Math 4056 **Name:** _____

Année scolaire 2017-2018

Examen d'estimation et d'analyse de données

Décembre 2017

Time Limit: 90 Minutes

Cet examen contient 8 pages (page de consignes incluse) et 23 questions.
Total des points : 46.
À l'exception de la question 12, entourer la (les) bonne(s) réponse(s).
Rendre le sujet après l'examen.

Total (SVP ne remplissez pas le tableau ci-dessous.)

Question	Points	Score
1	2	
2	2	
3	2	
4	2	
5	2	
6	2	
7	2	
8	2	
9	2	
10	2	
11	2	
12	2	
13	2	
14	2	
15	2	
16	2	
17	2	
18	2	
19	2	
20	2	
21	2	
22	2	
23	2	
Total:	46	

1. (2 points) À quoi correspond un objet de type `dataframe` sous le logiciel `R` ?
   A. Un vecteur.
   B. Un tableau de données.
   C. Un graphique.

2. (2 points) La fonction `summary()` permet de :
   A. donner des informations de type position sur toutes les variables quantitatives du jeu de données.
   B. réaliser une régression logistique.
   C. donner l'écart-type corrigé.

3. (2 points) Soit la commande suivante :

```
> shapiro.test( df$Global_Sales[ df$Genre == 'Adventure ']).
```

   Ce test statistique permet de :
   A. tester l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle une variable quantitative étudiée sur la population est distribuée normalement contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle cette même variable ne suit pas une loi normale.
   B. tester l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle une variable qualitative étudiée sur la population est distribuée normalement contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle cette même variable ne suit pas une loi normale.
   C. tester l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle deux variables quantitatives étudiées sur la population sont de la même distribution contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle ces variables ne suivent pas la même loi.

4. (2 points) Quelle est la différence entre la fonction `density()` et la fonction `hist()` ?
   A. La fonction `density()` est une méthode non paramétrique d'estimation de la densité de probabilité contrairement à l'histogramme.
   B. La fonction `hist()` renvoie un comptage, donc un nombre entier et par conséquent un résultat non continu.
   C. La fonction `hist()` utilise moins de données.

5. (2 points) La fonction `boxplot()` :
   A. fournit un graphique, présentant l'étendue, ainsi que les trois quartiles et les intervalles qui les séparent.
   B. tracent les variables utilisées pour une régression.
   C. trace la boîte à moustaches des données mises en argument dans la fonction.

6. (2 points) Soit la commande suivante :

```
> shapiro.test(df$Global_Sales[df$Genre=='Adventure'])
```

Ce test de Shapiro-Wilk est appliqué sur :

A. les ventes des jeux d'aventures.

B. les ventes globales des jeux d'aventures.

C. le nombre de jeux d'aventures de toutes les ventes globales.

7. (2 points) Soit la commande suivante :

```
> ks.test(dataf$rawpoll_clinton,x1)
```

Ce test permet de :

A. calculer l'écart-type entre `dataf$rawpoll_clinton` et `x1`.

B. tester la corrélation entre `dataf$rawpoll_clinton` et `x1`.

C. comparer la distribution de `dataf$rawpoll_clinton` et `x1`.

8. (2 points) Soit la commande suivante :

```
> kruskal.test( rawpoll_clinton ~ type , data = dataf )
```

Ce test est une alternative :

A. à l'ANOVA à un facteur.

B. au test de Shapiro-Wilk.

C. au test de Kruskal-Wallis.

9. (2 points) La fonction `pairs()` permet de :

A. produire une matrice de nuages de points.

B. renvoyer l'écart-type.

C. renvoyer la valeur maximale.

10. (2 points) La fonction `aes()` permet de :

A. renseigner les axes dans `ggplot2`.

B. faire une régression logistique.

C. donner des informations de type position sur toutes les variables quantitatives du jeu de données.

11. (2 points) Soit le code suivant :

```
lm(formula = mine ~ imdb, data = d)
Residuals:
Min 1Q Median 3Q Max
-5.2066 -0.7224 0.1808 0.7934 2.9871

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6387 0.6669 -0.958 0.339
imdb 0.9686 0.0884 10.957 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.254 on 420 degrees of freedom
Multiple R-squared: 0.2223,Adjusted R-squared: 0.2205
F-statistic: 120.1 on 1 and 420 DF, p-value: < 2.2e-16
```

Quelle modélisation avons-nous appliqué ? Justifiez votre réponse.

