Statistiques appliquées (L3 d'économie) - Cours de Patrick Sevestre - TD 2 - Corrigé

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

29 octobre 2007

Exercice 1 - Lien entre salaire et formation

Remarques préliminaires

Le tableau présentant les données de l'ennoncé nous fournit la loi du couple de variables aléatoires (niveau de formation ; salaire). On remarque donc que la somme de l'ensemble des valeurs est logiquement égale à 1. Soit F la variable aléatoire représentant la formation et S celle représentant le salaire. Par ailleurs, on peut considérer que le salaire est une variable continue : on comprendra donc que les inégalités strictes ou larges importent peu dans la lecture d'un tel tableau. Il faut cependant toujours prendre garde à la façon dont les données sont présentées pour ne pas faire d'erreurs de lecture.

Question 1 - Calcul de probabilités

Probabilité d'avoir une formation inférieure ou égale au bac et un salaire compris entre 1200 et 2400 euros

$P((F \leq Bac) \cap (1200 \leq S \leq 2400))$
$= P((F < Bac) \cap (1200 \leq S \leq 1600)) + P((F < Bac) \cap (1600 \leq S \leq 2400))$
$+P((F = Bac) \cap (1200 \leq S \leq 1600)) + P((F = Bac) \cap (1600 \leq S \leq 2400))$
$= 0,05 + 0,00 + 0,05 + 0,03 = 0,13$

Probabilité d'avoir une fomation supérieure à bac+3 et un salaire inférieur à 1600 euros

$P((F > Bac + 3) \cap (S < 1600))$
$= P((F > Bac + 3) \cap (S < 1200)) + P((F > Bac + 3) \cap (1200 \leq S < 1600)) = 0,03 + 0,05 = 0,08$

Question 2 - Distributions marginales

Rappel : Distribution marginale

La distribution marginale d'une variable aléatoire X est obtenue par la somme des probabilités conditionnelles de la variable X sachant Y.
$P(X = x_j) = \sum_{i=-\infty}^{+\infty} P((X = x_j) \cap (Y = y_i))$

Distribution marginale des salaires

On utilise maintenant les centres de classe plutôt que les intervalles : $P(S = 1100) = P(1000 < S < 1200)$.
$P(S = 1100) = P((S = 1100) \cap (F = 10)) + P((S = 1100) \cap (F = 12))$
$+P((S = 1100) \cap (F = 15)) + P((S = 1100) \cap (F = 18)) = 0,25 + 0,10 + 0,07 + 0,03 = 0,45$
En calculant de la même façon pour les autres valeurs du salaire, on obtient :

Salaire (S)	s=1100	s=1400	s=2000	s=3000
P(S=s)	0,45	0,25	0,15	0,15

Distribution marginale de la formation

On procède de la même façon, prenons l'exemple de F=Bac
$P(F=12) = P((F=12) \cap (S=1100)) + P((F=12) \cap (S=1400))$
$+P((F=12) \cap (S=2000)) + P((F=12) \cap (S=3000)) = 0,10 + 0,05 + 0,03 + 0,02 = 0,20$
En calculant de la même façon pour les autres valeurs de la formation, on obtient :

Formation (F)	f=10	f=12	f=15	f=18
P(F=f)	0,30	0,20	0,30	0,20

Question 3 - Distribution conditionnelle

Rappel : Distribution conditionnelle

Soit A et B deux événements :
$P(A \cap B) = P(A \mid B) * P(B) = P(B \mid A) * P(A)$

Distribution des salaires conditionnelle à une formation strictement supérieure au bac

$P(S=1100 \mid F>12) = \frac{P((S=1100) \cap (F>12))}{P(F>12)} = \frac{P((S=1100) \cap (F=15)) + P((S=1100) \cap (F=18))}{P(F=15)+P(F=18)} = \frac{0,07+0,03}{0,30+0,20} = \frac{0,10}{0,50} = 0,20$

