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 DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

BELAID ROUMAISSA

Titre :

Sur le Test Statistique "t"

Membres du Comité d'Examen :

Dr.	ABDELLI JIHANE	UMKB	Président
Dr.	**DHIABI SAMRA**	UMKB	Encadreur
Dr.	BERKANE HASSIBA	UMKB	Examinateur

September 2020

Dédicace

Je dédie ce travail à :

mon encadreur **DHIABI SAMRA** pour m'avoir guidé .

L'homme de ma vie et mon exemple éternel , à mon père que Dieu le garde dans son vaste paradis.

ma vie et mon bonheur à ma mère que Dieu la garde et la protège.

ma sœur SALSSABILE , et mes chers frères DJAMEL,IHAB ,MOHAMED et à ma deuxième famille MOKRANI.

Aux personnes qui m'ont toujours encouragés ,à mes amis.

Et finalement mes dédicaces pour l'ensemble de mes professeurs.

REMERCIEMENTS

Je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux,qui m'a donné la force et la patience d'accomplir ce mémoire.

En second lieu, je tiens à remercier mon encadreur Madame : **DHIABI SAMRA** pour ses précieux conseils et son aide durant toute la période du travail.

Mon vif remerciements vont également aux membres du jury pour l'intéret qu'il sont poté à ma recherche en acceptant d'examiner mon travail et de l'enrichir par leurs remarques.

Mes remerciements vont également à tous les enseignants et les responsables de ma faculté.

Enfin, je tiens également à remercier ma famille qui m'a toujours soutenue et toutes les personnes qui ont participé de prés ou de loin à la réalisation de ce travail. Merci.

Table des matières

Table des figures

Liste des tableaux

Introduction

La distribution **t** (**Student distribution**) est l'une des distributions d'échantillonnage importante qui ont de large applications dans les statistiques déductives, et surtout lorsque la taille de l'échantillon est petite. Tout le mérite en revient incontestablement aux publications du scientifique irlandais William Sealy Gosset 1908, quand il a publié ses recherches dans lesquelles cette distribution a été dérivée sous le nom **Student** et donc Connue sous le nom $\mathbf{t-Student}$. La distribution de t est similaire à la distribution normale , elle est également sous forme de cloche et symétrique autour de la moyenne arithmétique zéro, mais elle est plate ou plus plate que la distribution normale. Par conséquent, une plus grande partie de son surface est située sur les bords, et que la distribution de t est différente pour chaque taille d'échantillon n ,Mais avec l'augmentation de cela, la distribution de t est proche d'une distribution normale jusqu'à prèsque égale.

La présente étude vise à atteindre les objectifs suivants :

1. Définir la distribution de t et étudier certaines de ses propriétés théoriques.
2. Démontrer comment utiliser la distribution t dans un test autour d'une moyenne.
3. Expliquer comment la distribution t est utilisée dans le test de différence entre les deux moyennes de deux échantillons indépendants.
4. Démontrer comment la distribution t est utilisée pour tester la différence entre

les deux moyennes de deux échantillons dépendants.

5. Expliquez comment la distribution t est utilisée pour créer des intervalles de confiance pour une moyenne.

L'importance de cette recherche réside dans l'identification d'une des distributions d'échantillonnage importantes, qui est la distribution t, qui a de nombreuses utilisations pour identifier ses propriétés théoriques et comment elle est utilisée pour traiter les tests d'hypothèse et créer des intervalles de confiance en terme et comment choisir la formule appropriée pour un échantillon de variance inconnue et aussi pour deux échantillons indépendants. Ils sont liés à l'échantillon d'étude ainsi qu'à l'acquisition de l'habileté d'utiliser le programme statistique R, de le lire et d'interpréter ses résultats.

Ce méméroire comprend trois chapitres divisés comme suit :

– **Le premier chapitre :** Dans ce chapitre nous rappelons un certain nombre de généralités autour des tests d'hypothèse statistiques.
– **Le deuxième chapitre :** Dans ce chapitre nous énonçons les notions de base sur la distribution **t** et ses propriétés statistique. Nous détaillons dans le même chapitre le test **t** associé à cette distribution pour une seule moyenne, aussi pour la différence entre deux moyennes de deux échantillons indépendantes et dépendantes.
– **Le troisième chapitre :** On va présenter quelques tests pour affirmer les notions donnés en chapitre 2. Nous utilisons le logiciel R comme un logiciel statistique pour montrer les résultats obtenus.

Chapitre 1

Généralités

On s'intérèsse à la définition de la distribution t et de ses propriétés théoriques et de ses utilisations. Mais avant de discutersur cette distribution, il est nécessaire de présenter quelques concepts de base qui constituent le principal pilier de l'inférence statistique qui nous permet de l'utiliser et de le comprendre dans les prochaines sujets, que ce soit du côté théorique.ou pratique

Hypothèse statistique

Une hypothèse statistique est un énoncé concernant les caractéristiques(valeurs des paramètres, forme de la distribution des observations) d'une population.

1.1 Test d'hpothèse

Un test d'hypothèse (ou test statistisque) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultas d'échantillon, de faire un choisx entre deux hypothèses statistiques.

1.1.1 Hypothèse nulle (H_0) et hypothèse alternative (H_1)

L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle l'**hypothèse nulle** et erst notée H_0. N'importe qu'elle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle l'**hypothèse alternative**(ou contre-hypothèse) et est notée H_1.

1.1.2 Seuil de signification du test

Le risque, consenti à l'avance et que nous notons α de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie, s'appelle le **seuil de signification** du test et s'énonce en probabilité ainsi :

$$\alpha = P\left(\text{rejeter } H_0 \mid H_0 \text{ vraie}\right).$$

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une **région de rejet** de l'hpothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité α. Si par exemple, on choisit $\alpha = 0.05$, cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de rejet de H_0, bien que H_0 soit vraie et ceci uniquementd'après le hasard de l'echantillonnage.

