ECO 4272: Économétrie
Exercice 1: Réponses

Steve Ambler
Département des sciences économiques
École des sciences de la gestion
Université du Québec à Montréal

Hiver 2008

1 Espérance et variance d'une fonction linéaire d'une variable aléatoire (10 points)

Soit Y la variable aléatoire qui est la température en degrés Fahrenheit. Soit $\tilde{Y}$ la variable aléatoire qui est la température en degrés Celsius. Nous avons:

$$\tilde{Y} = \frac{5}{9}\left(Y - 32\right).$$

Donc, nous avons tout de suite que:

$$\mathrm{E}\left(\tilde{Y}\right) = \frac{5}{9}\mathrm{E}\left(Y\right) - \frac{5}{9}32 = \frac{5}{9}70 - \frac{5}{9}32 = 21.11,$$

$$\mathrm{Var}\left(\tilde{Y}\right) \equiv \sigma^2_{\tilde{Y}} = \left(\frac{5}{9}\right)^2 \mathrm{Var}\left(Y\right) = \left(\frac{5}{9}\right)^2 \, 7^2 = 15.12,$$

et

$$\sigma_{\tilde{Y}} = \sqrt{15.12} = 3.89.$$

2 Covariance entre deux variables aléatoires (10 points)

$$\mathrm{cov}(a+bX, c+dY) \equiv \mathrm{E}[(a+bX-(\mathrm{E}(a+bX)))(c+dY-(\mathrm{E}(c+dY)))].$$

Pour commencer, nous avons:

$$\mathrm{E}(a+bX) \equiv \sum_{i=1}^{m}(a+bX_i)\mathrm{Pr}(X=X_i)$$

$$= a + b\sum_{i=1}^{m} X_i \mathrm{Pr}(X=X_i) \equiv a + b\mu_X,$$

et de manière semblable,

$$\mathrm{E}(c+dY) = c + d\mu_Y.$$

Donc, nous avons:

$$\rightarrow \mathrm{E}[(a+bX-(\mathrm{E}(a+bX)))(c+dY-(\mathrm{E}(c+dY)))] = \mathrm{E}[(b(X-\mu_X))(d(Y-\mu_Y))]$$

$$\equiv \sum_{i=1}^{m}\sum_{j=1}^{n} b(X_i-\mu_X)d(Y_j-\mu_Y)\mathrm{Pr}(X=X_i, Y=Y_j)$$

$$= b\ d\ \sum_{i=1}^{m}\sum_{j=1}^{n}(X_i-\mu_X)(Y_j=\mu_Y)\mathrm{Pr}(X=X_i, Y=Y_j)$$

$$\equiv b\ d\ \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] \equiv b\ d\ \mathrm{Cov}(X,Y),$$

ce qui fut à démontrer.

3 Échantillonnage aléatoire (30 points)

Voici les éléments de réponse que je voulais voir pour cette question.

1. La variance échantillonnale devrait décroître avec la taille de l'échantillon.

2. La moyenne échantillonnale peut être différente de la vraie moyenne (4), mais lorsque l'échantillon devient grand il devient de plus en plus improbable d'obtenir une moyenne échantillonnale très éloignée de 4. Donc, je serai surpris de voir une moyenne échantillonnale très différente de 4 pour des tailles d'échantillon égales à 100, 1000 et 10 000.

3. Votre histogramme devrait ressembler de plus en plus à une cloche normale lorsque l'échantillon devient grand.

4 Convergence (10 points)

1. Nous avons:

   $$\mathrm{E}\left(\tilde{Y}\right) = \frac{1}{2}\mathrm{E}\left(Y_1\right) + \frac{1}{2}\frac{1}{(n-1)}\sum_{i=2}^{n}\mathrm{E}\left(Y_i\right) = \frac{1}{2}\mu_Y + \frac{1}{2}\frac{1}{(n-1)}(n-1)\mu_Y = \mu_Y,$$

   ce qui fut à démontrer.

2. Nous avons:

   $$\begin{aligned}\mathrm{Var}\left(\tilde{Y}\right) &= \mathrm{Var}\left(\frac{1}{2}Y_1 + \frac{1}{2}\frac{1}{(n-1)}\sum_{i=2}^{n}Y_i\right) \\ &= \frac{1}{4}\sigma_Y^2 + \frac{1}{4}\left(\frac{1}{(n-1)}\right)^2\sum_{i=2}^{n}\sigma_Y^2 \\ &= \frac{1}{4}\sigma_Y^2 + \frac{1}{4}\left(\frac{1}{(n-1)}\right)^2(n-1)\sigma_Y^2 = \frac{1}{4}\sigma_Y^2\left(1 + \frac{1}{(n-1)}\right).\end{aligned}$$

3. La variance diminue avec n, et tend vers $\frac{1}{4}\sigma_Y^2$ lorsque n tend vers l'infini.

4. L'estimateur $\tilde{Y}$ n'est pas convergent. Sa variance n'est jamais plus petite que $\frac{1}{4}\sigma_Y^2$, ce qui veut dire que la probabilité qu'un intervalle arbitrairement petit autour de μ_y contienne la valeur réalisée de la moyenne échantillonnale ne tend pas vers un.

