Examen de statistiques du 15/12/2020, M1AETPF

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 nécessaire. Une attention toute particulière sera portée quant à la **clarté** et à la **concision** de vos réponses. Le **seuil alpha** sera de **5%** pour tous les tests.

NB : La **calculatrice** et **documents** sous format **papier** sont autorisés.

Contexte :

Vous allez analyser une partie du jeu de données « Possum » issue de l'étude de Lindenmayer et al. (1995)* portant sur une espèce de d'opossum endémique d'Australie, le Phalanger De Montagne.

Ce jeu de données contient différentes mesures morphométriques réalisées sur 104 opossums piégés dans différents sites du sud de Victoria, au centre du Queensland.

Nous utiliserons ici un sous-échantillon du jeu de données complet, comportant les variables suivantes :

Pop : a factor which classifies the sites as **Vic** Victoria, **Other** New South Wales or Queensland
sex : a factor with levels **f** female, **m** male
age : age
totlngth : total body length
footlgth : foot length

```
Possum_sub <- subset (Possum, select = c(4,5,6,9,11))
```

```
> head(Possum_sub)
  Pop sex age totlngth footlgth
1 Vic   m   8     89.0     74.5
2 Vic   f   6     91.5     72.5
3 Vic   f   6     95.5     75.4
4 Vic   f   6     92.0     76.1
5 Vic   f   2     85.5     71.0
6 Vic   f   1     90.5     73.2
```

```
> str(Possum_sub)
'data.frame':   104 obs. of  5 variables:
 $ Pop     : chr  "Vic" "Vic" "Vic" "Vic" ...
 $ sex     : chr  "m" "f" "f" "f" ...
 $ age     : int  8 6 6 6 2 1 2 6 9 6 ...
 $ totlngth: num  89 91.5 95.5 92 85.5 90.5 89.5 91 91.5 89.5 ...
 $ footlgth: num  74.5 72.5 75.4 76.1 71 73.2 71.5 72.7 72.4 70.9 ...
```

```
> summary(Possum_sub)
     Pop                sex                  age           totlngth        footlgth
 Length:104         Length:104         Min.   :1.000   Min.   :75.00   Min.   :60.30
 Class :character   Class :character   1st Qu.:2.250   1st Qu.:84.00   1st Qu.:64.60
 Mode  :character   Mode  :character   Median :3.000   Median :88.00   Median :68.00
                                       Mean   :3.833   Mean   :87.09   Mean   :68.46
                                       3rd Qu.:5.000   3rd Qu.:90.00   3rd Qu.:72.50
                                       Max.   :9.000   Max.   :96.50   Max.   :77.90
                                       NA's   :2                       NA's   :1
```

Question 1 : Quelles vérifications (éventuellement corrections) doit-on faire avant toute analyse d'un jeu de données ? Appuyez-vous sur ce jeu de données pour illustrer votre réponse.

** Lindenmayer et al. (1995). Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala). Australian Journal of Zoology 43: 449-458.*

EXERCICE 1

Nous voulons savoir si la longueur totale du corps des opossums diffère selon le sexe des individus capturés uniquement dans la population de Victoria. Pour cela, nous devons sous-échantillonner le jeu de données Possum_sub à l'aide de la commande suivante :

```
Possum_Victoria <- subset (Possum_sub, Pop=="Vic", select = c (2, 4) )
```

	sex	totlngth
1	f	91.5
2	f	95.5
3	f	92
4	f	85.5
5	f	90.5
6	f	91
7	f	91.5
8	f	89.5
9	f	89.5
10	f	92
11	f	89.5
12	f	90.5
13	f	89
14	f	96.5
15	f	89
16	f	85
17	f	88
18	f	84
19	f	94
20	f	82.5
21	f	75
22	f	84.5
23	f	83
24	f	81
25	m	89
26	m	89.5
27	m	89.5
28	m	91.5
29	m	85.5
30	m	86
31	m	90
32	m	91
33	m	84
34	m	91.5
35	m	90
36	m	87
37	m	93
38	m	89
39	m	85.5
40	m	85
41	m	88
42	m	80.5
43	m	77
44	m	76
45	m	81
46	m	84

Question 2 : Quelles sont les variables dépendantes et indépendantes ?

