Examen de statistiques M1 AETPF, Session 1 (janvier 2024)

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 (12.5/20 points) : *(NB : Toute ressemblance avec des faits et des personnages existants ou ayant existé serait purement fortuite et ne pourrait être que le fruit d'une pure coïncidence)*

Partie A : Des enseignant.e.s de l'université d'Hamienvil préparent le plan d'échantillonnage de terrain qui sera réalisé par étudiants de Magister1 dans le cadre d'un module de Diagnostic agroécologique. Pour cela, sur ArcSIG, ils utilisent des couches d'occupation du sol couvrant la zone géographique étudiée ; la vallée de la ∑. Ils réalisent sous R le tirage au sort de 100 couples de coordonnées WGPS2024 localisant les points d'échantillonnage au sein de différents habitats d'intérêt : Prairie pâturée (Prai_Pat), prairie non pâturée (Prai_NP), bois jeune (Bois_J), bois vieux (Bois_V), lisière bois/prairie (LisR).

Ces 100 points d'échantillonnages sont ensuite distribués, en tenant compte des distances entre sites, à 2 groupes d'étudiants (les Abeilles Enragées : AE et les Ecureuils Belliqueux : EB) afin qu'ils puissent réaliser un relevé botanique (méthode du quadra) et poser un piège à arthropodes (barber) par point.

Après la phase de terrain et d'identification des arthropodes en laboratoire, les étudiants ont saisies leurs données dans un tableur, qui a ensuite été importé et stockées sous R dans un objet nommé Data. Il comporte les variables suivantes : La richesse taxonomique (S_Taxo_Arth) et l'abondance (Abdce_Arth) en arthropodes identifiés dans chaque barber, la richesse taxonomique de la flore relevée dans chaque quadra (S_Taxo_Vgtl), l'habitat correspondant à chaque site (Habitat) ainsi que le groupe l'ayant échantillonné (Groupe).

```
> str(Data)
'data.frame':   100 obs. of  5 variables:
 $ S_Taxo_Arth: int  11 11 12 14 4 5 6 6 7 7 ...
 $ Abdce_Arth : int  89 87 100 94 101 106 87 91 109 96 ...
 $ S_Taxo_Vgtl: int  27 26 26 27 8 13 15 17 18 20 ...
 $ Habitat    : Factor w/ 5 levels "Bois_J","Bois_V",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Groupe     : Factor w/ 2 levels "AE","EB": 1 2 1 1 2 2 2 1 1 2 ...
```

```
> summary(Data)
  S_Taxo_Arth      Abdce_Arth      S_Taxo_Vgtl        Habitat   Groupe
 Min.   : 3.00   Min.   : 55.00   Min.   : 8.00   Bois_J  :21   AE:47
 1st Qu.: 7.00   1st Qu.: 85.75   1st Qu.:19.75   Bois_V  :24   EB:53
 Median : 8.00   Median : 99.50   Median :23.00   LisR    :18
 Mean   : 8.14   Mean   :111.49   Mean   :22.36   Prai_NP :21
 3rd Qu.:10.00   3rd Qu.:143.00   3rd Qu.:26.00   Prai_Pat:16
 Max.   :14.00   Max.   :167.00   Max.   :37.00
```

Question 1. **a) S'agit-il d'une étude observationnelle ou expérimentale ? Justifiez brièvement.**

b) A quoi correspondent les individus statistiques dans cette étude ?

c) Quel type d'échantillonnage a été réalisé ? Justifiez brièvement.

d) Le plan d'échantillonnage de la biodiversité est-il équilibré ?

Partie B : Un des Ecureuil Belliqueux (que nous appellerons arbitrairement « Grincheux ») pense que son groupe a été désavantagé sur le terrain, en termes de distribution de nombre de sites à échantillonner par habitat, par rapport à l'autre groupe. Il exécute la ligne de code suivante sous R :

```
addmargins(table(Data$Habitat,Data$Groupe))

          AE  EB Sum
Bois_J     7  14  21
Bois_V    13  11  24
LisR       8  10  18
Prai_NP   10  11  21
Prai_Pat   9   7  16
Sum       47  53 100
```

Question 2. **a) Décrire succinctement à quoi sert cette commande R, ainsi que la matrice obtenue.**

b) Quel test statistique doit il utiliser pour tester son hypothèse avant d'aller se plaindre aux enseignant.e.s ? Justifiez brièvement.

c) A l'issue du test adéquat, il obtient une p-value = 0.6146. Que peut conclure Grincheux ?

Partie C : Les étudiants cherchent à savoir s'il existe un lien entre la richesse floristique et la richesse en arthropodes sur l'ensemble des sites. Ils réalisent les tests dont les résultats sous R sont les suivants :

```
        Shapiro-Wilk normality test

data:  Data$S_Taxo_Arth
W = 0.97565, p-value = 0.06067
```

```
        Shapiro-Wilk normality test

data:  Data$S_Taxo_Vgtl
W = 0.98407, p-value = 0.2716
```

```
data:  Data$S_Taxo_Arth and Data$S_Taxo_Vgtl
t = 16.653, df = 98, p-value ?
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7978758 0.9034622
sample estimates:
      cor
0.8595825
```

Question 3. **a) Quel est le test statistique approprié pour tester l'existence de ce lien ? Justifiez brièvement.**

b) Quelle ligne de code les étudiants doivent-ils exécuter sous R pour le réaliser ?

c) Que peuvent-ils conclure à l'issue de ce test ? Justifiez brièvement.

d) Concernant la question de recherche, qu'apporterait comme information(s) supplémentaire(s) l'utilisation d'1 régression linéaire ? Quelle(s) commande(s) R permettrait d'effectuer cette analyse ?

