Examen bases de statistiques

Petites questions (6 points)

À la sortie du métro Jussieu, on interroge un échantillon de n étudiants représentatifs de la population des étudiants de Jussieu. On souhaite confronter les données relevées aux affirmations énoncées ci-dessous. Dans chacun des cas, proposer un test statistique adapté. Présenter clairement le test : modèle, hypothèses confrontées et forme de la zone de rejet.

1. Un tiers des étudiants a relu ses notes hier soir avant d'aller en cours ce matin.
2. Les étudiants de Jussieu ont en moyenne 23 ans. Dans un premier temps on pourra supposer que la distributions des âges est gaussienne. Qu'en est-il si ce n'est plus le cas ?
3. La réussite universitaire d'un étudiant est indépendante du fait que celui-ci habite dans Paris ou en banlieue. Comme critère de réussite on prendra la moyenne obtenue au semestre précédent, en supposant que le sondage a lieu au printemps.

Maximum de vraisemblance (4 points)

On observe n variables aléatoires indépendantes et identiquement distribuées $X_1 \dots X_n$. La distribution commune des observations admet la densité suivante pour la mesure le Lebesgue :

$$x \mapsto e^{x-\theta}\ \mathbb{1}_{]-\infty,\theta]}(x)$$

avec $\theta > 0$.

1. Montrer que le maximum de vraisemblance de θ vaut $\hat{\theta} := \max_{i=1\dots n} X_i$.
2. Montrer que $n(\theta - \hat{\theta})$ suit une loi exponentielle de paramètre 1.
3. Montrer que $\hat{\theta}$ est asymptotiquement sans biais.
4. Pour $\alpha \in]0,1[$, trouver un intervalle de confiance de niveau $1-\alpha$ (on pourra s'appuyer sur les quantiles d'une loi adaptée, sans donner les valeurs de ces quantiles).

Modèle anova à un facteur (5 points)

On considère un modèle anova à 1 facteur possédant 3 niveaux (ou groupes), on se place dans le cadre gaussien du modèle linéaire (hypothèses H1 H2 H3 H4). On suppose que l'on dispose de n_j observations pour le niveau j, $j \in \{1,2,3\}$. On note $Y_{i,j}$ est la i-ème observation du niveau j. On note aussi θ_j l'espérance de $Y_{i,j}$.

1. Rappeler la loi de $Y_{i,j}$.
2. On note $Y = (Y_{1,1}, \dots Y_{n_1,1}, Y_{1,2}, \dots Y_{n_2,2}, Y_{1,3}, \dots Y_{n_3,3})'$. Rappeler l'écriture matricielle de ce modèle linéaire gaussien.
3. Retrouver l'expression de l'estimateur des moindres carrés $\hat{\theta}$ du paramètre θ.
4. Quelle est la loi de $\hat{\theta}$?
5. Donner un intervalle de confiance à 5% pour chacun des θ_j. Déduire de ces régions de confiance individuelle une région de confiance pour le vecteur θ.

6. On note $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$ trois nouvelles observations dans le même modèle anova, où $\tilde{Y}_j$ est une nouvelle observation du groupe j. Les $\tilde{Y}_j$ sont indépendants et de plus le vecteur $(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3)'$ est indépendant du vecteur des observations initiales Y. On pose $Z = \frac{1}{3}(\tilde{Y}_1 + \tilde{Y}_2 + \tilde{Y}_3)$. Proposer un intervalle de confiance à 5% pour $\mathbb{E}Z$ ainsi qu'un intervalle de prédiction à 5% pour Z.

Maximum de vraisemblance pour la loi d'un vecteur gaussien (Bonus)

Construire l'estimateur du maximum de vraisemblance pour n observations indépendantes d'un vecteur gaussien de loi $\mathcal{N}_p(\mu, \Sigma)$ où $\mu \in \mathbb{R}^p$ et Σ est une matrice symétrique $p \times p$ inversible.