Question 3 : Quels tests (paramétrique et non-paramétrique) pouvez-vous a priori utiliser pour effectuer cette comparaison ?

Question 4 : Réalisez le test approprié « à la main » en justifiant votre choix et votre conclusion.

NB : Vous disposez des informations suivantes :

```
> table(Possum_Victoria$sex)

 f  m
24 22
> round(mean(Possum_Victoria$totlngth[Possum_Victoria$sex=="f"]),2)
[1] 88.33
> round(var(Possum_Victoria$totlngth[Possum_Victoria$sex=="f"]),2)
[1] 24.58
> round(mean(Possum_Victoria$totlngth[Possum_Victoria$sex=="m"]),2)
[1] 86.52
> round(var(Possum_Victoria$totlngth[Possum_Victoria$sex=="m"]),2)
[1] 21.61


        Shapiro-Wilk normality test

data:  Possum_Victoria$totlngth[Possum_Victoria$sex == "f"]
W = 0.94962, p-value = 0.2659


        Shapiro-Wilk normality test

data:  Possum_Victoria$totlngth[Possum_Victoria$sex == "m"]
W = 0.92347, p-value = 0.08973
```

Question 5 : Ecrire la/les ligne(s) de code que vous utiliseriez pour réaliser ce(s) test(s) sous R.

EXERCICE 2

Pour évaluer si la longueur totale du corps des opossums (quelle que soit la population d'origine, Vic ou Other) diffère en fonction de la classe d'âge des individus, les 102 opossums du jeu de données ont été répartis dans les 4 groupes suivantes : (0,2] : 0 à 2 ans, (2,3] : 2 à 3 ans, (3,5] : 3 à 5 ans, (5,10] : 5 à 10 ans.

```
Possum_age <- na.omit( subset ( Possum_sub, select = c (3:4) ) )
Possum_age$age_groups <- cut ( Possum_age$age, c (0, 2, 3, 5, 10) )
summary( Possum_age )
```

```
      age           totlngth      age_groups
 Min.   :1.000   Min.   :75.00   (0,2] :26
 1st Qu.:2.250   1st Qu.:84.12   (2,3] :27
 Median :3.000   Median :88.00   (3,5] :27
 Mean   :3.833   Mean   :87.23   (5,10]:22
 3rd Qu.:5.000   3rd Qu.:90.00
 Max.   :9.000   Max.   :96.50
```

```
> M<-aggregate(Possum_age$totlngth,Possum_age["age_groups"],FUN=mean)
> M
  age_groups        x
1      (0,2] 85.59615
2      (2,3] 87.71111
3      (3,5] 86.98148
4     (5,10] 88.86364
```

Question 6 : Quelle est la nature des variables étudiées ? Sont-elles appariées ? Quelle variable correspond à la variable : a) dépendante et b) indépendante ?

Question 7 : Quels tests (paramétrique et non-paramétrique) pouvez-vous a priori utiliser pour comparer ces échantillons ? Quelles sont leurs conditions d'application ?

Question 8 : Le test de Shapiro-Wilk donne une p-value de 0.95 et le test de Bartlett une p-value de 0.15. Qu'en concluez-vous ? Ecrire la/les ligne(s) de code que vous utiliseriez pour réaliser ce(s) test(s) sous R.

Question 9 : En admettant que les conditions d'application du test paramétrique sont remplies, et au vu des données ci-dessous, peut-on dire que la longueur totale du corps des opossums est la même quelle que soit la classe d'âge ? Si non, lesquels diffèrent des autres et comment ?
Calculez les valeurs correspondant aux lettres A et B dans le tableau obtenu à l'aide de summary(aov_age), et indiquez si la lettre C correspond à une p-value significative ou non dans le tableau obtenu à l'aide de TukeyHSD(aov_age).
Justifiez vos réponses en vous appuyant sur les valeurs fournies dans les sorties R ci-après et sur vos connaissances du fonctionnement des tests utilisés.

