EISTI	**ING2-GI** **EXAMEN DE STATISTIQUES 2019-2020**
Durée : 2h	Calculatrice EISTI autorisée 4 feuilles manuscrites R/V autorisées

Barême (22 points)

~~**QCM : x points si bonne réponse / 0 point si pas de réponse / - x points si mauvaise réponse**~~

QCM 1 (~~à points négatifs~~) (3 points)

1) 0.5
2) 1
3) 0.5
4) 0.5
5) 0.5

QCM 2 (~~2 points~~)

~~0.5~~ par question

Exercice 1 (8 points)

1) 1
2) 0.5+1
3) 1 pour dessin + 1 pour erreur
4) 1.5
5) 0.5
6) 1.5

Exercice 2 (6 points)

1) 1
2) 1.5
3) 1.5
4) 1
5) 0.5 + 0.5

Exercice 3 (5 points)

1) 1
2) 1
3) 1
4) 1
5) 1

Le jeu de données utilisé dans cet examen croise des résultats de tests de personnalité avec différent type de drogues consommées. Il y a 1885 individus observés et 31 variables :

- 5 sont quantitatives à valeurs entre 12 et 60 et correspondent aux résultats de test de personnalité NEO-FFI-R (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousnes)
- 2 sont quantitavives à valeurs centrées-réduites et correspondent aux résultats des tests de personnalité BIS-11 et ImpSS (Impulsiveness, SensationSeeing)
- 24 sont qualitatives (Age, Gender, Education, Country, Ethnicity, Alcohol, Amphet, ...). Les niveaux pour la consommation des drogues sont :

CL0 = Never Used / CL1 = Used over a Decade Ago/ CL2 = Used in Last Decade / CL3 = Used in Last Year / CL4 = Used in Last Month / CL5 = Used in Last Week / CL6 = Used in Last Day.

Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	...	Alcohol	Amphet	...
45-54	0	left at 16	USA	White	12	43		CL5	CL0	
35-44	1	left at 18	Canada	Asian	13	43		CL4	CL0	
18-24	0	master degree	USA	White	14	47		CL5	CL3	
18-24	1	doctorat degree	UK	Black	14	51		CL6	CL0	

Tab. 1. Extrait du jeu de données

QCM

1) Soit un échantillon gaussien de taille 9. Sur lequel on a mesuré une moyenne $\bar{x}$=0.5 et une variance s^2=0.25. Pour un risque α=2%, la valeur lue dans la table de loi est 2.90. Répondre par vrai ou faux aux questions suivantes.
    a. On utilise la table de la loi normale N(0,1) F
    b. L'IDC est [0.02 ; 0.98] V
    c. Cela signifie que 98% des observations sont dans cet intervalle. F
    d. Si on choisit un risque α=5%, l'IDC sera moins grand. V
    e. Si on connait la variance, l'IDC sera moins grand. V
2) Pour chacune des questions suivantes, dire quel(s) test(s) vous pouvez utiliser (sans transformation des variables) dans la liste : Test de Student, Test du chi-deux, test de l'ANOVA, test de Wilcoxon, test de Kruskal-Wallis.
    a. Peut-on dire que la consommation de cannabis est liée au niveau d'étude ? Test d'indépendance du chi-deux
    b. Est-ce qu'il y a une différence significative de consommation d'alcool entre les hommes et les femmes ? test de comparaison de 2 échantillons : Student ou Wilcoxon
    c. Est-ce qu'on peut considérer les scores des tests de personnalité comme des variables gaussiennes ? Shapiro s'ils connaissent ou test d'ajustement du chi-deux.
    d. Est-ce que le niveau d'éducation a un impact sur la consommation de caféine ? Test de comparaison de plus de 2 échantillons, ANOVA ou Kruskal.

Exercice 1 : Test sur l'impulsivité

Le score moyen du test de personnalité sur l'impulsivité (Impulsiveness) est nul. Nous aimerions savoir si le score moyen pour les hommes est signifcativement supérieur à 0. Le jeu de données est constitué de 942 hommes. Sur l'échantillon on mesure la moyenne du score pour les hommes $\bar{x}$=0.264 et la variance s^2=0.140.

1) Ecrire les hypothèses H_0 et H_1 du test à mettre en place. N'oubliez pas de définir vos notations.