$P(S=1400 \mid F>12) = \frac{P((S=1400) \cap F(>12))}{P(F>12)} = \frac{P((S=1400) \cap (F=15)) + P((S=1400) \cap (F=18)}{P(F=15)+P(F=18)} = \frac{0,10+0,05}{0,30+0,20} = \frac{0,15}{0,50} = 0,30$
En poursuivant de la même façon, on obtient :

Salaire (S) sachant F>Bac	s=1100	s=1400	s=2000	s=3000
$P(S=s \mid F>12)$	0,20	0,30	0,24	0,26

Question 4 - Salaires moyens

Rappel : Loi des espérances itérées

L'espérance non-conditionnelle est égale à l'espérance de l'espérance conditionnelle.
$E(Y) = E_X(E(Y \mid X))$

Salaire moyen pour chaque niveau de formation

On calcule ici $E(S \mid F=f)$ pour l'ensemble des valeurs de la formation.
$E(S \mid F=10) = \sum_s s * P(S=s \mid F=10)$
$= 1100 * P(S=1100 \mid F=10) + 1400 * P(S=1400 \mid F=10)$
$+2000 * P(S=2000 \mid F=10) + 3000 * P(S=3000 \mid F=10)$
$= 1100 * \frac{P((S=1100) \cap (F=10))}{P(F=10)} + 1400 * \frac{P((S=1400) \cap (F=10))}{P(F=10)} + 2000 * \frac{P((S=2000) \cap (F=10))}{P(F=10)} + 3000 * \frac{P((S=3000) \cap (F=10))}{P(F=10)}$
$= 1100 * \frac{0,25}{0,30} + 1400 * \frac{0,05}{0,30} + 2000 * \frac{0,00}{0,30} + 3000 * \frac{0,00}{0,30} = 1150$
Le salaire moyen avec une formation inférieure au bac est donc de 1150 euros.
$E(S \mid F=12) = \sum_s s * P(S=s \mid F=12)$
$= 1100 * P(S=1100 \mid F=12) + 1400 * P(S=1400 \mid F=12)$
$+2000 * P(S=2000 \mid F=12) + 3000 * P(S=3000 \mid F=12)$
$= 1100 * \frac{P((S=1100) \cap (F=12))}{P(F=12)} + 1400 * \frac{P((S=1400) \cap (F=12))}{P(F=12)} + 2000 * \frac{P((S=2000) \cap (F=12))}{P(F=12)} + 3000 * \frac{P((S=3000) \cap (F=12))}{P(F=12)}$
$= 1100 * \frac{0,10}{0,20} + 1400 * \frac{0,05}{0,20} + 2000 * \frac{0,03}{0,20} + 3000 * \frac{0,02}{0,20} = 1500$
Le salaire moyen avec une formation égale au bac est donc de 1500 euros.
De la même façon, il vient :
$E(S \mid F=15) = 1756,67$
$E(S \mid F=18) = 2115$

Salaire moyen global

On utilise ici la loi des espérances itérées.
$E(S) = E_F(E(S \mid F))$
$= E(S \mid F = 10) * P(F = 10) + E(S \mid F = 12) * P(F = 12) + E(S \mid F = 15) * P(F = 15) + E(S \mid F = 18) * P(F = 18)$
$= 1150 * 0,30 + 1500 * 0,20 + 1756,67 * 0,30 + 2115 * 0,20 = 1595$
Le salaire moyen global est donc de 1595 euros.
On peut obtenir directement ce résultat à partir de la loi marginale du salaire en caluclant $E(S) = \sum_s s * P(S = s)$.

Question 5 - Variances des salaires

Rappel : Décomposition de la variance

La variance de X est égale à la moyenne des variances X sachant Y plus la variance des moyennes de X sachant Y.
$V(X) = E_Y(V(X \mid Y)) + V_Y(E(X \mid Y))$

