Examen de statistiques M1 STAAE, Session 1 (décembre 2022)

Durée : 120 min.

Consignes : Veillez à bien **expliquer votre démarche en justifiant vos choix en matière de test**, et à fournir les **résultats intermédiaires** qui vous permettent de conclure quand cela est nécessaire. Une attention toute particulière sera portée quant à la **clarté** et à la **concision** de vos réponses. Le **seuil alpha** est fixé à **5%**.

NB : La calculatrice et les **documents** sous format **papier** sont autorisés. Pas de téléphone ni de tablette/ordinateur.

EXERCICE 1 (10/20 points) : BIERES

A l'issue d'un test de dégustation, on a recueilli 8 notes d'acidité pour 4 marques de bières (B1 à B4).

Chaque bière est évaluée par un jury indépendant, et les bières sont notées sur une échelle de 0 (absence totale d'acidité) à 10 (acidité extrême). On souhaite savoir si les bières diffèrent par leur acidité.

Les données ont été saisies dans un tableur et stockées sous R dans un objet nommé DataNotesBieres.

```
> str(DataNotesBieres)
'data.frame':   32 obs. of  2 variables:
 $ Marque: Factor w/ 4 levels "B1","B2","B3",..: 1 1 1 1 1 1 1 1 2 2 ...
 $ Note  : num  5 5 5 6 7 7 8 10 0 1 ...
```

Questions :

1. Indiquez la nature, le rôle et le mode de sélection de chaque variable. Justifiez brièvement.

2. S'agit-il d'une étude observationnelle ou expérimentale ? le plan est-il équilibré ? Justifiez brièvement.

3. Quels tests pourriez-vous a priori utiliser pour répondre à la question posée ? Justifiez brièvement.

Malheureusement, le statisticien en charge de l'analyse des données a ramené chez lui 1 litre de chaque bière pour les tester, et a clairement abusé sur les quantités ingérées…. Il roupille actuellement, en boule sous son bureau….
Vous découvrez sur son bureau divers documents éparpillés, sur lesquels figurent différents tests statistiques qu'il a réalisé pour l'étude.
En tant que son stagiaire, il vous revient la mission de trier et d'interpréter correctement les bons tests pour fournir vos conclusions au comité scientifique de l'étude d'ici 1 heure.

4. Répondez à la question de recherche en indiquant quelles sont les sorties R (parmi les propositions numérotées suivantes) que vous avez utilisé, dans l'ordre, pour chaque étape du processus d'interprétation. Commentez-les brièvement (quelles informations vous apportent-elles ?) et concluez.

Vous disposez des informations suivantes :

1)

```
Analysis of Variance Table

Response: Note
           Df Sum Sq Mean Sq F value    Pr(>F)
Marque      3 184.84  61.615  17.145 1.623e-06 ***
Residuals 28 100.62   3.594
```

2)

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Note ~ Marque, data = DataNotesBieres)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
B2 - B1 == 0  -3.6250     0.9479  -3.824   0.0035 **
B3 - B1 == 0   1.0000     0.9479   1.055   0.7191
B4 - B1 == 0  -4.7500     0.9479  -5.011   <0.001 ***
B3 - B2 == 0   4.6250     0.9479   4.879   <0.001 ***
B4 - B2 == 0  -1.1250     0.9479  -1.187   0.6399
B4 - B3 == 0  -5.7500     0.9479  -6.066   <0.001 ***
```

3)

```
        Bartlett test of homogeneity of variances

data:  Note by Marque
Bartlett's K-squared = 0.64485, df = 3, p-value = 0.8861
```

4)

```
        Shapiro-Wilk normality test

data:  DataNotesBieres$Note
W = 0.95316, p-value = 0.1771
```

5)

```
        Kruskal-Wallis rank sum test

data:  Note by Marque
Kruskal-Wallis chi-squared = 20.672, df = 3, p-value = 0.0001232
```

6)

```
aggregate(Note~Marque,data=DataNotesBieres,FUN=mean)
Marque  Note
    B1 6.625
    B2 3.000
    B3 7.625
    B4 1.875
```

7)

