EISTI	**ING2-GI** **EXAMEN DE STATISTIQUES INFERENTIELLES 2018-2019**
Durée : 2h	Calculatrice EISTI autorisée 4 feuilles manuscrites R/V autorisées

Le jeu de données utilisé dans cet examen répertorie les 100 pays les plus peuplés du monde, caractérisés par 9 variables quantitative (population, densité, …) et une variable qualitative (Region).

	Population	Density	Net.migration	Infant.mortality	GDP	Literacy	Phones	Birthrate	Deathrate	Region
Afghanistan	31056997	48	23.06	163.07	700	36	3.2	46.6	20.34	ASIA
Algeria	32930091	13.8	-0.39	31	6000	70	78.1	17.14	4.61	AFRICA
Angola	12127071	9.7	0	191.19	1900	42	7.8	45.11	24.2	AFRICA
Argentina	39921833	14.4	0.61	15.18	11200	97.1	220.4	16.73	7.55	AMERICA
Australia	20264082	2.6	3.98	4.69	29000	100	565.5	12.14	7.51	OCEANIA
…..										

Tab. 1. Extrait du jeu de données

Exercice 1 : Test sur la mortalité infantile

Certains pays ont une mortalité infantile élevée. Cela a pour conséquence d'augmenter la mortalité infantile moyenne. Afin d'avoir un indicateur pour savoir si la distribution de cette variable est dissymétrique, nous effectuons un test pour déterminer si la mortalité infantile moyenne, μ, est significativement supérieure à la médiane de l'échantillon, med=30.3 .

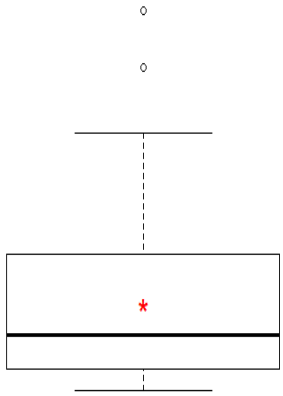

Fig. 1. Boxplot de la mortalité infantile. La croix rouge représente la moyenne

a) Ecrire les hypothèses nulle et alternative

$H_0 : \mu=30.3$ contre $H_1 : \mu>30.3$

b) Quelle est la statistique du test ? Quelle est sa loi ?

La moyenne de l'échantillon, $\overline{X}$ est la statistique du test. Etant donnée que l'échantillon est grand (n=100), on peut approcher la loi de $\overline{X}$ par une loi normale $N(\mu,\sigma^2/n)$. L'écart-type σ est inconnu mais l'échantillon étant grand, on peut le remplacer par son estimation.

c) Déterminer graphiquement l'allure de la région critique

La région critique est de la forme $W=\{\overline{X} > C\}$

d) Calculer le seuil de la région critique.

Sous l'hypothèses H_0, $\overline{X}$ suit une loi $N(30.3,s^2/n)$. D'où

$$\alpha = P(\overline{X} > C | H_0) = P(\frac{\overline{X}-30.3}{\sigma}\sqrt{n} > \frac{C-30.3}{\sigma}\sqrt{n}) = P(Z > C')$$

D'après la table des fractiles de la loi N(0,1), on obtient C'=1.6449 pour α=5%. D'où

$$C = 30.3 + 1.6449\frac{\sigma}{\sqrt{n}}.$$

d'où C=30.3+1.6449*39.7/10=36.8

e) La moyenne de l'échantillon est 45.3 et son écart-type est 39.7. Quelle est votre conclusion ?

La moyenne de l'échantillon est largement supérieure au seuil donc on accepte H_1 avec un rsique de 5%. On peut conclure que la moyenne de la mortalité infantile est significativement supérieure à la médiane.

Exercice 2 : Etude du lien entre la région et la mortalité infantile (1/2)

Dans cette première partie nous considérons la mortalité infantile sans transformation, c'est-à-dire comme une variable quantitative. On trouve ci-contre un résumé graphique de cette variable en fonction de la région. On a retiré l'Australie du jeu de données car c'est le seul pays représentant l'Océanie.