Sur la distribution d'echantillonnnage correspondra aussi une région complémentaire, dite **région d'acceptation** de H_0 (ou région de non-rejet) de probabilité $1 - \alpha$

Les seuils de signification les plus utilisés sont $\alpha = 0.05$ et $\alpha = 0.01$, dépendant des conséquences de rejeter à tort l'hypothèse H_0.

1.1.3 Exemple de formulation d'un Test

Supposons que nous affirmions que la valeur d'un paramètre θ d'une population est égale à la valeur θ_0. On s'interesse au changement possible du paramètre θ dans l'une ou l'autre direction (Soit $\theta > \theta_0$ soit $\theta < \theta_0$). On effectue un test bilatéral

Les hypothèses H_0 et H_1 sont alors :

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$$

On peut schématiser les régions de rejet et de non-rejet de H_0 comme suit :

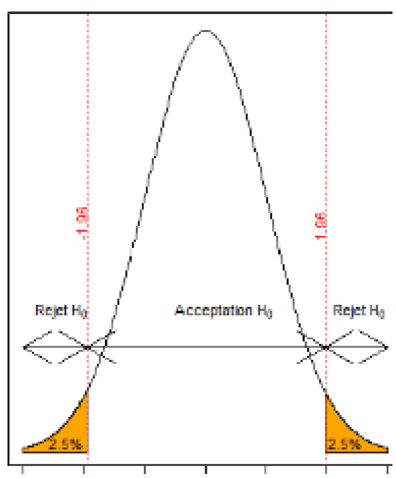

FIG. 1.1 – Le test bilatéral

Si, suite aux résultats de l'echantillon, la valeur de la statistique utilisée se situe dans l'intervalle $[\theta_{c_1}, \theta_{c_2}]$, on acceptera H_0 au seuil de signification choisi. Si, au contrraire, la valeur obtenue est supérieur à θ_{c_2} ou inférieur à θ_{c_1}, on rejette H_0 et on accepte H_1.

1.1.4 Risques de première et de deuxième espèce

Tous les règles de décision que nous avons déterminées acceptaient un risque α qui était le risque de rejeter à tort l'hypothèse H_0, c'est-à-dire le risque de rejeter l'hypothèse H_0, alors que H_0 est vraie. Ce risque s'appelle aussi le **risque de première espèce**.

La règle de décision du test comporte également un deuxième risque, à savoir de celui de ne pas rejeter l'hypothèse nulle H_0 alors que c'est l'hypothèse H_1 qui est vraie. C'est le **risque de deuxième espèce.**

Les deux risques peuvent se définir ainsi :

$\alpha = P($rejeter $H_0 \mid H_0$ vraie$)$=probabilité de commettre une erreur de première espèce.

$\beta = P($ne pas rejeter $H_0 \mid H_1$ vraie$)$=probabilité de commettre une erreur de deuxième espèce.

Le risque de première espèce α est choisi à priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative H_1 et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse H_1 que l'on suppose vraie.

Les risques liés aux tests d'hypothèses peuvent se résumer ainsi :

décision /réalité	H_0 est vraie	H_0 est fausse
accepter H_0	$1-\alpha$	erreur de type II (β)
rejeter H_0	erreur de type I (α)	$1-\beta$ (puissance)

Tab. 1.1 – Tableau résumant les espèces d'erreurs

La probabilité complémentaire du risque de deuxième espèce $(1-\beta)$ définit la **puissance du test** à l'égard de la valeur du paramètre dans l'hypothèse alternative H_1. La puissance du test représente la probabilité de rejeter l'hypothèse nulle H_0 lorsque l'hypothèse vrie est H_1. Plus β est petit, plus le test est puissant.

1.1.5 P-Valeur

La p-valeur est le plus petit réel $\alpha \in]0,1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $(100\alpha\%)$.Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejtant H_0 alors que H_0 est vrais.

1.1.6 Démarche d'un test statistique

Les étapes à suivre pour tester une hypothèse sont les suivantes :

(1)-définir l'hypothèse nulle (notée H_0) à contrôler.

(2)-choisir un test statistique ou une statistique pour contrôler H_0.

(3)-définir la distribution de la statistique sous l'hypothèse « H_0 est réalisée ».

(4)-définir le niveau de signification du test ou région critique notée α.

(5)-calculer à partir des données fournies par l'échantillon, la valeur de la statistique.

(6)-prendre une décision concernant l'hypothèse posée et conclure

Chapitre 2

Test de Student

Test-t de Student est un **test statistique** permettant de **comparer les moyennes** de deux groupes d'échantillons. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes au point de vue **statistique**.

Il existe plusieurs variants du **test-t de Student** :

– Le test-t de Student pour **échantillon unique**

Le test-t de Student comparant deux groupes **d'échantillons indépendants** (on parle de **test de Student non apparié**)

Le test-t de Student comparant deux groupes **d'échantillons dépendants** (on parle de test de **Student apparié**).

Ce chapitre a pour objectif de décrire les formules pour les différents types de **test de Student**. Le test de Student est dit **paramétrique** car, comme nous allons le voir, la formule dépend de la **moyenne** et de **l'écart-type** des observations à comparer.

2.1 Définition et propriétés de la loi de Student

Soit Z une variable aléatoire suit la loi normale standard $Z \sim N(0,1)$, et soit U une variable indépendante de Z et distribuée suivant la loi du χ^2 à k degrés de liberté Par définition, la variable :

$$T = \frac{Z}{\sqrt{U}}$$

suit une loi de Student à k degrés de liberté.