Variance su salaire pour chaque niveau de formation

On calcule ici $V(S \mid F = f)$ pour l'ensemble des valeurs de la formation.
$V(S \mid F = 10) = \sum_s [s - E(S \mid F = 10)]^2 * P(S = s \mid F = 10) = 12500$
La variance des salaires pour une formation inférieure au bac est donc de 12500.
On peut caculer l'écart-type des ces salaires : $\sigma(S \mid F = 10) = \sqrt{V(S \mid F = 10)} = 111,80$
L'écart-type des salaires pour une formation inférieure au bac est donc de 111,80 euros.
$V(S \mid F = 12) = \sum_s [s - E(S \mid F = 12)]^2 * P(S = s \mid F = 12)$
$= (1100 - 1500)^2 * P(S = 1100 \mid F = 12) + (1400 - 1500)^2 * P(S = 1400 \mid F = 12)$
$+(2000 - 1500)^2 * P(S = 2000 \mid F = 12) + (3000 - 1500)^2 * P(S = 3000 \mid F = 12)$
$= (-400)^2 * \frac{0,10}{0,20} + (-100)^2 * \frac{0,05}{0,20} + 500^2 * \frac{0,03}{0,20} + 1500^2 * \frac{0,02}{0,20} = 345000$
La variance des salaires pour une formation de niveau bac est donc de 345000.
L'écart-type est ici de 587,37 euros.
$V(S \mid F = 15) = \sum_s [s - E(S \mid F = 15)]^2 * P(S = s \mid F = 15) = 416455,56$
La variance des salaires pour une formation de niveau bac+1, bac+2 ou bac+3 est donc de 416455,56.
L'écart-type est ici de 645,33 euros.
$V(S \mid F = 18) = \sum_s [s - E(S \mid F = 18)]^2 * P(S = s \mid F = 18)$
$= (1100 - 2115)^2 * P(S = 1100 \mid F = 18) + (1400 - 2115)^2 * P(S = 1400 \mid F = 18)$
$+(2000 - 2115)^2 * P(S = 2000 \mid F = 18) + (3000 - 2115)^2 * P(S = 3000 \mid F = 18)$
$= (-1015)^2 * 0,15 + (-715)^2 * 0,25 + (-115)^2 * 0,20 + 885^2 * 0,40 = 598275$
La variance des salaires pour une formation supérieure à bac+3 est donc de 598275.
L'écart-type est ici de 773,48 euros.

Variance globale du salaire

On utilise ici la formule de décomposition de la variance.
$V(S) = E_F(V(S \mid F)) + V_F(E(S \mid F))$
Calculons tout d'abord $E_F(V(S \mid F))$
$E_F(V(S \mid F)) = \sum_f [V(S \mid F = f)] * P(F = f) = 317341,67$
Calculons maintenant $V_F(E(S \mid F))$
$V_F(E(S \mid F)) = \sum_f [E(S \mid F = f) - E(S)]^2 * P(F = f)$
$= [E(S \mid F = 10) - E(S)]^2 * P(F = 10) + [E(S \mid F = 12) - E(S)]^2 * P(F = 12)$
$+[E(S \mid F = 15) - E(S)]^2 * P(F = 15) + [E(S \mid F = 18) - E(S)]^2 * P(F = 18)$
$= [1150 - 1595]^2 * 0,30 + [1500 - 1595]^2 * 0,20 + [1756,67 - 1595]^2 * 0,30 + [2115 - 1595]^2 * 0,20$
$= (-445)^2 * 0,30 + (-95)^2 * 0,20 + 161,67^2 * 0,30 + 520^2 * 0,20 = 123133,33$
On peut maintenant calculer la variance des salaires $V(S)$
$V(S) = E_F(V(S \mid F)) + V_F(E(S \mid F)) = 317341,67 + 123133,33 = 440475$
La variance globale du salaire est donc de 440475.

L'écart-type des salaires est donc de 663,68.
Ce résultat peut être obtenu directement en calculant $V(S) = E(S^2) - E(S)^2$.

