Méthodes Statistiques pour l'Ingénieur (MSI)
Correction Examen

R. Billot, M. Canaud, R. Delhome, P.-A. Laharotte, C. Mintsa-Eya

18 Mars 2015

1 QCM

Les choix qu'il fallait retenir sont les suivants :

1. Réponse A.
2. Réponses A, B (on pense ici à la méthode des quotas, notamment) et C.
3. Réponse E : ici, il s'agit d'une méthode d'échantillonnage au jugé.
4. Réponse C.
5. Réponse C : on échantillonne par rapport aux groupes d'appartenance des individus, plutôt que par rapport aux individus eux-mêmes.
6. Réponse A et B: la taille de la population n'impacte pas la précision de l'échantillon (qui dépend de la taillle de l'échantillon lui-même).
7. Réponse D : les comtés représentent les strates retenues ici, alors que le mot-clé variabilité réfère à la stratification optimale.
8. Réponses A et C.
9. Réponse B, la borne de Cramer-Rao correspondant à l'inverse de l'information de Fisher.
10. Réponse E : la puissance est la valeur complémentaire de l'erreur de seconde espèce β, elle représente la probabilité de rejeter H0 à raison (H1 est vraie).

Le tableau 1 résume les réponses aux différents items du QCM.

Question	A	B	C	D	E
1	x				
2	x	x	x		
3					x
4			x		
5			x		
6	x	x			
7				x	
8	x		x		
9		x			
10					x

Table 1: Réponses au QCM

2 Problème : le rugby irlandais

2.1 Estimation

2.1.1 Question 1

Il est question ici d'estimer le paramètre p de la loi de Bernoulli suivant la méthode du maximum de vraisemblance. La fonction de masse de la loi de Bernoulli s'exprime de la manière suivante:

$$\mathbb{P}(X = x) = p^x.(1-p)^{(1-x)} \mathbb{1}_{\{0;1\}}(x) \tag{1}$$

où $\mathbb{1}_{\{0;1\}}(x)$ est la fonction indicatrice, elle signifie que la fonction n'est valide que pour des x valant 0 ou 1. Pour la suite des calculs, ce facteur n'apparaît pas dans les formules. La première étape consiste à calculer la fonction de vraisemblance $L(x_1, ..., x_n; p)$:

$$L(p|x_1, ..., x_n) = \prod_{i=1}^{n} (\mathbb{P}(X = x_i)) \tag{2}$$

$$L(p|x_1, ..., x_n) = \prod_{i=1}^{n} \left(p^{x_i}.(1-p)^{(1-x_i)}\right) \tag{3}$$

$$L(p|x_1, ..., x_n) = p^{\sum_{i=1}^{n}(x_i)}.(1-p)^{n-\sum_{i=1}^{n}(x_i)} \tag{4}$$

En appliquant la fonction $\ln()$ à l'équation 4, on obtient la fonction de log-vraisemblance:

$$\ln(L(p|x_1, ..., x_n)) = \sum_{i=1}^{n}(x_i).\ln(p) + \left(n - \sum_{i=1}^{n}(x_i)\right).\ln(1-p) \tag{5}$$

L'estimateur du maximum de vraisemblance $\hat{p}$ est obtenu lorsque la fonction 5 atteint son maximum, *i.e.* lorsque sa dérivée s'annule. Dans un premier temps, on exprime la dérivée de la fonction de log-vraisemblance par rapport à p:

$$\frac{\partial \ln(L(p|x_1, ..., x_n))}{\partial x} = \frac{1}{p}.\sum_{i=1}^{n}(x_i) - \frac{1}{1-p}.\left(n - \sum_{i=1}^{n}(x_i)\right) \tag{6}$$

