Statistiques appliquées (L3 d'économie) - Cours de Patrick Sevestre - TD 3 - Corrigé

Marc Sangnier - marc.sangnier@ens-cachan.fr

8 novembre 2007

Exercice 1 - Loyer d'un studio à Paris

Question A - Loyer moyen, loyer médian et variance des loyers

Studio	1	2	3	4	5	6
Loyer observé	390	460	650	410	270	780

Loyer moyen empirique

Soit X la variable décrivant le loyer d'un studio. Soit m_X le loyer moyen empirique. Soit n le nombre d'observation. Soit x_i la valeur observé pour le loyer du studio i.
$m_X = \sum_{i=1}^{6} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{6} x_i = \frac{2960}{6} = 493,33$

Loyer médian

Classons tout d'abord les obsrevations par ordre croissant du loyer.

Studio	5	1	4	2	3	6
Loyer observé	270	390	410	460	650	780

Le loyer médian correspond au loyer qui partage l'effectif est deux échantillons de même taille, l'un ayant des loyers inférieurs à la médiane, l'autre des loyers supérieurs à la médiane. Ici les deux sous-échantillons sont de taille égale à 3. Le loyer médian se situe donc entre 410 et 460 euros. Il est possible d'affiner la réponse en prenant le centre de cet intervalle, soit un loyer médian de 435 euros.

Variance des loyers

Calculons la variance empirique de notre ditribution.
$V_X = \sum_{i=1}^{6} (x_i - m_X)^2 \frac{1}{n} = 29222,22$
L'écart-type est quant à lui de 170,94 euros

Estimateur sans biais de la variance

L'estimateur $\hat{\sigma^2} = \sum_{i=1}^{6} (x_i - m_X)^2 \frac{1}{n-1}$ est un estimateur sans biais de la variance.
$\hat{\sigma^2} = 35066,67$

Question B - Variance de l'estimation du loyer moyen

On sait que la variance de la moyenne empirique m_X est $\frac{\sigma^2}{N}$.

On peut donc utiliser l'estimateur de la variance pour calculer la variance de notre estimation.

$\widehat{V(m_X)} = \frac{\hat{\sigma^2}}{n} = \frac{35066,67}{6} = 5844,44$

Question C - Echantillon plus large

$n' = 300$, $m'_X = 550$ et $\hat{\sigma'} = 300$

La variance de cette estimation du loyer est :

$\widehat{V(m'_X)} = \frac{\hat{\sigma'^2}}{n} = \frac{90000}{300} = 300$

On observe que cette variance est largement inférieure à celle obtenue précédemment. Les deux résultats sont extrémment différents. Ceci peut s'expliquer par deux raisons : soit les loyers ont été relevés dans deux populations différentes (deux groupes de studios différents) et les studios proposé en agence sont très différents de ceux proposés sur internet ; soit la première estimation est beaucoup moins précise que la seconde. La deuxième explication est la plus vraisemblable dans la mesure où les valeurs extrêmes ont beaucoup plus de poids dans un échantillon de petite taille ($n = 6$) que dans un échantillon de grande taille ($n = 300$).

La variance estimée du loyer moyen est plus faible dans le second échantillon en raison du plus grand nombre d'informations disponible. L'information accrôit la précision de l'estimation. Et ce, même si la variance estimée des loyers est plus faible dans le petit échantillon que dans le grand.

Question D - Moyenne arithmétique des deux estimations

Rappel : Variance d'une somme

$V(aX + bY) = a^2V(X) + b^2V(Y) + ab2cov(X;Y)$

Pondération 50/50

$l = 0,5m_X + 0,5m'_X$

$V(l) = 0,25V(m_X) + 0,25V(m'_X) + 0,25 * 2 * cov(m_X; m'_X)$

On admet ici que $cov(m_X; m'_X) = 0$. C'est à dire que les deux estimations sont indépentantes, ce qui est raisonnable dès lors que les annonces utilisées sont différentes.