Partie D : Les étudiants s'intéressent maintenant aux variations d'abondance des arthropodes dans les différents habitats échantillonnés. En raison d'un travail en équipe désordonné et d'un manque de communication, des sous-groupes d'étudiants se lancent chacun de leur côté dans différentes analyses, de façon non concertée.

Au moment de rassembler les résultats pour préparer l'oral de restitution du projet, personne n'est d'accord sur quels sont les bons tests à utiliser, ni comment les interpréter. Comme vous êtes la personne en charge de présenter et discuter ces résultats face aux enseignant.e.s, mais aussi de répondre à leurs questions sur les statistiques utilisées, vous devez prendre des décisions.

Les étudiants des différents sous-groupes vous remettent une liasse désordonnée de bouts de papiers sur lesquels sont imprimées les sorties R suivantes, sans plus d'explication…. Vous avez 1h pour vous préparer….

1)

```
Analysis of Variance Table

Response: Data$Abdce_Arth
             Df Sum Sq Mean Sq F value    Pr(>F)
Data$Habitat  4  90957 22739.2     ?    < 2.2e-16 ***
Residuals    95   6512    68.5     ?
```

2)

```
Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Abdce_Arth ~ Habitat, data = Data)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Bois_V - Bois_J == 0     43.571      2.474  17.612  < 1e-04 ***
LisR - Bois_J == 0       61.016      2.659  22.944  < 1e-04 ***
Prai_NP - Bois_J == 0   -10.000      2.555  -3.914  0.00156 **
Prai_Pat - Bois_J == 0  -16.324      2.747  -5.942  < 1e-04 ***
LisR - Bois_V == 0       17.444      2.582   6.757  < 1e-04 ***
Prai_NP - Bois_V == 0   -53.571      2.474 -21.654  < 1e-04 ***
Prai_Pat - Bois_V == 0  -59.896      2.672 -22.415  < 1e-04 ***
Prai_NP - LisR == 0     -71.016      2.659 -26.704  < 1e-04 ***
Prai_Pat - LisR == 0    -77.340      2.845 -27.187  < 1e-04 ***
Prai_Pat - Prai_NP == 0  -6.324      2.747  -2.302  0.15291
```

3)

```
        Bartlett test of homogeneity of variances

data:  Data$Abdce_Arth by Data$Habitat
Bartlett's K-squared = 8.5896, df = 4, p-value = 0.07222
```

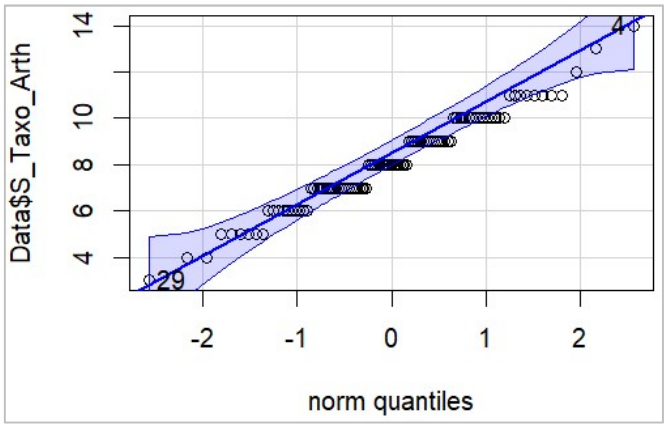
4)

```
        Shapiro-Wilk normality test

data:  Data$Abdce_Arth
W = 0.90241, p-value = 1.838e-06
```

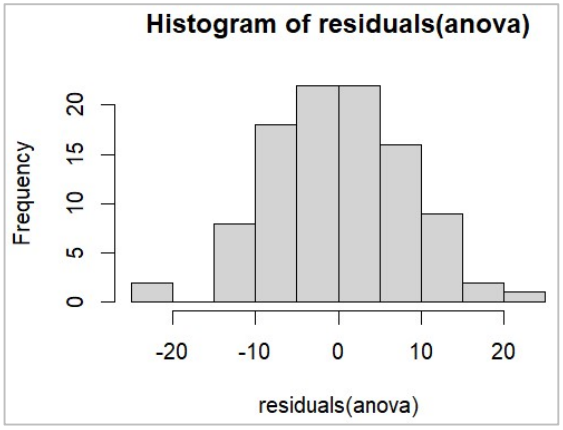
5)

```
        Kruskal-Wallis rank sum test

data:  Data$Abdce_Arth by Data$Habitat
Kruskal-Wallis chi-squared = 84.155, df = 4, p-value < 2.2e-16
```

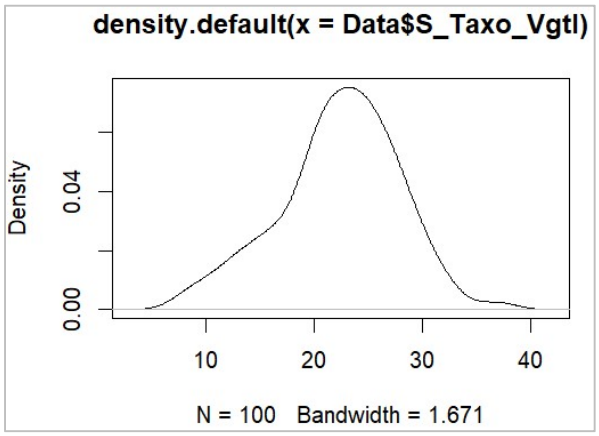
6)



7)



8)



