République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA
FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE
DÉPARTEMENT INFORMATIQUE

Polycopié du Cours

Analyse de Données

Analyse de Données en Informatique
Première Année Master (M1)

Préparé par :
Dr. AFROUN Faïrouz

Université de Biskra, 2024/2025

Table des matières

Table des figures

Chapitre 1

Introduction à la théorie de test d'hypothèses

Introduction

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Il serait important de chercher à présenter en détail l'ensemble des tests statistiques, mais la littérature est très abondante sur le sujet. Pour cela, dans ce chapitre nous allons se limiter aux tests classiques les plus simples et les plus usuels dans la pratique. En effet, Les tests présentés, sont concernés les tests à un seul échantillon, d'adéquation d'une loi, de comparaison de deux échantillons et enfin l'analyse de la variance à un seule facteur et analyse de la variance à deux facteurs.

1.1 Estimateur empirique de la moyenne et de la variance

1.1.1 Construction de l'estimateur

A partir d'un échantillon $(X_1, ..., X_n)$ de X, nous définirons la loi de probabilité empirique P_n tel que : $P_n = \frac{1}{n}\sum\limits_{i=1}^{i=n}\delta_{X_i}$ où δ_{X_i} est la masse de Dirac (fonction indicatrice) au point X_i. Cette loi de probabilité admet une fonction de répartition empirique, notée F_n :

$$F_n(x) = \frac{\text{nombre de } x_i \text{ inférieur ou égale à } x}{n} = P_n(]-\infty, x]) . \tag{1.1}$$

Définition 1.1 *(moyenne empirique)*
La moyenne empirique d'un n-échantillon (i.i.d) est la moyenne de la loi empirique notée :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i,$$

cette v.a. admet espérance et variance.

Définition 1.2 *(variance empirique) La variance empirique de la loi empirique est :*

$$\begin{cases} S_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2, & \textit{si } \mu \textit{ est connue ;} \\ S_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2, & \textit{si } \mu \textit{ est inconnue.} \end{cases}$$

Remarque 1.1 *La valeur moyenne de la variance empirique n'est pas exactement égale à la variance théorique σ^2 (estimateur avec biais), c'est pourquoi on introduit la variance empirique corrigée définie par :*

$$S_{n,c}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i-\overline{X}_n\right)^2.$$

1.1.2 Distribution de l'estimateur d'une moyenne et d'une variance

En général, une simple estimation d'une caractéristique (moyenne, variance, médiane,....) d'un échantillon ne suffit pas : il est nécessaire de connaître son degré d'imprécision. L'outil fondamental pour évaluer un estimateur et le comparer à d'autres, est bien que sa distribution d'échantillonnage. Par exemple, à égalité entre différents aspects, on préférera l'estimateur avec la plus petite variance. Cette section s'occupe de la présentation de la distribution de quelques estimateurs usuels (moyenne, variance).

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ, de variance σ^2, et un échantillon $X_1,\ X_2, ...,\ X_n$ de X de taille n.

1. Pour chaque échantillonnage on peut calculer la moyenne observée du caractère

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

   On démontre que $E(\overline{X}) = \mu$ (un estimateur sans biais de μ) et $Var(\overline{X}) = \frac{\sigma^2}{n}$.

2. Si la moyenne μ est **connue**, alors on considère la variance d'échantillon

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - \mu^2.$$

3. Si la moyenne μ est **inconnue** alors dans ce cas, l'estimateur sans biais de la variance de l'échantillon est définie comme suit :

$$\hat{\sigma_c}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{n}{n-1}\hat{S}^2.$$

   avec

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

Les lois de probabilité de l'estimateur d'une moyenne et d'une variance pour certaine situations peuvent être résumées dans ce qui suit :

1. Cas d'un petit échantillon gaussien $n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$
   - Si σ est connu alors, $U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale $N(0,1)$.
   - Si σ est inconnu, alors, $T = \frac{\overline{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté.