$U = \sum_{i=1}^{k} X_i^2$, alors $U \sim \chi_k^2$, où les X_i sont k variables aléatires i.i.d de loi normale centrrée réduite.

La densté de de T, notée f_T , est donnée par

$$f_T(t) = \left\{ \frac{\Gamma \frac{k+1}{2}}{\Gamma \frac{k}{2} \sqrt{k\pi}} \left[1 + \frac{t^2}{k} \right]^{-\frac{k+1}{2}} \right\}, \quad -\infty < t < +\infty$$

où Γ est la fonction Gammer d'Euler.

La densité f_T associée à la variable T est symétrique, centrée en 0 et en forme de cloche.

2.1.1 Espérance et variance de la loi de Student

$$E(T) = E\left[\frac{Z}{\sqrt{\frac{U}{k}}}\right] = E(Z).E\left[\frac{1}{\sqrt{\frac{U}{k}}}\right] = E(1).E(\sqrt{\frac{k}{U}}) = 0.E(\sqrt{\frac{k}{U}}) = 0$$

donc

$$E(T) = 0$$

$$E(T^2) = E\left[\frac{Z}{\sqrt{\frac{U}{k}}}\right]^2 = E(Z^2).E\left[\frac{1}{\sqrt{\frac{U}{k}}}\right]^2 = E(Z^2).E(\frac{k}{U}).$$

$$var(Z) = E(Z^2) - (E(Z))^2$$
$$E(Z^2) = 1 + 0 = 1$$

donc

$$E(T^2) = 1.E(\frac{k}{U}) = kE(\frac{1}{U})$$

$$\begin{aligned}
E(\frac{1}{U}) &= \int_0^{\infty} \frac{1}{u} f(u)du \\
&= \int_0^{\infty} \frac{1}{u} \frac{1}{\Gamma_{\frac{k}{2}} 2^{\frac{k}{2}}} x^{\frac{k}{2}-1} e^{-\frac{u}{2}} du \\
&= \frac{1}{\Gamma_{\frac{k}{2}} 2^{\frac{k}{2}}} \int_0^{\infty} x^{\frac{k}{2}-2} u^{-\frac{x}{2}} du \\
&= \frac{1}{\Gamma_{\frac{k}{2}} 2^{\frac{k}{2}}} \Gamma_{\frac{k}{2}-1} 2^{\frac{k}{2}-1} \\
&= \frac{2^{-1}}{\frac{k}{2}-1} = \frac{2^{-1}}{\frac{k-1}{2}} = \frac{1}{2(\frac{k-2}{2})} = \frac{1}{k-2}
\end{aligned}$$

Alors

$$E(T^2) = n.(\frac{1}{n-2})$$

Par conséquent

$$var(T) = E(T^2) - E(T) = \frac{n}{n-2} \quad n > 2$$

2.1.2 Dérivé première et seconde de la densité d'une distribution "t"

$$f_T(t) = \frac{\Gamma_{\frac{k+1}{2}}}{\sqrt{k\pi}\Gamma_{\frac{k}{2}}} \left[1 + \frac{t^2}{k}\right]^{-\left[\frac{k+1}{2}\right]} \qquad -\infty < t < +\infty$$

$$\Rightarrow Ln f_T(t) = Ln(c) - \frac{(n-1)}{2} Ln \left[1 + \frac{t^2}{n}\right],$$

où

$$c = \frac{\Gamma_{\frac{k+1}{2}}}{\sqrt{k\pi}\Gamma_{\frac{k}{2}}},$$

donc

$$\begin{aligned}
\frac{f\prime_{T(t)}}{f_T(t)} &= 0 - \frac{k+1}{2} \frac{1}{\left[1 + \frac{t^2}{k}\right]} . \frac{2t}{n} \\
&= 0 - n + 1 . \frac{k}{(t^2 k + 1)} . \frac{t}{k} \\
&\Rightarrow f\prime_T(t) = -f_T(t)(n+1)\frac{t}{(n+t^2)}
\end{aligned}$$

$$\begin{aligned}
f_T^{"}(t) &= -(k+1)\left[f_T(t)\frac{(k+t^2) - t(2t)}{(k+t^2)^2} + \frac{t}{(k+t^2)} f_T\prime(t)\right] \\
&= -(k+1)\left[f_T(t)\frac{(k+t^2) - t(2t)}{(k+t^2)^2} + \frac{-t}{(k+t^2)} f_T(t)(k+1)\frac{t}{k+t^2}\right] \\
&= -(k+1) f_T(t) \left[\frac{k - t^2}{(k+t^2)^2} + \frac{k}{(k+t^2)}\right] \\
&= -(k+1) f_T(t) \left[\frac{(k+t^2) - 2t^2 - t^2 k - t^2}{(k+t^2)^2}\right] \\
&= -(k+1) f_T(t) \left[\frac{k - 2t^2 - kt^2}{(k+t^2)^2}\right].
\end{aligned}$$

2.1.3 Comportement limite

Lorsque k est grand, la loi de Student peut être approchée par la loi normale centrée réduite. Une manière simple de le démontrer est d'utiliser le lemme de Scheffé.

2.1.4 La loi de student dans l'echantillonnage

Soient X_1,...,X_n, n variables mutuellement indépendantes et distribuées suivant une même loi normale $N(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 qui correspondent à un échantillon de taille n. Considérons la moyenne empirique

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

et l'estimateur sans biais de la variance

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

Par normalisation, la variable aléatoire

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

suit une loi normale standard(d'espérance 0 et de variance 1). La variable obtenue en rempalçant σ par S dans Z est

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}},$$

T sui la loi de student à $n - 1$ degrés de liberté. Ce résultat est utile pour trouver des intervalles de confiance quand σ^2 est inconnue, comme indiqué plus bas.

