DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Probabilidades e Estatística

LEAN, LEIC-A, LEIC-T, LEMat, LETI, MEBiol, MEBiom, MEEC, MEFT, MEMec, MEQ

1º semestre – 2015/2016
07/01/2016 – 11:00

Duração: 90 minutos **2º teste B**

Justifique convenientemente todas as respostas!

Grupo I	10 valores

1. Seja $\mathbf{X} = (X_1, \ldots, X_n)$, com $n > 2$, uma amostra aleatória de uma população com função de probabilidade

$$P(X = x) = (x-1)p^2(1-p)^{x-2},\ 0 < p \le 1 \text{ e } x \in \{2,3,\ldots\}$$

em que $E[X] = \frac{2}{p}$ e $Var[X] = \frac{2(1-p)}{p^2}$.

(a) Determine o estimador de máxima verosimilhança do parâmetro p. (3.0)

$\mathscr{L}(p, x_1, \ldots, x_n) \equiv f_{\mathbf{X}}(x_1, \ldots, x_n) \overset{iid}{=} \prod_{i=1}^n f_X(x_i) = \prod_{i=1}^n (x_i - 1)p^2(1-p)^{x_i-2} =$
$= p^{2n}(1-p)^{\sum_{i=1}^n x_i - 2n} \prod_{i=1}^n (x_i - 1)$
Para $p \in]0,1[$, $\log \mathscr{L}(p, x_1, \ldots, x_n) = 2n \log p + \left(\sum_{i=1}^n x_i - 2n\right) \log(1-p) + \sum_{i=1}^n \log(x_i - 1)$ (diferenciável em ordem a p)
$\frac{d\mathscr{L}}{dp} = 0 \iff \frac{2n}{p} - \frac{\sum_{i=1}^n x_i - 2n}{1-p} = 0 \iff p = \frac{2}{\bar{x}}$ e
$\frac{d^2\mathscr{L}}{dp^2} = -\frac{2n}{p^2} - \frac{\sum_{i=1}^n x_i - 2n}{(1-p)^2} < 0,\ \forall 0 < p < 1$ uma vez que $\sum_{i=1}^n x_i \ge 2n$
$\therefore \hat{p}_{MV} = \frac{2}{\bar{X}}$

(b) Tendo em vista a estimação do valor esperado de X, compare a eficiência do estimador $\bar{X}$ relativamente ao estimador $T = \frac{X_1 + X_n}{2}$. (1.5)

$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathrm{E}[X_i]}{n} = \mathrm{E}[X] = \frac{2}{p}$
$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathrm{Var}[X_i]}{n^2} = \frac{\mathrm{Var}[X]}{n} = \frac{2(1-p)}{np^2}$
$\mathrm{E}[T] = \mathrm{E}\left[\frac{X_1 + X_n}{2}\right] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_n]}{2} = \mathrm{E}[X] = \frac{2}{p}$
$\mathrm{Var}[T] = \mathrm{Var}\left[\frac{X_1 + X_n}{2}\right] = \frac{\mathrm{Var}[X_1] + \mathrm{Var}[X_n]}{4} = \frac{\mathrm{Var}[X]}{2} = \frac{(1-p)}{p^2}$
$$\frac{EQM[T]}{EQM[\bar{X}]} = \frac{\mathrm{Var}[T] + \left(\mathrm{E}[T] - 2/p\right)^2}{\mathrm{Var}[\bar{X}] + \left(\mathrm{E}[\bar{X}] - 2/p\right)^2} = \frac{n}{2} > 1, \text{ uma vez que } n > 2.$$
$\therefore \bar{X}$ é um estimador de $\mathrm{E}[X] = \frac{2}{p}$ mais eficiente que T.

2. Uma dada exploração leiteira tem por objetivo ter uma produção média diária de leite por vaca de 29 Kg. Sabe-se que a produção diária de leite por vaca é modelada por uma distribuição normal com variância igual a 2.25 Kg^2. Para analisar a produção média de leite por vaca nessa exploração, recolheu-se a produção diária de 12 vacas que conduziu a uma média por vaca de 28.0 Kg.

(a) Com base na amostra recolhida, obtenha um intervalo de confiança a 99% para o valor esperado da produção diária de leite por vaca. (3.0)

Sejam $Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ e $a = \Phi^{-1}(0.995) \approx 2.5758$.
$P(-a \le Z \le a) = 0.99 \iff P\left(\bar{X} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + a\frac{\sigma}{\sqrt{n}}\right) = 0.99$
$IAC_{0.99}(\mu) = \left[\bar{X} - a\frac{\sigma}{\sqrt{n}}, \bar{X} + a\frac{\sigma}{\sqrt{n}}\right] \implies IC_{0.99}(\mu) = [26.8847, 29.1154]$

(b) Teste ao nível de significância de 1% se o valor esperado da produção diária de leite por vaca é 29 Kg. O que pode concluir relativamente à satisfação do objetivo de a produção média diária de leite por vaca nessa exploração leiteira ser 29 Kg? (2.5)

Testar $H_0 : \mu = 29$ contra $H_1 : \mu \ne 29$ a um nível de significância α é equivalente a avaliar se $29 \in IC_{(1-\alpha)}(\mu)$. Uma vez que o valor 29 pertence ao intervalo de confiança determinado em (a) então não se deve rejeitar H_0 a um nível de significância de 0.01. Conclui-se assim, a esse nível de significância, que os dados não permitem afirmar que o objetivo pretendido não foi atingido.

