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

Eléments de corrigé

Exercice 1:

Nous avons un échantillon de loi Normale : $X_1, .., X_n \approx i.i.d.N\left[m; \sigma^2\right]$

1. Variance connue : $\sigma^2 = 4$

   (a) L'estimateur du maximum de vraisemblance de m est la moyenne empirique de l'échantillon, qui suit une loi Normale :

   $$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left[m; \frac{4}{n}\right]$$

   (b) La meilleure prévision de X_{n+1} est celle qui minimise l'erreur quadratique moyenne $E\left[\left(X_{n+1} - \widehat{x}\right)^2\right]$:.

   $$MSE = E\left[\left(X_{n+1} - \widehat{x}\right)^2\right] = V\left(X_{n+1}\right) + E\left[\left(m - \widehat{x}\right)^2\right]$$

   Puisque X_{n+1} est indépendant des observations déjà faites et d'espérance m inconnue, la meilleure prévision de X_{n+1} est l'estimateur de m qui minimise le risque quadratique moyen $E\left[\left(m - \widehat{x}\right)^2\right]$, c'est-à-dire ici $\overline{X}_n$.

   (c) L'erreur de prévision $Z = X_{n+1} - \overline{X}_n$ suit une loi Normale. L'espérance de Z est $E(Z) = m - m = 0$, la variance de Z est $V(Z) = V\left(X_{n+1}\right) + V\left(\overline{X}_n\right)$ car X_{n+1}et $\overline{X}_n$ sont indépendantes. Les expressions équivalentes pour $V(Z)$ sont :

   $$V(Z) = \sigma^2 + \frac{\sigma^2}{n} = 4 + \frac{4}{n} = 4\left(1 + \frac{1}{n}\right) = 4\frac{n+1}{n}$$

   Nous en déduisons la loi de l'erreur de prévision Z :

   $$Z = X_{n+1} - \overline{X}_n \approx N\left[0; \sigma^2 + \frac{\sigma^2}{n}\right] \text{ ou encore } N\left[0; 4\frac{n+1}{n}\right]$$

   (d) Intervalle de prévision pour X_{n+1} : la variable centrée réduite déduite de Z est

   $$U = \frac{X_{n+1} - \overline{X}_n}{\sqrt{\sigma^2 \frac{n+1}{n}}} \approx N[0; 1]$$

   L'intervalle bilatéral symétrique pour U est $[-u_\eta, u_\eta]$ tel que

   $$\begin{aligned} P\left[-u_\eta \leq U \leq u_\eta\right] &= 0,80 \\ P\left[U \leq u_\eta\right] &= 0,80 + \frac{1 - 0,80}{2} = 0,90 \end{aligned}$$

   Le calcul de u_η peut se faire de deux manières équivalentes (la seconde est plus rapide mais il faut savoir qu'une loi de Student de dl infini est la loi Normale).

   - Lecture dans la table de la loi de répartition de $N[0; 1]$:

     $$F(1.2816) = P\left[U \leq 1,2816\right] = 0,90$$

     en interpolant entre $F(1.28) = 0.8997$ et $F(1.29) = 0.9015$.

- Lecture dans la table de Student de d.l. infini, colonne $p = 1 - 0,80 = 0.20$

$$P\left[|U| > 1,28155\right] = 0.20$$

Nous en déduisons que $P\left[-1.2816 \leq \frac{X_{n+1}-\overline{X}_n}{\sqrt{\sigma^2 \frac{n+1}{n}}} \leq 1.2816\right] = 0,80$. La résolution des deux inégalités fournit un événement équivalent, de même probabilité :

$$P\left[\overline{X}_n - 1.2816\sqrt{\sigma^2\frac{n+1}{n}} \leq X_{n+1} \leq \overline{X}_n + 1.2816\sqrt{\sigma^2\frac{n+1}{n}}\right] = 0,80$$

Pour $n = 24$ et $\sigma^2 = 4$: $1.2816\sqrt{\sigma^2 \frac{n+1}{n}} = 2.616$. Cela conduit à l'intervalle de prévision 80% :

$$P\left[\overline{X}_n - 2.616 \leq X_{n+1} \leq \overline{X}_n + 2.616\right] = 0,80$$