2. Cas d'un grand échantillon ($n > 30$) et X de loi quelconque :
   - Dans ce cas, $U = \frac{\overline{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit approximativement la loi normale $N(0, 1)$.
3. Cas d'un échantillon gaussien (X de loi normale $N(\mu, \sigma^2)$) :
   - Si μ est connue alors la variable $Y^2 = n\frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi - Deux$ à n degrés de liberté.
   - Si μ est inconnue alors la variable $Y^2 = (n-1)\frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi - Deux$ à $n-1$ degrés de liberté.

1.2 Tests de conformité pour une moyenne

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ, d'écart-type σ, et un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n de X. La moyenne et la variance corrigée d'échantillon sont données respectivement par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } \hat{\sigma}_c^2 = \frac{n}{n-1}\hat{\sigma}^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

1.2.1 Cas d'un petit échantillon gaussien ($n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$)

Dans ce test deux cas sont envisageable. En effet, on peut distinguer le cas où l'écart-type est une quantité bien connue et le cas où l'écart-type n'est connue qu'approximativement à travers son estimateur.

1.2.1.1 Cas σ connu

Il s'agit de faire un choix entre plusieurs hypothèses possibles sur μ sans disposer d'informations suffisantes pour que ce choix soit sûr. On met en avant deux hypothèses privilégiées : l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Par exemple, on testera

$$H_0 :\ ''\mu = \mu_0'' \text{ contre } H_1 :\ ''\mu \neq \mu_0'',$$

avec μ_0 fixé arbitrairement. On veut savoir si l'on doit rejeter H_0 ou pas.

La résolution du présent problème consiste, en résumé, à réaliser les étapes suivantes :

1. Utilise une variable aléatoire dont on connaît la loi de probabilité lorsque H_0 est vraie. Par exemple, on prend $U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, en raison que lorsque H_0 est vraie, $U = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0, 1)$, et cela le fait que l'échantillon est issue d'une variable aléatoire d'une loi normale $X \rightsquigarrow N(\mu, \sigma^2)$.
2. Fixe une valeur $\alpha \in]0, 1[$. En général, on prend α (le risque) petit, le plus souvent
$$\alpha \in \{0.10, 0.05,\ 0.01,\ 0.01,\ 0.001\}.$$
3. Quantifier un réel u_α, tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$. Ce réel u_α peut être extrait de la table de la loi normale centrée et réduite (voir annexe A).
4. Comparer la moyenne empirique $\overline{X}$ de l'échantillon à la moyenne théorique $\mu = \mu_0$, sachant que l'hypothèse H_0 signifiera que les différences observées sont seulement dues aux fluctuations d'échantillonnage (i.e. ne sont pas significatives). En fin, on decide ce qui suit :

– On ne rejettera pas H_0 si les différences observées ne sont pas significatives, c'est-à-dire si U est "petite", ce que l'on peut formuler par $-u_\alpha < U < u_\alpha$, ou encore $|U| < u_\alpha$.
– On rejettera H_0 si les différences observées sont significatives, ce que l'on peut formuler par $U < -u_\alpha$ ou $U > u_\alpha$, c'est-à-dire $|U| > u_\alpha$. Par construction de u_α, on a $P(U > u_\alpha) = P(U < -u_\alpha) = \frac{\alpha}{2}$, soit encore $P(|U| > u_\alpha) = \alpha$ i.e. $P(U \notin]-u_\alpha, u_\alpha[) = \alpha$.

En pratique, on calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ et on décide

- de rejeter H_0 si $u \notin]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité forte de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ n'est pas conforme à la valeur théorique μ_0 mais on ne pourra pas donner de valeurs acceptable de μ ;
- de ne pas rejeter H_0 si $u \in]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité faible de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ est conforme à la valeur théorique μ_0 et que la valeur μ_0 ne peut être rejeter.
  Attention : d'autres valeurs μ_0', μ_0'', ... peuvent également convenir.

Erreurs de décision Il est à noter que, l'aspect aléatoire de l'échantillon (observations) peut nous faussé la décision finale (rejeter ou non l'hypothèse H_0). En effet, lorsque on rejette H_0 alors que H_0 est vraie, on commet une erreur. On a donc une probabilité α (car lorsque H_0 est vraie, on a $P(U \notin]-\mu_\alpha, \mu_\alpha[) = \alpha$) de se tromper : α est appelée ***erreur de première espèce***.