```
        Shapiro-Wilk normality test

data:  residuals(anova_biere)
W = 0.94593, p-value = 0.1104
```

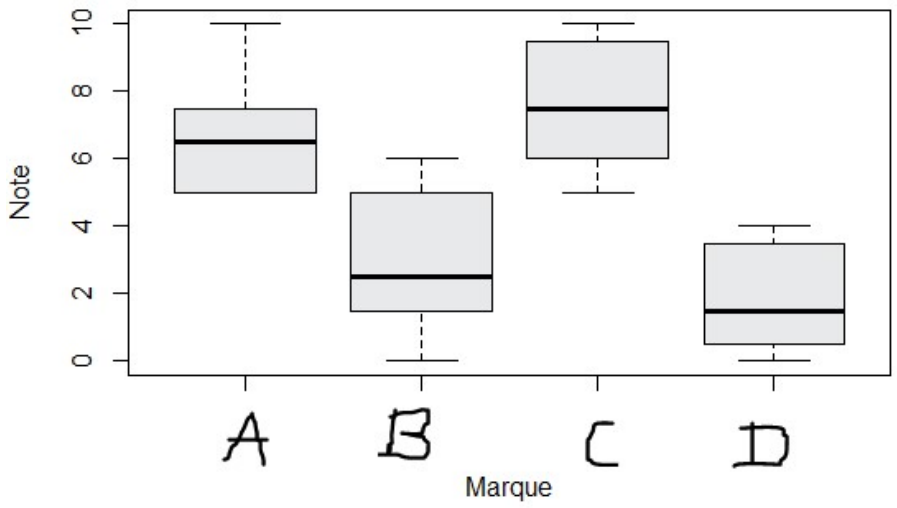
8)

```
Call:
lm(formula = Note ~ Marque, data = DataNotesBieres)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0000 -1.6250 -0.3125  1.3750  3.3750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6250     0.6702   9.885 1.24e-10 ***
MarqueB2     -3.6250     0.9479  -3.824 0.000672 ***
MarqueB3      1.0000     0.9479   1.055 0.300440
MarqueB4     -4.7500     0.9479  -5.011 2.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.896 on 28 degrees of freedom
Multiple R-squared:  0.6475,    Adjusted R-squared:  0.6097
F-statistic: 17.14 on 3 and 28 DF,  p-value: 1.623e-06
```

9)

