

Examen final
Estimation et Analyse de données
vendredi 20 novembre 2020
Myriam Bertrand/Mélanie Piot

Instructions :

- *Durée : 1h30min. Le sujet fait 7 pages. Rendre le sujet avec la copie sinon 0. NOM :*
- *Les documents sont autorisés. Les calculatrices sont autorisées.*
- *Vous trouverez à la fin du sujet une liste de quantiles.*

Exercice 1 (10 min)

Les 4 premières questions se rapportent aux figures suivantes.

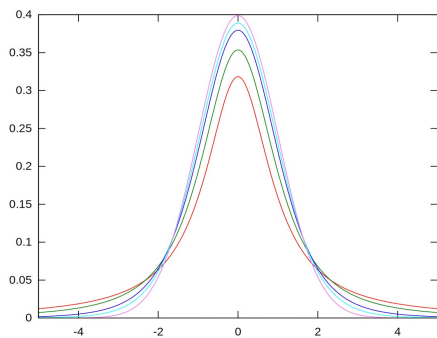


FIGURE 1 –

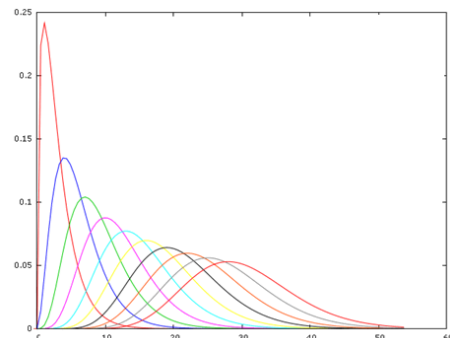


FIGURE 2 –

Question 1 Quelle(s) loi(s) peu(ven)t être représentée(s) par la Figure 1 :

- | | |
|--|---|
| ☐ Loi normale $\mathcal{N}$ | ☐ Loi de Student t |
| ☐ Loi de Fisher | ☐ Loi du khi-deux χ^2 |
| ☐ Loi exponentielle | |

Question 2 Quelle(s) loi(s) peu(ven)t être représentée(s) par la Figure 2 :

- | | |
|--|---|
| ☐ Loi normale $\mathcal{N}$ | ☐ Loi de Student t |
| ☐ Loi de Fisher | ☐ Loi du khi-deux χ^2 |
| ☐ Loi exponentielle | |

Question 3 Quel(s) test(s) utilise(nt) la Figure 1 :

- | | |
|---|---|
| ☐ test d'égalité de deux moyennes | ☐ test de normalité |
| ☐ test d'égalité de deux proportions | ☐ test d'égalité de deux variances |
| ☐ test d'indépendance de variables | |

Question 4 Quel(s) test(s) utilise(nt) la Figure 2 :

- | | |
|---|---|
| ☐ test d'égalité de deux moyennes | ☐ test de normalité |
| ☐ test d'égalité de deux proportions | ☐ test d'égalité de deux variances |
| ☐ test d'indépendance de variables | |

Question 5 Quelle(s) caractéristique(s) est(sont) sensible(s) aux valeurs extrêmes :

- | | | | |
|-------------------------------------|-------------------------------------|---------------------------------------|--|
| ☐ la moyenne | ☐ la médiane | ☐ l'écart-type | ☐ les quartiles |
|-------------------------------------|-------------------------------------|---------------------------------------|--|

Question 6 Si j'augmente mon niveau de confiance $1 - \alpha$ (*i.e.* je diminue α), alors je vais
☐ augmenter la zone de rejet ☐ diminuer la zone de rejet ☐ diminuer la puissance

Question 7 La p -valeur :
☐ donne l'erreur de deuxième espèce ☐ donne l'erreur de seconde espèce
☐ détermine si on rejette une hypothèse ☐ donne la puissance

Question 8 Considérons deux classes, de même taille, ayant la même moyenne de 11/20 à une évaluation et des écart-types respectifs de 1 et 3 points. Alors :

- | | |
|--|--|
| ☐ la variance inter modalités est non nulle | ☐ la variance intra modalités est nulle |
| ☐ la variance intra modalités est non nulle | ☐ la variance intra modalités est nulle |
| ☐ le rapport de corrélation est non nul | ☐ le rapport de corrélation est nul |

Exercice 2 (30 min)

Cet exercice se base sur le `dataframe` *Notes* de notes d'interro du cours "Estimation et Analyse de Données" d'apprentis 4A de l'ESIEA Laval. On répondra à l'aide des sorties R situées à la fin de l'exercice.

1. Remplir le texte à trous :

La taille de l'échantillon est de . Sur cet échantillon, on observe variable(s) quantitative(s) qui est/sont le(s) suivante(s) : et on observe variable(s) qualitative(s) nominale(s) qui est/sont le(s) suivante(s) :
 Chaque variable qualitative a modalités. La moyenne de la classe est de . Il y a 25% des étudiants qui ont une note supérieure à . Si on regarde les étudiants du début de l'alphabet, ils sont au nombre de et ils ont une moyenne de et un écart-type de ; si on regarde les étudiants de la fin de l'alphabet, ils sont au nombre de et ils ont une moyenne de et un écart-type de .