D'où, $\hat{p}$ est tel que:

$$\frac{1}{\hat{p}}.\sum_{i=1}^{n}(x_i) - \frac{1}{1-\hat{p}}.\left(n - \sum_{i=1}^{n}(x_i)\right) = 0 \tag{7}$$

$$\frac{1}{\hat{p}}.\sum_{i=1}^{n}(x_i) = \frac{1}{1-\hat{p}}.\left(n - \sum_{i=1}^{n}(x_i)\right) \tag{8}$$

$$(1-\hat{p}).\sum_{i=1}^{n}(x_i) = \hat{p}.\left(n - \sum_{i=1}^{n}(x_i)\right) \tag{9}$$

$$\sum_{i=1}^{n}(x_i) - \hat{p}.\sum_{i=1}^{n}(x_i) = \hat{p}.n - \hat{p}.\sum_{i=1}^{n}(x_i) \tag{10}$$

$$\sum_{i=1}^{n}(x_i) = \hat{p}.n \tag{11}$$

$$\hat{p} = \frac{1}{n}.\sum_{i=1}^{n}(x_i) = \bar{X} \tag{12}$$

D'après les informations fournies, on sait que l'Irlande a gagné 49 matches sur les 75 qu'elle a disputé. Ainsi, on a ici:

$$\hat{p} = \frac{49}{75} \approx 0.65 \tag{13}$$

2.1.2 Question 2

Chaque année, 5 expériences de Bernoulli (matches) supposément indépendantes sont répétées, chacune ayant une probabilité de $\hat{p} = 0.65$. On peut donc déduire que le nombre de victoires irlandaises lors d'un Tournoi suit une loi Binomiale, *i.e.* $Y \hookrightarrow \mathcal{B}(5, 0.65)$.

2.1.3 Question 3

Les estimateurs des paramètres de la loi Normale d'après la méthode du maximum de vraisemblance sont respectivement la moyenne empirique et l'écart-type empirique. Leur formule, ainsi que les valeurs numériques correspondant à la distribution du nombre d'essai inscrit par l'Irlande à chaque Tournoi, sont données par les équations suivantes:

$$\hat{m} = \frac{1}{n}.\sum_{i=1}^{n}(x_i) = \bar{X} \approx 12.53 \tag{14}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}(x_i - \bar{x})^2} = s \approx 3.38 \tag{15}$$

À noter ici que l'équation 15 renvoie à l'écart-type non corrigé, alors que le logiciel R renvoie par défaut l'écart-type corrigé (avec $n-1$ au dénominateur).

2.1.4 Question 4

Lorsqu'il est fait mention d'un niveau de confiance de 90% pour un test statistique, la marge d'erreur acceptable α vaut 0.1. La p-value représente la probabilité d'observer la valeur obtenue ou une valeur encore plus extrême sous H0 (*i.e.* si H0 est vraie). Si la p-value est inférieure au niveau de risque fixé, alors on rejette H0 car on considère que l'évènement observé était trop exceptionnel. Dans le cas présent $p.value = 0.17$: eu égard au niveau de confiance choisi, l'évènement observé ne paraît pas anormal ($p.value > \alpha$). On ne peut pas rejetter H0 avec un niveau de risque de 10%, on admet donc que le nombre d'essai inscrit par l'Irlande à chaque Tournoi des VI Nations suit une loi Normale.

L'histogramme ajusté montré dans l'énoncé porte pourtant à confusion, il semble que l'adéquation des données à la loi Normale n'est pas très convaincante. Même si la p-value nous pousse à accepter l'hypothèse nulle, elle reste assez proche du seuil de rejet, ce qui explique que la décision semble sujette à caution. Cependant, l'analyse statistique gagnerait très certainement en précision et en exactitude si le nombre d'individus de l'échantillon testé était plus important : seules 15 valeurs sont disponibles, ce qui paraît peu.

Pour aller plus loin, il pourrait être intéressant d'analyser plus en détail la dynamique temporelle (on voit que les premières années, plus d'essais furent marqués), ou encore les disparités selon les adversaires, pour comprendre la composition de ce phénomène.