```
        Pairwise comparisons using Dunn's all-pairs test

data: Note by Marque
alternative hypothesis: two.sided
P value adjustment method: holm
H0
              z value   Pr(>|z|)
B2 - B1 == 0   2.424 0.04601294   *
B3 - B1 == 0   0.616 0.84321265
B4 - B1 == 0   3.228 0.00623416  **
B3 - B2 == 0   3.040 0.00944966  **
B4 - B2 == 0   0.804 0.84321265
B4 - B3 == 0   3.844 0.00072607 ***
```

5. Sur la base de vos conclusions, indiquez sur votre copie à quelle marque de bière (B1 à B4) correspondent les lettres « A » à « D » sur le boxplot suivant.



EXERCICE 2 (10/20 points) : QCMs

*Pour **chaque question**, **reportez** sur votre **copie** la **lettre** de la/des réponses correctes (ne pas réécrire les réponses !).*

1. Parmi les propositions suivantes, laquelle ou lesquelles sont correctes :

a. Les variables dépendantes quantitatives permettent de comparer des fréquences observées et attendues.

b. Les tests de Student, de Pearson et ANOVA sont tous des tests de comparaison de moyennes.

c. Les tests non-paramétriques ont pour avantage d'être adaptés aux très petits échantillons et d'être les seuls tests permettant de traiter des données qualitatives.

d. L'intervalle de confiance est directement proportionnel à la moyenne de l'échantillon mesuré.

e. L'Homoscédasticité est un terme savant voulant signifier l'égalité des variances.

f. Si les conditions de l'ANOVA à 1 facteur ne sont pas remplies, et que l'énoncé concerne >3 races de vaches, on peut basculer sur un test de Krussel Jacob.

2. Quelles sont la/les proposition(s) fausse(s) :

a. Le calcul d'une moyenne est moins sensible aux valeurs extrêmes que le calcul d'une médiane.

b. Plus le niveau de validité interne est élevé, plus les résultats de l'étude pourront être généralisés.

c. Un échantillonnage de commodité correspond à un échantillonnage probabiliste.

d. Les statistiques inférentielles ont vocation à émettre des généralisations ou estimations concernant les paramètres de populations d'intérêt, à partir d'observations s'appuyant sur des échantillons.

3. Concernant les tests relevant la statistique du χ^2, quelles sont la/les proposition(s) correcte(s) :

a. Le test de conformité a pour but de comparer des distributions observées entre échantillons indépendants.

b. Les différences entre effectifs observés et attendues sont uniquement dues aux fluctuations d'échantillonnage.

c. Le test d'indépendance a pour but de tester l'existence d'une liaison entre deux variable aléatoire qualitatives issues du même échantillon.

d. Plus χ^2obs sera grand, moins H0 sera probable.

e. 2 variables sont dites indépendantes lorsque les variations de l'une des variables n'influencent pas les variations de l'autre.

4. Une équipe de recherche souhaite finaliser la mise au point de droïdes et s'intéresse à leur vitesse de pointe selon le type de pneu qu'ils portent (épais ou fins) et leur revêtement (fibre de carbone ou acier). Choisir les réponses vraies :

a. L'étude porte sur deux facteurs chacun comprenant deux modalités.

b. La vitesse de pointe des droïdes est une variable indépendante.

c. Le plan factoriel devrait contenir 4 conditions expérimentales.

d. La vitesse de pointe est une variable quantitative discrète.

5. Concernant la corrélation, quelles sont la/les proposition(s) fausse(s) :

a. Une intensité de corrélation forte se traduit par : une dispersion réduite du nuage de points et une pente de la relation linéaire forte.

b. Une intensité de corrélation forte se traduit par : une dispersion forte du nuage de points et une pente de la relation linéaire faible.

c. Une intensité de corrélation forte se traduit par : une dispersion réduite du nuage de points et une pente de la relation quelconque.

d. Une intensité de corrélation forte se traduit par : une dispersion forte du nuage de points et une pente de la relation quelconque.

e. La corrélation permet d'établir l'existence d'un lien entre 2VA (X,Y) jouant des rôles asymétriques.

f. Pour un coefficient de corrélation de Pearson = - 0.85, les points s'alignent sur une droite croissante.

g. Le coefficient de Sperman est adapté à l'étude de la monotocité du lien entre 2 variables qualitatives ordinales.

6. L'erreur de type β correspond à la :

a. Probabilité de ne pas rejeter H0 alors que celle-ci est fausse.

b. Probabilité de rejeter H0 alors qu'elle est vraie.

c. Probabilité de rejeter H1 alors qu'elle est vraie.

d. Probabilité d'accepter H1 alors qu'elle est fausse.

7. Concernant la puissance d'un test statistique, quelles sont les propositions correctes ?

a. Il diminue avec la taille de l'échantillon.

b. Il augmente avec la taille de l'échantillon.

c. La capacité de l'expérience à mettre en évidence une différence, même petite.

d. Il augmente lorsque α augmente.

e. Un test bilatéral est aussi efficace qu'un test unilatéral répété deux fois.

f. Un test non paramétrique est moins puissant qu'un test paramétrique.

8. Quelles sont la ou les bonnes réponses ?

a. Lorsqu'un échantillon est composé de plus de 30 individus, la formule utilisée dans le calcul de l'intervalle de confiance contient la valeur statistique Z, qui se lit dans la table de la loi normale centrée réduite.

b. Plus la variance est petite, plus les valeurs de la variable aléatoire sont proches de la moyenne.

c. Plus l'intervalle de confiance est grand, plus l'estimation est précise.

d. Dans le cas d'une distribution asymétrique positive, la valeur de la moyenne est inférieure à celle de la médiane, elle-même inférieure à celle du mode.

9. **Protocole et planification :** Un étudiant souhaite, par une approche expérimentale, déterminer l'impact d'une forte précipitation sur la richesse spécifique des macroinvertébrés à la surface du sol en forêt. Il définit des placettes de 3 m², et pour simuler une forte précipitation, il arrosera abondamment certaines placettes pendant plusieurs minutes. Après 30 min, il collectera et identifiera à l'espèce tous les macroinvertébrés pour déterminer la richesse spécifique de chaque placette (arrosées ou non).

Parmi les 4 protocoles expérimentaux suivants, lequel est le mieux adapté pour tester l'effet du facteur "précipitation" sur la richesse spécifique en macroinvertébrés du sol ?

a. Il définit 20 placettes qui seront arrosées et 20 placettes qui ne seront pas arrosées. Pour des raisons pratiques, il positionne les placettes arrosées à proximité des routes forestières pour pouvoir transporter le dispositif d'arrosage plus facilement. Il répartit les placettes non arrosées de manière aléatoire dans la forêt.

b. Il décide d'attendre qu'un orage éclate sur une partie de la forêt seulement et répartit ensuite aléatoirement 20 placettes sur la zone couverte par l'orage et 20 placettes sur la zone épargnée par l'orage.

c. Il répartit aléatoirement dans la forêt 10 paires de placettes. Dans chaque paire, 1 placette seulement est arrosée. Les deux placettes sont distantes de 4 mètres.

d. Il répartit aléatoirement dans la forêt 20 placettes qui seront arrosées et 20 autres placettes qui ne seront pas arrosées.

10. On retrouve des coccinelles asiatiques dans trois endroits différents du jardin de Marie Nkäfer. Elle compte le nombre de coccinelles de chaque couleur/motif sur différents supports dans son jardin et obtient le tableau de contingence suivant sur R :

	Bouleau	Fenêtre	Fusain	Sum
Rouge	39	34	36	109
Noir à points rouges	13	25	12	50
Noir à points jaunes	12	16	14	42
Jaune	21	22	31	74
Sum	85	97	93	275

Marie se demande si la répartition des effectifs de chaque type de coccinelle (couleur/motif) diffère selon le type de support, et décide de réaliser un test du Chi².

Question 10. 1. : Quel test du Chi² doit-elle utiliser ?

a. Chi² de conformité.
b. Chi² d'ajustement.
c. Chi² d'homogénéité.
d. Chi² d'indépendance.

Question 10. 2. : Quel est le nombre de degré de liberté associé au test ?

a. 12.
b. 6.
c. 7.
d. 20.
e. 5.
f. 9.

11. Un chercheur mesure le rendement d'une variété de blé dans 4 parcelles agricoles. Dans chaque parcelle, il répartit aléatoirement 5 quadrats dans lesquels il fait ses mesures de rendement. Il veut savoir si les 4 parcelles diffèrent en termes de rendement.
Quel test doit-il faire ?

a. Un test de Kruskal-Wallis.
b. Un test de student pour échantillons indépendants.
c. Un test de Wilcoxon-Mann-Whitney.
d. Un test de Student apparié.
e. Un test de Wilcoxon apparié.
f. Une Anova.

12. Dans une étude s'intéressant au microclimat forestier, des chercheurs effectuent une mesure de la température à 1.5m au-dessus du niveau du sol dans différents points d'une forêt. Pour chaque point, ils relèvent également la hauteur des arbres (hauteur_arbre), le pH du sol (pH) et le pourcentage de couverture de la canopée (couv_canop).
Ils réalisent ensuite une régression linéaire multiple pour connaître la relation entre la température et ces autres variables sur R. Voici le résultat obtenu :