Le tableau 2 donne un résumé numérique de la mortalité infantile par région et le tableau 3 est le tableau de l'analyse de la variance.

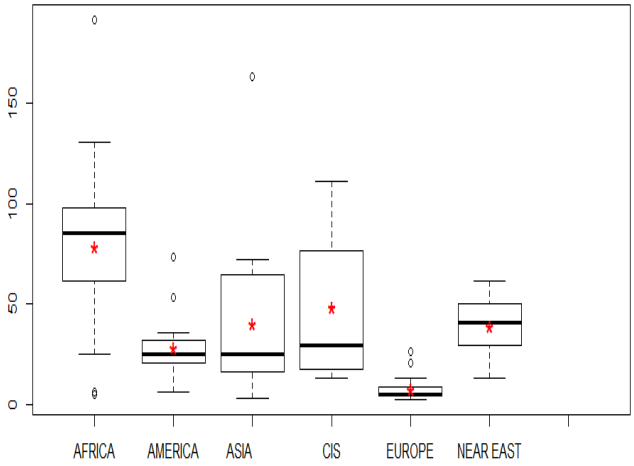

Fig. 2. Boxplot de la mortalité infantile en fonction de la région. Les croix rouges représentent les moyennes.

Région	Effectif	Moyenne	Variance
AFRICA	33	78.45	1410.42
AMERICA	16	28.41	264.28
ASIA	20	40.36	1415.85
CIS	7	48.84	1497.77
EUROPE	18	7.61	39.73
NEAR EAST	5	39.11	347.54
Total	*99*	*45.71*	*1573.48*

Tab. 2. Résumé numérique de la variable mortalité infantile par région

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*
Région	67150.28	**XXXXX**	13430.06
Résiduelle	**XXXXX**	93	**XXXXXX**
Total	154201.24		

Tab. 3. Tableau de l'analyse de la variance

a) Ecrire les hypothèses nulle et alternative du test.

Notons μ_i, i=1,…,6 les mortalités infantiles moyennes pour chaque région.

$H_0 : \mu_1=\mu_2=\ldots=\mu_6$

$H_1 : \exists i\neq j, \mu_i\neq\mu_j$

b) Donner les calculs ayant permis de trouver la somme des carrés de la région (67150,28)

$33x(78.45-45.71)^2+16x(28.41-45.71)^2+20x(40.36-45.71)+7x(48.84-45.71)^2+18x(7.61-45.71)^2+5x(39.11-45.71)^2$

c) Quels sont les degrés de liberté de la région ?

Nb modalités-1=5

d) Quelle est la somme des carrés résiduelle ?

Somme des carrés expliquée + Somme des carrés résiduelle = Somme des carrés totale

Somme des carrés résiduelle = 154201.24-67150.28=87050.96

e) Quelle est la valeur de la moyenne des carrés résiduelle ?

87050.96/93=936.03

f) Quelle est la valeur de la statistique du test de l'ANOVA sur cet échantillon ?

La statistique F=13430.06/936.03=14.35

g) Pour un risque à α=5%, le seuil de décision de ce test est 2.32 (lu dans la table de Fisher F(5 ;93)) . Quelle est votre conclusion ?

F>2.32 donc on accepte H_1 avec un risque de 5%. On considère donc que la région a un impact significatif sur la mortalité infantile moyenne.

h) Quel autre test auriez-vous pu utiliser ? Est-ce nécessaire ?

On aurait pu faire le test non paramétrique de Kruskall-Wallis dans le cas où les moyennes par région n'auraient pas été représentatives des échantillons. Sur la figure 2, on constate que les moyennes restent assez représentatives des échantillons, donc cela n'est pas nécessaire.