$V(l) = 0,25(5844 + 300) = 1536,1$

Question E - Amélioration de l'estimation

Soit a et $(1 - a)$ les pondérations optimales qui permettent d'améliorer la précision du résultat.

$\bar{l} = a * m_X + (1 - a) * m'_X$

Si on conserve l'hypothèse d'indépendance faite dans la question précédente, il vient :

$V(\bar{l}) = a^2V(m_X) + (1 - a)^2V(m'_X)$

Minimisons cette variance par rapport à $a \iff \frac{\partial V(l)}{\partial a} = 0$

$\iff 2aV(m_X) - 2(1 - a)V(m'_X) = 0 \iff a = \frac{V(m'_X)}{V(m_X)+V(m'_X)}$

Or, $V(m_X) = \frac{\sigma^2}{n}$ et $V(m'_X) = \frac{\sigma^2}{n'}$

Donc, $a = \frac{n}{n+n'}$

Le poids a accorder à la première estimation est donc $a = 0,0196$ et celui a accorder à la seconde $1 - a = 0,9804$.

Il est logique d'accorder davantage de poids à la seconde estimation qui est plus précise.

En appliquant ces pondérations, on obtient $V(\bar{l}) = 290,6$. La précision est donc améliorée.

On peut, à titre indicatif, appliquer ces pondérations pour calculer la nouvelle moyenne empirique des deux échantillons réunis :

$\bar{l} = a * m_X + (1 - a) * m'_X = 0,0196 * 493,33 + 0,9804 * 550 = 548,88$

Question F - Loi normale

Soit X la variable aléatoire représetant le loyer d'un sutido.

$X \rightarrow N(m; \sigma)$

L'estimateur le plus évident pour m est la moyenne empirique $\bar{l}$ calculée précédemment.

Puisqu'on a supposé que les loyers suivaient une loi Normale, leur moyenne suit aussi une loi Normale d'espérance m est d'écart-type $\frac{\sigma^2}{N}$, où N est le nombre d'observations.

Question G - Estimation du loyer d'un studio supplémentaire

Supposons que le loyer du studio en question s'écrit $X_{n+1} = m + \varepsilon_{n+1}$où $\varepsilon_{n+1} \rightarrow N(0; \sigma)$.

Cette écriture signifie qu'on suppose que le loyer en question peut diffèrer de la moyenne d'un écart-type.

$\widehat{E(X_{n+1})} = E(\widehat{m + \varepsilon_{n+1}}) = \hat{m} + E(\varepsilon_{n+1}) = \bar{l} + 0 = \bar{l}$

Le meilleur prédicteur du loyer de ce studio est donc la moyenne empirique calculée à partir des observations précédentes.

Question H - Loi de probabilité de l'écart à la moyenne des autres loyers

Soit $e = X_{n+1} - \bar{l}$ l'écart du loyer du studio à la moyenne des autres loyers observés.

e est donc la différence de deux variables aléatoires suivant des lois Normale.

e suit donc également une loi Normale. Donc $E(e) = \bar{l} - \bar{l} = 0$

Exprimons maintenant la variance de l'écart.

$V(e) = V(X_{n+1} - \bar{l}) = V(X_{n+1}) + V(\bar{l})$ puisqu'on a supposé que ce loyer était indépendant des obsrevations précédentes.

$V(e) = \sigma^2 + \frac{\sigma^2}{N} = \frac{N+1}{N}\sigma^2$

Question I - Probabilité de trouver un studio dont le loyer...

On a :

$\bar{l} = 548,88$

$\hat{\sigma^2} = 35066$

$\hat{\sigma'^2} = 90000$

On peut donc estimer σ^2 par $\hat{\sigma^2} = \frac{1}{n+n'-1}(5 * 35066 + 299 * 90000) = 88804$

Donc $\hat{\sigma} = 298$

Il vient, d'après la question précédente : $\widehat{V(e)} = \frac{307}{306}88804 = 89094$

De même : $\hat{\sigma}_e = \sqrt{\frac{307}{306}88804} = 298,5$

On cherche a calculer la probabilité de trouver un studio dont le loyer est inférieur de 50 euros au loyer moyen de notre échantillon.

