Université Paris 1, UFR 02, Licence de Sciences Economiques
STATISTIQUE, cours de Mme PRADEL
Partiel 6 février 2003

Eléments de corrigé

Exercice 1

1. $X_1,..,X_{n_1} \approx i.i.d.B(1;p_1)$ et $Y_1,..,Y_{n_2} \approx i.i.d.B(1;p_2)$, les variables prenant la valeur 1 si l'entreprise a une durée d'activité supérieure à 2 ans, 0 sinon.

   $i = 1,...,n_1 = 1839$ pour les entreprises créées sur une idée nouvelle, $j = 1,...,n_2 = 10371$ pour les autres entreprises, X_i et Y_j étant indépendantes.pour tout i et j.

2. Les hypothèses testées sont $\{p_1 = p_2\}$ contre $\{p_1 \neq p_2\}$.

3. test de seuil 10% :

   - La statistique utilisée est

$$\begin{aligned} Z &= \frac{F_1 - F_2}{\sqrt{F(1-F)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \\ F_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \ \ ,F_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j, \text{ et } \ F = \frac{n_1F_1 + n_2F_2}{n_1+n_2} \end{aligned}$$

   - Nous refuserons $\{p_1 = p_2\}$ si $|Z| > A$
   - Sous l'hypotèse de base $\{p_1 = p_2\}$, $Z \approx N(0;1)$
   - pour un seuil de 10%, nous lisons dans la table de la loi Normale centrée réduite : $P[U \leq 1,645] = 1 - 0,10/20 = 0,95$

     la règle est de décider que l'origine innovante d'une entreprise a une influence sur sa probabilité de survie à 2ans si

$$|Z| > 1,645$$

4. Les observations faites sont :

$$\begin{aligned} & f_1 = 0,7047 \quad f_2 = 0,6797 \quad f = 0,83465 \\ |z| &= 2,124 > 1,645 \end{aligned}$$

   Avec un risque d'erreur de 10%, nous constatons une différence significative entre les probabilités de survie à 2 ans des entreprises selon qu'elles sont créées sur une idée nouvelle ou non.

Exercice 2

$Y_1,...,Y_{20} \approx N[m;5]$

1. Le meilleur estimateur de m est la moyenne empirique de l'échantillon. Son espérance est égale à m, sa variance est $\frac{\sigma^2}{n} = \frac{5}{20} = \frac{1}{4}$

$$\overline{Y}_{20} = \frac{1}{20}\sum_{i=1}^{20} Y_i \approx N\left[m;\frac{1}{4}\right]$$

2. Intervalle de confiance 95% : fondé sur $\overline{Y}_{20}$: nous lisons dans la table de la loi normale centrée réduite :
   $P[U \leq 1,960] = 0,95 + (1-0,95)/2 = 0,975$
   $P\left[\left|\frac{\overline{Y}_{20}-m}{\sqrt{1/4}}\right| \leq 1,960\right] = O,95 \Longleftrightarrow P\left[\overline{Y}_{20} - \frac{1}{2}.1,960 \leq m \leq \overline{Y}_{20} + \frac{1}{2}.1,960\right] = O,95$

   $$P\left[\overline{Y}_{20} - 0,980 \leq m \leq \overline{Y}_{20} + 0,980\right] = O,95$$

   L'intervalle de confiance 95% est $I_{0,95} = \left\{\overline{Y}_{20} - 0,980 \leq m \leq \overline{Y}_{20} + 0,980\right\}$.

3. test de $H_o : \{m = 2\}$ contre $H_1 : \{m \neq 2\}$:

   (a) la statistique de test est $\overline{Y}_{20}$. La règle sera de rejeter $\{m = 2\}$ si $\overline{Y}_{20}$ est trop éloignée de 2 : $\left|\overline{Y}_{20} - 2\right| > A$
   Sous H_o, $\overline{Y}_{20} \approx N\left[2; \frac{1}{4}\right]$. et $\frac{\overline{Y}_{20}-2}{1/2} \approx N[0;1]$
   Pour un seuil de 5%, nous avons $P[U \leq 1,960] = 1 - 0,05/2 = 0,975$. La règle est de refuser $\{m = 2\}$ si $\left|\frac{\overline{Y}_{20}-2}{1/2}\right| > 1,960$, ou encore $\left\{\frac{\overline{Y}_{20}-2}{1/2} < -1,960 \quad \text{ou} \quad \frac{\overline{Y}_{20}-2}{1/2} > 1,960\right\}$

   $$\text{Refus de } \{m = 2\} \text{ si } \overline{Y}_{20} < 1,020 \text{ ou } \overline{Y}_{20} > 2,980$$

