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

Sont autorisés, outre une calculette et les tables statistiques : deux feuilles (recto verso) manuscrites, remplies par l'étudiant (à sa convenance et de sa main) des formules qu'il a choisies.

Il sera tenu compte de la justesse et de la pertinence des arguments justifiant les réponses

Les exercices sont indépendants les uns des autres

Exercice 1 (3 pts)

Nous observons la durée d'activité des 12210 entreprises créées en Ile de France en 1994. Parmi elles, il y en a 1839 qui ont été créées en exploitant une idée nouvelle et que nous désignons comme "innovantes". Les 10371 autres entreprises sont dites "non innovantes". Au terme de deux ans, il reste en activité 70.47% des entreprises innovantes et 67.97% des non innovantes. Nous voulons décider si le fait de créer sur une idée nouvelle influe sur la probabilité p de survie à deux ans pour les entreprises d'Ile de France.

1. Décrire avec précision le modèle statistique correspondant à ces observations : quelles sont les variables aléatoires observées et les lois de ces observations ?
2. Quelles sont les hypothèses testées ?
3. Quel est le test que vous proposez, pour un seuil de 10% ?
4. Appliquez ce test aux observations faites : à quelle conclusion vous conduit-il ?

Exercice 2 (9 pts)

Nous observons un échantillon de taille $n = 20$ d'une variable aléatoire Y Normale, d'espérance m et de variance $\sigma^2 = 5$:

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

1. Quel est le meilleur estimateur de m ? Quelle est la loi de cet estimateur ?
2. Construire un intervalle de confiance 95% pour m.
3. Nous désirons décider si, oui ou non, $m = 2$.
   (a) Construire le test de seuil 5%.
   (b) Calculer la puissance du test en $m = 3$
4. Votre chef de service vous dit :"moi, j'aurais regardé simplement si la valeur 2 tombait dans l'intervalle de confiance : si ce n'est pas le cas, je refuse l'hypothèse $\{m = 2\}$".
   (a) Quel est le seuil du test ainsi défini ?
   (b) Comparer la règle de votre chef de service avec la vôtre.
5. Application numérique. Les moments empiriques de l'échantillon observé sont :

$$\overline{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 3,1, \qquad ,s^2 = \frac{1}{19}\sum_{i=1}^{20} (y_i - \overline{y})^2 = 6,5$$

   Calculer l'intervalle de confiance correspondant et la décision prise au vu de ces observations.
6. Comment ces résultats sont-ils modifiés si la variance est inconnue (ne pas chercher à calculer la puissance du test)?

Exercice 3 (8 pts)

Nous étudions les variations du salaire (W) dans un secteur industriel en fonction de l'âge (AGE) du salarié, de la présence d'enfants (KID) dans son foyer et de son niveau d'éducation (EDU). Nous disposons de 239 observations Normales et indépendantes.

Les régressions séparées pour les hommes et pour les femmes fournissent les ajustements suivants :

Régression (I) Variable expliquée : W				
Nombre d'observations utilisées 123 hommes				
Variance résiduelle : 2.720 Somme des Carrés des résidus : 323.6376				
Variable	Coefficient	Ecart-type estimé	T-Statistique	PROB
AGE	0.021428	0.016356	1.310069	0.1927
KID	2.278292	0.349698	6.515029	0.0000
EDU	4.976235	0.067102	74.15916	0.0000
C	15.14876	1.253631	12.08391	0.0000
Statistique de Durbin-Watson : 1.901				

Régression (II) Variable expliquée : W				
Nombre d'observations utilisées 116 femmes				
Variance résiduelle : 4.817 Somme des Carrés des résidus : 539.4996				
Variable	Coefficient	Ecart-type estimé	T-Statistique	PROB
AGE	0.067825	0.031297	2.167146	0.0323
KID	−10.37956	0.536640	−19.34177	0.0000
EDU	5.116983	0.103131	49.61634	0.0000
C	10.53009	1.773137	5.938677	0.0000
Statistique de Durbin-Watson : 2.148				

1. Ces résultats sont-ils compatibles avec les hypothèses du modèle linéaire standard permettant d'utiliser toutes les statistiques des tableaux ?

   On supposera dans la suite que

$$\begin{aligned} W_i &= a_H AGE_i + b_H KID_i + c_H EDU_i + d_H + u_{Hi}, \text{ où } u_{Hi} \approx i.i.d.N\left[0; \sigma_H^2\right] \\ W_j &= a_F AGE_j + b_F KID_j + c_F EDU_j + d_F + u_{Fj}, \text{ où } u_{Fj} \approx i.i.d.N\left[0; \sigma_F^2\right] \\ i &= 1, ..123 \text{ hommes et } j = 1, ...116 \text{ femmes} : u_{Hi} \text{ et } u_{Fj} \text{ sont indépendants} \end{aligned}$$

2. Est-ce que l'âge est un facteur explicatif du salaire des hommes, au seuil de 5% ?

3. Peut-on accepter l'hypothèse que, sachant les explicatives, les salaires des hommes et des femmes ont la même variance ?