9)

```
> aggregate(Abdce_Arth~Habitat,data=Data, FUN=mean)
   Habitat Abdce_Arth
1   Bois_J    94.7619
2   Bois_V   138.3333
3     LisR   155.7778
4  Prai_NP    84.7619
5 Prai_Pat    78.4375
```

10)

```
        Shapiro-Wilk normality test

data:  residuals(anova)
W = 0.99312, p-value = 0.8955
```

11)

```
Call:
lm(formula = Data$Abdce_Arth ~ Data$Habitat)

Residuals:
     Min       1Q   Median       3Q      Max
-23.4375  -5.5186  -0.1076   5.2381  20.5625

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             94.762      1.807  52.450  < 2e-16 ***
Data$HabitatBois_V      43.571      2.474  17.612  < 2e-16 ***
Data$HabitatLisR        61.016      2.659  22.944  < 2e-16 ***
Data$HabitatPrai_NP    -10.000      2.555  -3.914 0.000171 ***
Data$HabitatPrai_Pat   -16.324      2.747  -5.942 4.63e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.279 on 95 degrees of freedom
Multiple R-squared:  0.9332,    Adjusted R-squared:  0.9304
F-statistic: 331.7 on 4 and 95 DF,  p-value: < 2.2e-16
```

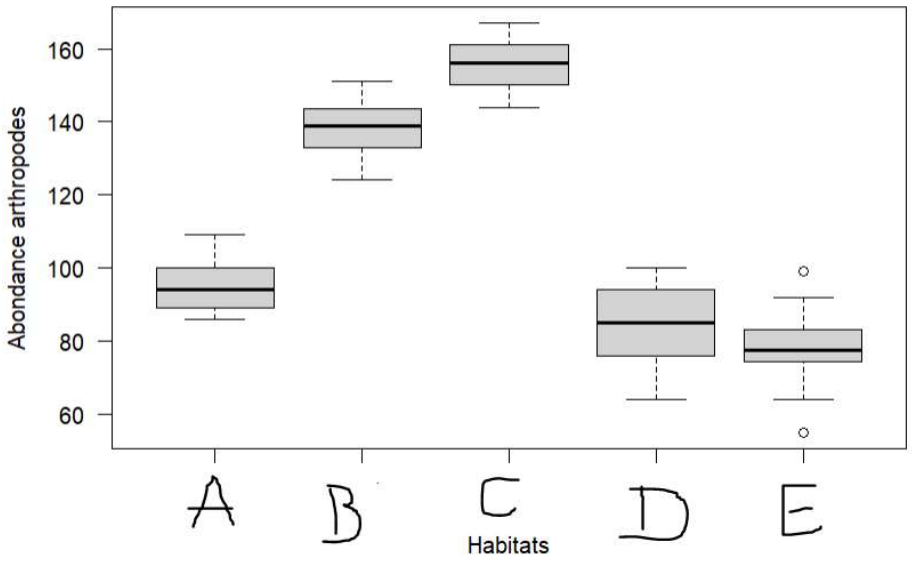
12)

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

data: Data$Abdce_Arth by Data$Habitat
alternative hypothesis: two.sided
P value adjustment method: holm
H0
                          z value   Pr(>|z|)
Bois_V - Bois_J == 0        3.337  0.0042354  **
LisR - Bois_J == 0          5.259 1.0127e-06 ***
Prai_NP - Bois_J == 0       1.785  0.1485552
Prai_Pat - Bois_J == 0      2.520  0.0468962   *
LisR - Bois_V == 0          2.220  0.0792607   .
Prai_NP - Bois_V == 0       5.180 1.3287e-06 ***
Prai_Pat - Bois_V == 0      5.681 1.0728e-07 ***
Prai_NP - LisR == 0         6.974 2.7708e-11 ***
Prai_Pat - LisR == 0        7.351 1.9712e-12 ***
Prai_Pat - Prai_NP == 0     0.860  0.3895574
```

Question 4. **a) Répondez à la question de recherche de la façon la plus complète, précise et synthétique possible en utilisant toutes les informations à votre disposition. Indiquez quelles sont les sorties R (parmi les propositions numérotées ci-dessus) que vous avez utilisé, et dans quel ordre, pour chaque étape du processus d'interprétation. Justifiez leur usage, commentez-les brièvement (quelles informations vous apportent-elles ?) et concluez d'un point de vu statistique ET ECOLOGIQUE.**

b) Quelle valeur remplace le point d'interrogation pour la F value de la sortie 1) ?

Partie E : représentation graphique



Question 5. **Sur la base de vos conclusions, indiquez sur votre copie à quels habitats correspondent les lettres « A » à « E » du boxplot ci-dessus. Justifiez brièvement.**

EXERCICE 2 (7.5/20 points) : QCMs

Répondre sur la grille fournie en dernière page de ce document, en y renseignant votre numéro étudiant, et la joindre à votre copie. Pour chaque question : le point est attribué si toutes les réponses sont correctes.

1. données sont correctes (aucune erreur) et la note minimale pour chaque question est de 0 (pas de points négatifs). **Parmi les propositions suivantes, laquelle ou lesquelles sont correctes :**

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

b. L'ANOVA à 1 facteur est l'alternative au test de Kruskal-Wallis si les conditions d'application de ce dernier ne sont pas vérifiées.

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. On finit par accepter H_0 si on a échoué à la faire rejeter.

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 plan expérimental peut être croisé et incomplet en même temps.

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.

e. La p-value correspond au risque de rejeter H1 à tort.

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 tests du χ^2 sont des tests unilatéraux.
c. Le test d'indépendance a pour but de tester l'existence d'une liaison entre deux variables aléatoires quantitatives issues du même échantillon.
d. Plus χ^2obs sera grand, moins H0 sera probable.
e. Ce sont des tests applicables à des échantillons de taille $n < 20$.

4. 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.

5. 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.
e. Aucune de ces propositions n'est correcte.

6. 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.
f. Un test non paramétrique est moins puissant qu'un test paramétrique.

7. Parmi les affirmations suivantes concernant la p-value, laquelle est (lesquelles sont) fausse(s) ?

a. Plus la p-value est grande, plus le risque de se tromper en rejetant H0 est faible.
b. Elle représente la probabilité, d'après les observations, d'accepter à tort l'hypothèse nulle H0.
c. Si H0 est vraie, c'est la probabilité d'obtenir le résultat observé sur l'échantillon sous l'effet seul du hasard.
d. Pour un seuil de significativité α donné, on rejette H0 si $\alpha > p$.

8. 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.

9. Parmi les propositions suivantes, laquelle est (lesquelles sont) correcte(s) ?

a. Plus l'intervalle de confiance est grand, plus l'estimation est précise.
b. 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.
c. La variance est un indicateur de dispersion toujours positif. Plus elle est petite, plus les valeurs de la variable aléatoire sont proches de la moyenne.
d. Un test bilatéral permet conclure à la supériorité d'un traitement.
e. L'unité d'un écart-type permet mieux de se rendre compte de la distribution des valeurs par rapport à la variance.

10. La prof de stats vous donne cette sortie R. Que pouvez-vous en déduire ?