2.2 Intervalles de confiance et tests

2.2.1 Question 5

On cherche ici à donner un intervalle de confiance pour la moyenne, alors que l'écart-type est inconnu (on n'utilisera donc l'estimateur du maximum de vraisemblance dans les développements à suivre). Il s'agira donc de lire dans la table des fractiles de la loi de Student avec $n - 1 = 14$ degrés de liberté. Le niveau de confiance du test est fixé à 90% ($\alpha = 10\%$), on a donc $1 - \alpha/2 = 0.95$. Le fractile qui nous intéresse est ainsi $t_{0.95;14} = 1.761$.

Le raisonnement part de la fonction pivotale de $\bar{X}$, définie par :

$$Z = \frac{\bar{X} - m}{\frac{s}{\sqrt{n-1}}} \tag{16}$$

Il est ici possible d'utiliser l'écart-type corrigé s^*, mais le terme sous la racine sera alors n, plutôt que n-1. A noter également que Z suit une loi de Student à $n-1$ degrés de liberté, du fait de l'utilisation de l'estimateur de l'écart-type.

L'intervalle de confiance est défini tel que :

$$IC_{90\%} = \left]\bar{X} - t_{0.95;14}\frac{s}{\sqrt{n-1}}; \bar{X} + t_{0.95;14}\frac{s}{\sqrt{n-1}}\right[\tag{17}$$

$$IC_{90\%} = \left]12.53 - 1.761\frac{3.38}{\sqrt{14}}; 12.53 + 1.761\frac{3.38}{\sqrt{14}}\right[\tag{18}$$

$$IC_{90\%} =]10.94; 14.13[\tag{19}$$

2.2.2 Question 6

On cherche à déterminer si le nombre d'essais marqués par l'Irlande au cours d'un Tournoi des VI Nations a bien une moyenne égale à 12, avec moins de 5% de chance de se tromper. On pose dans un premier temps:

$$H0 : m_0 = 12 \tag{20}$$

$$H1 : m_1 \neq m_0 \tag{21}$$

Nous sommes dans un cas de test d'hypothèse sur la moyenne avec variance inconnue. Le test est de nature bilatérale (cependant les calculs effectués avec une hypothèse alternative unilatérale sont également acceptés). La zone de rejet de l'hypothèse nulle est donc définie comme suit, pour $\alpha = 5\%$:

$$RC = \left\{\bar{X}|\bar{X} \leq m_0 - t_{0.975;n-1}\frac{s}{\sqrt{n-1}} \cup \bar{X} \geq m_0 + t_{0.975;n-1}\frac{s}{\sqrt{n-1}}\right\} \tag{22}$$

Après centrage-réduction, la zone de rejet peut également s'écrire:

$$RC = \{T_{obs}|T_{obs} \leq -t_{0.975;n-1} \cup T_{obs} \geq t_{0.975;n-1}\} \tag{23}$$

Ici, $T_{obs} = \frac{\bar{X}-m_0}{s^*/\sqrt{n}} = \frac{12.53-12}{3.38/\sqrt{14}} \approx 0.59$. On lit $t_{0.975;n-1}$ dans la table des fractiles de la loi de Student ($t_{0.975;n-1} = 2.145$). Comme $T_{obs} \notin RC$, il n'existe pas suffisamment d'évidence statistique pour rejeter l'hypothèse nulle avec moins de 5% de chance de se tromper : on admet que la vraie moyenne du nombre d'essais inscrits par l'Irlande vaut bien 12.

2.2.3 Question 7

Si on construit l'intervalle de confiance à 95% de la moyenne m, à l'image de ce qui a été fait à la question 5, on obtient :

$$IC_{95\%} = \left]\bar{X} - t_{0.975;14}\frac{s}{\sqrt{n-1}}; \bar{X} + t_{0.975;14}\frac{s}{\sqrt{n-1}}\right[\tag{24}$$

$$IC_{95\%} = \left]12.53 - 2.145\frac{3.38}{\sqrt{14}}; 12.53 + 2.145\frac{3.38}{\sqrt{14}}\right[\tag{25}$$

$$IC_{95\%} =]10.59; 14.47[\tag{26}$$

Cela nous permet directement de conclure sur le test d'hypothèse construit lors la question 6 : cet intervalle n'est ni plus ni moins que la région d'acceptation construite pendant le test (par opposition à la région de rejet). Il est statistiquement impossible de rejeter toute hypothèse nulle proposant comme moyenne une valeur incluse dans cet intervalle. En guise d'illustration, il est impossible de rejeter l'hypothèse nulle présentée par l'équation 20, où $m_0 = 12$, ce qui donne une réponse à la question 6.

