EISTI	**ING2-GI** **RATTRAPAGE DE STATISTIQUES INFERENTIELLES 2018-2019**
Durée : 2h Barème sur 20 (+4pts bonus) donné à titre indicatif	Calculatrice EISTI autorisée 4 feuilles manuscrites R/V autorisées

Questions diverses **(4 points)**

Pour les questions suivantes donner une réponse justifiée, claire et concise.

1) Dans un test d'hypothèses, expliquez pourquoi, en général, on ne peut pas calculer la valeur du risque de deuxième espèce.
2) Si on a la possibilité d'interroger toute la population, expliquer pourquoi les notions d'estimations ponctuelles et par intervalle de confiance n'ont plus de sens.
3) En prenant comme exemple la moyenne empirique justifié que les IDC sont plus étroits à mesure que la taille d'échantillon *n* est plus grande.
4) Expliciter la relation entre risque quadratique, variance et biais d'un estimateur. En déduire pourquoi, entre deux estimateurs sans biais, on choisira celui dont la variance est la plus petite.

La base de données

La base de données date de 1977 et représente des statistiques de criminalité dans 25 états américains. Les 25 états sont jugés représentatifs des 50 états américains à l'exception de Washington DC. Dans chaque état, sept types de crimes ou de délits sont repérés par leurs nombres annuels rapportés sur 100 000 habitants (variables quantitatives continues):

- Le nombre de meurtres (M)
- Le nombre de viols (VIOL)
- Le nombre de vols avec violence (VAV)
- Le nombre d'agressions (AGR)
- Le nombre de cambriolages (CAMB)
- Le nombre d'escroqueries(ESC)
- Le nombre de vols de voitures (VAUTO)

A cela s'ajoute la variable peine de mort prenant les modalités OUI/NON (variable qualitative nominale) suivant que la peine de mort était appliquée dans l'état en 1977 :

- La peine de mort (PMORT)

Les lignes suivantes présentent un échantillon du fichier de données :

ETATS	PMORT	M	VIOL	VAV	AGR	CAMB	ESC	VAUTO
Alabama	oui	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Alaska	non	10.8	51.6	96.8	284	1331.7	3369.8	753.3
Arizona	oui	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Arkansas	oui	8.8	27.6	83.2	203.4	972.6	1862.1	183.4

Table 0 – Extrait du fichier de données

On se propose de faire une étude complète de cet ensemble de données.

Résultats préliminaires

En utilisant l'ensemble des observations (25 états, 7 variables quantitatives, 1 variable qualitative), on donne les résultats suivant :

- Moyenne et écart-type non biaisé (table 1)

	M	*VIOL*	*VAV*	*AGR*	*CAMB*	*ESC*	*VAUTO*
Moyenne	7,44	25,73	124,09	211,30	1291,90	2671,29	377,53
Écart-type	3,87	10,76	88,35	100,25	432,46	725,91	193,39

Table 1 – Grandeurs statistiques de la base de données totale

- Matrice corrélations (table 2)

	M	*VIOL*	*VAV*	*AGR*	*CAMB*	*ESC*	*VAUTO*
M	1,00						
VIOL	0,60	1,00					
VAV	0,48	0,59	1,00				
AGR	0,65	0,74	0,56	1,00			
CAMB	0,39	0,71	0,64	0,62	1,00		
ESC	0,10	0,61	0,45	0,40	0,79	1,00	
VAUTO	0,07	0,35	0,59	0,28	0,56	0,44	1,00

Table 2 – Matrice des corrélations de la base de données totale

La table 3 fournit la moyenne et la variance (non biaisée) du nombre de meurtres (M) et de vols de voitures (VAUTO) suivant si la peine de mort (PMORT) est appliquée ou non.

		M		*VAUTO*	
PMORT	Effectif	Moyenne	Variance	Moyenne	Variance
OUI	16	8,56	14,46	340,15	16503,93
NON	9	5,27	9,27	450,08	73059,00

Table 3 – Grandeurs statistiques du nombre de meurtres et du nombre de vols de voitures calculées pour chaque modalité

Exercice 1 : Intervalle de confiance pour le nombre moyen de meurtres **(4 points)**

On considère que le nombre de meurtres (M) aux Etats-Unis (50 états) est une variable aléatoire X_M d'espérance μ_M et de variance σ_M^2.

