CORRECTION

EXAMEN DE STATISTIQUE INFERENTIELLE

janvier 2012

QUESTIONS DE REFLEXION

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.

1 pt

Car en général, on ne connaît pas la loi de l'estimateur sous l'hypothèse H_1 car la valeur du paramètre à tester n'est pas fixée.

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.

1 pt

La valeur calculée sur toute la population est la vraie valeur du paramètre et non une estimation. Toute l'ambigüité est de savoir ce que l'on considère comme la population.

3. En prenant comme exemple la moyenne empirique justifier que les IDC sont plus étroits à mesure que la taille d'échantillon *n* est plus grande.

1 pt

Les bornes de l'IDC sont fonction de $1/\sqrt{n}$ donc, plus n est grand, plus les bornes se « rapprochent ». Cela est tout à fait logique car plus l'échantillon est grand, plus on a d'information et plus l'estimation est précise.

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.

1 pt

Le risque quadratique est l'erreur quadratique moyenne entre l'estimation et la valeur théorique. Si on note T un estimateur de θ, on a :

$$R_\theta(T)=E[(T-\theta)^2]$$

et en développant cette expression on obtient,

$$R_\theta(T)=V(T)+B^2$$

où B est le biais et si l'estimateur est sans biais le risque quadratique correspond à la variance de l'estimateur. Donc plus la variance est petite et plus le risque quadratique, i.e l'erreur d'estimation, est petit.

5. En régression linéaire, pourquoi, dans le cas de p = 1 (une seule variable explicative), pour une même valeur du risque de 1ère espèce, le test (global) de Fisher et le test (de non nullité du coefficient) de Student donne le même résultat ?

1 pt

Test de Fisher p quelconque $\Leftrightarrow$ *Test de Fisher p=1*

$H_0 : a_1=a_2=\ldots=a_p=0$ $\qquad$ $H_0 : a_1=0$

H_1 : $\exists i,\ a_i \neq 0$ *H_1 : $a_1 \neq 0$*

Test de Student p quelconque ⇔ *Test de Fisher p=1*

Pour tout $i \in \{1,\dots,p\}$

H_0 : $a_i = 0$ *H_0 : $a_1 = 0$*

H_1 : $a_i \neq 0$ *H_1 : $a_1 \neq 0$*

Les deux tests sont donc identiques dans leurs hypothèses et on montre mathématiquement que ces deux tests sont semblables.

TRAVAIL A FAIRE

4 pt

1. On considère que le nombre de meurtres d'un état (M) est une variable aléatoire X_M d'espérance μ_M et de variance σ_M^2.

   (a) Quelle est la loi de l'estimateur que vous utilisez pour estimer μ_M? 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 ?

L'estimateur usuel de μ_M est la moyenne du nombre de meurtres des 50 états

Estimateur : 0.25 pt
Valeur estimée : 0.25 pt
Loi : 0.5 pt

$$\overline{X}_M = \frac{1}{50}\sum_{i=1}^{50} X_{M,i}$$

où $X_{M,i}$ est le nombre de morts de l'état n°i. Soit on suppose que l'échantillon est gaussien, i.e suit une loi normale $N(\mu_M,\sigma_M^2)$ *avec σ_M^2 inconnue est alors*

$$\sqrt{n}\frac{\overline{X}_M - \mu_M}{s_M^*} \sim t_{49}$$

Soit on n'émet aucun hypothèse sur la loi de l'échantillon mais on considère que l'échantillon est assez grand pour avoir une approximation précise de la variance σ_M^2 avec s_M^{*2} et on utilise le TCL pour avoir

$$\sqrt{n}\frac{\overline{X}_M - \mu_M}{s_M^*} \sim N(0,1).$$

On doit retrouver sensiblement les mêmes résultats.
La valeur estimée de μ_M sur l'échantillon est 7.44 meurtres pour 100 000 habitants.

(b) Calculer un intervalle de confiance avec un risque de α=0.05 pour le nombre moyen de meurtres (μ_M).

On cherche a et b tels que $P[a \leq \mu_M \leq b] = 0.95$.

Expression IDC : 1pt
Calcul IDC : 1pt

$$\Leftrightarrow P[\sqrt{n}\frac{\overline{x}_M - b}{s_M^*} \leq \sqrt{n}\frac{\overline{X}_M - \mu_M}{s_M^*} \leq \sqrt{n}\frac{\overline{x}_M - a}{s_M^*}] = 0.95$$

$$\Leftrightarrow P[b' \leq Z \leq a'] = 0.95$$

On suppose un risque symétrique. De plus la loi N(0,1) est symétrique par rapport à Oy donc

$$\Rightarrow \begin{cases} b' = -a' \\ P[Z \geq a'] = 0.025 \end{cases} \Rightarrow \begin{cases} a' = 1.96 \\ b' = -1.96 \end{cases} \Rightarrow \begin{cases} a = \bar{x}_M - 1.96\frac{s^*_M}{\sqrt{n}} \\ b = \bar{x}_M + 1.96\frac{s^*_M}{\sqrt{n}} \end{cases}$$

$$\Rightarrow \begin{cases} a = 7.44 - 1.96\frac{3.87}{\sqrt{50}} \approx 7.4 - \frac{2*3.9}{7} = 7.4 - \frac{7.8}{7} \approx 7.4 - 1.1 = 6.3 \\ b = 7.44 + 1.96\frac{3.87}{\sqrt{50}} \approx 7.4 + \frac{2*3.9}{7} = 7.4 + \frac{7.8}{7} \approx 7.4 + 1.1 = 8.5 \end{cases}$$

(avec Student on trouve a'=2)
Il y a donc 95% de chance que le nombre moyen de meurtres μ_M se trouve dans l'intervalle [6.3 ;8.5] meurtres pour 100 000 habitants.