Une autre situation où on peut commettre une erreur de décision est bien que celle lorsque on ne rejette pas H_0 alors que H_0 est fausse. Dans ce cas, on a une probabilité β de se tromper : β est appelée ***erreur de deuxième espèce***. Cette probabilité est difficilement calculable car dans la plupart des temps, on ne connaît pas la loi de U lorsque H_0 est fausse. La valeur $1-\beta$ est appelée la ***puissance du test***.

Finalement, ces déférentes situations peuvent être résumées par le schéma suivant :

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

Les différents tests usuels (formulation et décision) correspondant à la présente situation peuvent être résumés comme suit :

Test (bilatéral) H_0 : "$\mu = \mu_0$" **contre** H_1 : "$\mu \neq \mu_0$",
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :
- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : "$\mu = \mu_0$" **contre** H_1 : "$\mu > \mu_0$",
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U \geq u_\alpha) = 1-\alpha$ et on décide que :
- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : "$\mu = \mu_0$" **contre** H_1 : "$\mu < \mu_0$",
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U < u_\alpha) = 1-\alpha$ et on décide que :
- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

1.2.1.2 Cas σ inconnu

Par définition, on sait que $T = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté (voir section **??**). Alors, les différents tests précédents (bilateral et unilatéral) se font comme suit :

Test (bilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu \neq \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α sur la table de Student pour un degré de liberté $n-1$ tel que $P(-t_\alpha < T < t_\alpha) = 1-\alpha$ et on décide que :

- Si $t \in]-t_\alpha, t_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $t \notin]-t_\alpha, t_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu > \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T \geq t_\alpha) = 1-\alpha$ et on décide que :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu < \mu_0''$,

On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T < t_\alpha) = 1-\alpha$ et on décide que :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

1.2.2 Cas d'un grand échantillon : $n > 30$

Dans cette situation ($n > 30$), on se basons sur le TCL, on sait que la variable aléatoire $U = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit approximativement une loi normale centrée et réduite ($U \rightsquigarrow N(0,1)$).

Test (bilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu \neq \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :

- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu > \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U \geq u_\alpha) = 1-\alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $"\mu = \mu_0''$ **contre** H_1 : $"\mu < \mu_0''$,

On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U < u_\alpha) = 1-\alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

1.3 Tests d'homogénéité

Dans les différents tests présenté dans les sections précédentes on n'a considéré qu'un seule échantillon, pour lequel on s'intéresse si l'un de ses caractères (moyenne, variance, distribution) est conforme à une quantité fixée arbitrairement (cette dernière quantité représente généralement une norme du phénomène étudié). Cependant, dans la pratique, dans certaines situation on dispose de deux populations P_1 et P_2 ou voir même plus de deux populations, dont on étudie un même caractère et on désir comparer les populations quant à ce caractère, et donc à savoir si elles sont homogènes ou non. Dans cette section, nous se limitons au cas de test d'homogénéité de variance et de moyennes de deux populations indépendantes.

1.3.1 Comparaison de deux variances

Soient X et Y deux variables aléatoires indépendantes représentant le même caractère quantitative dans chacune des populations P_1 et P_2. On suppose que X et Y suivent des lois normales respectivement, $N(\mu_1; \sigma_1^2)$ et $N(\mu_2; \sigma_2^2)$.

De P_1, on extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille n_1 de X et de P_2, on extrait un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille n_2 de Y.