1) Quel est l'estimateur de μ_M? Quelle est sa loi (justifier) ? Faut-il émettre des hypothèses sur la variable aléatoire X_M ? Est-ce une loi exacte ou approchée ? Quelle est la valeur estimée de μ_M sur l'échantillon ?
2) Calculer un intervalle de confiance avec un risque de α=0.05 pour le nombre moyen de meurtres (μ_M) dans les 50 états.
3) En comparant cet intervalle avec la moyenne du nombre de meurtres calculée pour chaque modalité de la variable qualitative (cf. table 3), donner une première conclusion sur le lien entre le nombre de meurtres moyen (M) et la peine de mort (PMORT).

Exercice 2 : Test d'hypothèses sur le nombre de meurtres **(5 points)**

A la vue des résultats du tableau 3, nous aimerions savoir si le nombre de meurtres des états ne pratiquant pas la peine de mort est inférieur ou supérieur au seuil symbolique des 5 pour 100 000 habitants. Pour répondre à cette question, nous allons construire un test d'hypothèse avec un risque de 1%.

On considère que le nombre de meurtres des états ne pratiquant pas la peine de mort est une variable aléatoire X d'espérance μ et de variance σ^2. Nous souhaitons donc tester les hypothèses :

$H_0 : \mu=5$

$H_1 : \mu<5$ ou $\mu>5$

1) Quelle variable de décision allez-vous utiliser ? Quelle est sa loi ? Faut-il imposer des hypothèses supplémentaires sur la variable X ?
2) Déterminer graphiquement l'allure de la région critique.
3) Calculer le ou les seuils.
4) Etablir les règles de décision.
5) Que pouvez-vous en conclure ?

Exercice 3 : Modèle de prévision du nombre de meurtres **(7 points)**

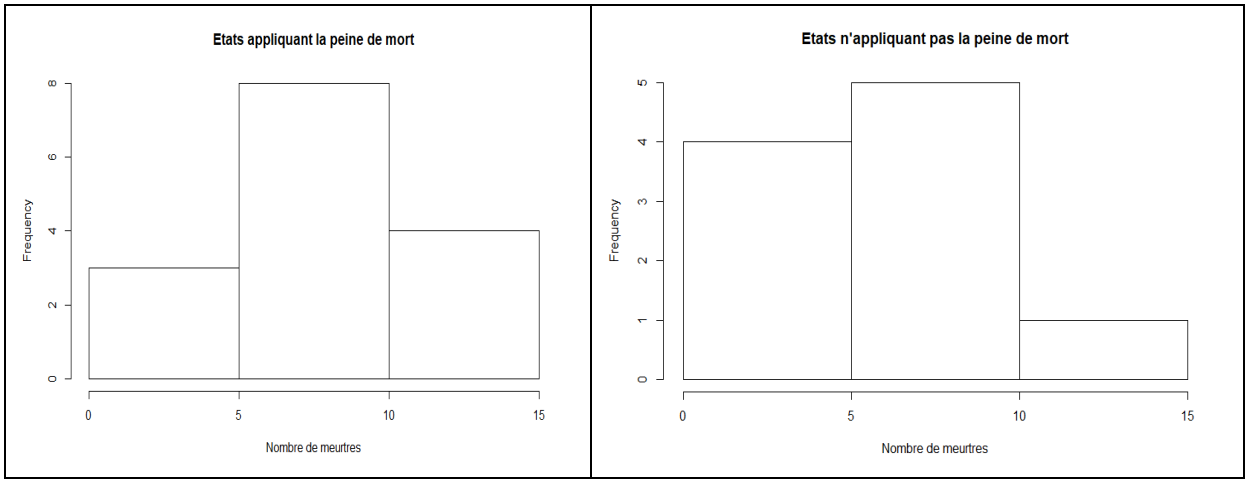
Nous aimerions définir un modèle permettant de prévoir le nombre de meurtres en fonction des autres variables quantitatives. Le listing ci-dessous fournit les résultats de trois régressions linéaires possibles calculées sur tout l'échantillon (peine de mort ou non).