```
  Welch Two Sample t-test

data:  groupe1 and groupe2
t = -2.0649, df = 57.49, p-value = 0.02268
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.4627194
sample estimates:
mean of x mean of y
 74.52957  79.84001
```

a. La moyenne du groupe 1 est significativement supérieure à la moyenne du groupe 2.
b. La moyenne du groupe 1 est significativement inférieure à la moyenne du groupe 2.
c. Il n'y a pas de différence significative entre les moyennes des deux groupes.
d. Les moyennes des deux groupes sont égales.
e. Il y a hétéroscédasticité.
f. La distribution des données est unimodale et symétrique.
g. Elle aurait aussi bien pu faire un test de Wilcoxon-Mann-Whitney.

11. Microclimat :

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 est (lesquelles sont) correcte(s) ?

a. La formule de la droite de régression de ce modèle est y = 2.6496 + 4.3927*x.
b. 15.88% 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. Ce modèle est pertinent pour expliquer les variations de température en forêt.
f. Dans le cas d'une régression multiple, on doit lire le R^2 correspondant au Adjusted R-Squared.
g. L'étape suivante sera de retirer du modèle le pH, puis recalculer les coefficients et les p-values associées.

12. A partir des données suivantes, pourquoi n'est-il pas possible d'effectuer un test du Chi² ?

Effectifs observés : 19 / 7 / 99 / 56 / 40 / 10 / 28 / 45 / 34
Effectifs théoriques : 12 / 3 / 4 / 15 / 0 / 4 / 10 / 28 / 0

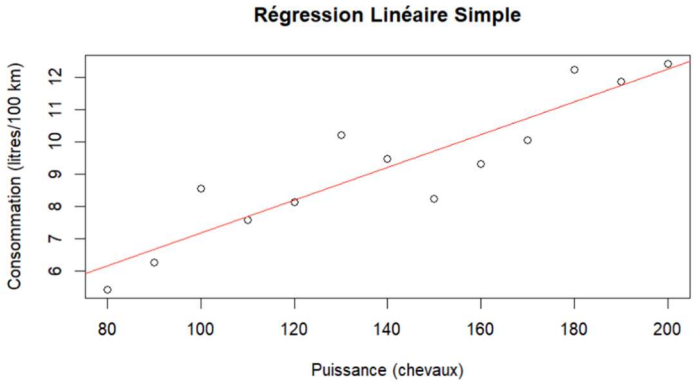
a. Plus de 20% des effectifs théoriques ont une valeur < 5.
b. Le nombre d'individus statistique est très faible (n<50).
c. La taille des effectifs théoriques n'est pas égale à celle des effectifs observés.
d. Il y a une valeur 0.
e. L'ensemble de ses propositions.

13. En roue libre :

Supposons que vous ayez collecté des données sur la consommation de carburant (en litres par 100 kilomètres) pour différentes voitures en fonction de leur puissance (en chevaux). Voici un jeu de données fictif en R :

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.88	0.52	3.62	0.003
Puissance	0.05	0.01	5.00	<0.001 *



Parmi les propositions suivantes, laquelle est (lesquelles sont) correcte(s) ?

a. La relation étudiée est asymétrique.
b. L'équation de la droite de régression est : Puissance = 1.88 + 0.05 x Consommation
c. Pour chaque cheval de puissance supplémentaire, la consommation moyenne de carburant augmente de 0.05 litre/100 km.
d. La variable "Puissance" est statistiquement significative à un niveau de confiance de 0.05.
e. L'intercept de 1.88 est interprété comme la consommation de carburant en litres pour une voiture de puissance nulle, sur un parcours de 100 km.

14. Parmi les propositions suivantes, laquelle est (lesquelles sont) correcte(s) ?

Une étude a été menée pour évaluer l'effet de trois différents types de fertilisants sur le rendement de trois variétés de plantes (A, B, C). Les rendements ont été mesurés en kilogrammes par hectare. Les données ont été organisées dans un tableau appelé "donnees_fertilisants" avec trois colonnes : "Fertilisant", "Variete" et "Rendement".

Pour réaliser une ANOVA sur ces données, quelles lignes de code vous semblent les plus pertinentes ?

a. Code1 <- lm(Rendement ~ Varieté, data = donnees_fertilisants)
   Res1<- summary(Code1)

b. Code2 <- lm(Rendement ~ Fertilisant + Variete, data = donnees_fertilisants)
   Res2<- anova(Code2)

c. Code3 <- lm(Rendement ~ Fertilisant * Variete, data = donnees_fertilisants)
   Res3 <- anova(Code3)

d. Code4 <- t.test(Rendement ~ Fertilisant, data = donnees_fertilisants)
   Res4<- summary(Code3)

e. Code5 <- Tukey(Rendement ~ Fertilisant * Variete, data = donnees_fertilisants)
   Res5<- anova(Code5)

15. Parmi les propositions suivantes concernant les résidus, laquelle est (lesquelles sont) correcte(s) ?

a. Sont étudiés pour l'applicabilité ou non de modèles, de type régression linéaire ou ANOVA
b. Représentent les différences entre les valeurs observées et les valeurs estimées par le modèle
c. Plus leur somme au carré est élevée, meilleur est l'ajustement du modèle aux données
d. Représentent la partie non expliquée par l'équation du modèle
e. Leur variance doit être corrélés linéairement avec les variables indépendantes

Questions	Réponses (cocher les cases correspondant à vos réponses)						
	A	B	C	D	E	F	G
Q1							
Q2							
Q3							
Q4							
Q5							
Q6							
Q7							
Q8							
Q9							
Q10							
Q11							
Q12							
Q13							
Q14							
Q15							
Q16							
Q17							
Q18							
Q19							
Q20							

(En cas d'erreur dans les réponses apportées sur la grille ci-dessus, vous pouvez les corriger en reportant l'ensemble de vos réponses correctes dans le tableau ci-dessous. Si ce dernier comporte des réponses, uniquement celles qu'il contient seront prises en considération pour l'évaluation)

Questions	Réponses (cocher les cases correspondant à vos réponses)						
	A	B	C	D	E	F	G
Q1							
Q2							
Q3							
Q4							
Q5							
Q6							
Q7							
Q8							
Q9							
Q10							
Q11							
Q12							
Q13							
Q14							
Q15							
Q16							
Q17							
Q18							
Q19							
Q20							