   (b) Lorsque $m = 3 : \overline{Y}_{20} \approx N\left[3; \frac{1}{4}\right]$, on a donc la puissance
   $\eta = P\left[\ \overline{Y}_{20} < 1,020 \text{ ou } \overline{Y}_{20} > 2,980\right] = P\left[U < \frac{1,020-3}{1/2} \text{ ou } U > \frac{2,980-3}{1/2}\right]$
   $\eta = P[U < -3,960] + P[U > -0,04] = 1 - P[U < 3,960] + P[U < 0,04]$
   $\eta = 1 - 1 + 0,5160 = 0.516$

   $$\eta = 0,516$$

4. La règle du chef de service est refuser $\{m = 2\}$ si l'intervalle ne recouvre pas la valeur 2.

   (a) Le seuil du test ainsi défini est la probabilité de cet événement lorsque $\{m = 2\}$.

   $$\alpha = P_{m=2}[2 \notin I_{0,95}] = 1 - 0,95 = 0,05$$

   On peut aussi le recalculer sachant que si $\{m = 2\}, \overline{Y}_{20} \approx N\left[2; \frac{1}{4}\right]$
   $\alpha = P\left[2 < \overline{Y}_{20} - 0,980 \text{ ou } \overline{Y}_{20} + 0,980 < 2\right] = 1 - P\left[\left|\overline{Y}_{20} - 2\right| < 0,980\right]$
   $\alpha = 1 - P\left[|U| < \frac{0,980}{1/2}\right] = 1 - P[|U| < 1,960] = 1 - 0,95 = 0,05$

   (b) Les deux tests sont identiques, puisque $[2 \notin I_{0,95}]$ est équivalent à $\left[2 < \overline{Y}_{20} - 0,980 \text{ ou } \overline{Y}_{20} + 0,980 < 2\right]$ ou encore $\left[\overline{Y}_{20} < 2 - 0,980 \quad \text{ou} \quad 2 + 0,980 < \overline{Y}_{20}\right]$, c'est-à-dire : $\left[\overline{Y}_{20} < 1,020 \text{ ou } 2,980 < \overline{Y}_{20}\right]$.

5. Application numérique : $\overline{y} = 3,1$ entraîne que

   $$I_{0,95} \quad : \quad 2,120 \leq m \leq 4,080$$
   avec un risque de 5% , nous décidons que $\{m \neq 2\}$

6. Si la variance est inconnue : nous remplaçons partout où elle intervient la variance σ^2 par son estimation sans biais

   $$S^2 = \frac{1}{19}\sum_{i=1}^{20}\left(Y_i - \overline{Y}_{20}\right)^2, \text{ et } Z = \frac{\overline{Y}_{20} - m}{\sqrt{S^2/20}} \approx STUDENT(19)$$

   On lit dans la table de student $P[|Z| > 2,093] = 0,05$.

L'intervalle de confiance devient $P\left[\overline{Y}_{20} - \frac{S}{\sqrt{20}}.2,093 \leq m \leq \overline{Y}_{20} + \frac{S}{\sqrt{20}}.2,093\right] = O,95$

$$P\left[\overline{Y}_{20} - 0,4680S \leq m \leq \overline{Y}_{20} + 0,4680S\right] = O,95$$

Le test de seuil 5% : refuser $\{m = 2\}$ si

$$\left|\frac{\overline{Y}_{20} - 2}{S}\right| > 0,4680$$

L'application numérique fournit ici : $I_{0,95} : 1,9068 \leq m \leq 4,2932$

La décision est ici que l'on ne peut, au risque de 5%, refuser l'hypothèse $\{m = 2\}$.

Exercice 3

1. La seule statistique ici disponible est celle de Durbin Watson qui teste que le modèle est linéaire standard (ce n'est qu'une condition nécessaire, mais pas suffisante).

   Dans la régression (I) : $k\prime = 3, n = 123 : d_L = 1,65$ et $d_U = 1,75$. L'observation $DW = 1,901$ est comprise entre d_U et 2 : au risque de 10%, nous ne rejetons pas l'hypothèse de modèle linéaire standard.

   Dans la régression (II), $k\prime = 3, n = 116 : d_L = 1,63$ et $d_U = 1,75$. L'observation de $4 - DW = 1,842$ est comprise entre d_U et 2 : au risque de 10%, nous ne refusons pas l'hypothèse de modèle linéaire standard.

2. Dans la régression (I), nous testons la nullité du coefficient de la variable AGE. Sa p-value associée est $0,1927 > 0,05$: au risque de 5%, nous considérons que l'âge n'est pas un facteur significatif pour expliquer le salaire des hommes.

3. Nous faisons les hypothèses suivantes :

$$\begin{aligned} u_{Hi} &= WH_i - a_H AGE_i - b_H KID_i - c_H EDU_i - d_H \approx i.i.d.N\left[0; \sigma_H^2\right] \\ u_{Fj} &= WF_j - a_F AGE_j - b_F KID_j - c_F EDU_j - d_F \approx i.i.d.N\left[0; \sigma_F^2\right] \\ & \quad u_{Hi} \text{ et } u_{Fj} \text{ sont indépendants} \end{aligned}$$

   Nous testons $\left\{\sigma_H^2 = \sigma_F^2\right\}$ contre $\left\{\sigma_H^2 \neq \sigma_F^2\right\}$.