Sorties R pour des données corporelles (6 points)

Nous étudions un jeu de données sur les dimensions corporelles d'un groupe d'individus.

```
> summary(Body)
```

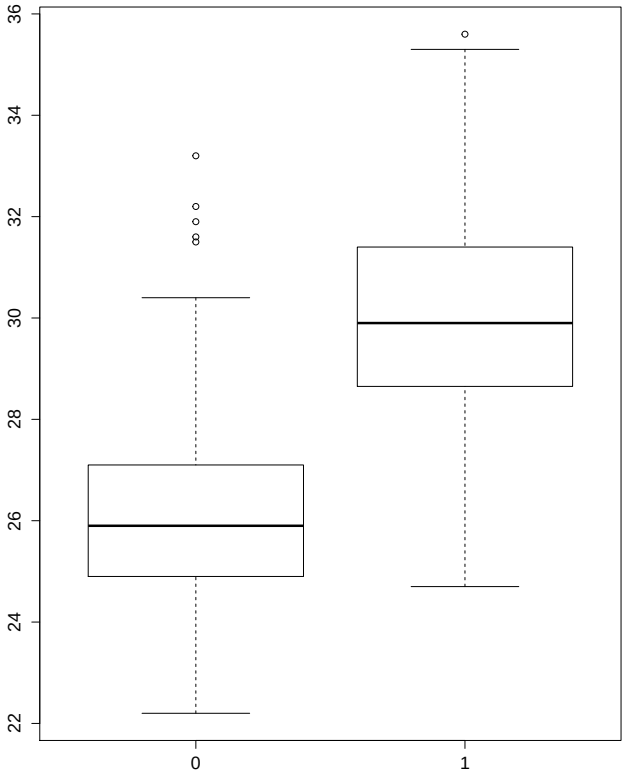
```
 DiametreCageThoracique DiametreCoudes  DiametrePoignets DiametreGenoux
 Min.   :22.20          Min.   : 9.90   Min.   : 8.10    Min.   :15.70
 1st Qu.:25.65          1st Qu.:12.40   1st Qu.: 9.80    1st Qu.:17.90
 Median :27.80          Median :13.30   Median :10.50    Median :18.70
 Mean   :27.97          Mean   :13.39   Mean   :10.54    Mean   :18.81
 3rd Qu.:29.95          3rd Qu.:14.40   3rd Qu.:11.20    3rd Qu.:19.60
 Max.   :35.60          Max.   :16.70   Max.   :13.30    Max.   :24.30
 DiametreChevilles      Age             Poids            Sexe
 Min.   : 9.90     Min.   :18.00   Min.   : 42.00   Min.   :0.0000
 1st Qu.:13.00     1st Qu.:23.00   1st Qu.: 58.40   1st Qu.:0.0000
 Median :13.80     Median :27.00   Median : 68.20   Median :0.0000
 Mean   :13.86     Mean   :30.18   Mean   : 69.15   Mean   :0.4872
 3rd Qu.:14.80     3rd Qu.:36.00   3rd Qu.: 78.85   3rd Qu.:1.0000
 Max.   :17.20     Max.   :67.00   Max.   :116.40   Max.   :1.0000
```

1. Quelle erreur dans l'importation des données est visible dans la sortie précédente ?
2. Vous trouverez ci-dessous une suite de codes R, des sorties et des graphiques. Pour chaque script, donner le numéro de la sortie ou du graphique qui a été obtenu.
3. Décrire et commenter le plus précisément possible les sorties C et G.