5 Tests d'hypothèse et intervalles de confiance (30 points)

Ici, il faut interpréter 108 comme étant l'estimé de l'écart type de la variable aléatoire d'intérêt, qui est le résultat du test. Autrement dit, nous avons:

$$\frac{1}{453-1}\sum_{i=1}^{453}\left(Y_i - \bar{Y}\right) = 108^2 \equiv s_Y^2$$

où Y est la variable aléatoire et $\bar{Y}$ est la moyenne échantillonnale. Donc, la variance de la moyenne échantillonnale est donnée par:

$$s_{\bar{Y}}^2 = \frac{s_Y^2}{453}.$$

- a) Pour construire un intervalle de confiance, nous faisons le raisonnement suivant:

$$\begin{aligned}
&\Pr\left(\mu_{Y_0} - c \leq \mu_Y \leq \mu_{Y_0} + c\right) \\
&= \Pr\left(-c \leq \mu_Y - \mu_{Y_0} \leq c\right) \\
&= \Pr\left(-\frac{c}{s_{\bar{Y}}} \leq \frac{\left(\mu_Y - \mu_{Y_0}\right)}{s_{\bar{Y}}} \leq \frac{c}{s_{\bar{Y}}}\right).
\end{aligned}$$

  Le terme $\frac{\left(\mu_Y - \mu_{Y_0}\right)}{s_{\bar{Y}}}$ est une variable aléatoire normalisée (on soustrait sa moyenne sous H_0 et on divise par son écart type estimé). Par hypothèse, la variable aléatoire Y est indépendamment distribuée. Donc, en grand échantillon, cette variable devrait être approximativement normale (standardisée). L'intervalle de confiance de 95% pour une normale standardisée est entre -1.96 et 1.96. Donc, la distance entre la moyenne sous l'hypothèse nulle et la borne inférieure de l'intervalle de confiance de 95% pour le résultat du test en Floride est donnée par la solution à:

$$1.96 = \frac{c}{s_{\bar{Y}}}$$

$$\rightarrow c = 1.96\ s_{\bar{Y}} = 1.96 \times 108/\sqrt{453} = 9.95.$$

  Donc, l'intervalle de confiance de 95% pour le résultat du test est entre $1013 - 9.95$ et $1013 + 9.95$.

- b) L'énoncé de la question nous dit que le résultat moyen pour les États Unis est 1000. Nous devons donc tester l'hypothèse nulle que le résultat moyen en Floride est égal à 1000. La statistique normalisée pour effectuer ce test est:

$$\frac{1013 - 1000}{(108/\sqrt{453})} = 2.56$$

  L'énoncé de la question parle de tester si le résultat moyen en Floride est “différent” par rapport au reste des États Unis. Donc, il serait logique de tester notre hypothèse nulle contre une alternative à deux extrémités. La valeur de la distribution normale (standardisée) cumulée pour -2.56 est 0.0055. Donc, la p-value de notre test est $2 \times 0.0055 = 0.011$. Nous rejetons l'hypothèse nulle avec un taux de significativité de 5%, mais nous l'acceptons à 1%.

- c)

- i) La différence entre les moyennes échantillonnales est donnée par:

$$1019 - 1013 = 6.$$

  Nous pouvons estimer l'écart type de cette différence entre moyennes échantillonnales sous l'hypothèse que les deux échantillons sont aléatoires et donc indépendants. La variance de la différence entre les moyennes échantillonnales peut-être estimée par:

$$\frac{s_b^2}{453} + \frac{s_a^2}{503} = \frac{108^2}{453} + \frac{95^2}{503} = 43.69$$

  La distance entre la moyenne sous l'hypothèse nulle et la borne inférieure de l'intervalle de confiance de 95% est donnée par la solution à:

$$1.96 = \frac{c}{\sqrt{43.96}} \to c = 12.96.$$

  Donc, l'intervalle de confiance de 95% est donné par $6 \ \pm \ 12.96$.

- ii) Sous l'hypothèse nulle que le cours n'a aucun effet sur le résultat moyen, la statistique normalisée suivante devrait approximativement suivre une distribution normale standardisée:

$$\frac{6}{\sqrt{43.69}} = 0.9077$$

  Nous devons tester l'hypothèse nulle contre l'hypothèse alternative que le cours a aidé à préparer le test. Donc, il s'agit d'une hypothèse alternative à une seule extrémité. La valeur de la distribution normale (standardisée) cumulée pour 0.91 est 0.8186. La p-value de notre test est 0.4443. Nous ne rejetons pas l'hypothèse nulle à des taux standard de significativité.

• d) Pour répondre à cette partie de la question, nous traitons le **changement** du résultat comme la variable aléatoire d'intérêt.

  - i) L'écart type de la moyenne échantillonnale du changement est donné par:

$$\frac{60}{\sqrt{453}} = 2.82.$$

    Donc, l'intervalle de confiance de 95% est donné par:

$$9 \ \pm \ 1.96 \times 2.82 = 9 \pm 5.53.$$

- ii) La statistique pour tester l'hypothèse nulle d'une absence d'amélioration est donnée par:

$$\frac{(9-0)}{2.82} = 3.95.$$

  L'hypothèse alternative est qu'il y a une amélioration, donc il s'agit d'une hypothèse alternative à une extrémité. La valeur de la normale standardisée cumulée pour 3.95 n'est pas dans le Tableau 1 à la fin du livre. Donc, la p-value du test est de loin inférieure à 0.002. Nous rejetons l'hypothèse nulle à tout niveau de significativité standard.

- iii) Il faudrait sélectionner n étudiants de manière aléatoire pour passer le test une première fois. Par la suite, il faudrait sélectionner une fraction (disons la moitié) de ces étudiants de manière aléatoire pour suivre le cours. Ensuite, on donne le test à tous les n étudiants et on compare les résultats de ceux qui on suivi le cours avec ceux n'ayant pas suivi le cours.