Université Mohamed Khider de Biskra
Département Informatique
Module: Analyse de données

Première Année Master
2024/2025

Corrigé des exercices de TD 02

Solution de l'Exercice 1

1. L'estimation ponctuelle de la moyenne et de l'écart-type de X sont données par :

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 102.7333 \\
\hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2 = 598.9238 (\text{ le fait que la vraie moyenne, } \mu, \text{ est inconnue}) \\
\hat{\sigma}_c &= \sqrt{\hat{\sigma}_c^2} = \sqrt{598.9238} = 24.4729
\end{aligned}$$

2. Dans cette question, la formulation du test à réaliser est la suivante :

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

Au vu de l'échantillon précédent (la taille de l'échantillon < 30, X suit une loi normale et la vraie variance est inconnue) c'est le test de Student qu'il faut réaliser. On a :
• d'une part la statistique du test $t = \frac{\overline{X}-\mu_0}{\hat{\sigma}_c/\sqrt{n}} = \frac{102.7333-110}{24.4729/\sqrt{15}} = -1.1500$,
• et d'autre part sur la table de la loi de Student $t_\alpha = t_{(n-1,1-\alpha/2)} = t_{(14,1-.02/2)} = 2.625$.
A cet effet, on constate que $t \in [-t_\alpha, t_\alpha](-1.1500 \in [-2.625, 2.625]) \Rightarrow H_0$ est vraie, c'est-à-dire à un seuil de risque 2% la taille moyenne des enfants est égale de 110cm.

Solution de l'Exercice 2 Afin de répondre aux questions de l'exercice on aura besoin des quantités suivantes :
La moyenne :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(83+96+99+110+130+95+74) = 98.1429,$$

et la variance :

$$\begin{aligned}
\hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2 \\
&= \frac{1}{6}\left((83-98.1429)^2+(96-98.1429)^2+...+(95-98.1429)^2+(74-98.1429)^2\right) \\
&= 330.4762.
\end{aligned}$$

1. La formulation du test à réaliser dans ce cas est :

$$H_0{}"\mu = \mu_0" \textit{ contre } H_1{}"\mu < \mu_0", \tag{1}$$

plus précisément :

$$H_0{}"\mu = 100" \textit{ contre } H_1{}"\mu < 100". \tag{2}$$

(a) La statistique du test est : $T = \frac{\overline{X}-\mu_0}{\sqrt{\hat{\sigma}_c^2/n}} = \frac{98.1429-100}{\sqrt{330.4762/7}} = -0.2703$,

(b) La valeur critique du test est $t_\alpha = t_{(n_1-1,1-\alpha)} = t_{(7-1,1-0.05)} = 1.943$ (de la table de la loi de Student),

(c) On remarque que $T \in]-t_\alpha, t_\alpha[$, alors on ne rejette pas H_0, c'est-à-dire le nombre des pouls est égale à 100 battements en moyenne, avec un risque 5% de se tromper.

2. Le test à réaliser dans ce cas est :

$$H_0"\sigma^2 = \sigma_0^2" \ contre \ H_1"\sigma^2 \neq \sigma_0^2", \tag{3}$$

plus précisément :

$$H_0"\sigma^2 = 300" \ contre \ H_1"\sigma^2 \neq 300". \tag{4}$$

(a) La statistique du test est : $Y = \frac{(n-1)\hat{\sigma}_c^2}{\sigma^2} = \frac{6\times 330.4762}{300} = 6.6095$,

(b) Les valeurs critiques du test sont :
$P\left(\chi^2_{(n-1)} > a_\alpha\right) = 1 - \alpha/2 \Rightarrow P\left(\chi^2_{(6)} > a_\alpha\right) = 0.975 \Rightarrow a_\alpha = 1.237$
$P\left(\chi^2_{(n-1)} > b_\alpha\right) = \alpha/2 \Rightarrow P\left(\chi^2_{(6)} > b_\alpha\right) = 0.025 \Rightarrow b_\alpha = 14.449.$
D'où l'intervalle du test de la variance σ^2 est [1.237 ; 14.449].

(c) On remarque que $Y \in]a_\alpha, b_\alpha[$, alors on ne rejette pas H_0, c'est-à-dire la variation des pouls est égale à 300, avec un risque 5% de se tromper.

Solution de l'Exercice 3 Soit les notation suivantes :
μ_1 est la vraie taille moyenne des garçons.
μ_2 est la vraie taille moyenne des filles.

