Université Paris 1, UFR 02, Licence de Sciences Economiques
STATISTIQUE, cours de Mme PRADEL
Partiel 14 juin 2002

Eléments de corrigé

Exercice 1

Nous avons envoyé 600 propositions d'assurance à des clients potentiels tirés dans une très large population et nous avons en retour obtenu 78 réponses favorables. Nous nous intéressons à la valeur de la probabilité p de succès pour chaque envoi et nous pouvons supposer que les comportements des individus sont indépendants les uns des autres.

1. modèle statistique correspondant à ces observations : la variable aléatoire observée est la variable de Bernoulli valant 1 si l'envoi est suivi d'une réponse favorable, 0 sinon. Nous disposons de 600 observations indépendantes et identiquement distribuées (tirage sans remise, mais dans une très large population, et comportements indépendants). On peut donc considérer que l'on a un échantillon de taille 600 d'une $Bernoulli(1,p)$.

2. L'estimateur optimal pour p est la fréquence observée :

$$\widehat{p} = \frac{1}{600}\sum_i X_i = F$$

   Les observations faites conduisent à $f = \frac{78}{600} = 0,13$

3. Loi de cet estimateur :

$$600F \sim Binomiale\,(600, p)$$

   Nous pouvons utiliser l'approximation normale pour les calculs numériques, car $np\,(1-p) = 600 * 0,13 * (1-0,13) = 67,4 > 15$

4. Intervalle bilatéral de confiance proche de 95% pour p:

   - loi de F : $F \# N\left(p, \frac{p(1-p)}{600}\right) \Longleftrightarrow \frac{F-p}{\sqrt{p(1-p)/600}} \# N\,(0,1)$
   - intervalle de probabilité 0,95 symétrique autour de p

     Lecture dans la table de la loi Normale :$P\,[|U| < 1,960] = 0,95$

     d'où :

$$P\left[\left|\frac{F-p}{\sqrt{p\,(1-p)\,/600}}\right| < 1,960\right] = 0,95$$

   - résolution approchée des inégalités : nous remplaçons au dénominateur p par son estimation F : Nous obtenons

$$-1,960 < \frac{F-p}{\sqrt{F\,(1-F)\,/600}} < 1,960$$

     soit :

$$P\left[F - 1,960\sqrt{F\,(1-F)\,/600} < p < F + 1,960\sqrt{F\,(1-F)\,/600}\right] \# 0,95$$

- Réalisation observée:

$$0,10309 < p < 0,15691$$

Nous vérifions que même pour la valeur la plus défavorable (ici la plus petite), l'approximation normale est toujours valable :

$$600 * 0,10309 * 0,89691 = 55,48 > 15$$

Exercice 2

Nous désirons proposer, par courrier, une assurance vie : notre fichier contient un grand nombre de clients potentiels dont nous connaissons l'âge, et nous nous demandons si les individus agés de "35 ou moins de 35 ans" ont moins de chance d'être intéressés par notre offre que ceux agés de "plus de 35 ans".

Nous envoyons $n_1 = 250$ propositions à des jeunes et $n_2 = 350$ propositions à des plus de 35 ans. Nous observons 29 retours favorables provenant des premiers et 49 retours favorables provenant des seconds.

1. Test d'égalité des probabilités de retour favorable provenant des deux classes d'âge : notons $p_1$et p_2 les probabilités de retour favorable provenant respectivement de la classe des "35 ou moins de 35 ans" et des "plus de 35 ans"

   (a) L'alternative (H_o contre H_1) que nous nous posons est $\{p_1 = p_2\}$ contre $\{p_1 < p_2\}$

   (b) Test de seuil 10%.