2. La variance σ^2 est inconnue. Cela ne change rien aux estimations et prévisions ponctuelles : m est estimé par $\overline{X}_n \approx N\left[m; \frac{\sigma^2}{n}\right]$, la meilleure prévision de X_{n+1} est $\overline{X}_n$ et l'erreur de prévision suit une loi Normale : $Z = X_{n+1} - \overline{X}_n \approx N\left[0; \sigma^2 \frac{n+1}{n}\right]$.
Par contre, l'intervalle de prévision doit être construit en remplaçant σ^2 par son estimateur sans biais :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Nous savons que $T = \frac{X_{n+1}-\overline{X}_n}{\sqrt{S^2 \frac{n+1}{n}}} \approx STUDENT\left(n-1\right)$.
Pour $n = 24$, nous lisons dans la table de Student, ligne 23, colonne $p = 1 - 0.80 = 0.20$:

$$P\left[|STUDENT\left(23\right)| \leq 1.319\right] = 0.80$$

L'intervalle de prévision devient :

$$P\left[\overline{X}_n - 1,319\sqrt{S^2\frac{25}{24}} \leq X_{n+1} \leq \overline{X}_n + 1,319\sqrt{S^2\frac{25}{24}}\right] = 0,80$$

$$P\left[\overline{X}_n - 1,346.S \leq X_{n+1} \leq \overline{X}_n + 1,346.S\right] = 0,80$$

car $1,319\sqrt{\frac{25}{24}} = 1,346$.

3. Les observations faites conduisent à :

m estimé par :	$\overline{x}_n = 143,82$
$IC_{80\%}$ si $\sigma^2 = 4$:	$141,20 \leq X_{n+1} \leq 146,44$
$IC_{80\%}$ si σ^2inconnue :	$140,93 \leq X_{n+1} \leq 146,71$

Exercice 2 :

1. Soit $X_i = 1$ si le client i apprécie le produit, 0 sinon. Nous notons $p =: P\left[X_i = 1\right]$
Nous avons un échantillon de taille $n = 500$ d'une loi de $BERNOULLI\left(1,p\right)$. Nous sommes placés devant l'alternative :$(0 \leq p \leq 0.5)$ (le produit n'est apprécié que par moins d'un client sur 2) ou $(0.5 < p \leq 1)$ (le produit est apprécié par plus d'un client sur 2).

2. L'hypothèse dont le rejet à tort a les conséquences les plus facheuses pour la responsable marketing est $0 \leq p \leq 0.5$, puisque refuser à tort cette hypothèse revient à continuer à développer un produit qui n'est pas apprécié par suffisamment de clients. Elle fait donc le test de $(p \leq 0.5)$ contre $(0.5 < p)$.

3. Test de deux hypothèses simples $H_o : (p = 0,5)$ contre $H_1 : (p = p_1)$ où $p_1 > 0,5$.

   - Le test de Neyman est fondé sur la fréquence observée de clients favorables au produit
   $$F_n = \frac{\sum X_i}{n}$$
   - Sous $H_o : F_n \approx N\left[0,5; \frac{0,25}{n}\right]$ car $np_o(1-p_o) = 500 * 0.5 * 0.5 = 125 > 15$: l'approximation Normale de la loi de F_n est largement justifiée.
   - Nous refusons $(p = 0,5)$ et décidons que $(p = p_1)$ si la fréquence observée est trop grande
   $$\text{Région critique : } F_n > A$$
   - pour un seuil de 5% : A doit vérifier
   $$\begin{aligned} P[F_n > A \mid p = 0,5] &= 0,05 \\ P\left[\frac{F_n - 0.5}{\sqrt{\frac{0,25}{500}}} > \frac{A - 0.5}{\sqrt{\frac{0,25}{500}}} \mid p = 0,5\right] &= 0,05 \end{aligned}$$
   lecture dans la table $N[0;1]$: $P[N[0;1] < 1.645] = 1 - 0.05 = 0.95$. Nous en déduisons que
   $$\begin{aligned} \frac{A - 0.5}{\sqrt{\frac{0,25}{500}}} &= 1.645 \\ A &= 0.5 + 1.645\sqrt{\frac{0,25}{500}} = 0.53678 \end{aligned}$$
   - Test de seuil 5% : nous refusons $(p = 0.5)$ et décidons que $(p = p_1)$ si $F_{observé} > 0.537$

   Le test ne dépend pas de la valeur prise par $p_1$tant que $(p_1 > 0.5)$: il est uniformément le plus puissant pour tester $(p = 0.5)$ contre $p > 0.5$.