	Loi de Student R	Loi de Student table (approximation)	Loi Normale table
$\mathbb{P}(. \leq -2.698)$	0.0087	0.01 ($= \mathbb{P}(. \leq -2.624)$)	0.0035
$\mathbb{P}(. \leq 1.592)$	0.9331	0.95 ($= \mathbb{P}(. \leq 1.761)$)	0.9441
$1-\beta$	**0.0755**	0.06	0.0592
β	**0.9245**	0.94	0.9408

Table 2: Calcul de la puissance du test de la question 8 (cas bilatéral)

2.2.4 Question 8

Les hypothèses décrivant ce nouveau problème sont les suivantes (α vaut 5%, *cf* question 6):

$$\begin{array}{l} H0 : m_0 = 12 \\ H1 : m_1 \neq m_0 \ (m_1 = 12.5) \end{array} \tag{27}$$

Dans le cas unilatéral, l'hypothèse alternative serait la suivante :

$$H1 : m_1 \geq m_0 \ (m_1 = 12.5) \tag{28}$$

La solution proposée développe le cas de l'équation 27. On se place dans un premier temps sous l'hypothèse H0, pour déterminer la région critique :

$$\mathbb{P}\left(-t_{0.975;n-1} \leq \frac{\bar{X} - m_0}{\frac{s}{\sqrt{n-1}}} \leq t_{0.975;n-1}\right) = 1-\alpha \tag{29}$$

$$\mathbb{P}\left(m_0 - t_{0.975;n-1}\frac{s}{\sqrt{n-1}} \leq \bar{X} \leq m_0 + t_{0.975;n-1}\frac{s}{\sqrt{n-1}}\right) = 1-\alpha \tag{30}$$

$$\mathbb{P}\left(\bar{X} \leq m_0 - t_{0.975;n-1}\frac{s}{\sqrt{n-1}}\right) + \mathbb{P}\left(\bar{X} \geq m_0 + t_{0.975;n-1}\frac{s}{\sqrt{n-1}}\right) = \alpha \tag{31}$$

L'équation 31 définit donc la région critique de ce test bilatéral. Désormais, on se place sous H1 pour évaluer l'erreur de seconde espèce β et la puissance du test $1-\beta$, à partir du résultat précédent :

$$\mathbb{P}\left(\frac{\bar{X} - m_1}{\frac{s}{\sqrt{n-1}}} \leq -t_{0.975;n-1} + \frac{m_0 - m_1}{\frac{s}{\sqrt{n-1}}}\right) + \mathbb{P}\left(\frac{\bar{X} - m_1}{\frac{s}{\sqrt{n-1}}} \geq t_{0.975;n-1} + \frac{m_0 - m_1}{\frac{s}{\sqrt{n-1}}}\right) = 1-\beta \tag{32}$$

$$\mathbb{P}\left(T_1 \leq -t_{0.975;n-1} + \frac{m_0 - m_1}{\frac{s}{\sqrt{n-1}}}\right) + \mathbb{P}\left(T_1 \geq t_{0.975;n-1} + \frac{m_0 - m_1}{\frac{s}{\sqrt{n-1}}}\right) = 1-\beta \tag{33}$$

$$\mathbb{P}\left(T_1 \leq -2.145 - \frac{0.5}{\frac{3.38}{\sqrt{14}}}\right) + \mathbb{P}\left(T_1 \geq 2.145 - \frac{0.5}{\frac{3.38}{\sqrt{14}}}\right) = 1-\beta \tag{34}$$

$$\mathbb{P}(T_1 \leq -2.698) + \mathbb{P}(T_1 \geq 1.592) = 1-\beta \tag{35}$$

$$\mathbb{P}(T_1 \leq -2.698) + 1 - \mathbb{P}(T_1 \leq 1.592) = 1-\beta \tag{36}$$