2. Que fait la fonction `shapiro.test()` ? Précisez les hypothèses nulle et alternative.
3. Quelles sont, ici, les conclusions de ce test sur les notes du début et de la fin de l'alphabet ?

Dans la suite, nous souhaitons comparer les notes des étudiants du début de l'alphabet avec ceux de la fin de l'alphabet. Pour cela, nous comparons les deux sous-échantillons de `note`.

4. On commence par comparer la dispersion des notes au sein de ces deux groupes.
 - (a) Quel test proposez-vous de faire ?
 - (b) Quelles sont les conditions d'applications et les hypothèses de test $\mathcal{H}_0$ et $\mathcal{H}_1$?
 - (c) Donner la statistique de test, la p -valeur et conclure.
 - (d) (*bonus*) Quelle quantité faudrait-il calculer pour préciser le risque d'erreur de la conclusion du test ?
5. Comparons maintenant les moyennes de ces deux groupes. En suivant le même procédé qu'à la question précédente, répondre à la problématique.

```
> Notes=as.data.frame(donnees_test_exam)

> str(Notes)
'data.frame': 53 obs. of 4 variables:
 $ note      : num  5 6 20 6 12 5 15 10 12 16 ...
 $ groupe    : Factor w/ 2 levels "2","1": 1 2 1 1 1 1 2 1 2 2 ...
 $ alphabet  : Factor w/ 2 levels "debut","fin": 1 1 1 1 1 1 1 1 1 1 ...
 $ genre     : Factor w/ 2 levels "H","F": 1 1 1 1 1 1 1 1 1 1

> summary(Notes)
      note      groupe  alphabet  genre
Min.   : 4.00    2:28    debut:30    H:46
1st Qu.: 8.00    1:25    fin  :23     F: 7
Median :10.00
Mean    :10.64
3rd Qu.:13.00
Max.    :20.00

> summary(note[alphabet=='debut'])
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      5.00   9.00   11.00   10.97   12.75   20.00

> summary(note[alphabet=='fin'])
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      4.00   7.50   9.00   10.22   12.50   17.00

> sd(note[alphabet=='debut'])
[1] 3.368038

> sd(note[alphabet=='fin'])
[1] 3.965264

> shapiro.test(note[alphabet=='debut'])
Shapiro-Wilk normality test
data:  note[alphabet == "debut"]
W = 0.96721, p-value = 0.466

> shapiro.test(note[alphabet=='fin'])
Shapiro-Wilk normality test
data:  note[alphabet == "fin"]
W = 0.91576, p-value = 0.05409

> shapiro.test(note)
Shapiro-Wilk normality test
data:  note
W = 0.97256, p-value = 0.2597
```

```
> var.test(note[alphabet=='debut'],note[alphabet=='fin'])
F test to compare two variances
data:  note[alphabet == "debut"] and note[alphabet == "fin"]
F = 0.72146, num df = 29, denom df = 22, p-value = 0.4059
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3163989 1.5736128
sample estimates:
ratio of variances
      0.7214556

> t.test(note[alphabet=='debut'],note[alphabet=='fin'],var.equal=TRUE)
Two Sample t-test
data:  note[alphabet == "debut"] and note[alphabet == "fin"]
t = 0.74319, df = 51, p-value = 0.4608
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.274746  2.773297
sample estimates:
mean of x mean of y
 10.96667  10.21739

> t.test(note[alphabet=='debut'],note[alphabet=='fin'],var.equal=FALSE)
Welch Two Sample t-test
data:  note[alphabet == "debut"] and note[alphabet == "fin"]
t = 0.72716, df = 43.071, p-value = 0.4711
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.328645  2.827196
sample estimates:
mean of x mean of y
 10.96667  10.21739
```

Exercice 3 (20 min)

On veut comparer le salaire des femmes et des hommes dans une grande entreprise. On considère 270 employés, on note leur genre puis on note le salaire selon "faible", "moyen" et "élevé". On observe le tableau ci-dessous.

genre / salaire	Faible	Moyen	Elevé
Homme	10	70	40
Femme	30	60	60

En utilisant les données R ci-dessous, répondez à la problématique avec les étapes suivantes :