4. Notons W la région de refus de $(p = 0.5)$: $W = \{F_n > 0.53678\}$. La probabilité de la région de refus de $(p = 0.5)$ est, pour tout p : $P[W \mid p]$ .Nous pouvons utiliser l'approximation normale pour la loi de $F_n$tant que $0.04 < p < 0.96$,car dans ce cas $np(1-p) > 0.04 * 0.96 * 500 = 19.2 > 15$. :
   $$F_n \approx N\left[p; \frac{p(1-p)}{n}\right] \quad \text{entraîne que} \quad \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N[0;1]$$
   $$P[W \mid p] = P\left[\frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.53678 - p}{\sqrt{\frac{p(1-p)}{n}}}\right] = 1 - \Phi\left(\frac{0.53678 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$
   ou $\Phi(.)$ est la fonction de répartition de la loi $N[0.1]$. La probabilité de refuser H_o est donc une fonction de p décroissante en fonction de
   $$u = \frac{0.53678 - p}{\sqrt{\frac{p(1-p)}{n}}} = \sqrt{n}\frac{0.53678 - p}{\sqrt{p(1-p)}}$$

la dérivée de u est

$$\begin{aligned}
\frac{du}{dp} &= \sqrt{n}\left[\frac{-1}{\sqrt{p(1-p)}} + \frac{(0.53678-p)}{p(1-p)\sqrt{p(1-p)}}\left(-\frac{1}{2}\right)(1-2p)\right] \\
\frac{du}{dp} &= -\frac{\sqrt{n}}{p(1-p)\sqrt{p(1-p)}}\left[p(1-p)+(0.53678-p)(0.5-p)\right] \\
\frac{du}{dp} &= -\frac{\sqrt{n}}{p(1-p)\sqrt{p(1-p)}}\left[0.26839-0.03678p\right] < 0
\end{aligned}$$

Nous voyons donc que u est une fonction décroissante de p sur l'intervalle $0 < p < 1$ et que $P[W \mid p]$ est donc une fonction croissante de p (tant que l'approximation normale est acceptable : nous avons fixé $0.04 < p < 0.96$).

Considérons maintenant le test de $(p \leq 0.5)$ contre $(p > 0.5)$ utilisant cette même région critique : la règle est de décider que $(p > 0.5)$ si $F_{observée} > 0.53678$ et de décider $(p \leq 0.5)$ sinon.

- Lorsque $p \leq 0.5$: $P[W \mid p]$ est un risque de première espèce. C'est une fonction croissante de p, son maximum est donc atteint pour $p = 0.5$. Le test construit est donc de seuil

$$\alpha = P[F_n > 0.53678 \mid p = 0.5] = 5\%$$

- Lorsque $p > 0.5$: $P(W \mid p)$ est une puissance et pour toute autre région W^* :

$$P(W^* \mid 0.05) \leq 0.05 \Longrightarrow P(W^* \mid p) \leq P(W \mid p)$$

puisque le test construit est u.m.p parmi les tests de seuil 5%.

Le test construit est donc u.m.p parmi les tests de seuil 5% pour tester $(p \leq 0.5)$ contre $(p > 0.5)$

5. Pour les observations faites , $F_{observée} = 287/500 = 0,574 > 0.53678$. La fréquence observée est dans la région critique, nous refusons l'hypothèse de base $(p \leq 0.5)$ et nous décidons que $(p > 0.5)$. La responsable marketing décide donc de poursuivre le développement du produit.

Exercice 3

1. Les hypothèses doivent être traduites en fonction des variables présentes dans notre étude :

$$\begin{aligned}
H1 &: E(TAILLE_i) = \beta_1 + \beta_2.NAISSANCE_i, \text{ pour } i = 1, ..., 70 \\
H2 &: cov(TAILLE_i, TAILLE_j) = 0 \text{ si } i \neq j \text{ et } var(TAILLE_i) = \sigma^2, \text{ pour } i, j = 1, ..., 70
\end{aligned}$$

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).