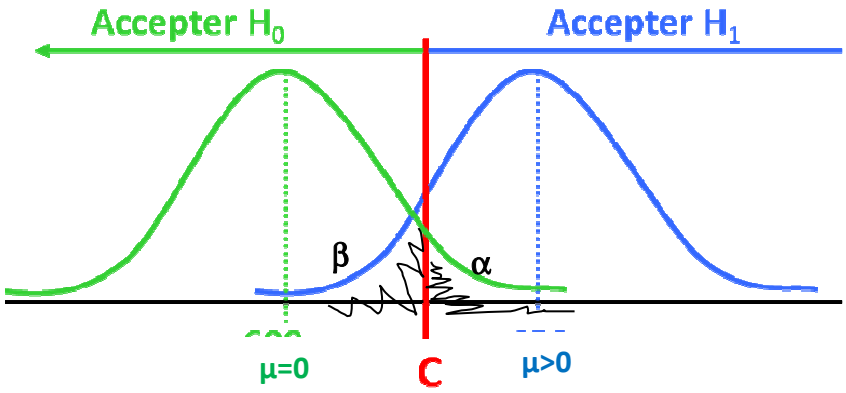
    $H_0 : \mu=0$ contre $H_1 : \mu>0$

    où μ est le score moyen d'impulsivité chez les hommes

2) Quelle est la statistique (variable de décision) du test ? Quelle est sa loi ? Justifiez votre réponse.

La moyenne de l'échantillon, $\bar{X}$ est la statistique du test. Etant donnée que l'échantillon est grand (n=100), on peut approcher la loi de $\bar{X}$ par une loi normale $N(\mu,\sigma^2/n)$. L'écart-type σ est inconnu mais l'échantillon étant grand, on peut le remplacer par son estimation.

3) Dessiner la région critique du test et représenter les erreurs de 1ère et 2ème espèces.



4) Déterminer le seuil de décision si on considère un risque à 5%

Sous l'hypothèses H_0, $\bar{X}$ suit une loi $N(0,s^2/n)$. D'où

$$\alpha = P(\bar{X} > C | H_0) = P(\frac{\bar{X}}{\sigma}\sqrt{n} > \frac{C}{\sigma}\sqrt{n}) = P(Z > C')$$

D'après la table des fractiles de la loi N(0,1), on obtient C'=1.6449 pour α=5%. D'où

$$C = 1.6449\frac{\sigma}{\sqrt{n}} = 1.6449\frac{\sqrt{0.140}}{\sqrt{942}} = 0.02\,.$$

en remplaçant σ^2 par son estimation s^2.

5) Quelle est votre conclusion ?

$\bar{x}$=0.264>>C donc on accepte l'hypothèse H_1 avec 5% de risque de se tromper. Le score d'impulsivité chez les hommes est significativement supérieur à 0.

6) Construire un intervalle de confiance avec un risque α=5% pour le score moyen d'impulivité des hommes.

On cherche a et b tel que P(a<μ<b)=0.95. Si on considère un risque symétrique, l'intervalle est de la forme

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Pour α=5%, on a $z_{1-\alpha/2}$=1.96, donc l'IDC est [0.24 ;0.288]

Exercice 2 : Peut-on prévoir l'impulsivité en fonction des résultats du test NEO-FFI-R

Les scores de personnalité s'obtiennent à partir de plusieurs tests. Le test NEO-FFI-R est un questionnaire de personnalité basé sur le modèle des Big Five : Névrosisme, Extraversion, Ouverture, Agréabilité, Conscience. L'impulsivité s'obtient avec le test BIS 11. L'idée de construire un modèle qui permet de prévoir l'impulsivité avec les scores obtenus au test des Big Five de façon à s'affranchir du test BIS 11.

Ci-dessous se trouvent les résultats d'une régression linéaire entre l'impulsivité et les Big Five.

```
Call:
lm(formula = Impulsiveness ~ Neuroticism + Extraversion + Openness +
    Agreeableness + Conscientiousness, data = Mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-3.05535 -0.54341 -0.00609  0.55236  2.69791

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)       -0.190119   0.268242  -0.709    0.479
Neuroticism        0.011805   0.002457   4.804 1.68e-06 ***
Extraversion       0.033515   0.003313  10.117  < 2e-16 ***
Openness           0.029447   0.003031   9.715  < 2e-16 ***
Agreeableness     -0.025588   0.003090  -8.282 2.28e-16 ***
Conscientiousness -0.043531   0.003099 -14.045  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8276 on 1879 degrees of freedom
Multiple R-squared:  0.2502, Adjusted R-squared:  0.2482
F-statistic: 125.4 on 5 and 1879 DF,  p-value: < 2.2e-16
```

1) Ecrire le modèle obtenu sans oublier le terme d'erreur.

