

**Math 4056**

**Name:** \_\_\_\_\_

**Novembre 2019**

**Examen Math 4056**

**11/2019**

**Time Limit: 90 Minutes**

---

Cet examen contient 11 pages (page de consignes incluse) et 23 questions.  
Total des points : 40.

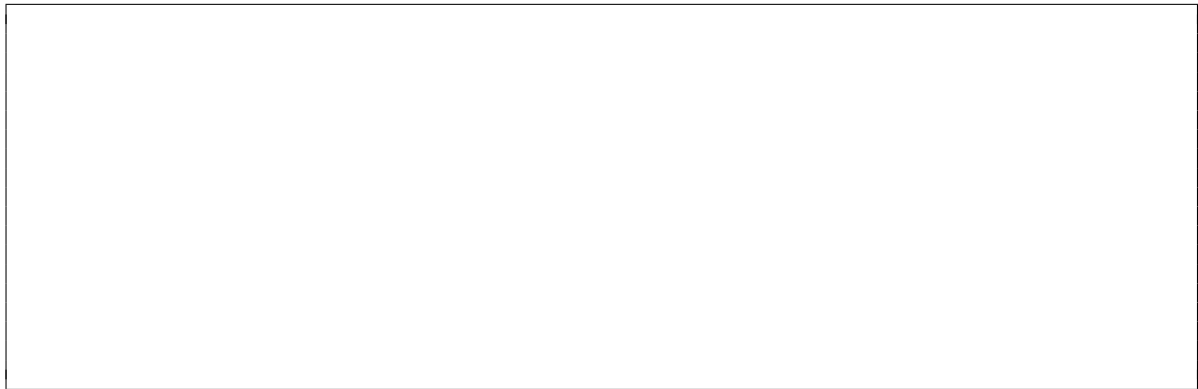
Pour les questions sous forme de Q.C.M. entourer la ou les bonne(s) réponse(s). Justifiez toute réponse écrite. Rendre le sujet après l'examen.

Total (ne rien remplir sur ce tableau)

Question	Points	Score
1	1	
2	1	
3	1	
4	1	
5	4	
6	1	
7	1	
8	1	
9	1	
10	1	
11	1	
12	1	
13	2	
14	1	
15	1	
16	1	
17	1	
18	4	
19	2	
20	1	
21	4	
22	4	
23	4	
Total:	40	

---

1. (1 point) La loi de Poisson permet de :
  - A. décrire le comportement d'un nombre d'évènements produits pendant une période fixée.
  - B. décrire la valeur d'évènements produite pendant une période fixée.
  - C. décrire le prix d'un poisson pendant une période de solde
  - D. décrire une somme de valeurs produites pendant une période fixée.
2. (1 point) Un modèle statistique paramétrique est noté :
  - A.  $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
  - B.  $(\Sigma, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$
  - C.  $(\Omega, \mathcal{B}, \mathbb{P}_\theta, \theta \in \Theta)$
  - D.  $(\Omega, \mathcal{A}, \mathbb{P}_\omega, \omega \in \Omega)$
3. (1 point) Quelle est la différence entre la fonction de densité de probabilité et la fonction de masse ?
  - A. la fonction de masse représente la probabilité d'une variable continue alors que la fonction de densité représente la probabilité d'une variable discrète.
  - B. la fonction de masse représente la probabilité d'une variable discrète alors que la fonction de densité représente la probabilité d'une variable continue.
  - C. la fonction de masse représente un estimateur par la méthode des moments alors que la fonction de densité est construite par la méthode du maximum de vraisemblance.
  - D. la fonction de masse représente un estimateur par la méthode du maximum de vraisemblance alors que la fonction de densité est construite par la méthode des moments.
4. (1 point) La fonction  $L_n : (x_1, \dots, x_n; \theta) \mapsto L_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\})$  pour des variables aléatoires  $X_i \sim \mathcal{L}(\theta)$  s'appelle
  - A. l'espérance de la loi  $\mathcal{L}$
  - B. l'histogramme de la loi  $\mathcal{L}$
  - C. le test de la loi  $\mathcal{L}$
  - D. la vraisemblance de la loi  $\mathcal{L}$
5. (4 points) Quelles est la différence entre la méthode des moments et la méthode du maximum de vraisemblance pour construire un estimateur ?