$k\prime = 1, n = 70 : d_L = 1.58$ et $d_U = 1.64$

L'observation $DW = 0.934$ est inférieure à d_L : au risque de 10%, nous rejetons l'hypothèse de modèle linéaire standard.

2. Nous ne pouvons effectuer le test de nullité de β_2 : la statistique calculée par l'ordinateur ne suit pas une loi de Student, les $PROB$ calculées et figurant dans le tableau de résultats ne sont pas utilisables. Il n'est donc pas possible d'affirmer que la taille à 25 ans dépend de la génération.

3. Nous avons deux modèles distincts :

$$\begin{array}{lll} H1\,(F) & : & E\,(TAILLE_i) = \beta_1 + \beta_2.NAISSANCE_i, \text{ pour } i = 1, ..., 35 \ (i \text{ est une femme}) \\ H2\,(F) & : & cov\,(TAILLE_i, TAILLE_j) = 0 \text{ si } i \neq j \text{ et } var\,(TAILLE_i) = \sigma_1^2, \text{ pour } i, j = 1, ..., 35 \end{array}$$

et

$$\begin{array}{lll} H1\,(H) & : & E\,(TAILLE_i) = \beta_3 + \beta_4.NAISSANCE_i, \text{ pour } i = 36, ..., 70 \ (i \text{ est un homme}) \\ H2\,(H) & : & cov\,(TAILLE_i, TAILLE_j) = 0 \text{ si } i \neq j \text{ et } var\,(TAILLE_i) = \sigma_2^2, \text{ pour } i, j = 36, ..., 70 \end{array}$$

   (a) Nous pouvons effectuer le test de Durbin-Watson sur chaque régression :
   $k\prime = 1, n = 35 : d_L = 1.40$ et $d_U = 1.52$

      i. Pour les femmes : $DW_F = 1.820$ est compris entre d_U et 2 : au seuil de 10%, nous pouvons accepter l'hypothèse de modèle linéaire standard

      ii. Pour les hommes : $DW_H = 2.066 > 2$, $4 - DW_H = 1.934$ est compris entre d_U et 2 : au seuil de 10%, nous pouvons accepter l'hypothèse de modèle linéaire standard.

   (b) test de nullité d'un coefficient de régression : la règle est de considérer le coefficient comme significativement différent de zéro au seuil de 5% si la PROB associée est inférieure à 0.05.

      i. Pour les femmes : $\widehat{\beta}_2 = 0.0943$ a pour $PROB$ associée $0.0919 > 0.05$: le coefficient n'est pas significativement différent de zéro et donc l'effet génération n'est pas significatif pour les femmes au seuil de 5%, mais il l'est au seuil de 10%.

      ii. Pour les hommes : $\widehat{\beta}_4 = 0.2248$ a pour $PROB$ associée $0.0137 < 0.05$: le coefficient est significativement différent de zéro et donc l'effet génération est significatif pour les hommes au seuil de 5% (a fortiori il l'est au seuil de 10%).

4. Nous pouvons rassembler les deux équations de régression (F) et (H) sous la forme globale :

$$\begin{array}{rcl} TAILLE_i & = & 1.548F_i + 0.0943NAISSANCEF_i + 1.599H_i + 0.2248NAISSANCEH_i + \widehat{u}_i \\ i & = & 1, .., 70 \end{array}$$

   La question est : ces coefficients obtenus sont-ils bien ceux que nous obtenons en régressant la taille sur les nouvelles variables ? Il faut vérifier que les équations normales sont bien satisfaites. Les coefficient (a, b, c, d) sont solution du système d'équations :

$$\begin{array}{rcl} \sum_{i=1}^{70} (TAILLE_i - aF_i - bNAISSANCEF_i - cH_i - dNAISSANCEH_i)\, F_i & = & 0 \\ \sum_{i=1}^{70} (TAILLE_i - aF_i - bNAISSANCEF_i - cH_i - dNAISSANCEH_i)\, NAISSANCEF_i & = & 0 \\ \sum_{i=1}^{70} (TAILLE_i - aF_i - bNAISSANCEF_i - cH_i - dNAISSANCEH_i)\, H_i & = & 0 \\ \sum_{i=1}^{70} (TAILLE_i - aF_i - bNAISSANCEF_i - cH_i - dNAISSANCEH_i)\, NAISSANCEH_i & = & 0 \end{array}$$