A. Une régression logistique.

B. Une régression de Markov.

C. Une régression linéaire simple.

12. (2 points) Soit le code suivant :

```
> summary(m2<-lm(mine~imdb+d$comedy +d$romance+d$mystery
+d$Stanley.Kubrick..+d$Lars.Von.Trier..+d$Darren.Aronofsky..+year.c,
data=d))
Call:
lm(formula = mine ~ imdb + d$comedy + d$romance + d$mystery +
d$Stanley.Kubrick.. + d$Lars.Von.Trier.. + d$Darren.Aronofsky.. +
year.c, data = d)
Residuals:
Min 1Q Median 3Q Max
-4.4265 -0.6212 0.1631 0.7760 2.5917
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.074930 0.651223 1.651 0.099574 .
imdb 0.727829 0.087238 8.343 1.10e-15 ***
d$comedy -0.598040 0.133533 -4.479 9.74e-06 ***
d$romance -0.411929 0.141274 -2.916 0.003741 **
d$mystery 0.315991 0.185906 1.700 0.089933 .
d$Stanley.Kubrick.. 1.066991 0.450826 2.367 0.018406 *
d$Lars.Von.Trier.. 2.117281 0.582790 3.633 0.000315 ***
d$Darren.Aronofsky.. 1.357664 0.584179 2.324 0.020607 *
year.c 0.016578 0.003693 4.488 9.32e-06 ***
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.156 on 413 degrees of freedom
Multiple R-squared: 0.3508,Adjusted R-squared: 0.3382
F-statistic: 27.89 on 8 and 413 DF, p-value: < 2.2e-16
```

Quelles sont les variables explicatives pertinentes dans notre modèle ?

13. (2 points) En théorie des probabilités, une densité de probabilité est :
    A. une loi particulière.
    B. une fonction qui permet de représenter une loi de probabilité sous forme d'intégrale.
    C. un écart-type.

14. (2 points)
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$
où σ représente l'écart-type et μ la moyenne. Cette équation correspond à :
    A. la densité de probabilité de la loi normale.
    B. la densité de probabilité de la loi de Poisson.
    C. un intervalle de confiance à 95%.

15. (2 points) La variance
    A. correspond au carré de l'écart-type.
    B. élevée indique que les valeurs sont très écartées les unes des autres.
    C. est nulle lorsque toutes les valeurs observées sur une série de données sont identiques.

16. (2 points) Le test de Kolmogorov-Smirnov :
    A. teste l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle une variable quantitative étudiée sur la population est distribuée normalement contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle cette même variable ne suit pas une loi normale.

B. teste l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle une variable qualitative étudiée sur la population est distribuée normalement contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle cette même variable ne suit pas une loi normale.

C. teste l'hypothèse nulle, notée $\mathcal{H}_0$, selon laquelle deux variables quantitatives étudiées sur la population sont de la même distribution contre l'hypothèse alternative, notée $\mathcal{H}_1$, selon laquelle ces variables ne suivent pas la même loi.

17. (2 points) À quoi correspond cette équation

$$\left[\overline{x} - 2\frac{\sigma_X}{\sqrt{n}}; \overline{x} + 2\frac{\sigma_X}{\sqrt{n}}\right]?$$

A. À la densité de probabilité de la loi normale.

B. À la densité de probabilité de la loi de Poisson.

C. À un intervalle de confiance à 95% pour une moyenne d'une variable aléatoire issue d'une population normale.

18. (2 points) Quel autre test ressemble à celui de Kolmogorov-Smirnov ?

A. Le test de Shapiro-Wilk si nous comparons la distribution avec la loi normale.

B. Le test de Hobs si nous comparons la distribution avec la loi de Poisson.

C. Le test de Hobs si nous comparons la distribution avec la loi de Rammstein.

19. (2 points) L'analyse de la variance (A.N.O.V.A) permet :

A. de calculer la moyenne.

B. peut être remplacer par un test de Kruskall-Wallis si les données ne suivent pas un loi normale.

C. de calculer l'écart-type.

20. (2 points) L'intervalle de confiance :

A. est une moyenne.

B. permet d'estimer un intervalle de valeurs probabilistes à partir d'un échantillon donné.

C. est une densité de probabilité.

21. (2 points) La probabilité de couverture est :

A. la probabilité de contenir des valeurs comprises entre un intervalle donné à partir d'estimations fournies.

B. une couette.

C. des valeurs comprises entre un intervalle donné à partir d'estimations fournies.

22. (2 points) Parmi les modèles ci-dessous, lesquels sont des modèles de régression ?

A. La régression de Rammstein.

B. La régression logistique.

C. La régression linéaire multiple.

23. (2 points) L'erreur quadratique moyenne se nomme aussi :

A. M.S.E.

B. M.S.T.

C. E.Q.M.