Probabilidades e Estatística

LEAN, LEE, LETI, MEAer, MEBiom, MEEC, MEMec

1º semestre – 2013/2014
11/06/2014 – 11:00

Duração: 90 minutos **2º teste B**

Justifique convenientemente todas as respostas!

Grupo I 10 valores

1. (a) Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória proveniente de uma população $X \sim \text{Bernoulli}(p)$, com $0 < p < 1$. Deduza o estimador de máxima verosimilhança do parâmetro p. (3.0)

Seja $(x_1, \ldots, x_n) \in \{0,1\}^n$ uma concretização de $(X_1, \ldots, X_n)$.
Função de verosimilhança:

$$L(p|x_1,\ldots,x_n) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|p) \overset{X_i\, ind.}{=} \prod_{i=1}^n f_{X_i}(x_i|p) \overset{X_i\, i.d.}{=} \prod_{i=1}^n f_X(x_i|p)$$
$$= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \quad 0 < p < 1.$$

Função de log-verosimilhança:

$$\ell(p|x_1,\ldots,x_n) \equiv \log L(p|x_1,\ldots,x_n) = \sum_{i=1}^n x_i \log p + (n - \sum_{i=1}^n x_i)\log(1-p)$$

diferenciável em ordem a p. Maximização de $\ell(p|x_1,\ldots,x_n)$:

$$\frac{d}{dp}\ell(p|x_1,\ldots,x_n) = 0 \Leftrightarrow \frac{\sum_{i=1}^n x_i}{p} - \frac{n-\sum_{i=1}^n x_i}{1-p} = 0 \Leftrightarrow \sum_{i=1}^n x_i - p\sum_{i=1}^n x_i - np + p\sum_{i=1}^n x_i = 0 \Leftrightarrow p = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.$$

Seja $\hat{p} = \bar{x}$, então $\hat{p}$ é maximizante de $\ell(p|x_1,\ldots,x_n)$ se e somente se (onde $\bar{x} \neq 0 \wedge \bar{x} \neq 1$) $\frac{d^2}{dp^2}\ell(p|x_1,\ldots,x_n)\big|_{p=\hat{p}} < 0$ e

$$\frac{d^2}{dp^2}\ell(p|x_1,\ldots,x_n)\big|_{p=\hat{p}} = \frac{d}{dp}\left(\frac{\sum_{i=1}^n x_i}{p} - \frac{n-\sum_{i=1}^n x_i}{1-p}\right)\Big|_{p=\hat{p}} = -\left(\frac{\sum_{i=1}^n x_i}{p^2} + \frac{n-\sum_{i=1}^n x_i}{(1-p)^2}\right)\Big|_{p=\bar{x}} < 0,$$

pois $0 < \bar{x} < 1$. Logo, o estimador de máxima verosimilhança de p é $\bar{X}$ e a sua estimativa é $\bar{x}$.

(b) Tendo sido seleccionadas ao acaso 1000 moedas de entre as moedas um euro em circulação em Portugal, constatou-se que 300 moedas da amostra não eram portuguesas. Com base nesta informação, determine um intervalo de confiança aproximado a 99% para a proporção p de moedas não portuguesas entre as moedas de um euro em circulação em Portugal. (3.0)

Seja $X_i = 1$, se a moeda não é portuguesa, e $X_i = 0$, se a moeda é portuguesa. $X_i \sim$ Bernoulli(p), onde $p = P$(moeda não ser portuguesa), $i = 1, \ldots, n$.
Se $X_1, \ldots, X_n$ são v.a. i.i.d. com $E(X_i) = p$ e $Var(X_i) = p(1-p) < \infty$, pelo T.L.C. (a aproximação válida porque $n = 1000 > 30$), então $\frac{\bar{X}-E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X}-p}{\sqrt{p(1-p)/n}} \overset{a}{\sim} N(0,1)$.
Variável fulcral: Para simplificar a dedução do intervalo de confiança considera-se outra aproximação ($\hat{p} = \bar{X}$ é estimador de máxima verosimilhança de p e a sua estimativa é $\frac{300}{1000} = 0.3$)

$$T = \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \overset{a}{\sim} N(0,1).$$

Quantis: $1-\alpha = 0.99 \approx P(-b < T < b|H_0)$, $\quad b : F_{N(0,1)}(b) = 1 - 0.01/2 = 0.995 \Rightarrow b = 2.5758$.
Intervalo aleatório de confiança:

$$-b \leq \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \leq b \Leftrightarrow \bar{X} - b\sqrt{\bar{X}(1-\bar{X})/n} \leq p \leq \bar{X} + b\sqrt{\bar{X}(1-\bar{X})/n}$$

$$\Rightarrow IAC_{0.90}(p) \approx \left[\bar{X} - b\sqrt{\bar{X}(1-\bar{X})/n}, \bar{X} + b\sqrt{\bar{X}(1-\bar{X})/n}\right].$$