   En tenant compte des définitions des nouvelles variables, la moitié des termes disparaissent dans chaque somme

et le système est équivalent à :

$$\begin{aligned}
\sum_{i=1}^{35} (TAILLE_i - a - bNAISSANCE_i) &= 0 \\
\sum_{i=1}^{35} (TAILLE_i - a - bNAISSANCE_i)\, NAISSANCE_i &= 0 \\
\sum_{i=36}^{70} (TAILLE_i - c - dNAISSANCE_i) &= 0 \\
\sum_{i=36}^{70} (TAILLE_i - c - dNAISSANCE_i)\, NAISSANCE_i &= 0
\end{aligned}$$

Les deux premières équations sont les équations normales de la régression (F), les deux dernières équations sont les équations normales de la régression (H). La solution est donc

$$\begin{aligned}
a = 1.548 \quad \text{et} \quad b = 0.0943 \\
c = 1.599 \quad \text{et} \quad d = 0.2248
\end{aligned}$$

5. Le modèle correspondant à la régression globale est :

$$\begin{aligned}
H1 \quad &: \quad E(TAILLE_i) = \beta_1 F_i + \beta_2 NAISSANCEF_i + \beta_3 H_i + \beta_4 NAISSANCEH_i \ \text{ pour } \ i = 1, .., 70 \\
H2 \quad &: \quad cov(TAILLE_i, TAILLE_j) = 0 \ \text{ si } \ i \neq j \ \text{ et } var(TAILLE_i) = \sigma^2, \ \text{ pour } i, j = 1, ..., 70
\end{aligned}$$

L'hypothèse H1 est bien une conséquence des hypothèses $H1(F)$ et $H1(H)$, et les observations sont sans corrélations. Mais il faut vérifier si $\sigma_1^2 = \sigma_2^2$.

Le test de $\left(\sigma_1^2 = \sigma_2^2\right)$ contre $\left(\sigma_1^2 \neq \sigma_2^2\right)$ est fondé sur la statistique

$$F = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} = \frac{SCR_F/(35-2)}{SCR_H/(35-2)} = \frac{SCR_F}{SCR_H}$$

Sous l'hypothèse de base H_o : $\left(\sigma_1^2 = \sigma_2^2\right)$, $F \approx FISHER(33, 33)$

Nous refusons H_o si $F > A$ ou $\frac{1}{F} > B$.

Pour un seuil de 10%, nous lisons A dans la table de Fisher 5%, et la valeur B sera la même, puisque les degrés de liberté sont les mêmes au numérateur et au dénominateur.

Lecture de table : $P[FISHER(33, 33) > 1.79] = 0.05$.

Règle de seuil 10% : refuser l'égalité des variances si $\frac{SCR_F}{SCR_H} > 1.79$ ou si $\frac{SCR_H}{SCR_F} > 1.79$.

Valeur observée pour la statistique de test :

$$\frac{SCR_H}{SCR_F} = \frac{0.188948}{0.064597} = 2.925 > 1.79$$

Conclusion du test : au seuil de 10% nous devons rejeter l'hypothèse que les tailles de femmes et les tailles de hommes sont de même variance.

6. Le calcul standard conduit aux coefficients corrects mais les variances de ces coefficients sont calculés en prenant comme hypothèse d'une variance commune pour toutes les observations.

   (a) Nous ne devons donc pas utiliser les PROB associées aux coefficients calculés dans la régression globale. Nous voyons ici que même au seuil 10% l'effet génération sur la taille des femmes n'est pas significatif, contrairement à ce que nous trouvons en faisant le test correct (dans la régression séparée).

(b) Nous devons garder les conclusions obtenues dans la régression séparée. Ici, la question est unilatérale : nous testons $\beta_2 = 0$ contre $\beta_2 > 0$. La p-value associée est $p - value = P(T > t_2)$. Or le listing standard donne $P(|T| > t_2) = 0.0919$. Nous en déduisons que $p - value = \frac{0.0919}{2} = 0.0459$.
Au seuil de 5%, nous pouvons conclure que la taille des femmes a augmenté depuis 1945.