4. Nous nous demandons si le niveau d'éducation a un impact différent pour les hommes et pour les femmes. On notera $\sigma^2_{c_H}$ et $\sigma^2_{c_F}$ les variances des estimateurs $\widehat{c_H}$ et $\widehat{c_F}$.

   (a) Quelles sont les lois des coefficients estimés de l'éducation dans chacune des deux régressions ? (il est inutile de chercher à calculer les variances $\sigma^2_{c_H}$ et $\sigma^2_{c_F}$). Quelles sont les estimations de c_H, c_F, $\sigma^2_{c_H}$ et $\sigma^2_{c_F}$ fournies par les tableaux de résultats ?

   (b) Démontrer que les deux estimateurs $\widehat{c_H}$ et $\widehat{c_F}$ sont indépendants l'un de l'autre.

   (c) En déduire la loi de leur différence.: écrire la variable centrée et réduite correspondante.

   (d) Le nombre d'observations est suffisament grand dans chacune des deux régressions pour que nous acceptions de considérer que lorsque l'on remplace, dans cette différence centrée et réduite, les variances par des estimateurs convergents, la variable obtenue suit une loi Normale centrée et réduite. En déduire un test d'égalité de c_H et c_F au seuil 10%.

(e) Compte tenu des observations, le niveau d'éducation a-t-il un impact sur le salaire différent pour les hommes et pour les femmes ?

5. Votre collègue étudie la même enquête, mais il a négligé de tenir compte du sexe des salariés, qui sont simplement rangés par ordre alphabétique. La régression MCO fournit le tableau de résultats suivants :

Régression (III) Variable expliquée : W				
Nombre d'observations utilisées 239				
Variance résiduelle : 36.314 Somme des Carrés des résidus : 8533.691				
Variable	Coefficient	Ecart-type estimé	T-Statistique	PROB
AGE	0.090390	0.047943	1.88536	0.0606
KID	-4.006715	0.928705	-4.314303	0.0000
EDU	6.098457	0.168563	36.17909	0.0000
C	-6.347135	2.998848	-2.116524	0.0354
Statistique de Durbin-Watson : 1.911				

Vous réordonnez les 239 observations selon l'identifiant INSEE (qui commence par 1 pour les hommes et par 2 pour les femmes) et vous refaites la régression, vous obtenez les résultats suivants :

Régression (IV) Variable expliquée : W				
Nombre d'observations utilisées 239				
Variance résiduelle : 36.314 Somme des Carrés des résidus : 8533.691				
Variable	Coefficient	Ecart-type estimé	T-Statistique	PROB
AGE	0.090390	0.047943	1.88536	0.0606
KID	-4.006715	0.928705	-4.314303	0.0000
EDU	6.098457	0.168563	36.17909	0.0000
C	-6.347135	2.998848	-2.116524	0.0354
Statistique de Durbin-Watson : 0.893				

(a) Que pensez vous du résultat obtenu par votre collègue : avait-il une chance de s'apercevoir de son erreur de spécification ?

(b) Comparez vos résultats avec ceux de votre collègue : pourquoi la statistique de Durbin et Watson est-elle la seule statistique qui ait changé ? Quelle règle d'utilisation du test de Durbin et Watson sur données individuelles pouvez-vous tirer de cet exemple ?

Extraits de tables statistiques

LOI NORMALE : $U \approx N(0,1)$:table de $F(u) = P(U \leq u)$

u	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

LOI DE STUDENT à ν degrés de liberté :
TABLE de t en fonction du degré de liberté et de la probabilité p, tels que $P(|T| > t) = p$

p	**0,90**	**0,70**	**0,50**	**0,40**	**0,30**	**0,20**	**0,10**	**0,05**	**0,02**	**0,01**
18	0,127	0,392	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,391	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,391	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
30	0,127	0,389	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,12566	0,38532	0,67449	0,84162	1,03643	1,28155	1,64485	1,95996	2,32634	2,57582

LOI DE FISHER : Table du quantile $F_{0,95}$en fonction de n_1 et n_2 : $P(F(n_1, n_2) > F_{0,95}) = 5\%$

$n_1 \backslash n_2$	**50**	**100**	**200**	∞
50	1,60	1,52	1,48	1,44
100	1,48	1,39	1,34	1,28
200	1,42	1,32	1,26	1,19
∞	1,35	1,24	1,17	1,00

TABLE de DURBIN-WATSON :
Test unilatéral de $\rho = 0$ contre $\rho > 0$, au seuil de 5% (test bilatéral : seuil de 10%)

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
n	dL	du	dL	du	dL	du	dL	du	dL	du
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,66	1,80
200	1,73	1,78	1,75	1,79	1,73	1,80	1,73	1,81	1,72	1,82