Université Paris 1, UFR 02, Licence de Sciences Economiques
STATISTIQUE, cours de Mme PRADEL
Partiel 21 janvier 2004

Eléments de corrigé

Exercice 1: intervalle de confiance pour une probabilité théorique

1. Nous avons un échantillon de BERNOULLI : $X_1, .., X_n \approx i.i.d.B(1;p)$, les variables prenant la valeur 1 si l'entreprise a bénéficié d'une aide publique à sa création, 0 sinon.

   L'estimateur du maximum de vraisemblance est la fréquence observée :

   $$F = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. Intervalle bilatéral de confiance 66,8% pour p. Tant que $np(1-p) > 15$ (5 à la rigueur si l'on n'est pas trop exigeant sur la précision des valeurs limitant l'intervalle), nous pouvons utiliser la loi limite de F :

   $$U = \frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \approx N[0;1]$$

   L'intervalle de probabilité pour cette statistique est donné par $P[|U| \leq a] = 0,668$. La valeur a vérifie donc

   $$P[U \leq a] = 0,668 + \frac{1-0,668}{2} = 0,834$$

   La lecture de la table Normale centrée réduite fournit $a = 0,97$. Nous en déduisons pour F et p :

   $$P\left[\left|\frac{F-p}{\sqrt{\frac{F(1-F)}{n}}}\right| \leq 0.97\right] = 0,668.$$

   La résolution en p de ces inégalités donne :

   $$P\left[F - 0.97\sqrt{\frac{F(1-F)}{n}} \leq p \leq F + 0.97\sqrt{\frac{F(1-F)}{n}}\right] = 0,668$$

   Avec un échantillon de taille $n = 343$, nous obtenons $\frac{0.97}{\sqrt{343}} = 0.0524$ l'intervalle de confiance :

   $$P\left[F - 0.0524\sqrt{F(1-F)} \leq p \leq F + 0.0524\sqrt{F(1-F)}\right] = 0,668$$

3. Les observations faites sont :

   $$\begin{aligned} f &= 0,36735 \\ 0.0524\sqrt{\frac{126}{343}\left(1-\frac{126}{343}\right)} &= 0.02526 \\ 0.3421 &\leq p \leq 0.3926 \\ 343 * 0,3421 * (1-0,3421) &= 77.2 > 15 : \text{l'approximation normale est acceptable} \end{aligned}$$

Exercice 2 : comparaison d'échantillons normaux

$$\begin{cases} \text{entreprises FT} & \overline{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = 141{,}46 \quad s_x^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i-\overline{x})^2 = 1973 \\ \text{entreprises HT} & \overline{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j = 160{,}86 \quad s_y^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(y_j-\overline{y})^2 = 9085 \end{cases}$$

1. Le modèle statistique est : $\begin{cases} \text{entreprises FT} & X_1,...,X_{n_1} \approx N\left[m_1;\sigma_1^2\right] \\ \text{entreprises HT} & Y_1,...,Y_{n_2} \approx N\left[m_2;\sigma_2^2\right] \end{cases}$,2 échantillons indépendants.

   Nous voulons tester $\{m_1 = m_2\}$ contre $\{m_1 \neq m_2\}$

2. Nous disposons de $n_1 = 14$ entreprises FT et $n_2 = 11$ entreprises HT :

   (a) les effectifs sont petits, nous ne pouvons faire le test que si en plus les variances sont égales $\left\{\sigma_1^2 = \sigma_2^2\right\}$

   (b) nous testons, au seuil de 10%, $\left\{\sigma_1^2 = \sigma_2^2\right\}$ contre $\left\{\sigma_1^2 \neq \sigma_2^2\right\}$
   statistique utilisée : le rapport des variances empiriques sans biais, qui suit, lorsque $\sigma_1^2 = \sigma_2^2$, une loi de FISHER :

   $$\begin{aligned} Z &= \frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}} = \frac{\frac{n_1}{n_1-1}s_1^2}{\frac{n_2}{n_2-1}s_2^2} \underset{Ho}{\sim} FISHER(n_1-1, n_2-1) \\ & Z \underset{Ho}{\sim} FISHER(13,10) \end{aligned}$$

   Nous déciderons que $\left\{\sigma_1^2 \neq \sigma_2^2\right\}$ si $Z > A$ ou $1/Z > B$. Lecture de table dans la loi de FISHER à 5% :

   $$P[FISHER(13,10) > 2{,}89] = 0.05 \text{ et } P[FISHER(10,13) > 2{,}67 = 0.05$$

   La règle de seuil 10% est donc de décider que les variances sont différentes si

   $$\frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}} > 2{,}89 \text{ ou } \frac{\widehat{\sigma_2^2}}{\widehat{\sigma_1^2}} > 2{,}67$$

   L'observation de $\widehat{\sigma_1^2} = \frac{14}{13} * 1973 = 2124.8$ et $\widehat{\sigma_2^2} = \frac{11}{10} * 9085 = 9993.5$ nous montre que

   $$\frac{\widehat{\sigma_2^2}}{\widehat{\sigma_1^2}} = \frac{9993.5}{2124.8} = 4{,}703 > 2{,}67$$

   Nous refusons donc, au seuil de 10%, l'hypothèse d'égalité des variances.