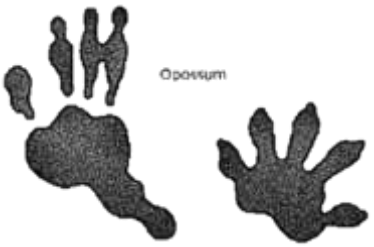
```
> aov_age<-aov(totlngth~age_groups, data=Possum_age)
> summary(aov_age)
             Df Sum Sq Mean Sq F value Pr(>F)
age_groups    3    136   45.35       B  0.0496 *
Residuals    98   1643       A
```

```
> TukeyHSD(aov_age)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = totlngth ~ age_groups, data = Possum_age)

$age_groups
                    diff        lwr      upr     p adj
(2,3]-(0,2]    2.1149573 -0.8258622 5.055777 0.2434749
(3,5]-(0,2]    1.3853276 -1.5554918 4.326147 0.6086403
(5,10]-(0,2]   3.2674825  0.1670550 6.367910 0.0347772
(3,5]-(2,3]   -0.7296296 -3.6425734 2.183314 0.9136733
(5,10]-(2,3]   1.1525253 -1.9214743 4.226525 0.7611831
(5,10]-(3,5]   1.8821549 -1.1918446 4.956154
```

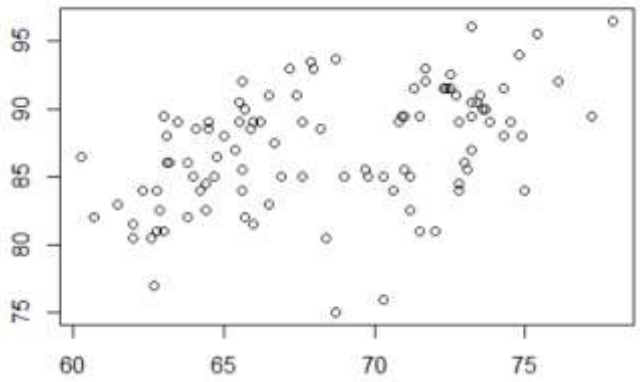
Question 10 : Quel type de graphique pourriez-vous utiliser afin de visualiser de façon synthétique la distribution des données de croissance au sein et entre les quatre classes d'âge d'opossums ?

EXERCICE 3

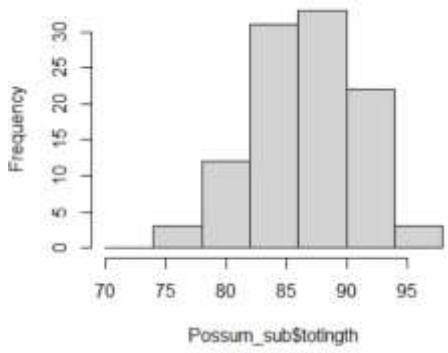
Dans le souci de minimiser le stress dû à la capture d'animaux, nous souhaiterions utiliser des tunnels à empreintes pour suivre ces populations d'opossums. Nous cherchons ici à évaluer est-ce que l'on peut prédire la taille d'un opossum (longueur totale du corps) à partir de ses empreintes (longueur du pied = foot length) ?

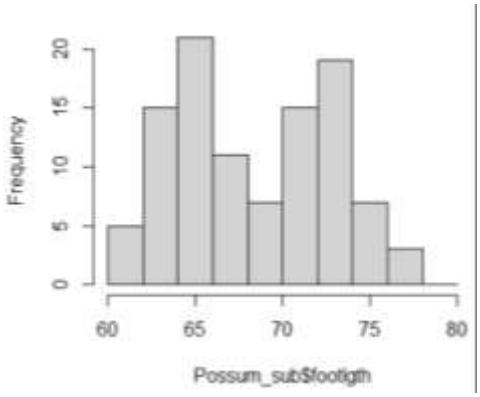


Question 11 : Est-ce qu'une relation linéaire entre ces 2 variables semble appropriée ? Quelle variable placer en abscisses et en ordonnées ?



