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

Première Année Master
2023/2024

Fiche des exercices de TD 02

Exercice 1 *(Estimation et Test de conformité de moyenne)* Un échantillon de 15 enfants d'une ville donnée à fourni les tailles suivants (en cm) :

70	85	93	99	101	105	110	121	138	166	74	85	93	99	102

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la tailles des enfants.

2. Au vu de l'échantillon, peut-on considérer, au seuil de signification 2%, que la taille moyenne des enfants est de 110 *cm*?

Exercice 2 *(Test de conformité de moyenne et de variance)*

On a mesuré, après une course de 400 mètres, le pouls (en battements par minute) de 7 étudiants suivants un cours d'éducation physique :

X	83	96	99	110	130	95	74

Supposons que l'accroissement du pouls est une variable aléatoire de loi normale $N(\mu, \sigma^2)$, alors à un risque $\alpha = 5\%$, peut-ont considérer que :

1. Le nombre des pouls est inférieur à 100 battements en moyenne.

2. Le variation des pouls est différente de 300.

Exercice 3 *(Test de Student : conformité et d'homogénéité)*

On dispose de deux échantillons d'étudiants de sexe masculin et féminin, dont on a relevé la taille. On se demande si les tailles observées peuvent être considérées comme différentes entre les deux groupes. Le tableau suivant résume les principaux caractéristiques des deux échantillons en question :

	Masculin	*Féminin*
Effectif du groupe	*11*	*10*
Moyenne	*182.43*	*168.80*
$\hat{\sigma}_c^2$	*54.95*	*26.20*

1. Peut-on dire, au seuil de signification 5%, que les variances des tailles observées peuvent être considérées les mêmes entre les deux groupes?

2. Peut-on dire, au seuil de signification 5%, que les tailles moyennes observées peuvent être considérées différentes entre les deux groupes?

3. Au seuil $\alpha = 5\%$, peut-on dire que la taille moyenne des garçons est supérieur a la taille moyenne des filles?

Exercice 4 *:* Afin de comparer deux types d'arbre, nous avons réalisés un recueil de hauteur de quelques arbres, dont les mesures sont rangées dans le tableau suivant.

							Somme
Arbre 1	23.3	24.0	24.3	24.5	25.0	25.9	147
Arbre 2	21.1	21.1	22.1	22.4	23.3		110

1. Déterminer la moyenne et la variance de chaque échantillons.

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

   a) Donner la forme du test à réalisé dans ce cas.

   b) Vérifier si les conditions du test sont satisfaites pour un seuil de risque $\alpha = 2\%$.

3. Si les conditions de 2.b) sont vérifiées alors :

   a) Donner la statistique du test donner dans 2.*a*) ainsi que sa réalisation.

   b) Donner la valeur critique associe à ce test, pour un seuil de risque $\alpha = 2\%$.

   c) Que peut-on conclure sur la hauteur moyenne des deux types d'arbres?

Exercice 5 *:* Une laiterie produit deux types du camemberts. La masse d'un camembert tiré au hasard dans la production, par la contrôle, est distribuée selon une loi normale de moyenne $\mu = 250$ et de variance σ^2. L'agent de contrôle a tiré un échantillon simple de chaque type, dont le tableau suivant fournit les masses mesurées en g :

X	*257*	*241*	*253*	*251*	*245*	*248*	*251*	*264*	*261*	$\times$	$\times$
Y	*235*	*252*	*243*	*240*	*243*	*239*	*240*	*246*	*246*	*246*	*243*

1. L'agent de contrôle indique que, les deux types des camemberts n'ont pas la même masse moyenne. Peut-on conclure, au seuil $\alpha = 5\%$, que l'agent de contrôle a raison?

2. Le responsable de production réclame et dit que si l'agent prend un seuil de risque 2%, alors il constatera que les masses moyennes des deux type des camemberts sont significativement les mêmes. Dans ce cas, est-ce-que l'agent de contrôle aura le droit de pénaliser l'entreprise?