```
Call:
lm(formula = temp ~ hauteur_arbre + pH + couv_canop, data = qcm)

Residuals:
    Min      1Q  Median      3Q     Max
-32.884 -11.382   0.154   7.662  44.409

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.6496    26.6710   0.099   0.9214
hauteur_arbre  4.3927     0.9794   4.485  7.9e-05 ***
pH            -0.5872     1.8166  -0.323   0.7485
couv_canop     2.4765     1.4029   1.765   0.0865 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.12 on 34 degrees of freedom
Multiple R-squared:  0.5836,    Adjusted R-squared:  0.5468
F-statistic: 15.88 on 3 and 34 DF,  p-value: 1.262e-06
```

Parmi les propositions suivantes, laquelle ou lesquelles sont correctes :

a. La formule de la droite de régression de ce modèle est y = 2.6496 + 4.3927*x.
b. 19,12% de la variation de la température est expliqué par ce modèle à 3 variables.
c. Le ph n'est pas lié significativement à la température dans la forêt.
d. La hauteur des arbres et la couverture de la canopée influencent significativement et positivement la température forestière.
e. Dans le cas d'une régression multiple, on doit lire le R^2 correspondant au Adjusted R-Squared.
f. On va retirer du modèle le ph, puis recalculer les coefficients de régression et les p-values associées.

13. En France un pommier de variété A produit en moyenne 150 kg de pommes par an. Nous cherchons à savoir si le verger d'un agriculteur produit davantage de kg de pommes que la population de référence. L'agriculteur a pesé la production de 10 de ses pommiers sur une année. Ses données sont stockées dans un objet R nommé Pommiers. On suppose ici que la variable Masse suit une loi normale.

Question 13. 1. : Quelles lignes de code R doit-on utiliser pour répondre à cette question ?

a.
```
Pommiers<-c(190,150,147,192,146,135,154,171,183,166)
t.test(Pommiers,mu=150,alternative = "less")
```

b.
```
Pommiers<-c(190,150,147,192,146,135,154,171,183,166)
t.test(Pommiers,mu=150,alternative = "greater")
```

c.
```
Pommiers<-c(190,150,147,192,146,135,154,171,183,166)
t.test(Pommiers,mu=150,alternative = "two.sided")
```

Question 13. 2. : Que peut-on conclure avec un t théorique de 1.833 et un t observé de 2.112 ?

a. Rejet H0 au seuil de 5%, acceptation H1.

b. Non rejet H0 au seuil de 5%, rejet H1.