Question 12 : Nous voulons mesurer la corrélation entre ces 2 variables. Nous disposons de informations suivantes. Quel test doit-on effectuer ? Interprétez la sortie R du test approprié en justifiant votre choix.






```
        Shapiro-Wilk normality test

data:  Possum_sub$totlngth
W = 0.98399, p-value = 0.2441
```

```
        Shapiro-Wilk normality test

data:  Possum_sub$footlgth
W = 0.95275, p-value = 0.001037
```

```
        Spearman's rank correlation rho

data:  Possum_sub$totlngth and Possum_sub$footlgth
S = 94024, p-value = 2.271e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4836822
```

```
        Pearson's product-moment correlation

data:  Possum_sub$totlngth and Possum_sub$footlgth
t = 4.9916, df = 101, p-value = 2.506e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2749789 0.5877587
sample estimates:
      cor
0.4448318
```

Question 13 : En admettant que les conditions d'application du modèle de régression linéaire simple sont remplies, et au vu des données ci-dessous, peut-on dire que la longueur du pied est un bon prédicteur de la taille (longueur totale du corps) des opossums ?

```
> mod_totL_footL<-lm(totlngth~footlgth, data=Possum_sub)
> summary(mod_totL_footL)

Call:
lm(formula = totlngth ~ footlgth, data = Possum_sub)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2332  -2.7252   0.6282   2.8954   6.8027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.24914    5.99812   9.545 9.05e-16 ***
footlgth     0.43645    0.08744   4.992 2.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.881 on 101 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1979,    Adjusted R-squared:  0.1899
F-statistic: 24.92 on 1 and 101 DF,  p-value: 2.506e-06
```

EXERCICE 4

Indiquez la ou les réponses exactes sur votre copie :

Question 14 : Un test non paramétrique :

a) nécessite d’exprimer les variables en rangs pour être utilisé.

b) est en général plus puissant qu’un test paramétrique.

c) ne nécessite pas que la distribution de la variable suive une loi normale dans les populations étudiées.

d) n’est soumis à aucune contrainte d’application.

Question 15 : Les tests non paramétriques de comparaison de rangs permettent de conclure sur :

a) la différence de moyenne.

b) la différence de mode.

c) la différence de position relative des rangs.

Question 16 : la p-value d'une différence de moyenne est :

a) La probabilité que la différence de moyenne soit réelle dans la population.

b) La probabilité que la différence dans la population soit au moins aussi importante que celle observée dans l'échantillon.

c) La probabilité d'obtenir, si le facteur étudié n'a pas d'effet, une différence au moins aussi importante que celle observée.

d) La probabilité d'obtenir une différence au moins aussi importante que celle observée sous le seul effet du hasard.

Question 17 : A propos des plans expérimentaux :

a) Un plan est équilibré lorsqu'il a le même nombre d'éléments dans chaque condition expérimentale.

b) Un plan est complet lorsqu'il a le même nombre d'éléments dans chaque condition expérimentale.

c) Un plan équilibré est nécessaire pour étudier une interaction.

d) Un plan complet est nécessaire pour étudier une interaction.

Question 18 : Quelles sont les affirmations vraies :

a) Une approche observationnelle est moins valide qu'une approche expérimentale.

b) Une approche expérimentale permet de tester des relations de causalité.

c) Les résultats d'une approche expérimentale sont facilement généralisables.

d) Les études observationnelles ont une forte validité interne.

e) Le contrôle des facteurs secondaires permet d'isoler l'effet des facteurs principaux sur la variable dépendante.