F_2 \sim \mathcal{N}\left(0, \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}\right)$$

- On va rejeter H_0 , si $\alpha = 0.05$:

$$|f_1 - f_2| > 1.96 \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Tests non paramétriques: Tests de Chi 2

On peut distinguer trois types de test du Chi 2:

- **Test du Chi 2 d'adéquation** (H_0 : le caractère X suit-il une loi particulière ?),

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- **Test du Chi 2 d'homogénéité** (H_0 : le caractère X suit-il la même loi dans deux populations données ?) ,
- **Test du Chi 2 d'indépendance** (H_0 : les caractères X et Y sont-ils indépendants ?).

Tests non paramétriques: Tests de Chi 2

- Soit une variable aléatoire X divisée en k classes de probabilités respectives p_i , $1 \leq i \leq k$. Soit un échantillon de cette variable fournissant les effectifs aléatoires N_i , $1 \leq i \leq k$, dans chacune des classes précédentes. On a $\mathbb{E}(N_i) = np_i$
- On considère la statistique

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

$$D^2 = \sum_{i=1}^k \frac{N_i - np_i}{np_i}$$

- Les éléments de D^2 ne sont pas indépendants ($\sum_{i=1}^k N_i = n$)
- Le résultat suivant est à admettre: pour n suffisamment grand $D^2 \sim \chi_{k-1}^2$.

Conduite du tests de Chi 2

- On considère une variable X pour laquelle on veut prouver que sa distribution est une distribution théorique particulière, de fonction de répartition F . Les hypothèses du test sont les suivantes :

H_0 : la distribution de X est la distribution théorique proposée.

H_1 : la distribution de X n'est pas cette distribution proposée.

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- Soit $(x_1, x_2, \dots, x_n)$ un échantillon de cette variable reparti en k classes, soit $N_1, \dots, N_k$ les effectifs empiriques de ces k classes et soit $(p_1, \dots, p_k)$ les probabilités théoriques.
- Chercher k dans la table de la loi de Chi 2 pour α et n donné tel que $P(D^2 \geq k) = \alpha$
- Calculez $d^2 = \sum_{i=1}^k \frac{n_i - np_i}{np_i}$, on rejette H_0 si $d^2 > k$. Sinon on garde H_0

Exemple: test de Chi 2

- **Exemple:** Un croisement entre roses rouges et blanches a donné en seconde génération des roses rouges, roses et blanches. Sur un échantillon de taille 600, on a trouvé les résultats suivants:

Couleur	effectifs observés n_i
rouges	141
roses	315
blanches	144

Document not to be used for teaching. All rights reserved by the author, Dr. Mohamed ASSELLAOU.

- L'exercice consiste à tester si les résultats obtenus sont conformes aux lois de Mendel: $p_{rouges} = 0.25$, $p_{roses} = 0.5$, $p_{blanches} = 0.25$ au risque $\alpha = 0.05$,
- On cherche t dans la table de la loi de χ^2_{k-1} pour $\alpha = 0.05$, tel que $P(D^2 > t) = 0.05$. on trouve $t = 5.991$.
- Puis on calcule $d^2 = \sum_{i=1}^3 \frac{n_i - 600p_i}{600p_i} = 1.53 \leq 5.991$. On conclut que l'hypothèse H_0 est adoptée.