Les probabilités exprimées dans l'équation 36 sont déterminées par la lecture des tables statistiques, ou par calcul informatique, comme indiqué dans le tableau 2. Le degré de précision de la table des fractiles de la loi de Student ne permet pas de déterminer le résultat autrement que par extrapolation. La lecture de la table des fractiles de la loi Normale est ici acceptée au titre d'approximation.

En somme, l'erreur de seconde espèce de ce test vaut 0.9245 (92.45%), alors que sa puissance ne s'élève qu'à 0.0755 (7.55%). Sa faiblesse vient du trop petit écart entre m_0 et m_1 fait que l'hypothèse alternative met en scène une moyenne trop proche de celle supposée par l'hypothèse nulle. En effet, l'écart $m_0 - m_1$ est trop petit pour arriver à détecter une différence. Cette situation est illustrée par la figure 1.

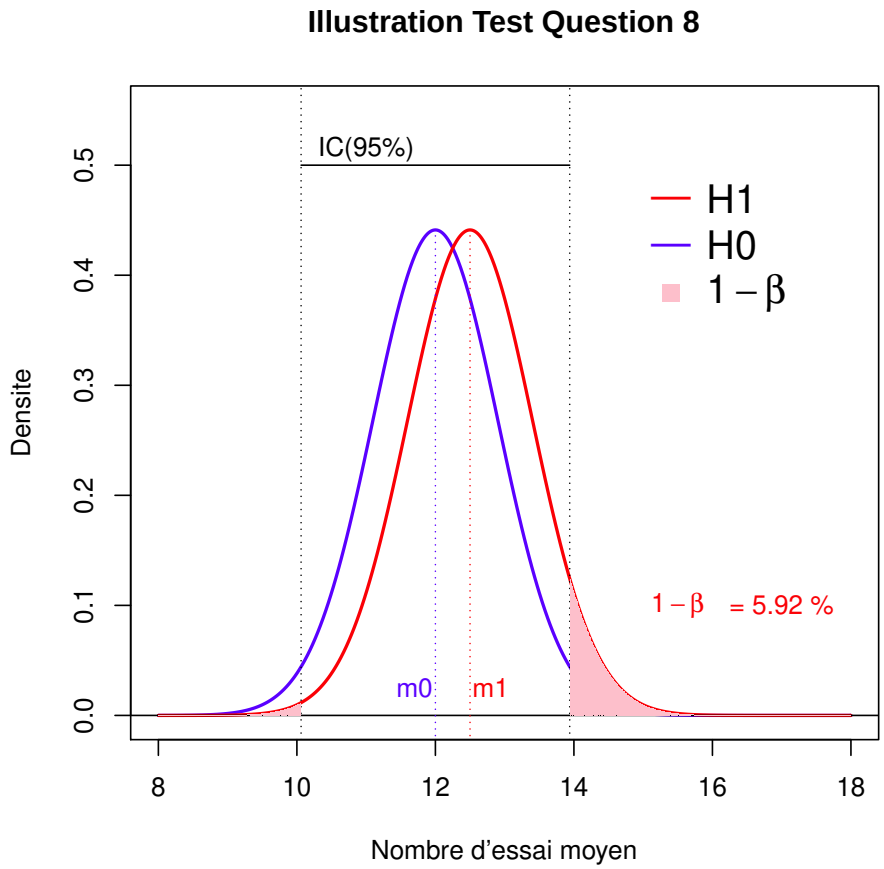

Figure 1: Calibrage des hypothèses de test de la question 8, et calcul de la puissance du test

2.3 Régression linéaire

2.3.1 Question 9

La droite considérée s'appelle la courbe de régression linéaire. Elle lie une variable explicative (le nombre d'essais marqués chaque année) et une variable expliquée (le classement final dans le Tournoi).