   (c) La taille des échantillons est trop petite pour utiliser le théorème central limite : nous ne pouvons faire le test d'égalité des capitaux moyens investis par les entreprises FT et HT :$_{observé}$

   $$\overline{x} \sim N\left[m_1; \frac{\sigma_1^2}{14}\right] \text{ et } \overline{y} \sim N\left[m_2; \frac{\sigma_2^2}{11}\right] \text{ sont bien indépendantes}$$

   mais il est impossible d'éliminer tous les paramètres pour obtenir une statistique de loi connue sous l'hypothèse $\{m_1 = m_2\}$.

3. Si nous disposons de $n_1 = 140$ entreprises FT et de $n_2 = 110$ entreprises HT : les variances étant encore mieux estimées, nous déciderons de la même manière qu'elles sont différentes. Nous pouvons utiliser directement l'écart réduit obtenu en remplaçant simplement les variances par leurs estimations :

$$\Delta = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{14} + \frac{s_2^2}{11}}} \approx N[0;1]$$

Nous refuserons $H_o = \{m_1 = m_2\}$ si $|\Delta| > A$. Pour un seuil de 10%, la valeur A doit vérifier :

$$P[\Delta \leq A] = 1 - \frac{0.10}{2} = 0,95$$

La lecture de la table Normale centrée réduite fournit $A = 1.645$. Nous devrons donc décider que $\{m_1 \neq m_2\}$ si

$$\left| \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{140} + \frac{s_2^2}{110}}} \right| > 1.65$$

Ici, $\Delta_{obs} = \frac{160.86 - 141.46}{\sqrt{1973/140 + 9085/110}} = 1.97 > 1.65$ (remarque : on trouve 1.96 lorsque l'on utilise les variances empiriques sans biais). La différence observée entre les moyennes empiriques est donc significative au seuil de 10% : le capital nécessaire à la création est en moyenne différent selon le degré technologique de l'activité de l'entreprise.

Exercice 3

1. $H1 : E(DUR) = aINV + bPRET + c$, $H2 : cov(DUR_i, DUR_j) = 0$ si $i \neq j$ et $var(DUR_i) = \sigma^2$

   La seule statistique ici disponible pour juger si le modèle est bien linéaire standard est celle de Durbin Watson; qui teste que la corrélation entre deux valeurs successives est nulle : c'est une condition nécessaire, mais pas suffisante, pour que toutes les corrélations soient nulles. Ici, sur données individuelles, c'est plutôt l'hétéroscédasticité qu'il aurait fallu pouvoir tester, mais nous nous contenterons du DW (qui peut tout de même alerter sur une erreur de spécification).

   $k\prime = 2, n = 145 : d_L = 1.65 + \frac{45}{50}(1.71 - 1.63) = 1.70$ et $d_U = 1.72 + \frac{45}{50}(1.76 - 1.72) = 1.756$

   L'observation $DW = 1,765$ est comprise entre d_U et 2 : au risque de 10%, nous ne rejetons pas l'hypothèse de modèle linéaire standard.

2. Le test global de $H_o : \{a = b = 0\}$ utilise la statistique de Fisher $F = \frac{SCE/2}{SCR/142} \underset{Ho}{\sim} FISHER(2;143)$

   Nous refusons H_o si $F > A$ ou si la p-value associée à la valeur observée est plus petite que le seuil choisi. Ici nous lisons que $P(FISHER(2;142) > 17,9) = 0.0000$

   Au seuil de 5% (par exemple), nous considérons que le capital investi et le montant du pret contracté sont globalement explicatives.

3. Tests individuels de $\{a = 0\}$ et $\{b = 0\}$: ce sont les tests de student pour chaque coefficient. On refuse la nullité du coefficient si la statistique de student associée est trop grande en valeur absolue : au seuil de 10% pour un degré de liberté de 142, la valeur critique est 1.65. Mais on peut aussi utiliser la p-value associée au coefficient : la règle est de refuser la nullité d'un coefficient dont la p-value est inférieure au seuil choisi. Ici, les deux p-values sont supérieures à 0.10 : aucune des deux variables n'est individuellement explicative.

4. Le paradoxe n'est qu'apparent. Le fait de refuser la nullité simultanée des coefficients veut dire que si nous éliminons l'une des deux variables, il faut garder l'autre. Le fait qu'en présence de l'autre chacune des variables soit négligeable est alors le signe qu'il existe une forte corrélation entre les deux et que les deux ensemble ne font pas mieux qu'une seule des deux : en effet, nous pouvons nous attendre à ce que le capital investi soit fortement corrélé (positivement) avec le montant du prêt qu'il a fallu contracter.