$P(X_{n+1} < \bar{l} - 50) = P(X_{n+1} - \bar{l} < -50) = P(e < -50) = P(e > 50) = 1 - P(e < 50)$

$= 1 - P(\frac{e}{298,5} < \frac{50}{298,5}) = 1 - P(Z < 0,168)$ où $Z \to N(0;1)$

$= 1 - 0,5675 = 0,4325$

Exercice 2 - Insertion professionnelle

Question A (texte de Patrick Sevestre)

Les salaires déclarés sont ceux perçus en euros (ou francs) courants au moment de l'entrée des personnes dans la vie active. Ils ne sont donc pas directement comparables car ils correspondaent à des périodes différentes. Or l'inflation a affecté l'évolution de ces salaires et un salaire de 1000 euros aujourdhui "ne vaut pas" (i.e. ne donne pas le même pouvoir d'achat) qu'un salaire de 6557 francs (équivalents à 1000 euros) il y a 20 ans. Il faut raisonner en valeur réelle, i.e. en éliminant leffet de l'inflation. Pour cela on déflate par lindice de prix :

$w_{rel} = \frac{w_{nominal}}{IPC/100}$ On divise l'IPC par 100 car celui est ex-primé en base 100 à la date de référence. Pour cela il faut un indice de prix couvrant l'ensemble de la période correspondant aux dates d'entrées des personnes dans l'emploi.

Question B

$X \to N(1450; 250)$

$P(X < 1350) = P(\frac{X-1450}{250} < \frac{1350-1450}{250}) = P(Y < -0,4)$ où $Y \to N(0;1)$

$= P(Y > 0,4) = 1 - P(Y < 0,4) = 1 - 0,6554 = 0,3446$

Question C

$P(\mid \bar{w} - E(w) \mid \leq 50) = P(\frac{|\bar{w} - E(w)|}{\sigma_{\bar{w}}} \leq \frac{50}{\sigma_{\bar{w}}})$

où $\sigma_{\bar{w}}$ est l'estimateur de la variance de la moyenne empirique : $\sigma_{\bar{w}} = \frac{250}{\sqrt{300}} = 14,43$

$= P(-3,5 \leq Y \leq 3,5) = P(Y \leq 3,5) - P(Y \leq -3,5) = P(Y \leq 3,5) - P(Y \geq 3,5)$

où $Y \to N(0;1)$

$= P(Y \leq 3,5) - [1 - P(Y \leq 3,5)] = 2P(Y \leq 3,5) - 1 = 2 * 0,9997 - 1 = 0,9994$

Question D

$P(\bar{w} < 1350) = P(\frac{\bar{w}-1450}{14,43} < \frac{1350-1450}{14,43}) = P(Y < -6,94) \approx 0$ où $Y \to N(0;1)$

Question E

$E(CDI = 1) = p = \frac{1}{N}\sum_{i=1}^{N} CDI_i$ où CDI_iest une variable indicatrice qui prend la valeur 1 lorsque l'individu observé a décroché un CDI en un an au plus.

La façon la plus simple d'estimer cette probabilité est d'utiliser la fréquence empirique de cet événement dans l'échantillon.

$\hat{p} = \frac{235}{300} = 0,783$

Question F

Pour modéliser la probabilité qu'un individu donné obtienne un CDI, on va utiliser la loi de Bernoulli.