Question 6 - Indépendance

Rappel : Indépendance

Si X et Y sont deux variables aléatoires indépendantes, alors $cov(X;Y) = 0$

Rappel : Logique

$p \Rightarrow q \Leftrightarrow (non-q) \Rightarrow (non-p)$

Rappel : Covariance

$cov(X;Y) = E[(X - E(X)) * (Y - E(Y))] = E(XY) - E(X)E(Y)$

Covariance du salaire (S) et de la formation (F)

$cov(S;F) = E[(S - E(S)) * (F - E(F))]$
Il nous faut tout d'abord calculer $E(F)$
On procède comme dans la question 4 en utilisant $E(F) = E_S[E(F \mid S)]$
$E(F \mid S = 1100) = 10 * \frac{P((F=10)\cap(S=1100))}{P(S=1100)} + 12 * \frac{P((F=12)\cap(S=1100))}{P(S=1100)} + 15 * \frac{P((F=15)\cap(S=1100))}{P(S=1100)} + 18 * \frac{P((F=18)\cap(S=1100))}{P(S=1100)}$
$= 10 * \frac{0{,}25}{0{,}45} + 12 * \frac{0{,}10}{0{,}45} + 15 * \frac{0{,}07}{0{,}145} + 18 * \frac{0{,}03}{0{,}45} = 11,75$
$E(F \mid S = 1400) = 10 * \frac{P((F=10)\cap(S=1400))}{P(S=1400)} + 12 * \frac{P((F=12)\cap(S=1400))}{P(S=1400)} + 15 * \frac{P((F=15)\cap(S=1400))}{P(S=1400)} + 18 * \frac{P((F=18)\cap(S=1400))}{P(S=1400)}$
$= 10 * \frac{0{,}05}{0{,}25} + 12 * \frac{0{,}05}{0{,}25} + 15 * \frac{0{,}10}{0{,}25} + 18 * \frac{0{,}05}{0{,}25} = 14$
$E(F \mid S = 2000) = 15,2$
$E(F \mid S = 3000) = 16,2$
$E(F) = 11,75 * 0,45 + 14 * 0,25 + 15,2 * 0,15 + 16,2 * 0,15 = 13,50$
On peut maintenant calculer $cov(S;F)$
$cov(S;F) = [1100 - E(S)][10 - E(F)] * P((S = 1100) \cap (F = 10)) + [1400 - E(S)][10 - E(F)] * P((S = 1400) \cap (F = 10))$
$+[2000 - E(S)][10 - E(F)] * P((S = 2000) \cap (F = 10)) + [3000 - E(S)][10 - E(F)] * P((S = 3000) \cap (F = 10))$
$+[1100 - E(S)][12 - E(F)] * P((S = 1100) \cap (F = 12)) + [1400 - E(S)][12 - E(F)] * P((S = 1400) \cap (F = 12))$
$+[2000 - E(S)][12 - E(F)] * P((S = 2000) \cap (F = 12)) + [3000 - E(S)][12 - E(F)] * P((S = 3000) \cap (F = 12))$
$+[1100 - E(S)][15 - E(F)] * P((S = 1100) \cap (F = 15)) + [1400 - E(S)][15 - E(F)] * P((S = 1400) \cap (F = 15))$
$+[2000 - E(S)][15 - E(F)] * P((S = 2000) \cap (F = 15)) + [3000 - E(S)][15 - E(F)] * P((S = 3000) \cap (F = 15))$
$+[1100 - E(S)][18 - E(F)] * P((S = 1100) \cap (F = 18)) + [1400 - E(S)][18 - E(F)] * P((S = 1400) \cap (F = 18))$
$+[2000 - E(S)][18 - E(F)] * P((S = 2000) \cap (F = 18)) + [3000 - E(S)][18 - E(F)] * P((S = 3000) \cap (F = 18))$
$= (-495) * (-3,5) * 0,25 + (-195) * (-3,5) * 0,05 + 405 * (-3,5) * 0,00 + 1405 * (-3,5) * 0,00$
$+(-495) * (-1,5) * 0,10 + (-195) * (-1,5) * 0,05 + 405 * (-1,5) * 0,03 + 1405 * (-1,5) * 0,02$
$+(-495) * 1,5 * 0,07 + (-195) * 1,5 * 0,10 + 405 * 1,5 * 0,08 + 1405 * 1,5 * 0,05$
$+(-495) * 4,5 * 0,03 + (-195) * 4,5 * 0,05 + 405 * 4,5 * 0,04 + 1405 * 4,5 * 0,08$
$= 1036,5$
On a donc $cov(S;F) \neq 0$
Le salaire n'est donc pas indépendant du niveau de formation.