   - Statistique utilisée :

$$Z = \frac{F_2 - F_1}{\sqrt{F\,(1-F)\,(1/250 + 1/350)}}$$

   où F_2 = proportions de retours favorables chez les "plus de 35 ans"
   F_1 = proportions de retours favorables chez les "35 ou moins de 35 ans"
   F = proportions de retours favorables dans l'ensembles des envois
   - Loi approchée sous l'hypothèse de base $Z \approx N\,[0;1]$
   - rejet de $\{p_1 = p_2\}$ au profit de $\{p_1 < p_2\}$ si $Z > A$
   - règle de décision correspondant au seuil de 10% choisi : lecture dans la table $N\,[0;1]$:

$$P\,\{U < 1.281\} = 0.90$$

   nous refuserons l'égalité des probabilités si $Z > 1.282$

   (c) Ici, $f_1 = 0.116, f_2 = 0.14, f = 0.13$ et $Z_{observé} = 0.862 < 1.282$:
   Au seuil de 10%, la proportion de retour favorable chez les plus de 35 ans n'est pas significativement supérieure à celle observée chez les moins de 35 ans.

2. Nous remarquons en fait que, parmi les plus de 35 ans, il peut être intéressant de distinguer les plus de 70 ans. Nos envois et les retours favorables sont répartis de la façon suivante :

   (a) Construire le tableau de contingence associé aux deux variables "âge" et "intérêt pour l'assurance proposée"

	$\hat{a}ge \leq 35$	$35 < \hat{a}ge \leq 70$	$70 < \hat{a}ge$	Total
retour favorable	29	35	14	78
non retour	221	165	136	522
nombre total d'envois	250	200	150	600

(b) Sous l'hypothèse d'indépendance entre l'âge et l'intérêt pour l'assurance proposée, l'effectif estimé de retours favorables que nous aurions du observer dans la classe des individus d'âge supérieur à 70 ans est

$$n\widehat{p}_{13} = \frac{n_{1.}n_{.3}}{n} = \frac{78 * 150}{300} = 19.5$$

On pourra utiliser le test du chi-deux pour tester l'hypothèse d'indépendance : l'effectif théorique minimum est celui que nous venons de calculer, il est supérieur à 5

(c) Pour calculer le chi-deux d'indépendance, la contribution de la classe des retours favorables parmi les plus de 70 ans est

$$\delta_{13} = \frac{(n_{13} - n\widehat{p}_{13})^2}{n\widehat{p}_{13}} = 1.5513$$

(d) la statistique du chi-deux a pour degré de liberté $\nu = (3-1) * (2-1) = 2$

(e) Un logiciel vous fournit la valeur $chi - deux = 5,797$, et la p-value associée : $PROB = 0,05510$. Inutile d'aller regarder la table du chi-deux : nous savons que

$$P\{\chi(2) > 5,797\} = 0,05510 < 0.10$$

Au seuil de 10%, nous refusons l'hypothèse d'indépendance entre l'âge et l'intérêt porté à notre produit.

Exercice 3

Nous disposons d'un échantillon de taille 9 d'une variable X suivant une loi Normale de variance connue . et dont l'espérance peut prendre des valeurs supérieures ou égales à 10. Nous désirons tester l'hypothèse $\{m = 10\}$ contre $\{m > 10\}$.

1. Dans un premier temps, nous nous fixons l'alternative $\{m = 10\}$ contre $\{m = 11\}$.

(a) Le test le plus puissant de seuil 10% pour tester $\{m = 10\}$ contre $\{m = 11\}$ est le test de Neyman, dont la région critique est définie par

$$\frac{L_o}{L_1} < k \quad \text{et} \quad P_{Ho}\left\{\frac{L_o}{L_1} < k\right\} = 0.10$$

- Nous savons que dans le cas d'un échantillon normal, puisque $11 > 10$, cela est équivalent à

$$\begin{aligned} \overline{X} &> A \\ P_{Ho}\{\overline{X} > A\} &= 0.10 \end{aligned}$$

- sous H_o, $\overline{X} \sim N\left[m_o = 10; \frac{\sigma^2}{n} = \frac{4}{9}\right]$
- Calcul de A : $P_{Ho}\{\overline{X} > A\} = P\left\{U > \frac{A-10}{2/3}\right\} = 0.10$
  lecture de table : $P\{U < 1.282\} = 0.90 \Longrightarrow \frac{A-10}{2/3} = 1.282 \Longrightarrow A = 10 + 1.282 * 2/3 = 10.855$
- Règle : nous refusons $m = 10$ si $\overline{X} > 10.855$