   - la statistique utilisée est le rapport des variances résiduelles.$\widehat{\sigma_H^2}$ et $\widehat{\sigma_F^2}$
   - Sous l'hypothèse $H_o = \left\{\sigma_H^2 = \sigma_F^2\right\}$:

$$F = \frac{\widehat{\sigma_F^2}}{\widehat{\sigma_H^2}} \approx FISHER\,(116 - 4, 123 - 4)$$

   - La règle est de refuser H_o, en décidant que $\left\{\sigma_H^2 \neq \sigma_F^2\right\}$si $\left\{F > A \text{ ou } \frac{1}{F} > C\right\}$
   - Pour un risque de 10%, on affectera une probabilité de 5% à chaque terme de la région critique.La lecture de A pour un $FISHER(112, 119)$ ou de C pour un $FISHER(119, 112)$ montre que ces deux nombres sont compris entre 1,26 et 1,39.
   - L'observation faite est $F = \frac{\widehat{\sigma_F^2}}{\widehat{\sigma_H^2}} = \frac{4,817}{2,720} = 1,771 > 1,39 > A$. Avec un risque de 10%, nous refusons l'égalité des variances des salaires des hommes et des femmes.

4. Sous les hypothèses déjà décrites précédemment :

(a) $\widehat{c_H} \approx N\left[c_H, \sigma^2_{cH}\right]$ et $\widehat{c_F} \approx N\left[c_F, \sigma^2_{cF}\right]$, les tableaux nous fournissent les estimations :

$$\begin{aligned} \widehat{c_H} &= 4,9762 \text{ et } \widehat{c_F} = 5,1170 \\ \widehat{\sigma^2_{cH}} &= (0,06710)^2 = 0,0045024 \text{ et } \widehat{\sigma^2_{cH}} = (0,10313)^2 = 0,010636 \end{aligned}$$

(b) Les deux estimateurs sont indépendants car le premier n'est fonction que des WH_i tandis que le second n'est fonction que des WF_i qui sont indépendants entre eux (observations indépendantes).

(c) On en déduit que $\widehat{c_H} - \widehat{c_F} \approx N\left[c_H - c_F, \sigma^2_{cH} + \sigma^2_{cF}\right]$ et

$$U = \frac{\widehat{c_H} - \widehat{c_F} - c_H + c_F}{\sqrt{\sigma^2_{cH} + \sigma^2_{cF}}} \approx N[0;1]$$

(d) Test de $H_o = \{c_H = c_F\}$ contre $\{c_H \neq c_F\}$
On considère que

$$\frac{\widehat{c_H} - \widehat{c_F} - c_H + c_F}{\sqrt{\widehat{\sigma^2_{cH}} + \widehat{\sigma^2_{cF}}}} \approx N[0;1]$$

- La statistique utilisée sera

$$Z = \frac{\widehat{c_H} - \widehat{c_F}}{\sqrt{\widehat{\sigma^2_{cH}} + \widehat{\sigma^2_{cF}}}}$$

- sous l'hypothèse H_o, $Z \approx N[0;1]$
- Nous refuserons l'égalité des coefficients si $|Z| > A$
- Pour un seuil $\alpha = 10\% : A = 1,645$: nous déciderons que $\{c_H \neq c_F\}$ si $\left|\frac{\widehat{c_H} - \widehat{c_F}}{\sqrt{\widehat{\sigma^2_{cH}} + \widehat{\sigma^2_{cF}}}}\right| > 1,645$

(e) ici, $|Z_{observé}| = \frac{5,1170 - 4,9762}{\sqrt{0,0045024 + 0,010636}} = 1,144 < 1,645$. Au seuil de 10%, les impacts de l'éducation sur le salaire ne diffèrent pas significativement entre hommes et femmes.

5. le test de Durbin et Watson sur données individuelles :

(a) le test donne une condition nécessaire mais pas suffisante. La statistique trouvée par le collègue ne signale aucun problème. Dans la mesure où les observations sont rangées par ordre alphabétique,cela ne risque pas de montrer une hétérogénéité homme/femmes.

(b) sur l'échantillon rangé avec les hommes d'abord, puis les femmes, il y a une erreur systématique. Seule DW fait intervenir les positions relatives des observations, avec les termes $\left(\widehat{u}_t - \widehat{u}_{t-1}\right)^2$. Toutes les autres statistiques sont des sommes, commutatives et indépendantes de l'ordre dans lequel les observations sont rangées.

Ici, le DW conduit à rejeter l'hypothèse du modèle linéaire standard. En fait, c'est le modèle "linéaire" qui n'est pas correctement spécifié.

Une règle à retenir, illustrée ici, est que sur données individuelles il faut effectuer le test de DW en réordonnant l'échantillon suivant toutes les variables pour lesquelles il risque d'y avoir une hétérogénéité de comportement.