Exercice 6 *:* Nous souhaitons comparer quatre traitements, notés A, B , C et D. Nous répartissons par tirage au sort les patients, et nous leur affectons l'un des quatre traitements. Nous mesurons sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Les mesures sont reportées dans le tableau ci-dessous:

Traitement A	*Traitement B*	*Traitement C*	*Traitement D*
36; 37; 35; 38; 41	*42 ; 38; 39; 42; 44*	*26; 26; 30 38; 34*	*42; 45; 50; 56; 58*

Pouvons-nous conclure, à un seuil de risque 1%, que le facteur traitement a une influence sur le critère retenue?

Exercice 7 *:* On s'intéresse au rendement d'orge pour quatre variétés différentes. On dispose de quatre parcelles avec une variété d'orge pour chacune. On répète cette expérience à des endroits différents. On a obtenu :

variété 1	variété 2	variété 3	variété 4
46 ; 43 ; 48	57 ; 53 ; 43 ; 54 ; 48	50 ; 41 ; 47 ; 42 ;	39 ; 51 ; 45 ; 43

Les quatre variétés sont-elles du même rendement en moyenne? pour un seuil de risque 5%

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

Première Année Master
2021/2022

Solution des exercices

Solution de l'Exercice 1 *(Estimation et Test de conformité de moyenne) L'estimation ponctuelle de la moyenne et de l'écart-type de X sont données par :*

$$
\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 102.73 \\
\hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2 = \frac{1}{14}[(70-102.73)^2+(85-102.73)^2+...+(102-102.73)^2] \\
&= 598.9238(\textit{ le fait que la vraie moyenne, } \mu\textit{, est inconnue}) \\
\hat{\sigma}_c &= \sqrt{\hat{\sigma}_c^2} = \sqrt{598.9238} = 24.4729
\end{aligned}
$$

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

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

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 *(Test de conformité de moyenne et de variance)*

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 :

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

   plus précisément :

   $$H_0\text{"}\mu = 100\text{"} \textit{ contre } H\text{"}\mu_1 < 100\text{"}. \tag{2}$$

   (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 *(Test de Student : conformité et d'homogénéité)* 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 :

$$H_0"\sigma_1^2 = \sigma_2^2" \ contre \ H_1"\sigma_1^2 \neq \sigma_2^2".$$

(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 peut calculer leur variance commune comme suit : $\hat{\sigma}_c^2 = \frac{(n_1-1)\hat{\sigma}^2_{c,1}+(n_2-1)\hat{\sigma}^2_{c,2}}{n_1+n_2-2} = \frac{10*54.95+9*26.20}{11+10-2} = 41.3316$.

2. Le test à réaliser dans ce cas est :

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

(a) la statistique de ce dernier est :

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

(b) 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),

(c) 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 :

$$H_0 "\mu_1 = \mu_2" \textit{ contre } H_1 "\mu_1 > \mu_2". \tag{5}$$

   (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 *(Test d'homogénéité de Student)*

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

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

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 :

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

   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 :

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

   La statistique de ce test (homogénéité de variance) est :
   $F = \frac{\hat{\sigma}^2_{c,2}}{\hat{\sigma}^2_{c,1}}$, 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 :

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

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 :

$$H_0 : \text{''}\ \mu_X = \mu_Y\ \text{''} \text{ contre } H_1 : \text{''}\ \mu_X \neq \mu_Y\ \text{''}, \tag{6}$$

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 (6) 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. Dans cette situation, la statistique du test est la même $t = 3.6594$ et la valeur critique est $t_\alpha = t_{(n_1+n_2-2,1-\alpha)} = t_{(20,1-0.02)} = ...$ On constate également que ....,

Solution de l'Exercice 6 *(ANOVA 1 à un plan équilibré)*

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	976.8	3	325.6	15.19	5.29
Intra-groupes	342.8	16	21.43		
Total	1324.55	19			

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