Coefficient de corrélation linéaire

On peut calculer le coefficient de corrélation linéaide de S et F.
$r(S;F) = \frac{cov(S;F)}{\sigma(S)*\sigma(F)}$
On calcule tout d'abord l'écart-type de F :
$\sigma(F) = \sqrt{V(F)}$
Avec $V(F) = (10-13,5)^2 * 0,3 + (12-13,5)^2 * 0,2 + (15-13,5)^2 * 0,30 + (18-13,5)^2 * 0,2 = 8,85$
Donc $\sigma(F) = \sqrt{8,85} = 2,97$
Or $\sigma(S) = 688,85$
D'où $r(S;F) = \frac{1036,5}{2,97*688,85} = 0,51$

Exercice 2 - Loi Normale bivariée

Notations

$(V;R) \to N(m_V; m_R; \sigma_V^2; \sigma_R^2; \rho)$ où ρ est le coefficient de corrélation linéaire.
$\rho = \frac{cov(V;R)}{\sigma_V * \sigma_R}$
Ici, on fait apparaître la variance dans les notations des paramètres de la loi Normale et non plus l'écart-type.
$V(X) = \sigma_X^2$ et $\sigma_X = \sqrt{V(X)}$

Question a - Distribution marginale des ventes

$V \to N(m_V; \sigma_V^2)$
Les ventes suivent une loi Normale d'espérances m_V et de variance σ_V^2.

Question b - Distribution conditionnelle des ventes

$(V \mid r) \to N[A + Br; \sigma_V^2(1-\rho^2)]$
Avec : $A = m_V - Bm_R$ et $B = \frac{cov(V;R)}{\sigma_R^2}$
Donc, avec une écriture plus lourde : $(V \mid R) \to N[m_V + \frac{cov(V;R)}{\sigma_R^2}(r - m_R); \sigma_V^2(1-\rho^2)]$

Question c - Espérance conditionnelle des ventes

$E(V \mid r) = A + Br = m_V + \frac{cov(V;R)}{\sigma_R^2}(r - m_R)$

Exercice 3 - Loi Normale multivariée

$X_n \to N(m; \sum)$
Où $X_n$est un vecteur aléatoire de dimension n et m le vecteur des espérances, également de dimension n.

$\sum$ est la matrice des variances-covariances. Elle est donc symétrique, régulière et définie positive. Il existe donc une matrice orthogonale S (Rappel : une matrice S est dite orthogonale si elle est telle que $S^{-1} =^t S$) telle que $S^{-1} \sum S = D$, où D est une matrice diagonale d'éléments λ_i, $1 \leq i \leq n$.

On peut donc définir un vecteur aléatoire Z de la façon suivante :
$Z =^t (X - m) \sum^{-1} (X - m) =^t (X - m) S D^{-1} S^{-1} (X - m)$
Posons $Y = S^{-1}(X - m) =^t S(X - m)$
Y suit donc une loi Normale centrée, dont la matrice des variances-covariances est $D = S^{-1} \sum S$.

Les composantes du vecteur Y, c'est à dire les variables aléatoires Y_i, sont donc indépendantes (puisque la matrice D est diagonale).

Par conséquent : $Y_i \to N(0; \sqrt{\lambda_i})$
Il vient donc : $Z =^t Y D^{-1} Y$
Ce qui peut s'écrire : $Z = \sum_{i=1}^{n} \frac{Y_i^2}{\lambda_i}$ qui est la somme des carrés de n variables aléatoire normales indépendantes centrées et réduites. Cette somme suit donc une loi du Khi-deux à n degrés de liberté.
$Z \to \chi_n^2$

Exercice 4 - Inégalité de Tchebychev (I)

Rappel : Inégalité de Markov

Si X est une variable aléatoire positive dont l'espérance existe, alors :
$\forall\lambda > 0$, $P[X \geq \lambda E(X)] \leq \frac{1}{\lambda}$ ou $P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$