Exercice 3 : Etude du lien entre la région et la mortalité infantile (2/2)

Dans cette deuxième partie, nous allons étudier le lien entre ces deux variables mais en ayant transformé la mortalité infantile en variable qualitative de la façon suivante :

- Prend la modalité *Low* si la mortalité infantile est <Q_1 où Q_1=13.05
- Prend la modalité *High* si la mortalité infantile est >Q_3 où Q_3=70.19
- Prend la modalité *Médium* entre Q_1 et Q_3

Pour la variable Région, on a regroupé les modalités ASIA et NEAR EAST et supprimé l'OCEANIE (car un seul pays).

	High	Low	Medium	*Tot*
ASIA & NEAR EAST	3	4	18	*25*
C.W. OF IND. STATES	3	0	4	*7*
AMER.	1	4	13	*18*
AFRICA	18	0	13	*31*
EUROPE	0	16	2	*18*
Tot	*25*	*24*	*50*	*99*

Tab. 4. Tableau de contingence

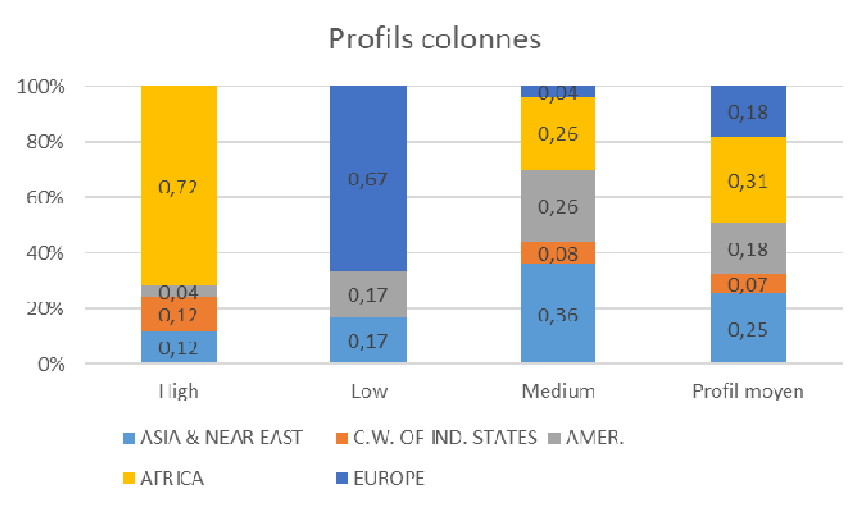

Fig. 3 Profils colonnes

a) Expliquer comment sont obtenus les profils colonnes de la figure 3 (Faire le calcul pour la modalité ASIA&NEAR EAST).

On divise chaque colonne par le total de la colonne. Par exemple le profil colonne de la case (EUROPE,Medium) est 2/50=0.04

b) Exprimer dans une phrase le chiffre 0,72 de la colonne *High*.

c) A quoi correspond le profil moyen de la figure 3 (Faire le calcul ayant permis de le représenter).

d) Exprimer dans une phrase le chiffre 0,07 de la colonne *Profil moyen*.

e) Commenter la figure 3.

On effectue le test du χ^2 pour savoir s'il y a un lien entre les deux variables.

```
        Pearson's Chi-squared test

data:  tab
X-squared = 74.109, df = 8, p-value =7.436E-13

Warning message:
In chisq.test(tab) : l'approximation du Chi-2 est peut-être incorrecte
```

f) Comment se calcule la statistique du test (donner un exemple de calcul sur une ou deux cases) ?

g) Pourquoi le logiciel affiche un warning ?

h) Quelle est votre conclusion ?

Exercice 4 : Modèle de prévision de la mortalité infantile

On considère les variable Net.migration, GDP, Literacy, Birthrate et Deathrate. La matrice des corrélations avec la mortalité infantile se trouve dans le tableau 5.