```
data:  Data$S_Taxo_Arth and Data$S_Taxo_Vgtl
t = 16.653, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7978758 0.9034622
sample estimates:
      cor
0.8595825
```

VRAI FAUX

2) Quelles affirmations sont vraies ?

- Un plan expérimental emboîté est un plan où chaque modalité du 1er facteur est croisée avec chaque modalités du 2nd facteur

- Lors d'un échantillonnage, il est possible de sélectionner délibérément certains individus par rapport à d'autres

- L'échantillon doit être représentatif de la population totale pour un échantillonnage probabiliste

- Un plan expérimental peut être croisé et incomplet en même temps

- La variabilité résiduelle est la part de variabilité due aux fluctuations d'échantillonnage (hasard)

- Lorsque la moyenne, la médiane et le mode sont identiques, la distribution est symétrique

- On utilise la loi de Student lorsque l'effectif de(s) l'échantillon(s) est (sont) grand(s) ($n \geq 30$)

- La p-value correspond au risque de rejeter H1 à tort

2) Quelles affirmations sont vraies ?

- La droite de régression passe par tous les points observés

- Il existe un point moyen par lequel passe la droite de régression

- Le test de Fisher permet de déterminer si au moins une VI influence la VD

- La réalisation d'une régression linéaire simple est possible uniquement si les résidus (observations) sont corrélés

Laquelle de ces affirmations est fausse ?

A. Le théorème central limite explique entre autres l'omniprésence de la loi normale dans la nature.
B. La précision d'une estimation est proportionnelle à la dispersion de l'échantillon et inversement proportionnel à la taille de l'échantillon.
C. La puissance d'un test correspond à la probabilité de rejeter H_0 quand celle-ci est fausse.
D. La statistique t de Student est utilisée dans le calcul de l'intervalle de confiance pour les grands échantillons. ***

Parmi les affirmations suivantes sur les tests de différences de moyenne, la ou lesquelles est/sont vraie(s) ?

A. Si l'on conclue à H_0 après un test de Kruskal-Wallis, il n'est pas nécessaire de réaliser un test post-hoc. ***
B. Pour interpréter les résultats d'une ANOVA, il faut d'abord vérifier l'indépendance, la normalité et l'homoscédasticité des résidus. ***
C. Pour comparer une moyenne observée à une moyenne théorique, on réalise une approximation par la loi de Student ou la loi normale centrée réduite. ***
D. Le test post-hoc à réaliser après une ANOVA significative est le test HDS de Tukey. ***

Qu'est ce qui est vrai à propos des tests d'hypothèse ?

A) On rejette H_0 si la p value est inférieure au seuil de risque α.
B) Si H_1 est acceptée, on dit que les différences observées sont significatives.
C) Si H_0 est acceptée, il y a un risque que les différences observées ne soient pas significatives.
D) On finit par accepter H_0 si on a échoué à la faire rejeter.

Réponses correctes : A, B, C et D

Question 1 :

Qu'est-ce qui est vrai à propos des tests du χ^2 en statistiques ?

a) Ils sont utilisés pour comparer les moyennes de deux groupes indépendants.

b) Ils permettent de comparer les distributions de fréquences observées avec celles attendues.

c) Ils ne peuvent être appliqué que lorsque les données sont normalement distribuées.

d) Ce sont des tests applicables à des échantillons de taille $n < 20$.

Réponse correcte : b) Ils permettent de comparer les distributions de fréquences observées avec celles attendues.

Question 2 :

Les tests du χ^2 sont des tests :

A) De corrélation
B) Paramétriques
C) Unilatéraux
D) Sans conditions d'application

Réponse correcte : C

Le test de corrélation entre les variables X et Y nous permet de savoir :

A) Si X et Y varient l'une en fonction de l'autre.
B) Quelle variable entre X et Y fait varier l'autre.
C) Si la corrélation qui est observée est positive ou négative.
D) Si X et Y suivent une distribution conforme à la loi normale.
E) Aucune des précédentes propositions n'est vraie.

Réponses : A et C

Dans quel.s cas utilise-t-on l'ANOVA 1 facteur ?

A) Pour comparer des variances
B) En tant qu'alternative au test de Kruskal-Wallis si les conditions d'application de ce dernier ne sont pas vérifiées
C) Pour comparer des échantillons appariés
D) Pour comparer les moyennes de 2 échantillons
E) Si les conditions d'un test paramétriques ne sont pas vérifiées
F) Aucune des précédentes propositions n'est vraie

Réponse correcte : F

Lors d'un test paramétrique de comparaison de plus de 2 échantillons indépendants:
Les échantillons doivent absolument être distribués normalement
Les résidus des échantillons doivent être indépendants
On peut s'affranchir de l'homoscédasticité des résidus si les effectifs sont égaux et grands
Les résidus des échantillons doivent être appariés

Les propositionssuivantessont-ellesvraies?

Une investigation de type expérimentale est adaptée pour révéler des relations causales

Une investigation de type observationnelle cache de nombreuses interactions potentiellement à l'œuvre dans la nature

Dans un plan expérimental emboité, toutes les combinaisons sont testées en croisant les 2 facteurs

La précision des estimations issues d'un échantillonnage aléatoire peut être estimée

Un échantillonde volontaire est représentatif de la population source

Les résultats issus d'un échantillon de convenance ontune faible validité externe

Un échantillonnagesystématique se prête bien aux analyses statistiques

Un échantillonnage aléatoire simple est facile à déployer et à disposer dans l'espace

Un test statistique paramétrique…

Bilatéral permet de conclure à la supériorité d'un traitement

Permet de traiter des données qualitatives

Permet de comparer des échantillons issus de populations ayant des distributions de données différentes

Ne nécessitent pas forcément une distribution normale de la variable mesurée au sein de(s) échantillon(s)si ces derniers sont grands

2 ECH

4. Dans quel/s exemples le test T de Student peut il être utilisé ?

A. Comparaison des niveaux de pression artérielle avant et après l'administration d'un médicament chez un groupe de patients.
B. Comparaison des notes moyennes entre deux groupes d'étudiants, l'un participant à des cours en ligne et l'autre à des cours en présentiel.
C. Comparaison des niveaux de douleur entre trois groupes de patients, deux recevant deux dose différentes de traitement expérimental et le dernier recevant un placebo.
D. Comparaison des scores de satisfaction de trois groupes de clients après avoir utilisé différentes versions d'un même produit.

7. Vous venez d'obtenir cette sortie R : Que concluez vous ?