Pour justifier cela, onintroduit la variable aléatoire

$$U = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2,$$

qui permet d'ecrire

$$\frac{S^2}{\sigma^2} = \frac{U}{n-1}$$

et

$$T = \frac{\overline{X} - \mu}{\sigma \diagup \sqrt{n}} . \frac{1}{S \diagup \sigma} = \frac{Z}{\sqrt{U \diagup n-1}}.$$

Pour terminer il faut montrer que Z et U sont indépendantes et que U suit une loi du χ^2 à $n-1$ degrés de liberté.

2.2 Test t-student de moyennes

2.2.1 Test de Student pour échantillon unique

Nous sommes danns le cadre d'un échantillon issu d'une loi $N(\mu, \sigma^2)$. Cette fois l'écart-type est inconnu. On veut tester les hypothèses (version unilatérale ici) :

$$H_0 : \mu \leq \mu_0$$

contre

$$H_1 : \mu > \mu_0.$$

Le test repose sur le calcul de la moyenne $\overline{X}$ et de l'écart-type S de l'chantillon puis de la statistique

$$T = \frac{\overline{X} - \mu_0}{S \diagup \sqrt{n}}.$$

Où S est l'estimateur de la variance définit précédement.

Sous l'hypothèse nulle, cette statistique Z suit une loi de Student à $n-1$ degrés de liberté. On peut donc définir la région critique comme étant

$$W = \{T > qt(1-\alpha, n-1)\}$$

où $qt(1-\alpha, n-1)$ est le quantile correspondant.

On a par exemple pour $n = 25$ et le niveau conventionnel $\alpha = 0.05$. La région critique $=\{T > 1.711\}$. On rejettera l'hypothèse nulle si la valeur de la statistique T calculée sur l'échantillon appartient à cet ensemble.

2.2.2 Test "t" de Student pour échantillons indépendants

Dans ce cas de figure, il s'agit de comparer deux moyennes observées. Lorsque les deux groupes d'échantillons à comparer n'ont aucun lien, on utilise le **test t de Student indépendant** (ou non **apparié**).

A titre d'exemple, nous avons un groupe de 100 individus (50 femmes et 50 hommes) pris au hasard au sein de la population. On se pose la question à savoir si le poids moyen des femmes est significativement différent de celui des hommes ?

Dans cet exemple on parle de test de Student non apparié car les deux groupes à comparer n'ont aucun lien. Il s'agit donc de calculer le poids moyen des femmes et de celui des hommes et d'évaluer si la différence est significative au point de vue statistique.

Soient X_1 et X_2deux variables aléatoires indépendantes de même lois normales de moyennes μ_1 et μ_2 et d'écart types σ_1 et σ_2. On dispose de deux échantillons indépendants $\left\{X_1^{(1)}, ..., X_{n1}^{(1)}\right\}$ et $\left\{X_2^{(2)}, ..., X_{n2}^{(2)}\right\}$ tels que $X_i^{(1)}$ (resp. $X_i^{(2)}$) suit la même loi que X_1 (resp. X_2).

Sachant les échantillons, on cherche à décider si les moyennes μ_1 et μ_2 sont significativement différentes ou non. On test alors

$$H_0 : \mu_1 = \mu_2$$

contre

$$H_1 : \mu_1 \neq \mu_2$$

au risque α.

Si les écart types σ_1 et σ_2 sont connus :

On calcule

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

On rejette H_0 au risque α si $T \notin \left[-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}\right]$ où la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans la table de la loi normales centrée réduite.

Si les écart types σ_1 et σ_2 sont inconnus et $(n_1 > 30$,$n_2 > 30)$:

On calcule

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}}}$$

On rejette H_0 au risque α si $T \notin \left[-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}\right]$ où la valeur $t_{1-\frac{\alpha}{2}}$ est lue dans la table de la loi normales centrée réduite.

Si n_1 ou n_2 est inférieur à 30 et $\sigma_1 = \sigma_2$:

On calcule

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\widehat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où

$$\widehat{\sigma} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

On rejette H_0 au risque α si $T \notin \left[-t_{1-\frac{\alpha}{2};n_1+n_2-2}, t_{1-\frac{\alpha}{2};n_1+n_2-2}\right]$ où la valeur $t_{1-\frac{\alpha}{2};n_1+n_2-2}$ est lue dans la table de Student à $n_1 + n_2 - 2$ degrés de liberté.

2.2.3 Test-t de Student pour séries appariés

Le **test de Student apparié** permet de comparer la moyenne de deux séries de valeurs ayant un lien.

Par exemple, 20 souris ont reçu un traitement X pendant 3 mois. On se pose la question à savoir si le traitement X a un impact sur le poids des souris au bout des 3 mois. Le poids des 20 souris a donc été mesuré avant et après traitement. Ce qui nous donne 20 séries de valeurs avant traitement et 20 autres séries de valeurs après traitement provenant de la mesure du poids des mêmes souris.

Il s'agit bien dans cet exemple, d'un **test de Student apparié** car les deux séries de valeurs ont un lien (les souris). Pour chaque souris, on a deux mesures (l'une avant et l'autre après traitement).

Notons par :

d : différence individuelle entre les deux valeurs d'une paire.

μ_d : Valeur moyenne des différences d pour la opulation de tiutes les paires.

$\overline{d}$: valeur moyenne des différences.

S_d : ecart type des différences d pour les données appariées de l'échantillon.

n : nombre de paires.