1. Le test à réaliser dans ce cas est :

$$H_0"\mu_1 = \mu_0" \ contre \ H_1"\mu_1 \neq \mu_0", \tag{5}$$

plus précisément :

$$H_0"\mu_1 = 180" \ contre \ H_1"\mu_1 \neq 180". \tag{6}$$

(a) La statistique du test est : $T_1 = \frac{\overline{X}-\mu_0}{\sqrt{\hat{\sigma}_c^2/n_1}} = \frac{182.43-180}{\sqrt{54.95/11}} = 1.0872$,

(b) La valeur critique du test est $t_\alpha = t_{(n_1-1,1-\alpha/2)} = t_{(10,1-0.05/2)} = 2.228$ (de la table de la loi de Student),

(c) On remarque que $T_1 \in]-t_\alpha, t_\alpha[$ alors on rejette pas H_0, c'est-à-dire on admet que la taille moyenne des garçon est de 180cm avec un risque 5% de se tromper.

2. Le test à réaliser dans ce cas est :

$$H_0"\mu_1 = \mu_2" \ contre \ H_1"\mu_1 \neq \mu_2"$$

mais pour réaliser ce test on est contraint à réaliser d'abord le test suivant :

$$H_0"\hat{\sigma}^2_{c,1} = \hat{\sigma}^2_{c,2}" \ contre \ H_1"\hat{\sigma}^2_{c,1} \neq \hat{\sigma}^2_{c,2}".$$

(a) Sachant que $\hat{\sigma}^2_{c,1} > \hat{\sigma}^2_{c,2}$ alors, la statistique du test est : $F = \frac{\hat{\sigma}^2_{c,1}}{\hat{\sigma}^2_{c,2}} = \frac{54.95}{26.20} = 2.0973$,

(b) La valeur critique du test est $f_\alpha = f_{(n_1-1,n_2-1,1-\alpha/2)} = f_{(10,9,0.975)} = 3.96$ (de la table de la loi de Fisher),

(c) On remarque que $F \in [1, f_\alpha[$ alors on admet que les deux échantillons ont la même variance avec un risque 5% de se tromper.

(d) Le fait que les deux échantillons ont la même variance, alors on calcule la variance commune : $\hat{\sigma}_c^2 = \frac{(n_1-1)\hat{\sigma}^2_{c,1}+(n_2-1)\hat{\sigma}^2_{c,2}}{n_1+n_2-2} = \frac{10*54.95+9*26.20}{11+10-2} = 41.3316.$

(e) Revenant maintenant au test (5), la statistique de ce dernier est :

$$T_2 = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}_c^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{182.43 - 168.80}{\sqrt{41.3316(\frac{1}{11} + \frac{1}{10})}} = 4.8522,$$

(f) La valeur critique du test est $t_\alpha = t_{(n_1+n_2-2,1-\alpha/2)} = t_{(19,1-0.05/2)} = 2.093$ (de la table de la loi de Student),

(g) On remarque que $T_2 \notin] - t_\alpha, t_\alpha[$ alors on rejette H_0, c'est-à-dire on admet que la taille moyenne des garçon est significativement différente de celle des filles.

3. Le test à réaliser dans ce cas est :

$$H_0"\mu_1 = \mu_2" \text{ } contre \text{ } H_1"\mu_1 > \mu_2"$$

(a) Les conditions de ce test sont vérifier dans la deuxième question (homogénéité de variance et calcul de la variance commune), de plus la statistique correspondante à ce test est la même que celle du test (5), c'est-à-dire $T_3 = T_2 = 4.8522$.

(b) la valeur critique du test est définie cette fois-ci par :
$t_\alpha = t_{(n_1+n_2-2,1-\alpha)} = t_{(19,1-0.05)} = 1.729$ (de la table de la loi de Student),

(c) On remarque que $T_3 > t_\alpha$ alors on rejette H_0, c'est-à-dire on admet que la taille moyenne des garçon est supérieur à celle des filles.

Solution de l'Exercice 4

1. Déterminer une estimation ponctuelle de la moyenne et de la variance de chaque échantillons.
On a :

$$\begin{aligned} \overline{X} &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = 24.5 \quad \text{et} \quad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i = 22 \\ \hat{\sigma}_{c,1}^2 &= \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \overline{X})^2 = 0.7880 \text{ et } \hat{\sigma}_{c,2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \overline{Y})^2 = 0.8700 \end{aligned}$$

2. Supposons qu'on désir savoir si les deux types d'arbres ont la même hauteur en moyenne.

a) Le test à réaliser est bien que le test d'homogénéité de moyennes, qu'on peut formulé comme suit :

$$H_0 : \text{ }''\mu_1 = \mu_2'' \text{ contre } H_1 : \text{ }''\mu_1 \neq \mu_2'', \qquad (*)$$

b) La condition nécessaire pour la réalisation de ce test est que les deux échantillons ont la même variance, c'est-à-dire il faut réaliser d'abord le test suivant :

