INSTITUT NATIONAL des SCIENCES APPLIQUÉES — INSA ROUEN

Statistiques

Final 2005/2006 – Corrigé indicatif

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION
3ème année
Resp: Ph. Leray

1 Estimation de paramètres... (8 pts)

Soit $X_1 \dots X_n$ un échantillon de taille n dont la v.a. parente X suit une loi normale de paramètres μ_X et σ^2. Soit $Y_1 \dots Y_m$ un échantillon de taille m dont la v.a. parente Y suit une loi normale de paramètres μ_Y et σ^2. Le but de cet exercice est de construire des estimateurs non biaisés des paramètres μ_X, μ_Y et σ^2.

1. (4 points) Déterminer les estimateurs du maximum de vraisemblance de ces trois paramètres, en fonction des échantillons $X_1 \dots X_n$ et $Y_1 \dots Y_m$.

$L(X_1 \dots X_n, Y_1 \dots Y_m; \mu_X, \mu_y, \sigma^2) = (2\pi\sigma^2)^{-(n+m)/2} e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^n (x_i-\mu_X)^2 + \sum_{j=1}^m (y_i-\mu_Y)^2)}$
$\ln L(X_1 \dots Y_m; \mu_X, \mu_y, \sigma^2) = -\frac{n+m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\sum_{i=1}^n (x_i-\mu_X)^2 + \sum_{j=1}^m (y_i-\mu_Y)^2)$

Puisqu'il faut trouver les estimateurs du max. de vraisemblances des trois paramètres à la fois, il faut résoudre le système suivant :
$\frac{\partial \ln L}{\partial \mu_X} = 0$; $\frac{\partial \ln L}{\partial \mu_Y} = 0$ et $\frac{\partial \ln L}{\partial \sigma^2} = 0$
La première équation donne : $-\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i-\mu_X)(-1) = 0$, soit après simplification $\boxed{\hat{\mu}_X = \overline{X}}$.
La seconde équation donne : $-\frac{1}{2\sigma^2} \sum_{j=1}^m 2(y_j-\mu_Y)(-1) = 0$, soit après simplification $\boxed{\hat{\mu}_Y = \overline{Y}}$.
La dernière équation donne : $-\frac{n+m}{2\sigma^2} + \frac{1}{2\sigma^4}(\sum_{i=1}^n (x_i-\mu_X)^2 + \sum_{j=1}^m (y_i-\mu_Y)^2)$,
soit : $\hat{\sigma}^2 = \frac{1}{n+m}(\sum_{i=1}^n (x_i-\mu_X)^2 + \sum_{j=1}^m (y_i-\mu_Y)^2)$ qui se simplifie grâce aux résultats des 2 premières équations : $\boxed{\hat{\sigma}^2 = \frac{1}{n+m}(\sum_{i=1}^n (x_i-\overline{X})^2 + \sum_{j=1}^m (y_i-\overline{Y})^2)}$.
Notons que les deux sommes peuvent se ré-écrire en fonction de $S_X^{*^2}$ et $S_Y^{*^2}$:
$\boxed{\hat{\sigma}^2 = \frac{1}{n+m}((n-1)S_X^{*^2} + (m-1)S_Y^{*^2})}$.

2. (2 points) Calculer le biais de ces trois estimateurs, et proposer si nécessaire de nouveaux estimateurs non biaisés.

$E(\hat{\mu}_X) = E(\overline{X}) = E(X) = \mu_X$: estimateur non biaisé
$E(\hat{\mu}_Y) = E(\overline{Y}) = E(Y) = \mu_Y$: estimateur non biaisé

$E(\hat{\sigma}^2) = \frac{1}{n+m} E((n-1)S_X^{*^2} + (m-1)S_Y^{*^2}) = \frac{1}{n+m}((n-1)E(S_X^{*^2}) + (m-1)E(S_Y^{*^2}))$ $\boxed{= \frac{n+m-2}{n+m}\sigma^2}$
(il faut se souvenir que $S_X^{*^2}$ et $S_Y^{*^2}$ sont tous les deux des estimateurs sans biais de σ^2)
$\hat{\sigma}^2$ est donc un estimateur biaisé qu'il faut corriger ... ce qui nous donne comme estimateur non biaisé
$\boxed{\hat{\sigma}^{*^2} = \frac{1}{n+m-2}((n-1)S_X^{*^2} + (m-1)S_Y^{*^2})}$

3. (2 points) Ces estimateurs sont-ils efficaces ?