	Net.migration	Infant.mortality	GDP	Literacy	Birthrate
Net.migration	1.000	0.091	0.32	-0.076	-0.014
Infant.mortality	0.091	1.000	-0.64	-0.755	0.866
GDP	0.316	-0.642	1.00	0.557	-0.672
Literacy	-0.076	-0.755	0.56	1.000	-0.811
Birthrate	-0.014	0.866	-0.67	-0.811	1.000
Deathrate	0.240	0.667	-0.23	-0.387	0.476

Tab. 5. Matrice des corrélations

On effectue une régression linéaire entre la mortalité infantile et les autres variables :

```
Coefficients:
               Estimate Std. Error t value       Pr(>|t|)
(Intercept)    6.7706779 14.5007336   0.467       0.641638
Net.migration  0.9117590  0.5599709   1.628       0.106824
GDP           -0.0008225  0.0002327  -3.534       0.000637 ***
Literacy      -0.2297816  0.1215690  -1.890       0.061821 .
Birthrate      1.4792275  0.2463283   6.005       3.5787E-8 ***
Deathrate      2.7270539  0.3778287   7.218       1.35E-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.22 on 94 degrees of freedom
Multiple R-squared:  0.8603,   Adjusted R-squared:  0.8529
F-statistic: 115.8 on 5 and 94 DF, p-value: < 2.2E-16
```

a) Ecrire le modèle obtenu.

Infant.mortality=6.77+0.91Net.migration-0.0008GDP-0.23Literacy+1.48Birthrate+2.73Deathrate+ε

b) Quelles sont les hypothèses sur les résidus ?

On suppose que les résidus sont centrés de variance constante gaussiens ($\varepsilon \sim N(0,\sigma^2)$) et non corrélés.

c) A quoi correspond `Multiple R-squared:  0.8603` (Faites une phrase pour l'exprimer) ?

Il s'agit du coefficient de détermination. On peut dire que 86% de la variabilité de la mortalité infantile est expliquée par ce modèle.

d) A quel test correspond la dernière ligne ? Ecrire les hypothèses nulle et alternative. Conclusion.

Il s'agit du test de Fisher :

$H_0 : a_1=a_2=\ldots=a_5=0$ contre H_1 : au moins un des coefficient a_i est non nul.

où a_i sont les coefficients devant les variables explicatives.

La p-valeur du test étant très petite (2.2^{E}-16) on peut conclure qu'au moins un des 5 coefficient est non nul. Le modèle est donc pertinent.

e) A quel test correspond la dernière colonne du tableau ? Ecrire les hypothèses nulle et alternative. Conclusions.

Il s'agit du test de Student :

$H_0 : a_i=0$ contre $H_1 : a_i \neq 0$

Pour un risque α=5%, les variables ayant un coefficient significativement non nul sont : GDP, Birthrate, Deathrate. On peut envisager retirer les autres variables dans une méthode pas-à-pas.

f) Que faudrait-il vérifier et/ou modifier avant de pouvoir utiliser ce modèle ?

Avant de valider le modèle, il faut

- Vérifier que les variables explicatives sont non corrélées
- Retirer les variables non significatives du modèle dans une procédure pas-à-pas
- Vérifier les hypothèses sur les résidus
- Vérifier s'il n'y a pas d'outliers avec les résidus standardisés.

On utilise ce modèle pour prédire la mortalité infantile de l'Afghanistan :

```
predict(modele,Mydata[1,],interval = "confidence",alpha=0.05)
                  fit      lwr     upr
Afghanistan  143.3482 118.6985 167.998
```

g) Comment est obtenue la valeur `fit=143.3482` ?

On applique le modèle de la question a) avec les valeurs des variables explicatives de l'Afghanistan :

Infant.mortality=6.77+0.91x23.06-0.0008x70-0.23x36+1.48x46.6+2.73x20.34

h) A quoi correspondent les valeurs de `lwr` et `upr`? Faites une phrase pour l'exprimer.

Il s'agit des bornes de l'intervalle de confiance pour la prévision. On peut donc dire qu'il y a 95% de chance que la mortalité infantile de l'Afghanistan se trouve entre 118.70 et 168. Effectivement la mortalité infantile de l'Afghanistan est 163.07$\in$[118.70 ;168].