Intervalo de confiança:

$$\begin{aligned} IC_{0.99}(p) &\approx \left[\bar{x} - b\sqrt{\bar{x}(1-\bar{x})/n}, \bar{x} + b\sqrt{\bar{x}(1-\bar{x})/n}\right] \\ &= \left[0.3 - 2.5758\sqrt{0.3(1-0.3)/1000}, 0.3 + 2.5758\sqrt{0.3(1-0.3)/1000}\right] \\ &= [0.263, 0.337]. \end{aligned}$$

2. Uma determinada fábrica emite diariamente para a atmosfera elevadas quantidades de certo poluente. A direcção de ambiente dessa fábrica afirma que a mesma está a cumprir o limite imposto por lei, ou seja, em média a fábrica emite no máximo 16 unidades/dia desse poluente. Uma entidade independente mediu durante 40 dias, escolhidos ao acaso, a quantidade diária de poluente emitido pela fábrica, tendo obtido para essas observações uma média de 19 unidades. Com base na amostra, e supondo que a quantidade diária de poluente emitido pela fábrica tem distribuição normal com desvio padrão de 5 unidades, responda às seguintes questões:

(a) Pode-se concluir, para um nível de significância de 5%, que a fábrica não está a cumprir o limite imposto por lei? (3.0)

Seja X = quantidade diária de poluente emitido pela fábrica. $(X_1, \ldots, X_n)$ uma a.a. de $X \sim N(\mu, \sigma = 5)$.
Hipóteses:
$H_0 : \mu \leq 16$ *versus* $H_1 : \mu > 16$.
Estatística do teste:

$$T_0 = \frac{\bar{X}-16}{\sigma/\sqrt{n}} \overset{\mu=16}{\sim} N(0,1),$$

cujo valor observado é $t_0 = \frac{19-16}{5/\sqrt{40}} = 3.795$.
Região crítica:
$\alpha = 0.05 = P(\text{Rejeitar } H_0|\mu = 16) \Leftrightarrow 0.05 = P(T_0 > b|\mu = 16) \Leftrightarrow F_{N(0,1)}(b) = 0.95 \Leftrightarrow b = 1.6449$.
$RC =]1.6449, +\infty[$.
Decisão: Se $t_0 > 1.6449 \Rightarrow$ Rejeita-se H_0 e se $t_0 \leq 1.6449 \Rightarrow$ Não se rejeita H_0. Logo, como $t_0 = 3.795 > 1.6449$, rejeita-se a hipótese de a fábrica estar a cumprir o limite imposto por lei (emissão média de no máximo 16 unidades/dia desse poluente) ao nível de significância de 5%.

(b) Ao efectuar o teste da alínea anterior, qual será a probabilidade de, correctamente, se concluir que a que a fábrica não está a cumprir o limite imposto por lei, caso o valor médio da quantidade de poluente emitido pela fábrica seja de 18 unidades/dia? (1.0)

$$P(\text{Tomar decisão correcta}|\mu = 18) = P(\text{Rejeitar } H_0|\mu = 18) = P\Big(\frac{\bar{X}-16}{\sigma/\sqrt{n}} > 1.6449\Big|\mu = 18\Big)$$

$$\text{Se } \mu = 18, \text{então } \frac{\bar{X}-18}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$= P\Big(\frac{\bar{X}-18}{5/\sqrt{40}} > \frac{16}{5/\sqrt{40}} + 1.6449 - \frac{18}{5/\sqrt{40}}\Big|\mu = 18\Big)$$

$$= 1 - F_{N(0,1)}(-0.885) = F_{N(0,1)}(-0.885) \approx 0.8119.$$

Grupo II 10 valores

1. Segundo as hipóteses de herança de Mendel, filhos de pais em que ambos os progenitores possuem sangue do grupo AB podem ter grupo sanguíneo AA, AB e BB com probabilidades, respectivamente, 0.25, 0.50 e 0.25. Recolhida uma amostra de 284 crianças com ambos os progenitores possuindo sangue do grupo AB, obtiveram-se os seguintes resultados: (4.0)

Grupo sanguíneo	AA	AB	BB
Número de crianças	65	152	67

Será que estes resultados são consistentes com as hipóteses de herança de Mendel? Decida através do valor-p do teste, tendo em consideração os níveis de significância usuais.