```
  Welch Two Sample t-test

data:  groupe1 and groupe2
t = -2.0649, df = 57.49, p-value = 0.02268
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.4627194
sample estimates:
mean of x mean of y
 74.52957  79.84001
```

A. La moyenne du groupe 1 est significativement supérieure à la moyenne du groupe 2.
B. La moyenne du groupe 1 est significativement inférieure à la moyenne du groupe 2.
C. Il n'y a pas de différence significative entre les moyennes des deux groupes.
D. Les moyennes des deux groupes sont égales
Il y a hétéroscédasticité
la distribution des données est unimodale et symétrique

8. Vous venez d'obtenir cette sortie R :

```
Wilcoxon signed rank test with continuity correction

data:  avant_traitement and apres_traitement
V = 16, p-value = 0.06836
alternative hypothesis: true location shift is not equal to 0
```

Que concluez vous ?

A. On rejette l'hypothèse nulle, indiquant un changement significatif.
B. On ne rejette pas l'hypothèse nulle, suggérant l'absence de changement significatif.
C. La p-value n'a aucune signification dans le test de Wilcoxon.
D. La p-value indique la taille de l'effet entre les mesures avant et après le traitement.

```
        One Sample t-test

data:  musaraignes
t = 2.7846, df = 15, p-value = 0.01389
alternative hypothesis: true mean is not equal to 23.3
95 percent confidence interval:
 24.28511 30.71489
sample estimates:
mean of x
     27.5
```

1) Quelles affirmations sont vraies ?

- La ligne de code est : `t.test(musaraignes, mu=23.3, alternative="greater")`

- t = 2,7846 représente la distribution théorique

- C'est un test bilatéral

- On ne peut pas rejeter H0