   Impulsiveness=-0.19+0.1×Neuroticism+0.03×Extraversion+0.03×Openness-0.25×Agreeableness-0.4×Conscientiousness+ε

   où $\varepsilon \sim N(0,\sigma^2)$. (-0.25 si oubli de l'erreur)

2) A quel test correspond la dernière ligne ? Ecrire les hypothèses nulle et alternative. Conclusion.

Il s'agit du test de Fisher

$$H_0 : a_1=a_2=a_3=a_4=a_5=0 \qquad H_1 : \exists i \in \{1,\dots,5\}, a_i=0$$

La p valeur est très petite donc on accepte H_1 *.i.e.*au moins une des variables explicatives à un impact significatif sur l'impulsivité dans ce modèle.

3) A quel test correspond la ligne « Intercept » du tableau ? Ecrire les hypothèses nulle et alternative. Conclusions.

Il s'agit du test de Student pour savoir si la constante est nulle ou pas

$$H_0 : a_0=0 \quad H_1 : a_0 \neq 0$$

La p-valeur est grande (0.479) donc on accepte H_0. La constante est considérée comme nulle dans ce modèle.

4) Même question avec la ligne « Openness ».

Il s'agit du test de Student pour savoir si le coefficient devant cette variablae est nul ou pas

$$H_0 : a_i=0 \quad H_1 : a_i \neq 0$$

La p-valeur est très petite (10^{-16}) donc on accepte H_1. Le coefficient devant Openness est significativement non nul. Cela signifie que cette variable a un impact linéaire sur la variable Impulsiness.

5) Quel pourcentage de la variabilité des scores d'impulsivité observés est expliqué par ce modèle ? Quelle est votre conclusion par rapport à la problématique initiale ?

Le $R^2=0.25$ dont seulement 25% de la variabilité de l'impulsivité est expliqué par ce modèle. On ne peut donc pas s'affranchir du test BIS 11 pour connaitre l'impulsivité d'une personne.

Exercice 3 : Y-a-t-il un lien entre l'âge et la consommation de LSD ?

Afin de répondre à cette question, nous allons effectuer un test du chi-deux. Pour cela on remplace les classes de consommation CL0,…,CL6 par deux classes :

C0 : jamais consommé

C1 : déjà consommé

On obtient le tableau de contingence suivant.

	18-24	25-34	35-44	45-54	55+
C0	319	288	211	180	71
C1	324	193	145	114	40

1) Quelles sont les hypothèses H_0 et H_1 de test du chi-deux ?

   H_0: les variables sont indépendantes

   H_1 : les variables sont liées

2) Quels seraient les effectifs théoriques des 18-24 ans ne consommant pas d'alcool si les deux variables sont indépendantes ?

Effectifs obs

	18-24	25-34	35-44	45-54	55+	tot
C0	319	288	211	180	71	1069
C1	324	193	145	114	40	816
	643	481	356	294	111	1885

Fréquences obs

	18-24	25-34	35-44	45-54	55+	tot
C0	0,17	0,15	0,11	0,1	0,04	0,57
C1	0,17	0,1	0,08	0,06	0,02	0,43
	0,34	0,26	0,19	0,16	0,06	1

Eff. th

	18-24	25-34	35-44	45-54	55+	tot
C0	**365**	273	202	167	62,9	1069
C1	278	208	154	127	48,1	816
	643	481	356	294	111	1885

Fréq. th

	18-24	25-34	35-44	45-54	55+	tot
C0	0,19	0,14	0,11	0,09	0,03	0,57
C1	0,15	0,11	0,08	0,07	0,03	0,43
	0,34	0,26	0,19	0,16	0,06	1

3) Quelle est la loi suivie par la statistique du test ?

Une loi du chi-deux à (5-1)*(2-1)=4 d.d.l.

On trouve les résultats suivants

```
        Pearson's Chi-squared test

data:  tab
X-squared = 20.932, df = 4, p-value = 0.0003267
```

4) A quoi correspond `X-squared = 20.932`?

Cela correspond à la valeur de la statistique du test sur l'échantillon, c'est-à-dire la distance du chi-deux entre le tableau des effectifs observés et celui des effectifs théoriques sous H_0.

5) Pensez-vous que la consommation de LSD soit liée avec l'âge ?

La p valeur est très petite donc on accepte H_1 *.i.e.* que la consommation de LSD est liée à l'âge.