(b) La puissance de ce test est :

$$\begin{aligned} \eta &= P_{H1}\{\overline{X} > 10.855\} = P\left\{U > \frac{10 + 1.282 * 2/3 - 11}{2/3}\right\} \\ \frac{10 + 1.282 * 2/3 - 11}{2/3} &= 1.282 - 1.5 = -0.2184 \\ \eta &= P\{U > -0.2184\} = P\{U < 0.2184\} = 0.5865 \end{aligned}$$

2. Nous revenons à notre problème initial, qui est de tester $\{m = 10\}$ contre $\{m > 10\}$

   (a) Le test construit à la première question est uniformément le plus puissant pour tester $\{m = 10\}$ contre $\{m > 10\}$ puisque le test de Neyman ne dépend de m_1 que par sa position par rapport à 10.

   (b) Pour que la puissance soit supérieure ou égale à 0,90, il faut que

$$\begin{aligned} \eta(m) &= P_m\left\{\overline{X} > 10.855\right\} = P\left\{U > \frac{10 + 1.282 * 2/3 - m}{2/3}\right\} \geq 0.90 \\ 1.282 + 15 - 3m/2 &\leq -1.282 \\ m &\geq (2 * 1.282 + 15)\, 2/3 = 11.709 \end{aligned}$$

   La puissance sera supérieure à 90% dès que $m \geq 11.71$.

Exercice 4

Nous examinons la production de 25 entreprises d'un même secteur en fonction du capital et du travail mis en oeuvre : nous notons

$$\begin{cases} Q_i \text{ la production de l'entreprise} \\ W_i \text{ le nombre de salariés (équivalents temps plein)} \\ K_i \text{ le capital investi} \end{cases}$$

1. La régression de Q sur W et K fournit les résultats suivants

AJUSTEMENT	1	DEPENDENT VARIABLE	Q
SMPL 1 25		TOTAL OBSERVATIONS	25
FISHER global	4427,470	SSR	7,14 E+12
DURBIN-WATSON	2.902	SE of regression	569731,7

VARIABLE	COEFFICIENT	STAND. ERROR	T-STATISTIC	PROB
W	126,9178	4,972678	25,52302	0,0000
K	0,029528	0,049432	0,597351	0,5564
constante	65784,78	130468,2	0,504221	0,6191

   (a) Hypothèses de modèle linéaire standard pour l'écart type résiduel et Normal pour la "PROB" figurant sur le listing ci-dessus.

   (b) $DW = 2;902$, $4 - DW = 1,098$ Lecture de la table : $k\prime = 2, n = 25 : d_L = 1,21, d_U = 1,55$ nous constatons que $1,098 < 1,21 = d_L$: le test de DW conduit à rejeter les hypothèses de modèle linéaire standard.

   (c) Le voisin fait un test de studend : il lit $P\left[|STUDENT(22)| > 0,597351\right] = 0,5564 > 0,10$.Il en déduit que le coefficient de K n'est pas significativement différent de zéro. Cette affirmation est statistiquement infondée car nous venons de rejeter les hypothèses du modèle linéaire standard. Il n'est donc plus correct de dire que la "T-stat" suive une loi de student.

2. Nous considérons maintenant une production de type Cobb-Douglas :

$$Q = \lambda W^{\alpha} K^{\beta} \qquad \text{(modele theorique)}$$

   et les variables prises en logarithmes :

$$LQ_i = \ln(Q_i)\,; LW_i = \ln(W_i)\,; LK_i = \ln(K_i)$$

Nous supposons que :

$$\begin{array}{rcll} LQ_i & = & aLW_i + bLK_i + c + u_i & \text{(modele statistique)} \\ & & (a,b,c) \text{ réels non contraints} & \\ u_1, u_2, ..., u_{25} & \sim & i.i.d.N\left[0, \sigma^2\right] & \end{array}$$

modèle théorique entraîne : $LQ_i = \ln\lambda + \alpha LW_i + \beta LK_i$

comparé au modèle statistique, cela nous conduit à :

$$a = \alpha, b = \beta, c = \ln\lambda$$

3. La régression de LQ sur LW et LK fournit l'ajustement suivant (ce sont les t-de student qui figurent sous les coefficients):

$$\begin{array}{rcl} LQ_i & = & \underset{[14,42]}{0.8620} LW_i + \underset{[2,31]}{0.1384} LK_i + \underset{[12,29]}{4.246} + \widehat{u}_i \\ DW & = & 2,394, SCR = 0,8652 \end{array}$$

   (a) $DW = 2,394 : 4 - DW = 1,606 > 1,55 = d_U$. Les hypothèses de modèle linéaire standard sont acceptables.