6. (1 point) Le moment centré d'ordre deux d'une variable aléatoire  $X$  :  $\mathbb{E}[(X - \mathbb{E}[X])^2]$  correspond à
- A. l'espérance de  $x$
  - B. la variance de  $X$
  - C. l'écart-type de  $X$
  - D. l'erreur standard résiduelle de  $X$
7. (1 point) Soit la commande suivante :

```
> shapiro.test( df$Vente[ df$Genre == 'Action'] )
```

ce test permet de :

- A. tester l'hypothèse nulle, notée  $H_0$ , selon laquelle une variable quantitative étudiée sur la population entière est distribuée normalement contre l'hypothèse alternative, notée  $H_1$ , selon laquelle cette même variable ne suit pas une loi normale.
  - B. tester l'hypothèse nulle, notée  $H_0$ , selon laquelle une variable qualitative étudiée sur la population entière est distribuée normalement contre l'hypothèse alternative, notée  $H_1$ , selon laquelle cette même variable ne suit pas une loi normale.
  - C. tester l'hypothèse nulle, notée  $H_0$ , selon laquelle deux variables quantitatives étudiées sur la population entière sont de la même distribution contre l'hypothèse alternative, notée  $H_1$ , selon laquelle ces variables ne suivent pas la même loi.
8. (1 point) Soit la commande suivante :

```
> cor(dataf$rawpoll_clinton,x1, method ="pearson")
```

Cette commande permet de :

- A. calculer l'écart-type entre `dataf$rawpoll_clinton` et `x1`.
- B. calculer le coefficient de corrélation entre `dataf$rawpoll_clinton` et `x1`.

- C. comparer la distribution de `dataf$rawpoll_clinton` et `x1`.
9. (1 point) Soit la commande suivante :
- ```
> kruskal.test( rawpoll_clinton ~ type , data = dataf )
```
- Ce test est une alternative :
- A. à l'ANOVA.
  - B. au test de Shapiro-Wilk.
  - C. au test de Kruskal-Wallis.
10. (1 point) La fonction `ks()` permet de :
- A. comparer la distribution de deux variables qualitative.
  - B. comparer la distribution de deux variables quantitative.
  - C. renvoyer la valeur maximale.
11. (1 point) 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.
12. (1 point) Soit le code suivant :
- ```
> summary(lm(formula = mine ~ imdb, data = d))
```
- Call:  
`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.

13. (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 cinq variables explicatives les plus pertinentes dans notre modèle ?  
Comment font-elles varier la variable *mine* ? Justifiez bien vos réponses.



14. (1 point) Soit la fonction :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

où  $\sigma$  représente l'écart-type et  $\mu$  la moyenne. Cette fonction 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. (1 point) Le test de Kolmogorov-Smirnov :
- A. teste l'hypothèse nulle, notée  $H_0$ , selon laquelle une variable quantitative étudiée sur la population entière est distribuée normalement contre l'hypothèse alternative, notée  $H_1$ , selon laquelle cette même variable ne suit pas une loi normale.
  - B. teste l'hypothèse nulle, notée  $H_0$ , selon laquelle une variable qualitative étudiée sur la population entière est distribuée normalement contre l'hypothèse alternative, notée  $H_1$ , selon laquelle cette même variable ne suit pas une loi normale.
  - C. teste l'hypothèse nulle, notée  $H_0$ , selon laquelle deux variables quantitatives étudiées sur la population entière sont de la même distribution contre l'hypothèse alternative, notée  $H_1$ , selon laquelle ces variables ne suivent pas la même loi.
16. (1 point) À quoi correspond cet intervalle :

$$\left[\bar{x} - 1.95\frac{\sigma}{\sqrt{n}}; \bar{x} + 1.95\frac{\sigma}{\sqrt{n}}\right]?$$

- A. à un intervalle de confiance à  $\approx 95\%$  sur la moyenne d'une population suivant la loi normale dont on connaît la variance.
- B. à un intervalle de confiance à  $\approx 95\%$  sur la moyenne d'une population suivant la loi normale dont on ne connaît pas la variance.