Application du théorème de Cramer-Rao aux trois estimateurs $\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}^{*^2}$

2 Questions pour un champion... (5 pts)

Considérons que le temps passé quotidiennement, par un individu, devant un écran quelconque à regarder ou jouer à *Questions pour un champion* est une loi normale. Sur un échantillon de 20 personnes, on a observé que ce temps était en moyenne de 3h et 13 minutes avec un écart-type s de 46 minutes.

1. (1 point) Estimer ponctuellement (et sans biais) la moyenne et l'écart-type du temps passé devant la télévision pour l'ensemble de la population française.

L'énoncé indique que pour une réalisation d'un échantillon de taille $n = 20$, on a $\overline{x} = 193$ min. et $s = 46$ min.
$\overline{X}$ est un estimateur non biaisé de μ, nous pouvons donc estimer ponctuellement μ par $\boxed{\hat{\mu} = \overline{x} = 193}$.
Par contre S^2 est un estimateur **biaisé** de σ^2. Il faut donc estimer ponctuellement σ^2 par l'estimateur non biaisé $s^{*^2} = \frac{n}{n-1}s^2$, soit $\boxed{s^* = 47,195}$

2. (2 points) Estimer cette même moyenne par un intervalle avec un coefficient de confiance de 95%.

Il nous faut ici estimer l'intervalle de confiance d'une moyenne d'une population normale d'écart-type σ **inconnu**.
La statistique intéressante est $\sqrt{n}\frac{\overline{X}-\mu}{S}$ qui suit une loi de Student à $n-1$ degrés de liberté.
Nous cherchons donc $P(A < \sqrt{n}\frac{\overline{X}-\mu}{S} < B) = 1-\alpha = 0,95$.
La table des fractiles de la loi de Student nous donnent $B = 2,093 = -A$, il faut donc maintenant ré-écrire les inégalités $A < \sqrt{n}\frac{\overline{X}-\mu}{S} < B$ pour encadrer μ, et obtenir l'intervalle de confiance
[170,91 ; 215,09]

3. (2 points) Sur un échantillon plus important de 100 personnes on a observé un écart-type corrigé de 40 minutes. Tester l'hypothèse que la variance est de 1500 contre l'hypothèse qu'elle est plus importante (avec un seuil de 10%).

Il s'agit d'un test sur une variance d'une loi normale de moyenne inconnue.
Nos hypothèses sont $H_0 = [\sigma^2 = \sigma_0^2 = 1500]$ contre $H_1 = [\sigma^2 > \sigma_0^2]$
La statistique intéressante ici est $(n-1)\frac{{S^*}^2}{\sigma^2}$ qui suit une loi du χ^2 à $n-1$ d.d.l.
La région critique est de la forme ${s^*}^2 > h$
Le calcul de h s'effectue en utilisant la définition de α :
$\alpha = P(\text{décider } H_1|H_0) = P({S^*}^2 > h|\sigma^2 = \sigma_0^2) = P((n-1)\frac{{S^*}^2}{\sigma_0^2} > (n-1)\frac{h}{\sigma_0^2})$,
d'où $(n-1)\frac{h}{\sigma_0^2} = \chi^2(ddl = 99, p = 0.9) = 117.406$ et $h = 1778.9$
L'énoncé nous indique que ${s^*}^2 = 40^2 = 1600 < h$ ... on accepte donc H_0

3 Bowling ... (7 pts)

Les ASI4 affirment être meilleurs au bowling que les ASI3 ... Pour cela, ils s'appuient sur les résultats de certains étudiants pris au hasard lors la soirée Bowling de cette année (d'après ces chiffres, le plus petit score, 34 a été obtenu par un ASI3, le meilleur score 143 par un ASI4) :

ASI3	x_i	112	103	81	84	114	72	124	118	76	93	65	34	130	85	110
ASI4	y_i	123	105	69	139	107	143	133	78	60	96					

$\sum x_i = 1401$ $\sum x_i^2 = 140281$ $\sum y_i = 1053$ $\sum y_i^2 = 118723$

1. (0,5 points) Poser les hypothèses H_0 et H_1.