2.3.2 Question 10

Les résidus empiriques sont les différences entre les valeurs empiriques et les valeurs théoriques de la variable expliquée, d'après le modèle de régression (correspondant respectivement à l'ordonnée de chaque point du nuage et à leur projection verticale sur la droite de régression). Ces résidus empiriques sont représentés par les pointillés rouges sur la figure 2.

2.3.3 Question 11

Le coefficient de détermination R^2 est le carré du coefficient de corrélation, il représente la part de la variance de la variable expliquée qui est effectivement expliquée par la variance de la variable explicative. Ainsi, plus R^2 est proche de 1 (cette mesure étant comprise entre 0 et 1), plus le modèle de régression est de bonne qualité. Le résidu non expliqué est inclus dans l'erreur du modèle.

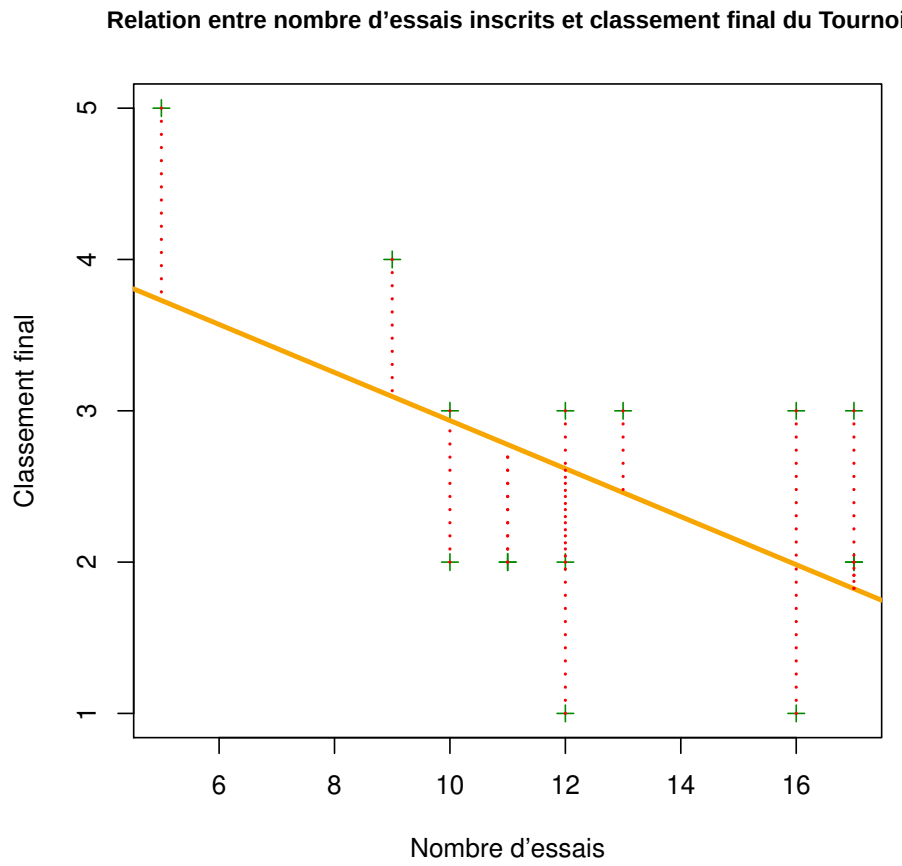

Figure 2: Régression linéaire entre le nombre d'essais marqués par l'Irlande au Tournoi des VI Nations et son classement final

On a:

$$R^2 = \frac{cov^2_{X,Y}}{s^2_X . s^2_Y} \tag{37}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{38}$$

L'application numérique donne:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \approx -29.214 \tag{39}$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 184 \tag{40}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \approx 16.857 \tag{41}$$

$$R^2 \approx \frac{(-29.214)^2}{184 * 16.857} \approx 0.275 \tag{42}$$

Le modèle de régression est donc de qualité relativement moyenne.