On calcule

$$T = \frac{\overline{d}}{\frac{S_d}{\sqrt{n}}}$$

On rejette H_0 au risque α si $T \notin \left[-t_{1-\frac{\alpha}{2};n_1+n_2-2}, t_{1-\frac{\alpha}{2};n_1+n_2-2}\right]$ où la valeur $t_{1-\frac{\alpha}{2};n_1+n_2-2}$ est lue dans la table de Student à $n_1 + n_2 - 2$ degrés de liberté.

2.3 Application : intervalle de confiance associé à l'espérance d'une variable de loi normale de variance inconnue

Dans cette partie, on présente une méthode pour déterminer l'**intrvalle de confiance** de l'espérance μ d'une loi normale. Notons que si la variance est connue, il vaut mieux utiliser directement la loi normale avec la moyenne $\overline{X}$.

étant donné un risque α

$$P\left(-t_{\frac{\alpha}{2}}^{n-1} < \frac{\overline{X} - \mu}{S \diagup \sqrt{n}} < t_{\frac{\alpha}{2}}^{n-1}\right) = 1 - \alpha.$$

L'intervalle de confiance bilatéral de μ au niveau de confiance $1 - \alpha$ est donnée par

$$\left[\overline{X} - t_{\frac{\alpha}{2}}^{n-1}, \overline{X} + t_{\frac{\alpha}{2}}^{n-1}\right]$$

avec $\overline{X}$, l'estimateur ponctuel de l'espérance et S, l'estimateur non biaisé de la variance définis ci-dessus. t_{γ}^{k} est le quantile d'ordre $1-\gamma$ de la loi de student à k degrés de liberté, c'est l'unique nombre qui vérifie

$$P(T \leq t_{\gamma}^{k}) = 1 - \gamma$$

lorsque T suit la loi de student à k degrés de liberté.

Démonstration :

Posons à nouveau

$$T = \frac{\overline{X} - \mu}{S \diagup \sqrt{n}}$$

Nous avons vu que T suit une loi de Student à $n-1$ degrés de liberté. Avec la symétrie et la continuité de la loi nous avons

$$\begin{aligned}
P(-t \leq T \leq t) &= P(T \leq t) - P(T < -t) \\
&= P(T \leq t) - P\,(T > t) \\
&= 2P(T \leq t) - 1.
\end{aligned}$$

En particulier

$$P\left(-t_{\frac{\alpha}{2}}^{k} < \frac{\overline{X} - \mu}{S \diagup \sqrt{n}} < t_{\frac{\alpha}{2}}^{k}\right) = 2(1 - \frac{\alpha}{2}) - 1 = 1 - \alpha$$

cela donne la probabilité cherchée. L'intervalle est donné par

$$-t_{\frac{\alpha}{2}}^{n-1} < \frac{\overline{X} - \mu}{S \diagup \sqrt{n}} < t_{\frac{\alpha}{2}}^{n-1} \Leftrightarrow \overline{X} - t_{\frac{\alpha}{2}}^{n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{\alpha}{2}}^{n-1} \frac{S}{\sqrt{n}}$$

Voici les tailles mésurées en cm sur un échantillon de 8 personnes :

On en calcule la moyenne statistique $\overline{x}$ et la variance sans biais S^2

i	1	2	3	4	5	6	7	8
x_i	155	160	161	167	171	177	180	181

TAB. 2.1 – les tailles mésurées en cm

$$\overline{X} = 169, \ S^2 = 96,857142$$

Prenons un risque $\alpha = 5\%$, donc un niveau de confiance $1 - \alpha = 95\%$. Aux arrondis près, le tableau des quantiles ci dessous donne $t^7_{5\%} = 1,895$, et l'intervalle de confiance est

$$\left[\overline{X} - t^7_{5\%}\frac{S}{\sqrt{8}}, \overline{X} + t^7_{5\%}\frac{S}{\sqrt{8}}\right] = [162,4; 175,6] \, .$$

La probabilité que la taille moyenne de la population soit dans cet intervalle est de 95%. Or la taille moyenne des français est de 177 cm, mais 177 n'appartient pas à cet intervalle de confiance, on peut alors dire que cet échantillon ne correspond pas à la population française, avec 10% d'erreur. C'est un exemple d'application du test de Student.

Chapitre 3

Application du test t avec R

Ce chapitre décrit comment faire un **test t** à un seul échantillon et à deux échantillons indépendantes et appariés dans R (ou dans R studio). Pour vérifier si la distribution des données est normale, on utilise soit le Test de Kolmogorov-Smirnov ou bien le test de Shapiro-Wilk.

3.1 Test de dépendance de données pour une distribution normale

3.1.1 Test Kolmogorov - Smirnov

Le test de Kolmogorov-Smirnov est un test d'hypothèse utilisé pour décider si un échantillon suit une loi de probabilité donnée ou si deux échantillons suivent la même loi.

Sous R on peut réaliser ce test avec la fonction :

$$ks.test()$$

Exemple 3.1.1

Un test a été étalonné sur une population A de manière que sa distribution suive une loi normale de moyenne 13 et d'écart type 3. sur un échantillon de taille $n = 10$ issu d'une population B, on a observé les valeurs suivantes :

8.43	8.70	11.27	12.92	13.05	13.05	13.17	13.44	13.89	18.90

TAB. 3.1 – Les données d'un échantillon de taille n= 10

Ces valeurs sont-elles compatibles avec l'hypothèse selon laquelle la variable sous-jacente est distribuée selon une loi normale de moyenne 13 et d'écart type 3 ?