Seja X a v.a. que representa o grupo sanguíneo de uma criança com ambos os progenitores possuindo sangue do grupo AB: $X = 1$, se AA; $X = 2$, se AB; $X = 3$, se BB.
Hipóteses:
$H_0 : P(X = 1) = P(X = 3) = 0.25, P(X = 2) = 0.5$ *versus* $H_1 : X$ tem outra função de probabilidade.
Estatística do teste:

$$T_0 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\overset{a}{\sim}} \chi^2_{(k-\beta-1)} = \chi^2_{(2)},$$

cujo valor observado é dado por

classe	o_i	p_i	$e_i = np_i$	$\frac{(o_i-e_i)^2}{e_i}$
{X=1}	65	0.25	71 (≥ 5)	0.5070
{X=2}	152	0.50	142 (≥ 5)	0.7042
{X=3}	67	0.25	71 (≥ 5)	0.2254
	$n = 284$	1	284	$t_0 = 1.4366$

Sejam $p_1 = P(X = 1|H_0 \text{ verdadeiro}) = 0.25$, $p_2 = P(X = 2|H_0 \text{ verdadeiro}) = 0.50$ e $p_3 = P(X = 3|H_0 \text{ verdadeiro}) = 0.25$. Note-se que $k = 3$ ($e_i \geq 5, \forall i$) e $\beta = 0$ (não foram estimados parâmetros).
Valor-p:
$p = P(T_0 \geq t_0|H_0 \text{ verdadeiro}) = 1 - F_{\chi^2_{(2)}}(1.4366)) = 1 - 0.5124 = 0.4876.$
Decisão: Se $\alpha \geq 0.4876 \Rightarrow$ Rejeita-se H_0 e se $\alpha < 0.4876 \Rightarrow$ Não se rejeita H_0. Logo, aos níveis de significância usuais $(0.01, 0.05, 0.10)$, não se rejeita a hipótese de estes resultados serem consistentes com as hipóteses de herança de Mendel.

2. Para estudar a relação entre a altura das ondas (X, em metros) e o montante Y (em milhares de euros) dos estragos causados na orla costeira em dias de forte agitação marítima, foram obtidas observações relativas a 30 dias com forte agitação marítima, que conduziram a:

$$\sum_{i=1}^{30} x_i = 212 \qquad \sum_{i=1}^{30} x_i^2 = 1824 \qquad \sum_{i=1}^{30} x_i y_i = 6410 \qquad \sum_{i=1}^{30} y_i = 630 \qquad \sum_{i=1}^{30} y_i^2 = 28230$$

(a) Ajuste um modelo de regressão linear simples de Y sobre x e obtenha uma estimativa para o incremento esperado no montante dos estragos causados na orla costeira por um acréscimo de meio metro (2.0) na altura das ondas.

$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{6410-30\times7.0667\times21}{1824-30\times7.0667^2} = 6.009$ e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 21 - 6.009\times7.0667 = -21.461$

Equação de regressão ajustada: $\hat{E}(Y|x) = -21.461 + 6.006\,x$.

Seja $\Delta = E(Y|x+1/2) - E(Y|x) = \beta_1/2$. Logo, $\hat{\Delta} = \frac{\hat{\beta}_1}{2} = \frac{6.009}{2} = 3.0045$.

(b) Assumindo as hipóteses de trabalho habituais, teste ao nível de significância de 5% a hipótese de a altura das ondas não influenciar o montante dos estragos causados na orla costeira. (4.0)

Hipóteses:

$H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$.

Estatística do teste:

$$T_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}} \overset{H_0}{\sim} t_{(n-2)} = t_{(28)},$$

cujo valor observado é $t_0 = 6.009 \Big/ \sqrt{\frac{115.542}{1824-30\times7.0667^2}} = 10.091$,

onde $\hat{\sigma}^2 = \frac{1}{n-2}[(\sum_{i=1}^n y_i^2 - n\bar{y}^2) - (\hat{\beta}_1)^2(\sum_{i=1}^n x_i^2 - n\bar{x}^2)] = \frac{(28230-30\times21^2)-6.009^2(1824-30\times7.0667^2)}{28} = 115.542$.

Região crítica:

$\alpha = 0.05 = P(\text{Rejeitar } H_0|H_0 \text{ verdadeiro}) \Leftrightarrow 0.05 = P(|T_0| > b|H_0) \Leftrightarrow F_{t_{(28)}}(b) = 0.975 \Leftrightarrow b = 2.048$.

$RC =]-\infty, -2.048[\cup]2.048, +\infty[$.

Decisão: Se $|t_0| > 2.048 \Rightarrow$ Rejeita-se H_0 e se $|t_0| \leq 2.048 \Rightarrow$ Não se rejeita H_0. Logo, como $t_0 = 10.091 > 2.048$, rejeita-se a hipótese de a altura das ondas não influenciar o montante dos estragos causados na orla costeira ao nível de significância de 5%.