Les moyennes empiriques des deux échantillons sont alors

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i;$$

et leurs variances corrigées sont :

$$\hat{\sigma}_{c,1}^2 = \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2 \text{ avec } \hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2 - \overline{X}^2,$$

$$\hat{\sigma}_{c,2}^2 = \frac{n_2}{n_2 - 1}\hat{\sigma}_2^2 \text{ avec } \hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2 - \overline{Y}^2.$$

On veut réaliser le test bilatéral suivant :

$$H_0 : \text{"}\sigma_1^2 = \sigma_2^2\text{" contre } H_1 : \text{"}\sigma_1^2 \neq \sigma_2^2\text{"}.$$

Les étapes de la réalisation de ce test peuvent être résumées comme suit :

1. On calcule la réalisation $f_c = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$. Si nécessaire, on permute les échantillons de sorte que $f_c \geq 1$ $\left(\text{c'est-à-dire } f_c = \frac{\max\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}{\min\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}\right)$.
2. Sachant que sous l'hypothèse H_0, la statistique (variable aléatoire) $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$ suit une loi de Fisher à $(n_1 - 1; n_2 - 1)$ degrés de liberté, alors à partir de la table de Fisher on détermine f_α tel que : $P(F \geq f_\alpha) = \frac{\alpha}{2}$(ou encore $P(F \leq f_\alpha) = 1 - \frac{\alpha}{2}$).
3. La règle de décision se fait comme suit :
   - si $f_c < f_\alpha$, alors on ne peut rejeter H_0 (H_0 est vraie).
   - si $f_c \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Avec le même raisonnement on va trouver la zone de non rejet de l'hypothèse nulle dans les tests unilatéral. Les résultats des différents tests sont résumés dans le tableau suivant :

Hypothèse	Zone de non-rejet H_0
H_0 : "$\sigma_1^2 = \sigma_2^2$ " contre H_1 : "$\sigma_1^2 \neq \sigma_2^2$ "	$[1; f(n_1 - 1, n_2 - 1, 1 - \frac{\alpha}{2})]$
H_0 : "$\sigma_1^2 = \sigma_2^2$ " contre H_1 : "$\sigma_1^2 > \sigma_2^2$ "	$[1; f(n_1 - 1, n_2 - 1, 1 - \alpha)]$
H_0 : "$\sigma_1^2 = \sigma_2^2$ " contre H_1 : "$\sigma_1^2 < \sigma_2^2$ "	$[1; f(n_2 - 1, n_1 - 1, 1 - \alpha)]$, avec $f_c = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$

tel que $f(n, m, 1 - \alpha)$ est lu dans la table de loi Fisher-Snedecor $(1 - \alpha)$ à colonne n, ligne m, de plus on ne rejettera pas H_0 si f_c appartient à la zone de non-rejet de H_0 et on rejettera H_0 sinon.

1.3.2 Comparaison de deux moyennes

Dans cette section, nous allons intéresser à l'homogénéité de deux populations par rapport à la moyenne. Notons que, le test de comparaison de deux moyennes dépend de la distribution des échantillons dont on dispose. Dans le cadre de ce document, nous allons se focalisé sur le cas où les deux échantillons sont de grand taille issues d'une loi quelconque et le cas où les deux échantillons sont gaussien et de petite taille.

1.3.2.1 Cas des grands échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère qualitative étudié dans chaque population. On suppose que X et Y suivent une loi quelconque de moyennes respectives μ_1 et μ_2 et d'écart-types respectifs σ_1 et σ_2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 > 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 > 30$ de Y.

Soit la statistique

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{1.2}$$

et u sa réalisation.

Test (bilatéral) H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 \neq \mu_2''$,

Sous l'hypothèse H_0, la statistique U définie par (1.2) suit approximativement la loi normale centrée réduite $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}^2_{c,1}}{n_1}+\frac{\hat{\sigma}^2_{c,2}}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que :

$$P(-u_\alpha < U < u_\alpha) = 1 - \alpha,$$

c'est-à-dire

$$P(U < u_\alpha) = 1 - \frac{\alpha}{2},$$

et on décide :

- de ne pas rejeter H_0 si $u \in\ ] - u_\alpha, u_\alpha[$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \notin\ ] - u_\alpha, u_\alpha[$.

Test (unilatéral) de H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}^2_{c,1}}{n_1}+\frac{\hat{\sigma}^2_{c,2}}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que : $P(U \geq u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u < u_\alpha$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \geq u_\alpha$.

Test (unilatéral) H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}^2_{c,1}}{n_1}+\frac{\hat{\sigma}^2_{c,2}}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que $P(U < u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u > -u_\alpha$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \leq -u_\alpha$.