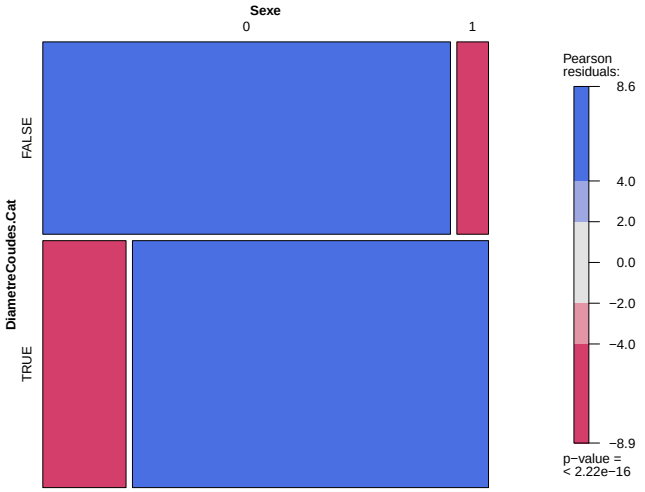
```
# Script 1 ########
> t.test(DiametreCoudes, mu = 13,conf.level=0.86)

# Script 2 ########
> DiametreCoudes.Cat <- DiametreCoudes > 13
> DiametreCoudes.Cat <- factor(DiametreCoudes.Cat)
> table(DiametreCoudes.Cat)

# Script 3 ########
> table(DiametreCoudes.Cat,Sexe)

# Script 4 ########
> barplot(table(DiametreCoudes.Cat))

# Script 5 ########
> plot(Age,DiametreCoudes,xlab= "",ylab = "")

# Script 6 ########
> cor(Body[,1:4])

# Script 7 ########
> cor.test(Body[,1],Body[,2])

# Script 8 ########
> boxplot(Body[,1]~ Sexe,xlab = "",ylab = "")

# Script  9 ########
> A <- DiametreCoudes[Sexe == "1"]
> B <- DiametreCoudes[Sexe == "0"]
> var.test(A,B)

# Script  10 ########
t.test(A,B,var.equal =TRUE)

# Script   11########
> chisq.test(DiametreCoudes.Cat,Sexe)

# Script   12 ########
> library(vcd)
> matable <- table(DiametreCoudes.Cat,Sexe)
> mosaic(matable,shade = TRUE)

# Script   13 ########
> body.reg <- lm(DiametreGenoux ~DiametrePoignets)
> summary(body.reg)
```

```
# Sortie   A #########
        Pearson's Chi-squared test with Yates' continuity correction

data:  DiametreCoudes.Cat and Sexe
X-squared = 269.4369, df = 1, p-value < 2.2e-16

# Sortie   B  #########
                      DiametreCageThoracique DiametreCoudes DiametrePoignets
DiametreCageThoracique              1.0000000      0.7588682        0.7308643
DiametreCoudes                      0.7588682      1.0000000        0.8399305
DiametrePoignets                    0.7308643      0.8399305        1.0000000
DiametreGenoux                      0.6590648      0.7315042        0.7124844
                      DiametreGenoux
DiametreCageThoracique      0.6590648
DiametreCoudes              0.7315042
DiametrePoignets            0.7124844
DiametreGenoux              1.0000000

# Sortie   C  #########
Call:
lm(formula = DiametreGenoux ~ DiametrePoignets)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6991 -0.5790 -0.0590  0.5452  3.8043

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.09189    0.47163   17.16   <2e-16 ***
DiametrePoignets  1.01671    0.04456   22.82   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.9465 on 505 degrees of freedom
Multiple R-squared:  0.5076,        Adjusted R-squared:  0.5067
F-statistic: 520.7 on 1 and 505 DF,  p-value: < 2.2e-16

# Sortie D  #########
DiametreCoudes.Cat
FALSE  TRUE
  222   285

# Sortie E  #########
                  Sexe
DiametreCoudes.Cat   0   1
             FALSE 206  16
             TRUE   54 231
```