On peut utiliser la commande ks.test du package "stats" .

code R :

$$> X1 < -c(8.43, 8.70, 11.27, 12.92, 13.05, 13.05, 13.17, 13.44, 13.89, 18.90)$$

$$> ks.test(X1, "pnorm", mean = 13, sd = 3)$$

Résultat de la commande :

```
One-sample Kolmogorov-Smirnov test

data : X1

D = 0.28336, p-value = 0.3982

alternative hypothesis : two-sided

Warning message :

In ks.test(X1, "pnorm", mean = 13, sd = 3) :

 aucun ex-aequo ne devrait être présent pour le test de Kolmogorov-Smirnov
```

Le commentaire :

Le $p - value$ non significative car $(p - value = 0.3982 > 0.05)$.

L'échantillon est distribuée selon une loi normale .

3.1.2 Test de Shapiro-Wilk

Le test de Shapiro-Wilk est un test permettant de savoir si une série de données suit une loi normale.si $p - value > 0.05$ donc Les données sont réparties selon la distribution normale.

Exemple 3.1.2

Sur un échantillon de taille $n = 10$, on a observé les valeurs suivantes d'une VD numérique :

8	9	9	10	10	10	11	13	14	14

TAB. 3.2 – les valeurs d'une VD numérique

Est-il légitime de supposer que la distribution de la VD dans la population parente suit une loi normale ?

Le test de Shapiro-Wilk est disponible dans le package "stats". La fonction correspondante est "shapiro.test".

<u>code R :</u>

```
>X2 <- c(8, 9, 9, 10, 10, 10, 11, 13, 14, 14)

>shapiro.test(X2)
```

<u>Résultat de la commande :</u>

```
Shapiro-Wilk normality test

data : X2

W = 0.88491, p-value = 0.1485
```

Le commentaire :

L'exemple ci-dessus renvoie une $p - value$ non significative car $(p - value = 0.1485 > 0.05)$.L'échantillon suit donc une loi normale.

Exemple 3.1.3

Code R :

```
>shapiro.test(runif(100,min=2,max=4))
```

Résultat de la commande :

```
shapiro-wilk normality test

data : runif(100,min=2,max=4)

w=0.94719,p-value=0.0005443
```

Le commentaire :

Dans l'exemple 2, ci-dessus, la p-value est significative car $(p - value = 0.0005443 < 0.05)$.

L'échantillon ne suit donc pas une loi normale.

3.2 Test t-student de moyennes

3.2.1 Test " t "pour un échantillon

Le test de Student pour échantillon unique (appelé $one - samplet - test$en anglais) permet de comparer une moyenne observée à une moyenne théorique.

Soit x un vecteur contenant une série de valeurs. Pour comparer la moyenne de x à une moyenne théorique mu, le code R est de la forme :

$$t.test(x, mu = 0)$$

Exemple 3.2.1

On a pesé une dizaine de souris, On se pose la question à savoir si leur poids moyen est significativement différent de 200g ?

L'hypothèse H_0 est : la moyenne $m = 200$

Leurs poids en gramme est affiché ci-dessous :

les souris	Poids (g)
souris_1	355.20
souris_2	533.30
souris_3	630.10
souris_4	218.50
souris_5	551.50
souris_6	560.70
souris_7	431.00
souris_8	434.40
souris_9	432.30
souris_10	393.20

TAB. 3.3 – Poids des souris en gramme

Code R :

```
> # poids des souris

> x<-c(442.7, 380.2, 406.8, 507.7, 615.1, 486.8, 438.7, 390.7, 399.5, 789.9)

> # test de student pour échantillon unique

> res<-t.test(x, mu=200)

> res # Affichage du résultat du test
```

Résultat :

```
One Sample t-test

data : x

t = 7.0567, df = 9, p-value = 5.942e-05

alternative hypothesis : true mean is not equal to 200

95 percent confidence interval :

 394.1884 577.4316

sample estimates :

mean of x

 485.81
```

Dans le résultat ci-dessus : t est la statistique de Student ($t = 7.0567$), df est le degré de liberté ($df = 9$), $p - value$ est le degré de significativité du test ($p - value = 5.942 \times 10 - 5$).

L'intervalle de confiance de la moyenne à 95%est également montrée (intervalle de confiance= $[394.1884, 577.4316]$) ; et enfin, on a la valeur moyenne de la série x ($moyenne = 485.81$).

Le commentaire :

La $p - value$ du test est de 5.942×10^{-5}. Ce qui est largement inférieur à 0.05. On rejette l'hypothèse H_0 et on conclut que le poids moyen des souris est significativement différent de 200g avec une $p - value = 5.942 \times 10^{-5}$.

3.2.2 Test "t "pour deux échantillons indépendants(non-apparié)

Le test de student non-apparié permet de comparer deux groupes d'échantillons qui n'ont aucun lien (donc indépendants).Un format simplifié de la fonction R à utiliser est :

$$t.test(x, y)$$

Exemple 3.2.2

Pour illustrer l'utilisation du test de student non-apparié, on s'intéresse à savoir si le poids moyen des femmes est différents de celui des hommes. Pour cela, à titre d'exemple, on a pesé 10 femmes et 10 hommes pris au hasard dans la population. Leurs poids en kilogramme (kg) est montré dans le tableau ci-dessous. Le nombre d'individus considérés ici est bien évidemment faible.

C'est juste pour illustrer le test de student.

Le poids moyen des femmes est-il significativement différent de celui des hommes ?

A partir de la table de données ci-dessus vous avez deux méthodes pour faire le test de student en fonction de la structure des données.