La démarche et les résultats des trois tests ci-dessus restent valable si on remplace σ_1^2 ou σ_2^2 par leurs estimations $\hat{\sigma}_{c,1}^2$, le fait que $U = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}}$ suit aussi une loi normale centrée réduite (on peut le justifier par le TCL).

1.3.2.2 Cas de petits échantillons ($n \leq 30$)

Soient X et Y des variables aléatoires indépendantes représentant le caractère dans chaque population. On suppose que X et Y suivent une loi normal de moyennes respectives μ_1 et μ_2, de variance respectives σ_1^2 et σ_2^2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 \leq 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 \leq 30$ de Y.

Test (bilatéral) H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 \neq \mu_2''$,

Afin de réaliser ce test, nous définissons la statistique suivante :

$$T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}}. \tag{1.3}$$

Sous l'hypothèse H_0 et l'hypothèse $\sigma_1 = \sigma_2$ la statistique du test définie dans (1.3) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Cependant, dans la pratique on ne sait pas si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ou non. A cet effet, on doit d'abord tester l'égalité des deux variances, $\sigma_1^2 = \sigma_2^2$ (Voir section 1.3.1).

Si cette dernière hypothèse est retenue, alors la valeur commune σ^2 peut être estimer par $\hat{\sigma}_c^2 = \frac{(n_1-1)\sigma_{c,1}^2+(n_2-1)\sigma_{c,2}^2}{n_1+n_2-2}$. Ensuite, on calcule la réalisation de la statistique T, c'est-à-dire $t = \frac{\overline{x}-\overline{y}}{\hat{\sigma}_c \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ et on détermine sur la table de la loi de Student la valeurs critique, t_α, du test tel que : $P(-t_\alpha < T < t_\alpha) = 1-\alpha$. Finalement, on décide que :

- On ne peut rejeter H_0 si $t \in]-t_\alpha, t_\alpha[$;
- On rejette H_0 si $\notin]-t_\alpha, t_\alpha[$, avec une probabilité α de se tromper dans la décision.

Test (unilatéral) H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (1.3) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Ainsi, on détermine t_α sur la table de la loi de Student pour un $n = n_1 + n_2 - 2$ et qui vérifié l'égalité $P(T \geq t_\alpha) = 1 - \alpha$ et on décide :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la décision.

Test (unilatéral) H_0 : $''\mu_1 = \mu_2''$ **contre** H_1 : $''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (1.3) suit encore approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour prendre la décision sur le rejet de l'hypothèse H_0, il suffit de déterminer sur la table de Student pour un *ddl* $n = n_1 + n_2 - 2$ la valeur critique t_α tel que $P(T < t_\alpha) = 1 - \alpha$ et on décide :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la decision.

1.4 Analyse de la variance à un facteur (ANOVA 1)

Dans cette section, nous allons intéresser à un cas plus générale pour la comparaison de moyennes et cela lorsque le nombre d'échantillon est supérieur strictement à deux. Plus précisément nous allons intéresser à la technique d'analyse de la variance à un seul facteur qui est la plus adéquate avec la situation.

1.4.1 Position du problème

Supposons que nous ayons 3 bases de données contenant un type d'information bien déterminé où nous désirons savoir si ces bases de données ont une influence sur l'espace mémoire occupé, sur le disque dur, par cette information ou non. À cet effet, nous avons réalisés un recueil de capacité mémoire occupé par six (06) fichiers de chacune des trois bases de données, dont les mesures sont rangées dans le tableau suivant.

N°	Base 1	Base 2	Base 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

TABLE 1.1: *Espace mémoire occupé selon la base de données*

Soit les notions et les notations suivantes :
- La Base de données : Variable qualitative contenant trois modalités, appelée facteur.
- Espace mémoire occupé : Réponse, notée X, et μ_i l'espace mémoire moyen occupé dans la $i^{\text{ème}}$ base de données ($i = \overline{1,3}$).