1. Définir les variables étudiées. Donner leurs types.
2. Donner le nom du tableau précédent.
3. Donner le nom du test à effectuer.
4. Donner les deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ associées au test.
5. Vérifier les conditions d'applications.
6. Donner les effectifs marginaux.
7. Donner les effectifs sous hypothèse d'indépendance.
8. Donner la valeur de la statistique de ce test et la conclusion.
9. Donner le risque d'erreur avec lequel vous pouvez affirmer votre décision.

```
> Salaire<-matrix(c(10,70,40,30,60,60),byrow=T,nrow=2,  
dimnames=list(c("Homme","Femme"),c("Faible","Moyen","Elevé")))
```

```
> Salaire  
      Faible  Moyen  Elevé  
Homme      10     70    40  
Femme      30     60    60  
> chisq.test(Salaire)
```

Pearson's Chi-squared test

```
data:  Salaire  
X-squared = 11.579, df = 2, p-value = 0.00306
```

```
> chisq.test(Salaire)$expected  
      Faible  Moyen  Elevé  
Homme 17.77778     ?  44.44444  
Femme 22.22222 72.22222     ?
```

Exercice 4 (20 min)

Un certain composant électronique est fabriqué une fois par mois par l'entreprise Micro-Systèmes. La quantité fabriquée varie avec la demande du marché. Dans le but de planifier la production et d'établir certaines normes sur le nombre d'hommes-minutes exigés pour la production de différents lots de ce composant électronique, le responsable de la production a relevé l'information suivante pour 15 groupes de production. Le nombre d'hommes-minutes est identifié par Y et la quantité fabriquée (en unités) par X.

Groupe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_i	150	192	264	371	300	358	192	134	242	238	226	302	340	182	169
x_i	35	42	64	88	70	85	40	30	55	60	51	72	80	44	39

En utilisant les données R ci-dessous, répondez aux questions suivantes :

1. Quel est le premier coefficient à estimer pour étudier la liaison de ces variables ? Commenter cette valeur.
2. On envisage un modèle linéaire. Donner l'équation de la droite de régression estimée.
3. D'après l'équation de régression, si le nombre d'unités à fabriquer augmente de 10, quelle sera l'augmentation correspondante du nombre moyen d'hommes-minutes requis ?
4. Donner la variation qui est expliquée par la droite de régression et la variation qui est inexpliquée par la droite.
5. Déterminer le pourcentage de variation qui est expliqué par la droite de régression.
6. Donner un intervalle de confiance à 95% pour la pente de la droite.
7. (*bonus*) Tester si l'ordonnée à l'origine est nulle.

```
> quantite_X=c(35,42,64,88,70,85,40,30,55,60,51,72,80,44,39)
> hom_min_Y=c(150,192,264,371,300,358,192,134,242,238,226,302,340,182,169)

> cor(hom_min_Y,quantite_X)
[1] 0.9939843

> model=lm(hom_min_Y~quantite_X)

> summary(model)
Call:
lm(formula = hom_min_Y ~ quantite_X)
Residuals:
    Min       1Q   Median       3Q      Max
-18.050  -4.167   1.534   4.908  16.283
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.0511     7.3458   2.049 0.0612 .
quantite_X    4.0166     0.1227  32.723 7.12e-14
Residual standard error: 8.667 on 13 degrees of freedom
Multiple R-squared:  0.988, Adjusted R-squared:  0.9871
F-statistic: 1071 on 1 and 13 DF, p-value: 7.119e-14
```

```
> shapiro.test(residuals(model))
Shapiro-Wilk normality test
data: residuals(model)
W = 0.98085, p-value = 0.975

> anova(model)
Analysis of Variance Table
Response: hom_min_Y
      Df Sum Sq Mean Sq F value    Pr(>F)
quantite_X  1  80441    80441  1070.8 7.119e-14
Residuals 13    977         75

> confint(model)
              2.5 %    97.5 %
(Intercept) -0.8185637 30.920850
quantite_X   3.7514656  4.281828
```

Liste de quantiles

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.05)
[1] -1.644854
> qnorm(0.95)
[1] 1.644854
> qnorm(0.975)
[1] 1.959964

> qt(0.95,51)
[1] 1.675285
> qt(0.975,51)
[1] 2.007584
> qt(0.95,29)
[1] 1.699127
> qt(0.975,29)
[1] 2.04523
> qt(0.95,22)
[1] 1.717144
> qt(0.975,22)
[1] 2.073873
> qt(0.95,13)
[1] 1.770933
> qt(0.975,13)
[1] 2.160369

> qf(0.025,29,22)
[1] 0.4584709
> qf(0.05,29,22)
[1] 0.5206113
> qf(0.95,29,22)
[1] 1.990396
> qf(0.975,29,22)
[1] 2.280209
> qf(0.025,53,53)
[1] 0.5802607
> qf(0.05,53,53)
[1] 0.6338422
> qf(0.95,53,53)
[1] 1.57768
> qf(0.975,53,53)
[1] 1.723363

> qchisq(0.95,2)
[1] 5.991465
> qchisq(0.975,2)
[1] 7.377759
> qchisq(0.95,270)
[1] 309.3258
> qchisq(0.975,270)
[1] 317.4092
```