/	Group	Poids (kg)
1	Femme	42.10
2	Femme	53.80
3	Femme	30.00
4	Femme	45.80
5	Femme	57.70
6	Femme	59.20
7	Femme	82.40
8	Femme	66.20
9	Femme	66.90
10	Femme	51.20
11	Homme	80.70
12	Homme	85.10
13	Homme	88.60
14	Homme	81.70
15	Homme	69.80
16	Homme	79.50
17	Homme	107.20
18	Homme	69.30
19	Homme	80.90
20	Homme	63.00

TAB. 3.4 – Poids moyen des femmes est différents de celui des hommes

1) Méthode 1 - Les données sont stockées dans deux vecteurs différents (x et y) :

Code R :

```
> # poids des femmes

> x<- round(rnorm(10, mean=57, sd=15), 1)

> x

 [1]53.9 24.0 55.9 27.8 37.5 56.8 47.0 31.7 76.2 49.1

> # Poids des hommes

> y <- round(rnorm(10, mean=75, sd=15), 1)

> y

 [1] 49.6 56.1 63.4 91.5 85.7 79.7 94.0 79.2 53.5 64.3
```

Dans ce cas le test de student peut être effectué avec le code R suivant :

$$res < -t.test(x, y)$$

Résultat de la commande :

```
Welch Two Sample t-test

data : x and y

t = -3.564, df = 17.99, p-value = 0.00222

alternative hypothesis : true difference in means is not equal to 0

95 percent confidence interval :

 -40.87 -10.55

sample estimates :

mean of x mean of y

 45.99 71.70
```

2) **Méthode 2 - Les données sont stockées dans une table de type data.frame :**

Code R :

```
> d<-as.data.frame(list(
+                group=c(rep("Femme", 10), rep("Homme", 10)),
+                poids=c(x, y)
+                ))
> head(d)
```

Résultat de la commande :

```
  group poids
1 Femme 63.8
2 Femme 30.5
3 Femme 60.2
4 Femme 60.8
5 Femme 32.1
6 Femme 45.3
```

Dans cette configuration, le test de student peut être effectué en utilisant le code R suivant :

```
> #res<-t.test(d$poids ~d$group)
> res<-t.test(poids ~group, data=d)
> res
```

Résultat de la commande :

```
Welch Two Sample t-test
data : poids by group
t = -3.564, df = 17.99, p-value = 0.00222
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
 -40.87 -10.55
sample estimates :
mean in group Femme mean in group Homme
              45.99               71.70
```

Dans le résultat ci-dessus : t est la statistique de student ($t = -3.564$), df est le degré de liberté ($df = 17.99$), p-value est le degré de significativité du test ($p - value = 0.0022$).

L'intervalle de confiance de la différence des moyennes à 95% est également montrée (intervalle de confiance= $[-40.87, -10.55]$).

et enfin, on a la valeur moyenne des deux groupes (poids moyen des femmes = 45.99, poids moyen des hommes =71.70).

Le commentaire :

La p-value du test est de 0.0022. Ce qui est largement inférieur à0.05.

On conclut que le poids moyen des femmes est significativement différent de celui des hommes avec une p-value = 0.0022.

3.2.3 Test "t" pour deux échantillons appariés

Le test de Student apparié permet de comparer les moyennes de deux séries de valeurs présentant un lien.

Un format simplifié de la fonction R à utiliser est :

$$t.test(x, y, paired = TRUE)$$

Exemple 3.2.3

Le tableau ci-dessous présentée 10 souris ont reçu un traitement X grossissant pendant 3 mois. Leur poids a été mesuré avant et après traitement. Pour chaque souris on a donc deux valeurs(l'une avant et l'autre après traitement).

/	poids_avant	poids_aprés
Souris_1	235.20	464.40
Souris_2	221.80	351.60
Souris_3	200.30	416.90
Souris_4	183.90	403.80
Souris_5	185.90	415.90
Souris_6	202.90	386.40
Souris_7	202.60	398.50
Souris_8	193.60	423.40
Souris_9	208.80	362.40
Souris_10	179.40	372.30

TAB. 3.5 – Poids a été mesuré avant et après traitement.

La question est de savoir si le poids des souris a significativement changé après les 3 mois de traitement ?

Il s'agit bien d'un test de Student apparié car les deux mesures à comparer proviennent des même souris.

Code R :

```
># Poids des souris avant traitement
>x<-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
># Poids des souris après traitement
>y<-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2)
>res<-t.test(x, y, paired=TRUE)
>res
```

Résultat :

```
Paired t-test
data : x and y
t = -20.88, df = 9, p-value = 6.2e-09
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
 -215.6 -173.4
sample estimates :
mean of the differences
             -194.5
```

Dans le résultat ci-dessus : t est la statistique de Student ($t = -20.88$), df est le degré de liberté ($df = 9$), p-value est le degré de significativité du test ($p - value = 6.2 \times 10^{-9}$). L'intervalle de confiance de la différence des moyennes à 95% est également montrée (intervalle de confiance= $[-215.56, -173.4]$) ; et enfin, on a la valeur moyenne de la différence des deux séries (moyenne de la différence $= -194.5$).

Le commentaire :

La p-value du test est de 6.2003×10^{-9}. Ce qui est largement inférieur à 0.05. On conclut que le poids moyen des souris avant traitement est significativement différent de celui après traitement avec une p-value $= 6.2003 \times 10^{-9}$.

Bibliographie

[1] Abdul Hamid Al-Baddawi. (2008). Méthodes statistiques appliquées.Dar Al-Shorouk pour publication et distribution.

[2] Ali Abdul Salam Al-Ammari et Ali Hassan Al-Ajili.(2000).Statistiques. propabilités théoriques et application - Dar EIGA.

[3] Cottet-Emard, F. (2014). Probabilités et tests d'hypothèses. De Boeck Superieur.