Répondre à notre objectif consiste à la réalisation du test suivant :

$$H_0 : \text{"}\mu_1 = \mu_2 = \mu_3 = \mu\text{"} \text{ contre } H_1 : \text{"}\exists\ i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j;\text{"}.$$

Pour réaliser ce test nous pourrons le décomposer en trois sous-tests où nous comparons l'espace mémoire moyen occupé par l'information de l'attribut en question deux à deux selon les bases de données. Mais afin de contourner le problème d'erreur α gonflé, le fait elle ne réalise qu'une seule comparaison à la fois, nous utilisons la technique statistique connue sous le nom d'analyse de variance (en anglais : Analyse Of Variance (ANOVA)) plutôt que des tests de Student t (voir section 1.3.2) multiples. Remarquez que l'ANOVA peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t.

1.4.2 Analyse de la variance à un seul facteur

L'identification de l'$ANOVA$ 1 au sens littéraire peut être résumée dans la définition suivante :

Définition 1.3 *(ANOVA 1)*
L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités (groupes) sur les moyennes d'une variable quantitative X.

Les problèmes concernés par la technique $ANOVA$ 1 se présente sous leurs formes générale suivante :

N°	*groupe* 1	*groupe* 2		*groupe* p
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	$\cdots$	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	$\cdots$	$X_{4,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_j	$X_{n_1,1}$	$X_{n_2,2}$	$\cdots$	$X_{n_p,p}$

et le modèle mathématique leurs associés est donné par :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1,n}, j = \overline{1,p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0,\sigma^2), \tag{1.4}$$

où X_{ij} est la $j^{\text{ième}}$ réalisation de la variable quantitative X dans le $i^{\text{ième}}$ échantillon et ϵ_{ij} sont les erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par :

$$H_0 : {''}\mu_1 = \mu_2 = ... = \mu_p = \mu{''} \text{ contre } H_1 : {''}\exists\ i,j \in \{1,2,...,p\} \text{ tel que } \mu_i \neq \mu_j{''}. \tag{1.5}$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en oeuvre de l'ANOVA 1 qui nous permet de réaliser ce test.

1.4.3 Les étapes de l'$ANOVA$ 1

Afin de réaliser le test définie dans (1.5), principalement trois conditions doit être vérifiées préalablement, à savoir :

- Les p échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont la même variance : *Homogénéité* des variances ou *homoscédasticité*.

Si ces trois dernières conditions sont vérifiées alors, on peut utiliser la technique $ANOVA$ 1 pour réaliser le test (1.5), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

- La moyenne de toutes les observations : $\overline{X} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} X_{ij}$ avec $n = \sum_{j=1}^{p} n_j$;
- Moyenne de chaque échantillon : $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$, pour $j = \overline{1,p}$;
- Variance de chaque échantillon : $\hat{\sigma}_i^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$, pour $j = \overline{1,p}$;
- La variance de toutes les observations : $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2$ avec $n = \sum_{j=1}^{p} n_j$.

On peut monter facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2, \tag{1.6}$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j-\overline{X})^2. \tag{1.7}$$

On multipliant (1.7), par n on obtient :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j-\overline{X})^2}_{SC_{Fac}}, \tag{1.8}$$

où,

SC_{Tot} : est la variation totale qui représente la dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur (variation inter-groupes) qui représente la dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle (variation intra-groupes) qui représente la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

L'idée la plus naturelle est de dire que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation intra-groupes (résiduelle) associée au caractère, c'est-à-dire,

- Si H_0 est vraie, alors la variation SC_{Fac} due au facteur doit être petite par rapport à la variation résiduelle SC_{Res}.
- Par contre, si H_1 est vraie alors la variation SC_{Fac} due au facteur doit être grande par rapport à la quantité SC_{Res}.

Pour comparer ces quantités, Fisher a considéré le rapport des carrées moyennes associées au facteur CM_{Fac} et les carrées moyennes résiduelles CM_{Res}, où

le carré moyen associé au facteur est : $CM_{Fac} = \frac{SC_{Fac}}{p-1}$.

le carré moyen résiduel est : $CM_{Res} = \frac{SC_{Res}}{n-p}$.