```
        F test to compare two variances

data:  souris$Masse by souris$Sexe
F = 1.8318, num df = 11, denom df = 9, p-value = 0.0026
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4682299 6.5721346
sample estimates:
ratio of variances
           1.83175
```

À la suite de cette sortie RStudio, pour comparer deux échantillons indépendant,on pourrait:
Réaliser un test paramétrique de Welch
Réaliser un test paramétrique de Student
Réaliser un test non paramétrique de Wilcoxon-Mann-Whitney
Réaliser un test non paramétrique de Wilcoxon-Mann-Whitney apparié

REG

Pour pouvoir appliquer une régression linéaire multiple :

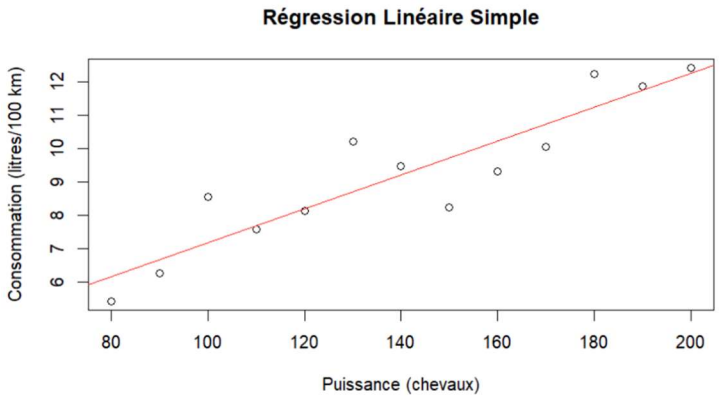
- il faut avoir une homoscédasticité des résidus
- les variables indépendantes ne doivent pas être corrélée linéairement avec les variables dépendantes
- le nombre d'observation n doit être strictement égal au nombre de paramètres à estimer
- il ne faut pas avoir de fortes corrélations entre les variables indépendantes

Les résidus :

A. Sont les différences entre les valeurs observées et les valeurs estimées par un modèle de régression
B. Représentent la partie non expliquée par l'équation de régression
C. Sont étudiés pour l'applicabilité ou non de la régression linéaire

D. Peuvent être studentisés en les divisant par leur variance

Supposons que vous ayez collecté des données sur la consommation de carburant (en litres par 100 kilomètres) pour différentes voitures en fonction de leur puissance (en chevaux). Voici un jeu de données fictif en R :



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.88	0.52	3.62	0.003
Puissance	0.05	0.01	5.00	<0.001 *

(répondre par VRAI ou FAUX)

a) La consommation moyenne de carburant est de 1.88 litre/100 km lorsque la puissance de la voiture est nulle.

VRAI

b) Pour chaque cheval de puissance supplémentaire, la consommation moyenne de carburant augmente de 0.05 litre/100 km.

VRAI

c) La variable "Puissance" est statistiquement significative à un niveau de confiance de 0.05.

VRAI

d) L'intercept de 1.88 litre/100 km est interprété comme la consommation de carburant lorsque la puissance de la voiture est nulle.

FAUX

On réalise sur RStudio un modèle de régression linéaire pour étudier les variations d'une variable Y en fonction d'une variable X. On a vérifié les différentes conditions (linéarité, indépendance, normalité et homoscédasticité) puis on appelle les résultats du modèle de régression (une partie est affichée ci-dessous). Avec l'ensemble de ces informations, quelle(s) affirmation(s) pouvez-vous considérer comme fausse(s) ?

```
Residual standard error: 33.06 on 12 degrees of freedom
Multiple R-squared: 0.731, Adjusted R-squared: 0,6861
F-statistic: 16.3 on 2 and 12 DF, p-value: 0.0003792
```

A. D'après le test de Fisher, le modèle de régression est hautement significatif.
B. Le modèle de régression linéaire est simple, on peut se fier au R^2 multiple plutôt qu'au R^2 ajusté.
C. D'après le modèle, 73% de la variation de X est expliquée par la variation de Y. ***
D. `F-statistic` correspond au coefficient de détermination. ***

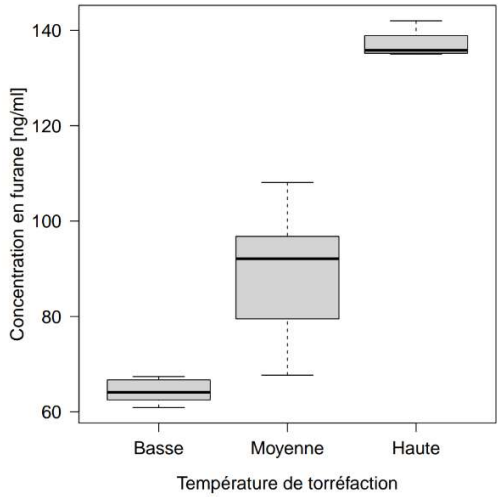
ANOVA KW

Une dernière tasse.

Le furane étant généré lors de la torréfaction du café (étape durant laquelle les grains de café sont grillés pour que les sucres et les acides présents à l'intérieur donnent des arômes), l'équipe de chercheurs a cherché à déterminer si différents processus de torréfaction conduisaient à des concentrations en furane différentes. Trois conditions de torréfaction différentes ont été étudiées, avec des températures variables. Les données sont présentées ici :

Temp. Basse	Temp. Moyenne	Temp. Haute
66.7	79.5	135.8
64.1	92.1	142
67.4	67.7	138.9
62.5	96.8	135
60.9	108.1	135.2

Concentration en furane, en ng/ml, pour trois torréfactions différentes



Voici l'analyse de ces données qui a été effectuée sous R_{studio}:

summary.aov(lm(furane~temperature))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
temperature	2	13825	6913	79.31	1.21e-07
Residuals	12	1046	87		

kruskal.test(furane~temperature)

Kruskal-Wallis rank sum test

data: furane by temperature

Kruskal-Wallis chi-squared = 12.5, df = 2, p-value = 0.00193

Quel est le test approprié selon le respect des conditions d'applications et quelle en est l'interprétation ?

A. ANOVA 1

B. Kruskal-Wallis

C. H_0 : La température de torréfaction n'a pas d'effet sur la concentration en furane des cafés

D. H_1 : Au moins un mode de torréfaction est caractérisé par une moyenne de concentration en furane dans les cafés différente des autres

√ B et D

Un étudiant souhaite réaliser une ANOVA sur deux variables quantitatives X et Y appartenant à un même dataframe. Il écrit dans RStudio le script suivant, en ayant chargé au préalable les éventuels packages nécessaires.