Grupo II	10 valores

1. Num estudo realizado pela autoridade de segurança rodoviária foram registados para os dias úteis da semana os seguintes dados: (4.0)

Dia da semana	segunda	terça	quarta	quinta	sexta	TOTAL
Nº de acidentes	87	75	81	66	91	400

Teste a hipótese de os acidentes se distribuírem uniformemente pelos vários dias úteis da semana. Decida com base no valor-p.

Numerando os dias úteis da semana de 1 a 5, seja X ="dia da semana em que ocorre um acidente rodoviário". Pretende-se testar $H_0 : X \sim U(\{1,\ldots,5\})$ contra $H_1 : X \not\sim U(\{1,\ldots,5\})$.
Seja $p_i^0 = P(X = i \mid H_0) = 1/5$, $i = 1,\ldots,5$.

i	o_i	p_i^0	$e_i = np_i^0$
1	87	0.2	80
2	75	0.2	80
3	81	0.2	80
4	66	0.2	80
5	91	0.2	80
	$n = 400$		

Como todas as classes têm uma frequência esperada superior a 5 não é necessário agrupar classes ($k = 5$) e, não havendo qualquer parâmetro estimado ($\beta = 0$), a estatística de teste é $Q_0 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\overset{a}{\sim}} \chi^2_{(4)}$.
Tem-se $q_0 \approx 4.9$ e valor$-p = P(Q_0 > q_0 \mid H_0) = 1 - F_{\chi^2_{(4)}}(4.9) \approx 0.298$. Deve-se rejeitar H_0 para níveis de significância ≥ 0.298 e não rejeitar no caso contrário. Para os níveis de significância usuais os dados não fornecem evidência para a rejeição de H_0.

2. Para testar um instrumento que mede a concentração de ácido lático no sangue foram utilizadas 16 amostras sanguíneas para as quais se conhece essa concentração e registou-se o valor da concentração fornecido pelo instrumento. Seja x a concentração conhecida de ácido lático e Y a concentração de ácido lático medida pelo instrumento. Admitindo a validade do modelo de regressão linear simples, foram calculadas as seguintes estatísticas, incluindo estimativas de mínimos quadrados de parâmetros do modelo de regressão:

$$\sum_{i=1}^{16} x_i = 117,\ \sum_{i=1}^{16} x_i^2 = 1305,\ \sum_{i=1}^{16} y_i = 147.1,\ \sum_{i=1}^{16} y_i^2 = 2045.75,\ \hat{\beta}_1 = 1.22, \hat{\sigma} = 1.18$$

(a) Obtenha a estimativa pontual para o incremento no valor esperado fornecido pelo instrumento para a concentração de ácido lático no sangue provocada pelo aumento de 5 unidades da concentração conhecida. (1.0)

Pretende-se estimar $\delta = E[Y \mid x+5] - E[Y \mid x] = \beta_0 + \beta_1(x+5) - (\beta_0 + \beta_1 x) = 5\beta_1$. Como $\hat{\beta}_1 = 1.22$ tem-se $\hat{\delta} = 6.1$.

(b) Indicando hipóteses de trabalho convenientes, teste a significância do modelo de regressão linear ao nível de significância de 5%. (3.0)

Admitindo que $Y_i \mid x_i \overset{ind}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1,\ldots,16$, quer-se testar $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$.
Seja $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 16\bar{x}^2}}} \sim t_{(14)}$. Sob H_0, tem-se a estatística de teste $T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 16\bar{x}^2}}} \sim t_{(14)}$.
Para o nível de significância de 0.05 deve-se rejeitar H_0 se $|T_0| > a$ em que $a = F^{-1}_{t_{(18)}}(0.975) \approx 2.1448$.
Como $t_0 \approx 21.92$ rejeita-se H_0 para o nível de significância de 0.05, ou seja, os dados não permitem afirmar que x não influencia Y sob o modelo de regressão adotado.

(c) Tirando partido da relação entre o coeficiente de determinação e a estimativa de mínimos quadrados de β_1, calcule o coeficiente de determinação e interprete o valor obtido. (2.0)

$$R^2 = \frac{\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \times \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)} = \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2} \approx 0.965.$$

Conclui-se que 96.5% da variabilidade observada nas medições efetuadas é explicada pelo MRLS o que evidencia o bom ajustamento desse modelo aos dados.