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \rightsquigarrow f_{(p-1,n-p)}.$$

Décision : Pour un seuil de risque donné α, les tables de Fisher nous fournissent une valeur critique f_α tel que :

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1-\alpha,$$

- si $f_{obs} < f_\alpha \implies$ on ne peut pas rejeter H_0 (le facteur n'a aucune influence sur le caractère étudier),
- si $f_{obs} \geq f_\alpha \implies$ on rejette H_0 (le facteur influe sur le caractère étudié),

avec f_{obs} est la réalisation de la variable (statistique) F_{obs}.

Les résultats d'une ANOVA 1 sont souvent présentés dans un tableau ayant la forme suivante :

	Somme des carrés	Degrés de libertés	Carré moyen	ratio	Fischer
source de variation	SC	ddl	CM	F_{obs}	c
Inter-groupe (Fac)	SC_{Fac}	$p-1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-groupe (Rés)	SC_{Res}	$n-p$	CM_{Res}		
Total	SC_{Tot}	$n-1$			

1.4.4 Exemple d'application

Reprenant l'exemple présenté dans la section 1.4.1. Les étapes qu'on doit suivre pour réaliser le test

$$H_0 : \text{"}\ \mu_1 = \mu_2 = \mu_3 = \mu\ \text{"} \text{ contre } H_1 : \text{"}\ \exists\ i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j\ \text{"},$$

à l'aide de la technique ANOVA 1, sont les suivantes :

- Calculer les moyennes des différents échantillons : $\overline{X}_1 = 24.73$, $\overline{X}_2 = 21.53$ et $\overline{X}_3 = 23.60$.
- Calculer la moyenne globale de toutes les observations : $\overline{X} = \frac{1}{n}(n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3) = 23.2889$.
- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe	31.5911	2	15.7956	12.02	3.6823
Intra-groupe	19.7067	15	1.3138		
Total	51.2978	17			

- Décision : on constate que $f_{obs} = 12.02 > f_\alpha = 3.6823$ (pour un risque de $\alpha = 5\%$), donc les espaces moyens occupés par les informations sont significativement différents d'une bases de données à une autre. Cela signifié que le facteur bases de données influe sur l'espace mémoire occupé par les informations stockées.

Conclusion

A partir les différentes notions et différents tests exposés dans ce chapitre on peut conclure que :

Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations.

Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard, ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

Les tests d'hypothèses font appel à un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, indépendance, etc.) et qui doivent être vérifié préalablement.

Bibliographie

[1] J. Bass, *Eleménts de calcul de probabilités.* Masson, 1974.

[2] N. Ben Righi, M. Cherfaoui *Estimation paramétrique : Intervalle et région de confiance.* Mémoire Master en Mathématique Option Statistique, Université de Biskra, 2016.

[3] G. Calot, *Cours de calcul des probabilités.* Dunod, 1967.

[4] D. Foudrinier, *Statistique inférentielle : Cours et exercices.* Dunod, Paris 2002.

[5] H. Gudeida, A. Roubi *Tests de comparaison.* Mémoire Master en Mathématique Option Statistique, Université de Biskra, 2016.

[6] J. Guégand, J. P. Gavini, *Probabilités.* 1998.

[7] K. Khaldi, *Méthodes statistique et Probabilités.* Casbah, 2000.

[8] A. Krief, S. Levy, *Calcul des probabilités.* Hermann, 1972.

[9] M. Laviéville, *Statistique et Probabilités : Rapepels de cours et exercuces résolus.* Dunod, 1996.

[10] J.P. Lecoutre, S. Legait, P. Tassi, *Statistique : Exercices corrigés et rappels de cours.* Masson, 1987.

[11] J.P. Lecoutre, *Statistique et probabilité, manuel et exercices corrigés.* quatrième édition. Masson, 2009.

[12] M. Sheldon, M. Ross, *Initiation aux probabilités.* Presses polytechniques et universitaires normandes, 1994.

[13] G. Saporta, *Probabilité, analyse des données et statistique.* Editions Technip, 1990.

[14] P. Tassi, *Méthodes statistiques.* Edition Economica, 2004.