Mon hypothèse par défaut H_0 est que les résultats au Bowling des ASI3 et des ASI4 suivent la même loi, et ainsi que $ASI_4 = ASI_3$.
Mon hypothèse alternative H_1 correpond à l'affirmation des ASI4, c'est à dire que la variable $ASI_4 > ASI_3$

2. (1,5 points) Proposer deux méthodes différentes permettant de prendre une décision, en précisant (et en vérifiant) les circonstances vous permettant de les appliquer.

Deux tests statistiques pourraient être appliqués :
- le test du rang, complétement indépendant du type de distribution des deux v.a.
- le test d'égalité des moyennes (Student), mais qui ne peut marcher que si les deux v.a. suivent des lois normales de même variance ... il faut donc vérifier (de manière approximative et graphique par exemple) la normalité des deux variables, sur les réalisations données dans l'énoncé, puis faire le test d'égalité des variances (Fisher) avant de réaliser le test de Student ...

3. (5 points) Appliquer ces tests, avec un coefficient de confiance de 98%.

– **Test du Rang**
Nos hypothèses sont $H_0 = [F(ASI_3) = F(ASI_4)]$ contre $H_1 = [F(ASI_3) < F(ASI_4)]$.
La région critique est de la forme $W_3 < n(n+m+1)/2 - u_{1-\alpha}\sqrt{nm(n+m+1)/12} = 171$

Il faut donc calculer le rang des échantillons des ASI3, ce qui nous donne $w_3 = 179$, en dehors de la région critique ... nous décidons donc H_0, i.e. les résultats en bowling des ASI3 et des ASI4 suivent la même loi !

– **Tests de Student + Fisher**
Nos hypothèses pour le test d'égalité des variances sont $H_0 = [\sigma_4^2 = \sigma_3^2]$ contre $H_1 = [\sigma_4^2 \neq \sigma_3^2]$

Sous H_0, la statistique $\frac{S_3^{*^2}}{S_4^{*^2}}$ suit une loi de Fisher de d.d.l ($n-1=14$, $m-1=9$). La région critique est de la forme $\frac{S_3^{*^2}}{S_4^{*^2}} < A \ \cup \ \frac{S_3^{*^2}}{S_4^{*^2}} > B$.
A et B sont donnés par les fractiles de la loi de Fisher, en répartissant l'erreur sur les deux parties de la région critique : $B = \mathcal{F}(d_1 = 14; d_2 = 9; p = 0.99) = 4.03$ et $A = 1/B \simeq 0.25$.
Avec les données de l'énoncé, on calcule $\frac{s_3^{*^2}}{s_4^{*^2}} = 0.77$, valeur n'appartenant pas à la région critique ... nous décidons donc H_0, i.e. les deux v.a. ont la même variance.

Cette information nous permet de calculer une meilleure estimation de cette variance $\sigma^{*^2} = \frac{1}{n+m}((n-1)S_3^{*^2} + (m-1)S_4^{*^2}) = 750.8$

Puisque nous savons maintenant que les deux v.a. ont la même variance, il suffirait de montrer que $\mu_4 > \mu_3$ pour prouver l'assertion des ASI4... Nos hypothèses pour le test d'égalité des moyennes sont donc $H_0 = [\mu_4 = \mu_3]$ contre $H_1 = [\mu_4 > \mu_3]$ (attention, test uni-latéral ... H_1 ne concerne pas la différence ...)
Sous H_0, la statistique $\frac{\overline{ASI_4} - \overline{ASI_3}}{S^*\sqrt{1/n+1/m}}$ suit une loi de Student de d.d.l ($n+m-2=23$). La région critique est de la forme $\frac{\overline{ASI_4} - \overline{ASI_3}}{S^*\sqrt{1/n+1/m}} > A$ et A est donné par les fractiles de la loi de Student $A = \mathcal{T}(ddl = 23; p = 0.04) \simeq 2.1$ (attention, vous ne pouvez pas lire directement la valeur pour $p = 0.02$ (correspondant à une confiance de 98%) dans la table du support de cours qui correspond à un test bilatéral symétrique ... il faut donc chercher la valeur de t telle que toute l'erreur est commise "à droite", soit $p/2 = 0.02$ avec les notations de la table.
Avec les données de l'énoncé, on calcule $\frac{\overline{ASI_4} - \overline{ASI_3}}{s^*\sqrt{1/n+1/m}} = 1.06$, valeur n'appartenant pas à la région critique ... nous décidons donc H_0, i.e. les résultats en bowling des ASI3 et des ASI4 suivent la même loi !