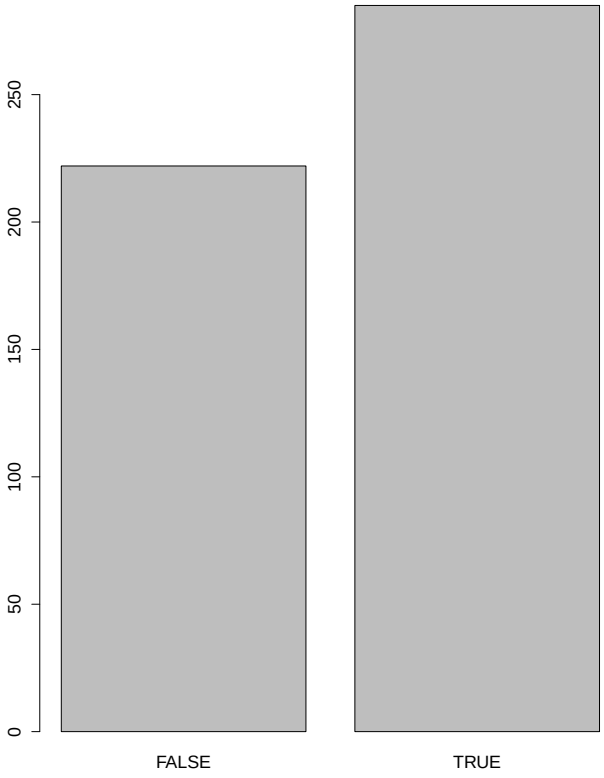
```
# Sortie F  #########
        One Sample t-test

data:  DiametreCoudes
t = 6.4111, df = 506, p-value = 3.316e-10
alternative hypothesis: true mean is not equal to 13
86 percent confidence interval:
 13.29640 13.47402
sample estimates:
mean of x
 13.38521

# Sortie    G #########
        Two Sample t-test

data:  A and B
t = 27.3798, df = 505, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.940180 2.240144
sample estimates:
mean of x mean of y
 14.45709  12.36692

# Sortie   H  #########
        Pearson's product-moment correlation

data:  Body[, 1] and Body[, 2]
t = 26.1859, df = 505, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7193214 0.7935122
sample estimates:
      cor
0.7588682

# Sortie   I  #########
        F test to compare two variances

data:  A and B
F = 1.1135, num df = 246, denom df = 259, p-value = 0.3931
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8698255 1.4264567
sample estimates:
ratio of variances
          1.113454
```



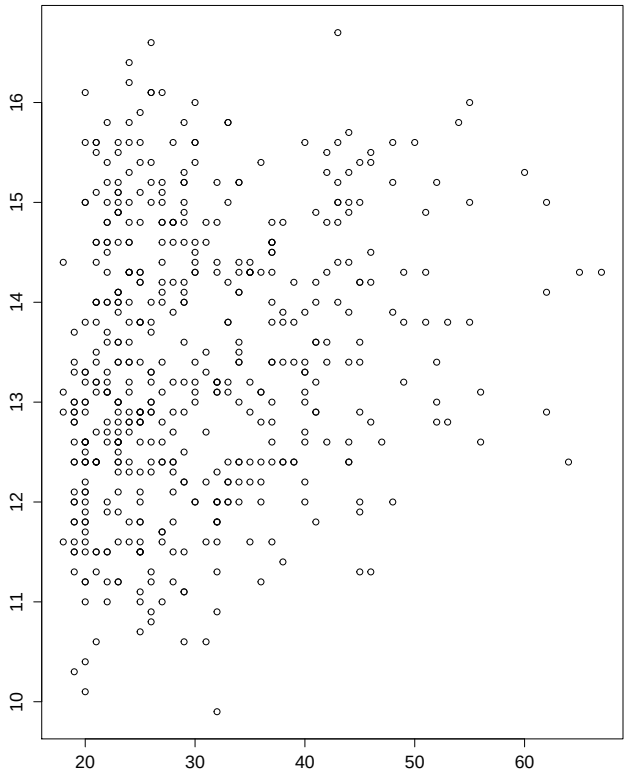
(a) Figure A



(b) Figure B



(c) Figure C



(d) Figure D

FIGURE 1 – Sorties graphiques pour les données corporelles