(c) En comparant cet intervalle avec la moyenne du nombre de meurtres calculée pour chaque modalité de la variable qualitative (cf. table 3), quelle est la conclusion que vous pouvez établir concernant le lien entre le nombre de meurtres (M) et la peine de mort (PMORT).

1 pt

On remarque que les valeurs moyennes du nombre de meurtres suivant que les états appliquent ou non la peine de mort n'appartiennent pas à l'IDC. Quand l'état applique la peine de mort, la valeur moyenne est supérieure ($\bar{x}_M = 8.56$) et quand l'état n'applique pas la peine de mort cette valeur est inférieure ($\bar{x}_M = 5.27$). Il semble donc qu'il y ait un lien entre la modalité de la variable qualitative et le nombre de meurtres.

2 pt

2. Nous allons maintenant étudier l'influence de la peine de mort (PMORT) sur les meurtres (M) et sur les vols de voitures (VAUTO).

(a) Quelle hypothèse devait vous faire sur les variables considérées pour pouvoir faire une analyse de la variance ?

0.5 pt

Pour pouvoir utiliser le test d'hypothèses de l'analyse de la variance, on doit supposer que les variables quantitatives « nombres de meurtres » et « vols de voitures » sont gaussiennes.

(b) Que pouvez-vous conclure à partir du tableau de l'analyse de la variance suivant ?

Décision : 0.25 pt
Risque : 0.25 pt

La valeur du test calculée sur l'échantillon est supérieure (9.56) à la valeur critique de la loi de Fisher à 5% (4.04), donc on accepte l'hypothèse H_1, c-a-d on considère que la variable qualitative «peine de mort » a une influence significative sur le nombre de meurtres avec 5% de chance de se tromper.

(c) Que pouvez-vous conclure à partir du tableau de l'analyse de la variance suivant ?

Décision : 0.25 pt
Risque : 0.25 pt

La valeur du test calculée sur l'échantillon est inférieure (3.84) à la valeur critique de la loi de Fisher à 5% (4.04), donc on rejette l'hypothèse H_1, c-a-d on considère que la variable qualitative «peine de mort » n'a pas d'influence significative sur les vols de voitures et on ne connaît pas le risque de se tromper.

(d) Quelle conclusion générale pouvez-vous en tirer. Quelle(s) autre(s) analyse(s) statistique(s) pourriez-vous faire pour confirmer ou démentir votre conclusion ?

0.5 pt

Par exemple : La peine de mort semble avoir une influence sur les crimes « graves » mais peu sur les délits mineurs, peut-être parce que les délits mineurs ne sont pas sanctionnés par la peine de mort. Il faudrait procéder à une analyse de la variance sur toutes les autres variables quantitatives pour confirmer.

5 pt

3. 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 :

$$H_0 : \mu = 5$$
$$H_1 : \mu < 5 \text{ ou } \mu > 5$$

(a) Quelle variable de décision allez-vous utiliser ? Quelle est sa loi ? Faut-il imposer des hypothèses supplémentaires sur la variable X ?

L'estimateur usuel de μ est la moyenne du nombre de meurtres des 17 états ne pratiquant pas la peine de mort

$$\overline{X} = \frac{1}{17}\sum_{i=1}^{17} X_i$$

1 pt

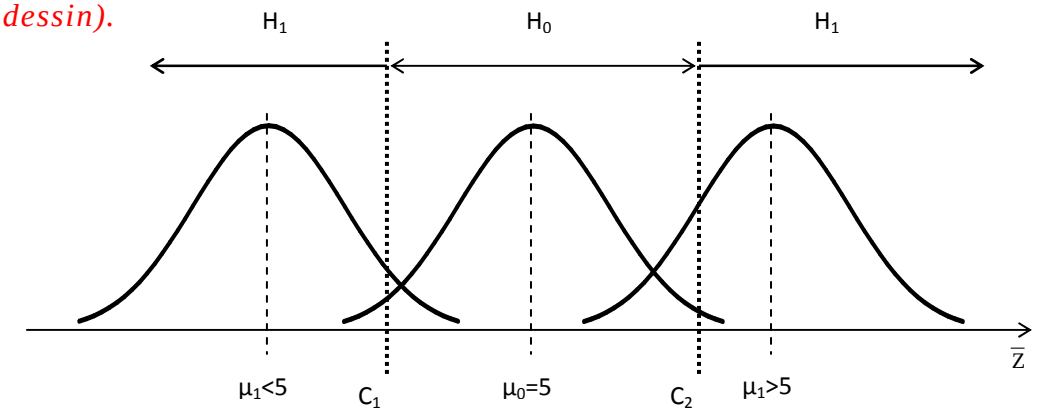
où X_i est le nombre de morts de l'état n°i. L'échantillon (n=17) n'est pas suffisamment grand pour approcher la loi de l'estimateur avec une loi normale ou pour remplacer la variance théorique par la variance estimée. On doit donc obligatoirement supposer que l'échantillon est gaussien, i.e suit une loi normale $N(\mu, \sigma^2)$ avec σ^2 inconnue est alors