```
1. boxplot(variable_Y~variable_X, data_frame)
2. anova<-lm(variable_Y~variable_X, data = data_frame)
3. shapiro.test(residuals(anova))
4. hist(residuals(anova))
5. bartlett.test(variable_Y~variable_X, data = data_frame)
6. summary(anova)
```

Quelle erreur a-t-il réalisé ?

A. Il a oublié une parenthèse à la fin des lignes 3 et 4.
B. La fonction utilisée à la ligne 5 n'existe pas.
C. A la ligne 6, il doit utiliser la fonction `View()` au lieu de la fonction `summary()`.
D. Il n'a réalisé aucune erreur, son script est a priori correct. ***

Une étude a été menée pour évaluer l'effet de trois différents types de fertilisants sur le rendement de trois variétés de plantes (A, B, C). Les rendements ont été mesurés en kilogrammes par hectare. Les données ont été organisées dans un tableau appelé "donnees_fertilisants" avec trois colonnes : "Fertilisant", "Variete" et "Rendement". Pour réaliser une ANOVA sur ces données, choisissez le script R approprié.

```
# Chargement des données

donnees_fertilisants <- read.table("donnees_fertilisants.csv", header = TRUE, sep = ",")


# Script 1: ANOVA avec la fonction aov()

resultats_anova_correct <- aov(Rendement ~ Fertilisant * Variete, data = donnees_fertilisants)


# Script 2: ANOVA avec une erreur dans la spécification du modèle

modele_lm_incorrect <- lm(Rendement ~ Fertilisant + Variete, data = donnees_fertilisants)
```

```
resultats_anova_incorrect <- anova(modele_lm_incorrect)


# Script 3: ANOVA avec la fonction lm() et anova()

modele_lm_correct <- lm(Rendement ~ Fertilisant * Variete, data = donnees_fertilisants)

resultats_anova_lm_correct <- anova(modele_lm_correct)


# Script 4: ANOVA avec une fonction totalement à côté de la plaque

resultats_anova_totalement_incorrecte <- t.test(Rendement ~ Fertilisant, data =
donnees_fertilisants)
```

Parmi ces scripts R, quels sont ceux corrects pour réaliser une ANOVA sur les données "donnees_fertilisants"?

a. Script 1 et Script 2

b. Script 1 et Script 3

c. Script 2 et Script 3

d. Script 3 et Script 4

Réponses : a. Script 1 et Script 3

COR

Quels sont les termes correspondants à:

Corrélation :

A. Etablir l'existence d'un lien entre 2 VA (X, Y) jouant des rôles **symétriques**

B. lien **asymétrique** : il y a un sens dans la relation, et c'est ce qu'on cherche à modéliser

C. On s'intéresse à la dispersion ET à la pente du nuage de point

D. On s'intéresse seulement à la dispersion du nuage de point, pas à sa pente

√ A et D

Question 1 : On s'intéresse à la relation entre le nombre d'heures d'étude par semaine et les résultats de 30 étudiants. On collecte les données et souhaite utiliser des tests de corrélation pour analyser les résultats. Qu'est ce qui est vrai à propos des tests de corrélation de Pearson et de Spearman dans ce contexte ?

a) Le test de corrélation de Pearson est plus approprié si la relation entre les heures d'étude et les performances académiques n'est pas linéaire.

b) Le test de corrélation de Pearson doit être préféré à celui de Spearman si la distribution des deux variables ne suit pas de loi normale.

c) Les tests de corrélation mesurent la forme, le sens et la direction d'une relation entre deux variables qualitatives.

d) Si la relation entre les heures d'étude et les performances académiques est monotone, le test de corrélation de Spearman est à utiliser.

Réponse correcte : D

CHI2

4. Ces lignes de codent permettent-elles de réaliser un test du Khi² :

eff_obs<-c(1790,547,548,213)

prop_theo<-c(9/16,3/16,3/16,1/16)

chisq.test(eff_obs,p=eff_theo,rescale.p = T)

-Oui cela est correct ←bonne réponse
-Oui et on aurait pu aussi utiliser cette formule ; chisq.test(eff_obs,p=prop_theo) ←bonne réponse
-Oui et on aurait pu aussi utiliser cette formule ; chisq(eff_obs,p=prop_obs)
-Ces lignes sont bonnes mais permettent de faire un test de shapiro

1) Le jeu de donnée de mes effectifs observeé est le suivant:

19 / 7 / 99 / 56 / 40 / 10 / 28 / 45 / 34

Le jeu de donnée de mes effectifs théoriques est le suivant:

12 / 3 / 4 / 15 / 0 / 4 / 10 / 28 / 0

Pourquoi ne puis-je pas effectuer un test du Chi^2 ?

A) Plus de 20% des effectifs théoriques ont une valeur < 5
B) Le nombre d'individus statistique est très faible (n<50)
C) La taille des effectifs théoriques n'est pas égale à celle des effectifs observés
D) Il y a une valeur 0
E) L'ensemble de ses propositions

Réponse : A D

R

1. Dans cette matrice, on veut savoir par position directe, la valeur de Jean-Bernard qui est dans une matrice appelé « taille_tous_m », on peut utiliser la formule :

```
      Marc       Sophie        Julie     François jean_bernard
      1.76         1.56         1.64         2.00         1.47
   bernard    catherine
      1.73         1.90
```

- taille_tous_m[5]

-taille_tous_m [-c (1,2,3,4,6, 7)]

-taille_tous_m [-c (5)]