1) Pour chaque modèle, expliquer si celui-ci est significatif et peut être utilisé pour faire de la prévision.
2) Supposons que la meilleure relation linéaire soit celle avec toutes les variables (analyse n°1).
   a) Ecrire le modèle obtenu.
   b) Quelles sont les hypothèses sur les résidus ?
   c) Quel est le pourcentage de variabilité du nombre de meurtres expliqué par ce modèle ?
   d) A quel test correspond la p-valeur de chacune des 7 lignes du tableau ? Ecrire les hypothèses nulle et alternative. Conclusions.
   e) Que faudrait-il vérifier et/ou modifier avant de pouvoir utiliser ce modèle ?
   f) Quelle est la valeur prédite du nombre de meurtres si le nombre de viols = 30, le nombre de vols avec violence = 120, le nombre d'agressions = 200, le nombre de cambriolages = 1300, le nombre d'escroqueries = 2700, le nombre de vols de voitures = 400 ?

Exercice bonus : Impact de la peine de mort **(4 points)**

Nous allons maintenant étudier l'influence de la peine de mort (PMORT) sur les meurtres (M).

1) Quel test d'hypothèses met-on en place ?
2) Justifier que les conditions d'application du test sont vérifiées.
3) Quelle est l'hypothèse nulle de ce test ?
4) La p-valeur du test est p=3,31^{E}-03. Que pouvez-vous conclure ?
5) On effectue le même test pour déterminer l'impact de la peine de mort sur le vol de voiture. On obtient une p-valeur=0,15. Quelle est votre conclusion ? Pouvez-vous l'expliquer.



Distribution du nombre de meurtres

ANNEXE

==========

CRIME : Analyse 1

==========

Variables explicatives : TOUTES : VIOL, VAV, AGR, CAMB, ESC, VAUTO

```
Call:
lm(formula = M ~ 1 + Viol + VaV + Agr + Camb + Esc + Vauto, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1920 -1.6564 -0.1161  1.6928  5.2239

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.5703379  1.4737188   3.780 0.000479 ***
Viol         0.1597384  0.0612715   2.607 0.012502 *
VaV          0.0117729  0.0063080   1.866 0.068822 .
Agr          0.0108238  0.0059503   1.819 0.075877 .
Camb         0.0022453  0.0018504   1.213 0.231576
Esc         -0.0026414  0.0008851  -2.984 0.004674 **
Vauto       -0.0048470  0.0025538  -1.898 0.064434 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.581 on 18 degrees of freedom
Multiple R-squared: 0.609,      Adjusted R-squared: 0.5545
F-statistic: 11.16 on 6 and 18 DF,  p-value: 1.769e-07
```

==========

CRIME : Analyse 2

==========

Variable explicative : ESC, VAUTO

```
Call:
lm(formula = M ~ 1 + Esc + Vauto, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0153 -3.3350 -0.3255  2.6769  8.1747

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.9577254  2.1466509   2.775  0.00789 **
Esc         0.0004735  0.0008624   0.549  0.58557
Vauto       0.0005864  0.0032371   0.181  0.85702
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.926 on 22 degrees of freedom
Multiple R-squared: 0.01108,    Adjusted R-squared: -0.031
F-statistic: 0.2633 on 2 and 22 DF,  p-value: 0.7697
```

==========

CRIME : Analyse 3

==========

Variable explicative : VIOL, AGR, ESC

```
Call:
lm(formula = M ~ Agr, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8518 -2.2835 -0.5756  1.5642  7.4113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.158405   0.989237   2.182   0.0340 *
Agr         0.025015   0.004238   5.903 3.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.974 on 23 degrees of freedom
Multiple R-squared: 0.4206,     Adjusted R-squared: 0.4085
F-statistic: 34.85 on 1 and 23 DF,  p-value: 3.524e-07
```