Rappel : Inégalité de Bienaymé-Tchebytchev

En appliquant l'inégalité de Markov à la variable aléatoire $X - E(X)$ pour $k = 2$, on obtient :
$\forall\varepsilon > 0$, $P(\mid X - E(X) \mid \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}$

Exercice

On veut montrer, en utilisant l'inégalité de Bienaymé-Tchebytchev que : $\forall\varepsilon > 0$, $lim_{n\to+\infty}P(\mid Z_n - X \mid < \varepsilon) = 1$
$Z_n = X + Y_n \iff Z_n - X = Y_n$
Montrons donc que $\forall\varepsilon > 0$, $lim_{n\to+\infty}P(\mid Y_n \mid > \varepsilon) = 0$
D'après l'énoncé : $E(Y_n) = 0$
On peut écrire l'inégalité de Bienaymé-Tchebytchev pour Y_n :
$\forall\varepsilon > 0$, $P(\mid Y_n - E(Y_n) \mid \geq \varepsilon) \leq \frac{V(Y_n)}{\varepsilon^2}$
Or, d'après l'énoncé : $E(Y_n) = 0$ et $V(Y_n) = \frac{1}{n}$, donc $lim_{n\to+\infty}V(Y_n) = 0$
Or, on sait qu'une probabilité est toujours supérieure ou égale à 0. On peut donc écire, que lorsque n tend vers l'infini :
$0 \leq P(\mid Y_n \mid \geq \varepsilon) \leq 0$
$P(\mid Y_n \mid \geq \varepsilon)$ converge donc vers 0 lorsque n tend vers l'infini.
Donc, $\forall\varepsilon > 0$, $lim_{n\to+\infty}P(\mid Y_n \mid > \varepsilon) = 0$, ce qui est ce que nous voulions démontrer.
Z_n converge donc en probabilité vers X.

Exercice 5 - Inégalité de Tchebychev (II)

On peut écrire l'inégalité de Bienaymé-Tchebytchev pour X :
$\forall\varepsilon > 0$, $P(\mid X - m \mid \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$
Posons $\varepsilon = \lambda\sigma$.
$P(\mid X - m \mid \geq \lambda\sigma) \leq \frac{1}{\lambda^2} \iff 1 - P(\mid X - m \mid < \lambda\sigma) \leq \frac{1}{\lambda^2}$
$\iff P(\mid X - m \mid < \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}$, c'est à dire ce que nous voulions démontrer.

Exercice 6 - Notions de convergence

Convergence en probabilité

On dit que la suite de variables aléatoires (X_n) converge en probabilité vers une variable aléatoire X, si $\forall\varepsilon > 0$:
$P(\mid X_n - X \mid < \varepsilon) \to 1$ quand $n \to +\infty$
ou, de façon équivalente :
$P(\mid X_n - X \mid > \varepsilon) \to 0$ quand $n \to +\infty$

Convergence en loi

On dit que la suite de variables aléatoires (X_n) de fonction de répartition $F_n(.)$ converge en loi vers une variable aléatoire X de fonction de répartition $F(.)$ si la suite $F_n(x)$ converge vers $F(x)$ en tout point x où $F(.)$ est continue.
$\forall x/F(.)$ est continue en x :
$lim_{n\to+\infty} \mid F_n(x) - F(x) \mid = 0$

Convergence presque sûre

On dit que la suite de variables aléatoires (X_n) converge presque sûrement vers une variable aléatoire X, si $\forall \varepsilon > 0$:

$lim_{n \to +\infty} P(\mid X_i - X \mid > \varepsilon, \forall i \geq n) = 0$

Convergence en moyenne quadratique

On dit que la suite de variables aléatoires (X_n) converge en moyenne quadratique vers une variable aléatoire X, si :

$lim_{n \to +\infty} E(\parallel X_n - X \parallel^2) = 0$

Liens entre les différentes convergences

Convergence Presque Sûre $\Rightarrow$Convergence en Probabilité $\Rightarrow$Convergence en Loi