Rappel : Loi de Bernouilli

Si la variable aléatoire X suit une loi de Bernouilli de paramètre p, on note : $X \rightarrow B(p)$

Avec : $P(x = 1) = p$ et $P(x = 0) = 1 - p$

On a alors : $E(X) = p$ et $V(X) = p(1 - p)$

Modélisation de la proportion d'individus qui décrochent un CDI (texte de Patrick Sevestre)

Pour modéliser la probabilité la proportion dindividus obtenant un CDI, on peut raisonner de deux manières :

1) considérer que l'on étudie la moyenne de n variables aléatoires indépendantes distribuées selon une loi de Bernoulli. Dans ce cas, on sait que la moyenne empirique va converger (puisque$n = 300$) vers la vraie moyenne (p) et que la variance de cette moyenne est égale à $\frac{\sigma^2}{n} = \frac{p(1-p)}{n}$

2) considérer que la proportion étudiée peut s'interpréter comme le nombre dindividus obtenant un CDI parmi n, divisé par n . Soit X la variable aléatoire décrivant le nombre dindividu décrochant un CDI parmi n, elle suit une loi binomiale d'espérance np et de variance $np(1-p)$. La proportion étudiée peut être représentée par une variable aléatoire $Y = \frac{X}{n}$ d'espérance p et de variance $\frac{p(1-p)}{n}$. Puisque n est grand, on peut soit recourir au théorème central limite, soit utiliser la convergence de la loi binomiale vers la loi normale. La proportion étudiée peut donc ici être modélisée par une loi normale despérance p et de variance $\frac{p(1-p)}{n}$.

Question G

Soit T la variable aléatoire décrivant le taux observé de CDI. Adptons l'approximation suivante : $T \rightarrow N(p; \frac{p(1-p)}{n})$

$P(T > 0,9) = P(\frac{T-p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0,9-p}{\sqrt{\frac{p(1-p)}{n}}}) = P(Y > \frac{0,9-p}{\sqrt{\frac{p(1-p)}{n}}})$ où $Y \rightarrow N(0; 1)$

Cette probabilité dépend donc de la vraie valeur de p.

Par exemple, si on choisit $p = 0,7$, il vient $P(T > 0,9) \approx 0$

Ceci montre que dès lors qu'on travaille avec un échantillon suffisamment grand, la moyenne de l'échantillon a une probabilité très faible de s'écarter sensiblement de la

vraie moyenne ; ce qui est positif. Cela suppose seulement que les observations soient indépendamment et identiquement distribuées.

Exercice 3 - Exercice supplémentaire

On se propose de montrer que la variance empirique est un estimateur biaisé de la variance.

Soit une suite de variables aléatoires (X_n).

$\forall i, E(X_i) = m$

$\forall i, V(X_i) = \sigma^2$

On dispose de n observations.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est la moyenne empirique, un estimateur sans biais de la moyenne : $E(\bar{X}) = m$

$\bar{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est la variance empirique.

Réécrivons la variance empirique :

$\bar{V} = \frac{1}{n}\sum_{i=1}^{n}(X_i - m + m - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}[(X_i - m) - (\bar{X} - m)]^2$

$= \frac{1}{n}[\sum(X_i - m)^2 - \sum 2(X_i - m)(\bar{X} - m) + \sum(\bar{X} - m)^2]$

$= \frac{1}{n}\sum(X_i - m)^2 - \frac{2}{n}(\bar{X} - m)\sum(X_i - m) + (\bar{X} - m)^2$

$= \frac{1}{n}\sum(X_i - m)^2 - 2(\bar{X} - m)^2 + (\bar{X} - m)^2 = \frac{1}{n}\sum(X_i - m)^2 - (\bar{X} - m)^2$

Passons à l'espérance :

$E(\bar{V}) = \frac{1}{n}\sum E(X_i - m)^2 - E(\bar{X} - m)^2$

$= E(X_i - m)^2 - E(\bar{X} - m)^2 = V(X_i) - V(\bar{X})$

Or : $V(\bar{X}) = \frac{\sigma^2}{n}$ et $V(X_i) = \sigma^2$

Donc :

$E(\bar{V}) = \sigma^2 - \frac{\sigma^2}{n} \neq \sigma^2$

La variance empirique est donc un estimateur biaisé de la variance.