   (b) lecture de table : $P\left[|STUDENT\,(22)| > 2,074\right] = 0,05$
   Nous pouvons faire le test de Student : nous rejetterons l'hypothèse $(b = 0)$ si $|T - stat| > 2,074$. Ici, nous constatons que $2,31 > 2,074$ Nous rejetons l'hypothèse $(b = 0)$ avec un risque de 5% : le capital est bien une variable significative.
   L'erreur du voisin est d'avoir appliqué un test alors que les hypothèses structurelles de ce test ne sont pas satisfaites. Sa conclusion étant infondée, c'est le résultat trouvé ici qui est valable..

4. Remarquant que $0.8620 + 0.1384 = 1,0004$, nous désirons tester l'hypothèse $H_o = \{a + b = 1\}$.

   (a) Sous cette contrainte, la fonction $\lambda W^\alpha K^\beta$ est homogène de degré 1 : L'équation $Q = \lambda W^\alpha K^\beta$ reste valable si on divise toutes les trois variables par un même nombre.

   (b) Le modèle contraint par H_o est toujours un modèle linéaire standard, car la contrainte envisagée est linéaire. Il reste 2 paramètres libres, mais le modèle reste linéaire. Le fait qu'il soit standard n'a aucune raison d'être affecté, puisque cela concerne les moments centrés d'ordre 2.

5. Nous définissons les nouvelles variables "capital par tête" et "production par tête" en divisant respectivement K et Q par W. En logarithmes, cela nous conduit aux variables :

$$LQpt = Ln\left(\frac{Q}{W}\right) \quad \text{et} \quad LKpt = Ln\left(\frac{K}{W}\right)$$

   (a) Il suffit de remplacer LQ_i par $Ln\left[\frac{Q_i}{W_i}\right] + LW_i$, LK_i par $Ln\left[\frac{K_i}{W_i}\right] + LW_i$, L'équation du modèle statistique en fonction de ces nouvelles variables, de LW et des paramètres (a, b, c) non contraints est :

$$\begin{array}{rcl} Ln\left[\frac{Q_i}{W_i}\right] + LW_i & = & aLW_i + bLn\left[\frac{K_i}{W_i}\right] + bLW_i + c + u_i \\ LQpt_i & = & (a + b - 1)\, LW_i + bKpt_i + c + u_i \end{array}$$

   (b) La régression de $LQpt$ sur LW et $LKpt$ fournit l'ajustement suivant (ce sont les t-de student qui figurent sous les coefficients):

$$\begin{array}{rcl} LQpt_i & = & \underset{[0,0220]}{0,0004} LW_i + \underset{[2,31]}{0.1384} LKpt_i + \underset{[12,29]}{4.246} + \widehat{v}_i \\ DW & = & 2,394, SCR = 0,8652 \end{array}$$

car le coefficient de LW_i est $\left(\widehat{a}+\widehat{b}-1\right)$ et celui de $LKpt_i$ est $\widehat{b}$. : dans l'équation de régression, le changement de variable devient :

$$LQpt_i = (0.8620 + 0.1384 - 1)\, LW_i + 0.1384 LKpt_i + 4.246 + \widehat{u}_i$$

les résidus sont inchangés, ni les statistiques déduites des résidus DW et SCR

(c) Pour risque de 10%, nous lisons $P\left[|STUDENT\,(22)| > 1,717\right] = 0,10$.Evidemment, $0,0220 < 1,717$: le t de student est tellement petit qu'il est pratiquement inutile d'aller regarder la table! Nous acceptons donc la contrainte $\{a + b = 1\}$.