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

La statistique de ce test (homogénéité de variance) est :
$F = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$, car la deuxième variance est plus grande que la première.
La réalisation de cette statistique est : $f = \frac{0.8700}{0.7880} = 1.1041$.
La valeur critique du test, f_α, correspond au fractile d'ordre $1 - \alpha/2 = 1 - 0.02/2 = 0.99$ d'une loi de Fisher de degrés de liberté $(n_2 - 1, n_1 - 1) = (4, 5)$ d'où par la lecture sur la table de la loi de Fisher on obtient $f_\alpha = 11.39$. On constate que $f < f_\alpha$, alors les deux échantillons ont les mêmes variances.

3. On a les deux échantillons sont issus d'une loi normale, mutuellement indépendants de plus ils ont la même variance (voir réponse 2.b)), donc le test $(*)$ sera réaliser à l'aide du test d'homogénéité de moyenne de Student.

   a) La statistique du test $(*)$ est :

$$T = \frac{\overline{X} - \overline{Y}}{\hat{\sigma}_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

   où $\hat{\sigma}_c^2$ est la variance globale des deux échantillons définie par :

$$\hat{\sigma}_c^2 = \frac{(n_1 - 1)\hat{\sigma}_{c,1}^2 + (n_2 - 1)\hat{\sigma}_{c,2}^2}{n_1 + n_2 - 2} = \frac{(6-1)0.7880 + (5-1)0.8700}{6+5-2} = 0.8244.$$

   Donc, la réalisation de la statistique T est :

$$t = \frac{24.5 - 22}{0.9080\sqrt{\frac{1}{6} + \frac{1}{5}}} = 4.5469.$$

   b) La valeur critique du test est le fractile d'ordre $1 - \alpha/2 = 1 - .02/2$ d'une loi de Student de degré de liberté $n_1 + n_2 - 2 = 9$, d'où $t_\alpha = 2.821$.

   c) On à $t \notin [-t_\alpha, t_\alpha]$, donc H_0 est fausse, cela signifier que c'est l'hypothèse H_1 qui est vraie c'est-à-dire les deux type d'arbres ont des hauteurs moyennes significativement différentes.

Solution de l'Exercice 5

1. Afin de confirmer ou de démentir ce que l'agent indique, nous devons réaliser le test d'homogénéité de moyenne des deux échantillons, dont la formulation du test est donné par :

$$H_0 : \text{''} \mu_X = \mu_Y \text{''} \text{ contre } H_1 : \text{''} \mu_X \neq \mu_Y \text{''}, \tag{7}$$

   mais on doit vérifier d'abord l'homogénéité de leurs variances, et cela en réalisant le test suivant :

$$H_0 : \text{''} \sigma_X^2 = \sigma_Y^2 \text{''} \text{ contre } H_1 : \text{''} \sigma_X^2 \neq \sigma_Y^2 \text{''},$$

   Notons que les moyennes est les variances des deux échantillons sont données par :
   $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{9} X_i = \frac{1}{9}\sum_{i=1}^{9} X_i = 252.3333$ et $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{11} Y_i = \frac{1}{11}\sum_{i=1}^{11} Y_i = 243$
   Sachant que, dans cet exercice, la vraie moyenne est connue où $\mu = 250$ alors
   $\hat{\sigma}_X^2 = \frac{1}{n_1}\sum_{i=1}^{9}(X_i - \mu)^2 = \frac{1}{9}\sum_{i=1}^{9}(X_i - 250)^2 = 48.6667.$
   $\hat{\sigma}_Y^2 = \frac{1}{n_2}\sum_{i=1}^{11}(Y_i - \mu)^2 = \frac{1}{11}\sum_{i=1}^{11}(Y_i - 250)^2 = 18.7273.$

   a) La statistique du test d'homogénéité de variance des deux échantillons est donnée par : $F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$ (le fait que $\hat{\sigma}_X^2 > \hat{\sigma}_Y^2$) est sa réalisation est $f = 2.5987$.

   b) La valeur critique du test, pour $\alpha = 5\%$, est : $f_\alpha = f_{(n_1, n_2, 1-\alpha/2)} = f_{(9,11,0.975)} = 3.59$.

   c) On constate que : $f < f_\alpha (2.5987 < 3.59)$, cela signifier que les deux échantillons ont la même variance.

   d) Le fait que les deux échantillons ont la même variance donc on calcule la variance commune définie par :

$$\hat{\sigma}^2 = \frac{n_1\hat{\sigma}_X^2 + n_2\hat{\sigma}_Y^2}{n_1 + n_2} = 32.2000$$

e) Ainsi, la statistique du test d'homogénéité de moyenne (7) est donnée par :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{et sa réalisation est} \quad t = 3.6594.$$

f) La valeur critique du test est : $t_\alpha = t(n_1 + n_2, 1 - \alpha/2) = t(20, 1 - 0.05/2) = 2.086$.

g) On constate que, $t \notin [-t_\alpha, t_\alpha]$, cela signifier que l'agent à raison une fois de plus, les deux types du camembert n'ont pas la même masse moyenne.