$$\sqrt{n}\frac{\overline{X}-\mu}{s^*} \sim t_{16}$$

(b) Déterminer graphiquement l'allure de la région critique.

La région critique W est la région d'acceptation de H_1 d'où $W=\{\overline{X} < C_1$ ou $\overline{X} > C_2\}$ et $\overline{w}$ est la région d'acceptation de H_0 d'où $\overline{w} = \{C_1 \leq \overline{X} \leq C_2\}$ (cf. dessin).

1 pt



(c) Calculer le ou les seuils.

On sait que $\alpha=P(W/H_0 \text{ vraie})=P(\overline{X}<C_1 \text{ ou } \overline{X}>C_2 / H_0 \text{ vraie})$. Afin de simplifier le calcul de probabilité, on passe à $1-\alpha=P(\overline{w}/H_0 \text{ vraie})=P(C_1\leq\overline{X}\leq C_2/H_0 \text{ vraie})$

Expression : 1pt
Calcul : 1pt

Supposons H_0 vraie alors $\mu=5$ et $T=\sqrt{n}\frac{\overline{X}-5}{s^*}\sim t_{16}$ *où $n=17$, d'où*

$$1-\alpha=P(C_1\leq\overline{X}\leq C_2)=P\left(\frac{C_1-5}{s^*}\sqrt{n}\leq\frac{\overline{X}-5}{s^*}\sqrt{n}\leq\frac{C_2-5}{s^*}\sqrt{n}\right) \Leftrightarrow P(C'_1\leq T\leq C_2')=0.99$$

On suppose que le risque est symétrique, or la loi de Student est aussi symétrique par rapport à 0, on a donc $C'_1=-C'_2$ et $P(T\geq C_2')=0.005 \Leftrightarrow C'_2=2.921 \Rightarrow$

$$C_2=5+\frac{2.921*s^*}{\sqrt{n}}=5+\frac{2.921*\sqrt{9.27}}{\sqrt{17}}\approx 5+\frac{9}{4}=7.25 \quad \textit{et}$$

$$C_1=5-\frac{2.921*s^*}{\sqrt{n}}=5-\frac{2.921*\sqrt{9.27}}{\sqrt{17}}\approx 5-\frac{9}{4}=2.75$$

(d) Etablir les règles de décision.

(e) Que pouvez-vous en conclure ?

Si $\overline{x}<2.75$ ou $\overline{x}>7.25$, on accepte H_1, i.e on considère les échantillons différents avec 5% de chance de se tromper. Dans le cas contraire, on accepte H_0 mais on ne connaît pas le risque encouru car on ne connaît pas la loi sous H_1. Or ici $\overline{x}=5.27$ donc on considère que les états n'appliquant pas la peine de mort respectent le seuil de 5 meurtres pour 100 000 habitants.

1 pt

2 pt

4. Nous aimerions définir un modèle permettant de calculer 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).

(a) Pour chaque modèle, expliquer si celui-ci est correct et peut être utilisé pour faire de la prévision. Rendre le tableau ci-joint sans oublier de mettre votre nom.

<u>Modèle 1</u> : le test de Fisher est OK ce qui signifie que le modèle est globalement bon. Cependant, le modèle a un R^2 assez faible et les tests de Student montrent que certaines variables sont inutiles dans ce modèle (CAMB,VaV,..)

<u>Modèle 2</u> : le test de Fisher montre qu'aucune variable n'est influente dans le modèle donc le modèle n'a pas lieu d'être.

1 pt

<u>Modèle 3</u> : Le modèle n'implique qu'une seule variable, les tests de Fisher et de Student sont donc équivalents et montrent que la variable est

influente dans le modèle. Cependant le R^2 est très faible. Cette variable peut donc être retenue dans le modèle mais ne suffit pas à expliquer le nombre de meurtres.

Conclusion : Aucun des trois modèles n'est vraiment satisfaisant. Le modèle 1 semble le plus acceptable

(b) Supposons que la meilleure relation linéaire soit celle avec toutes les variables (analyse n°1). Donner alors 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.

Le modèle 1 donne :

$$M=5.57+0.160Viol+0.012*VaV+0.011*Agr+0.002*Camb-0.003*Esc-0.005*Vauto+\varepsilon$$

1 pt

On obtient donc la valeur prédite :

$$M=5.57+0.160*30+0.012*120+0.011*200+0.002*1300-0.003*2700-0.005*400=6.51$$