- C. à un intervalle de confiance à  $\approx 95\%$  sur la moyenne d'une population suivant la loi de Poisson dont on connaît la variance.
- D. à un intervalle de confiance à  $\approx 95\%$  sur la moyenne d'une population suivant la loi de Poisson dont on ne connaît pas la variance.
17. (1 point) Quel autre test ressemble à celui de Shapiro-Wilk ?
- A. Le test de Kolmogorov-Smirnov si nous comparons la distribution avec la loi normale.
- B. Le test de Kruskal Wallis si nous comparons la distribution avec la loi de Poisson.
- C. Le test de A.N.O.V.A. si nous comparons la distribution avec la loi de Rammstein.
18. (4 points) L'analyse de la variance (A.N.O.V.A) permet :
- 
19. (2 points) L'intervalle de confiance :
- A. est une moyenne.
- B. permet d'estimer un intervalle de valeur probabiliste à partir d'un échantillon donné.
- C. est une densité de probabilité.
20. (1 point) Pour calculer le quantile d'ordre associé à un intervalle de confiance autour de la moyenne d'une variable aléatoire suivant la loi normale dont on ne connaît pas la variance nous utiliserons la table
- A. de Student
- B. du Khi2
- C. de Mann Whitney
21. (4 points) Nous disposons d'un jeu de données qui associe à différentes années, les ventes réalisées.



```
> annees <- c(1:6)
> ventes <- c(350,420,570,690,920,710)
> df <- data.frame(annees = annees,
                   ventes = ventes)
> df$x_iy_i = annees * ventes
> df$x_iy_i
> df$x_i_carre = (annees)^2
> df$x_i_carre

[1] 350 840 1710 2760 4600 4260
[1] 1 4 9 16 25 36

> sum_xi = sum(df$annees)
> sum_xi
> sum_yi = sum(df$ventes)
> sum_yi
> sum_xi_yi = sum(df$x_iy_i )
> sum_xi_yi
> sum_xi_carre = sum(df$x_i_carre)
> sum_xi_carre

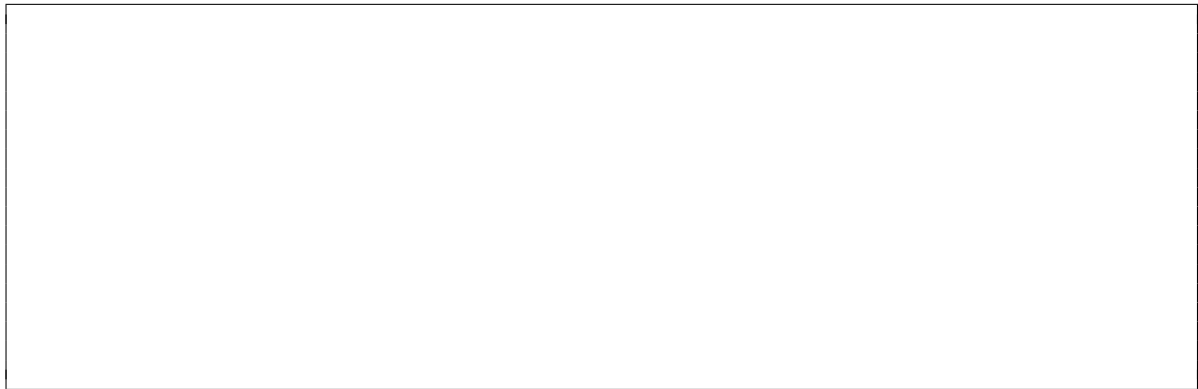
[1] 21
[1] 3660
[1] 14520
[1] 91

mean(df$annees)
mean(df$ventes)

> a = (sum_xi_yi - n * mean(df$annees) * mean(df$ventes) )
/ (sum_xi_carre - n * (mean(df$annees))^2 )

[1] 3.5
[1] 610
```

Expliquez le principe de la méthode M.S.E. et son utilité



22. (4 points) En accord avec les commandes de la question précédente, quel est l'équation de la droite de régression permettant de modéliser l'évolution des ventes en fonction des années.



23. (4 points) Détaillez les quatre graphiques ci-dessous en précisant leur utilité.

```
#cas équilibrée
> ep_3_mois <- c(43,40,41)
> ep_6_mois <- c(36,40,39)
> ep_12_mois <- c(28,24,33)
> ep_24_mois <- c(32,29,32)

> df <- data.frame(valeur=c(ep_3_mois,ep_6_mois,ep_12_mois,ep_24_mois),
                   echantillon = c(replicate(3,"ep_3_mois"),replicate(3,"ep_6_mois"),
                                   replicate(3,"ep_12_mois"),replicate(3,"ep_24_mois")))

> df$echantillon <- as.factor(df$echantillon)
> fit <- aov(valeur ~ echantillon, data=df)
> summary(fit)
              Df Sum Sq Mean Sq F value    Pr(>F)
echantillon    3   334.3    111.4    14.86 0.00124 **
Residuals      8    60.0      7.5
---
```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
> plot(fit)