2. Le test à réaliser dans ce cas est bien que :

$$H_0 : \text{"}\mu_Y = 250\text{"} \text{ contre } H_1 : \text{"}\mu_Y < 250\text{"}.$$

a) La réalisation de statistique du test est : $t = \frac{\overline{Y} - 250}{\sqrt{\hat{\sigma}_Y^2 / n_2}} = -5.3648$

b) La valeur critique du test est : $t_\alpha = t_{(n_2, 1-\alpha)} = t_{(11, 1-0.05)} = 1.796$.

c) On constate que $t < -t_\alpha (-5.3648 < -1.796)$, donc l'agent a le droit de pénaliser l'entreprise.

3. Dans cette situation, la statistique du test est la même $t = -5.3648$ et la valeur critique est $t_\alpha = t_{(n_2, 1-\alpha)} = t_{(11, 1-0.02)} = 2.3281$. On constate également que $t < -t_\alpha$ $(-5.3648 < -2.3281)$, ce qui confirme que l'agent à le droit de pénalisé l'entreprise le fait que la masse moyenne du deuxième type de camembert est significativement inférieur à la norme (<250g).

Solution de l'Exercice 6 Les différentes moyennes (de chaque échantillon et globale) sont données par :

$$\overline{X}_1 = 37.40,\ \overline{X}_2 = 41.00,\ \overline{X}_3 = 30.80,\ \overline{X}_4 = 50.20 \text{ et } \overline{X} = 39.85.$$

En exploitant ces dernières quantités pour le calcul des différentes variations on obtient :

	SC	ddl	CM	f	f_α
Inter-groupes	961.2667	3	320.4222	14.1123	5.29
Intra-groupes	363.2833	16	22.7052		
Total	1324.55	19			

On constate que $f > f_\alpha$ cela signifie qu'on doit rejeter H_0. C'est-à-dire le facteur traitement a une influence significative sur les durées séparant deux crise d'asthme.

Solution de l'Exercice 7

1. Les estimations des différentes moyennes μ_1, μ_2, μ_3, μ_4 et m sont données respectivement par :

$\triangleright$ $\overline{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} = \frac{1}{3}(137) = 45.67$, $\triangleright$ $\overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} = \frac{1}{5}(255) = 51.00$,

$\triangleright$ $\overline{X}_3 = \frac{1}{n_3}\sum_{j=1}^{n_3} x_{3j} = \frac{1}{4}(180) = 45.00$, $\triangleright$ $\overline{X}_4 = \frac{1}{n_4}\sum_{j=1}^{n_4} x_{4j} = \frac{1}{4}(178) = 44.50$,

et $\overline{X} = \frac{1}{n}\sum_{i=1}^{4} n_i \overline{X}_i = \frac{1}{16}(3 * 45.67 + 5 * 51.00 + 4 * 45.00 + 4 * 44.50) = 46.8750$.

2. Considérons l'hypothèse (H_0) : les rendements moyens de chaque variété sont égaux.

a) Dans ce cas, le problème est le même que celui de l'exercice précedent. Contrairement à ce dernier, on remarque que les tailles des échantillons ne sont pas les mêmes, mais le raisonnement de l'ANOVA ne changera pas . En effet, la décomposition de la variation totale dans cette situation sera :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p} n_j(\overline{X}_j-\overline{X})^2}_{SC_{Fac}}, \tag{8}$$

où, n_j : est la taille du $j^{\text{ième}}$ échantillon (groupe).

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

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

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

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}
Inter-groupe	126.0833	3	42.0278	1.9128
Intra-groupe	263.6667	12	21.9722	
Total	389.7500	15		

b) On a d'une part $F_{obs} = 1.9128$ et d'autre part $f_\alpha = f(3, 12, 1 - 0.01) = 5.9525$, alors on rejette pas H_0 car $F_{obs} < f_\alpha$, cela signifier qu'il y a pas une différence significatives entre les rendement des différentes variétés d'orge.