[4] CHAMPELY, S. (2006). Tests statistiques paramétriques : Puissance, taille d'e et et taille d'échantillon (sous R).

[5] JOLION, J. M. (2006). Probabilités et Statistique.

[6] LABARERE, J. Tests paramétriques de comparaison de 2 moyennes.

[7] Lejeune, M. (2004). Statistique : La théorie et ses applications. Springer Science & Business Media.

[8] Monbet, V. (2009). Tests statistiques Notes de cours. Cité en, 133.

[9] RIVOT, E. (2017). Notions d'Echantillonnage et incertitudes d'échantillonnage.

[10] Ruch,J.J.(2013). Statistique : tests d'hypothèses. Préparation à l'Agrégation Bordeaux, 11.

[11] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions Technip.

[12] Veysseyre, R. (2014). Aide-mémoire-Statistique et probabilités pour les ingénieurs. Dunod.

[13] Ycart, B. (2002). Estimation paramétrique, tests statistiques. Centtre de Publication Universitaire.

Annexe A : Logiciel *R*

3.3 Qu'est-ce-que le langage R ?

• Le langage **R** est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

• **R** a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team.

L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et Robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

R est un logiciel de traitement statistique des données. Il fonctionne sous la forme d'un interpréteur de commandes. Il dispose d'une bibliothèque très large de fonctions statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des "packages", des modules externes compilés (sous forme de DLL sous Windows) que l'on peut télécharger gratuitement sur internet. R propose également une palette étendue de fonctionnalités graphiques. Il est possible d'utiliser R en mode intercatif sans jamais avoir à programmer.

R est un logiciel dans lequel de nombreuses techniques statistiques modernes et classiques ont été implémentées. Les méthodes les plus courantes permettant de réaliser une analyse statistique telles que :

– statistique descriptive .

– tests d'hypothèses .

– analyse de la variance.

– méthodes de régression linéaire (simple et multiple) .

– etc.

sont enchâssées directement dans le cœur du système. Notez également que la plupart des méthodes avancées de statistique sont aussi disponibles au travers de modules externes appelés packages. Ceux-ci sont faciles à installer directement à partir d'un menu du logiciel. Ils sont tous regroupés sur le site internet du Comprehensive R Archive Network (CRAN)

(http ://cran.r-project. org) sur lequel vous pouvez les consulter. Ce site fournit aussi, pour certains grands domaines d'étude, une liste commentée des packages associés à ces thèmes (appelée Task View), ce qui facilite ainsi la recherche d'une méthode statistique particulière. Par ailleurs, une documentation détaillée en anglais de chaque package est disponible sur le CRAN.

3.4 R et les graphiques

Une des grandes forces de R réside dans ses capacités, bien supérieures à celles des autres logiciels courants du marché, à combiner un langage de programmation avec la possibilité de réaliser des graphiques de qualité. Les graphiques usuels s'obtiennent aisément au moyen de fonctions prédéfinies.Ces dernières possèdent de trés nombreux paramétres permettant par exemple d'ajouter des titres, des légendes, des couleurs, etc. Mais il est également possible d'effectuer des graphiques plus sophistiqués permettant de représenter des données complexes telles que des courbes de surface ou de niveau, des volumes affichés avec un effet 3D, des courbes de densité, et bien d'autres choses encore.Il vous est également possible d'y ajouter des formules mathématiques. Vous pouvez aussi agencer ou superposer plusieurs graphiques sur une même fenêtre,et utiliser de nombreuses palettes de couleur.

Vous pouvez obtenir une d´emonstration des possibilit´es graphiques de R en tapant successivement les commande suivantes :

```
demo(image)
example(contour)
demo(graphics)
demo(persp)
demo(plotmath)
demo(Hershey)
require(lattice)  # Charge le package que vous devez avoir
                            # préalablement installé en passant par le
                            # menu Packages/Installer le(s) package(s).
demo(lattice)
example(wireframe)
require(rgl) # Même remarque que ci-dessus.
demo(rgl)    # Possibilité d'interaction avec la souris.
example(persp3d)
```

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$i.i.d$	:	Indépendantes et identiquement déstribuées
p	:	Probabilité
$E(.)$	:	Espérance mathématique
$var(.)$	:	Variance
σ	:	écart type
$\overline{X}$	:	Moyenne empirique
s^2	:	Variance empirique
dll	:	Degré de liberté
$f(.)$	:	Densité de probabilité
$\sim$	:	Suit en loi
$N(0,1)$	:	Loi normale centrée et réduite
χ_k^2	:	Loi de khi-deux à k degrés de liberté

Résumé

Le test-t de student est un test paramétrique qui compare la moyenne observée d'un échantillon statistique à une valeur fixée,il permet aussi de comparer les moyennes de deux échantilllons gaussiens.L'objectif de ce mémoire est de savoir si les moyennes des deux groupes sont significativement différentes au point de vue statistique.

Mots clés:loi de student,test sur la moyenne,p-valeur.

الملخص

اختبار "ت" هو اختبار معلمي يقارن المتوسط الملاحظ للعينة الاحصائية بقيمة ثابتة.يسمح ايضا بمقارنة متوسطي عينتين طبيعتين.الهدف من هذه المذكرة هو معرفة اذا كانت متوسطات هاتين العينتين تختلفان اختلافا كبيرا من الناحية الاحصائية.

الكلمات المفتاحية :توزيع ستيودانت ،اختبار على المتوسط،القيمة -p.

Abstract

Student's t- test is a parametric test that compares the observed mean of a statistical sample with a fixed value, it also allows to compare the means of two Gaussian samples. The objective of this thesis is to know if the means of the two groups are statistically significantly different

Keys words:student law,test on the mean,p-value.