Probabilidades e Estatística

LEE, LEGI, LEGM, LMAC, MEAmbi, MEAer, MEC

1º semestre – 2015/2016
07/01/2016 – 09:00

Duração: 90 minutos **2º teste A**

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

$$P(X = x) = p(1-p)^x,\ 0 < p < 1 \text{ e } x \in \mathbb{N}_0,$$

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

(a) Determine a estimativa de máxima verosimilhança do parâmetro p sabendo que a realização $(x_1, \ldots, x_n)$ da amostra aleatória referida conduziu a um valor positivo para $\sum_{i=1}^n x_i$. (3.0)

$\mathscr{L}(p, x_1, \ldots, x_n) \equiv f_{\mathbf{X}}(x_1, \ldots, x_n) \overset{iid}{=} \prod_{i=1}^n f_X(x_i) = \prod_{i=1}^n p(1-p)^{x_i} = p^n(1-p)^{\sum_{i=1}^n x_i}$

Para $p \in]0,1[$, $\log \mathscr{L}(p, x_1, \ldots, x_n) = n \log p + \left(\sum_{i=1}^n x_i\right) \log(1-p)$ (diferenciável em ordem a p)

Para $\sum_{i=1}^n x_i > 0$, $\frac{d\mathscr{L}}{dp} = 0 \iff \frac{n}{p} - \frac{\sum_{i=1}^n x_i}{1-p} = 0 \iff p = \frac{1}{1+\bar{x}}$ e

$\frac{d^2\mathscr{L}}{dp^2} = -\frac{n}{p^2} - \frac{\sum_{i=1}^n x_i}{(1-p)^2} < 0, \forall 0 < p < 1$ uma vez que $\sum_{i=1}^n x_i > 0$

$\therefore \hat{p}_{MV} = \frac{1}{1+\bar{x}}$

(b) Tendo em vista a estimação do valor esperado de X, compare a eficiência do estimador $\bar{X}$ relativamente ao estimador $T = \frac{X_1 + (n-1)X_n}{n}$. (1.5)

$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathrm{E}[X_i]}{n} = \mathrm{E}[X] = \frac{1}{p} - 1$

$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n \mathrm{Var}[X_i]}{n^2} = \frac{\mathrm{Var}[X]}{n} = \frac{(1-p)}{np^2}$

$\mathrm{E}[T] = \mathrm{E}\left[\frac{X_1 + (n-1)X_n}{n}\right] = \frac{\mathrm{E}[X_1] + (n-1)\,\mathrm{E}[X_n]}{n} = \mathrm{E}[X] = \frac{1}{p} - 1$

$\mathrm{Var}[T] = \mathrm{Var}\left[\frac{X_1 + (n-1)X_n}{n}\right] = \frac{\mathrm{Var}[X_1] + (n-1)^2\,\mathrm{Var}[X_n]}{n^2} = \frac{1 + (n-1)^2}{n^2}\mathrm{Var}[X] = \frac{1 + (n-1)^2}{n^2}\frac{(1-p)}{p^2}$

$$\frac{EQM[T]}{EQM[\bar{X}]} = \frac{\mathrm{Var}[T] + \left(\mathrm{E}[T] - (1/p - 1)\right)^2}{\mathrm{Var}[\bar{X}] + \left(\mathrm{E}[\bar{X}] - (1/p - 1)\right)^2} = \frac{1 + (n-1)^2}{n} = n - 2 + \frac{2}{n} > 1, \text{ uma vez que } n > 2.$$

$\therefore \bar{X}$ é um estimador de $\mathrm{E}[X] = \frac{1}{p} - 1$ mais eficiente que T.

2. Num certo processo químico é muito importante que uma dada solução tenha um pH de exatamente 8.20. O método utilizado na determinação do pH fornece medições que se admite terem distribuição normal de valor esperado igual ao valor do pH da solução e desvio padrão de 0.1. Para avaliar o pH de uma solução, efetuaram-se 10 medições de modo independente, tendo-se obtido uma soma das observações igual a 81.79.

(a) Com base na amostra recolhida, obtenha um intervalo de confiança para o valor do pH da solução com nível de confiança de 90%. (3.0)

Sejam $Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ e $a = \Phi^{-1}(0.95) \approx 1.6449$.

$P(-a \le Z \le a) = 0.90 \iff P\left(\bar{X} - a\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + a\frac{\sigma}{\sqrt{n}}\right) = 0.90$

$IAC_{0.90}(\mu) = \left[\bar{X} - a\frac{\sigma}{\sqrt{n}}, \bar{X} + a\frac{\sigma}{\sqrt{n}}\right] \implies IC_{0.90}(\mu) = [8.1270, 8.2310]$

(b) Teste ao nível de significância de 10% se o valor esperado do método de determinação do PH é 8.20. O que pode concluir relativamente à utilização desta solução no referido processo químico? (2.5)

Testar $H_0 : \mu = 8.20$ contra $H_1 : \mu \neq 8.20$ a um nível de significância α é equivalente a avaliar se $8.20 \in IC_{(1-\alpha)}(\mu)$. Uma vez que o valor 8.20 pertence ao intervalo de confiança determinado em (a) então não se deve rejeitar H_0 a um nível de significância de 0.10. Conclui-se assim, a esse nível de significância, que os dados não permitem afirmar que a solução em causa não deve ser usada no processo químico referido.

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	86	74	80	65	90	395

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	86	0.2	79
2	74	0.2	79
3	80	0.2	79
4	65	0.2	79
5	90	0.2	79
	$n = 395$		

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 $q_0 \approx 4.962$ e valor$-p = P(Q_0 > q_0 \mid H_0) = 1 - F_{\chi^2_{(4)}}(4.962) \approx 0.29$. Deve-se rejeitar H_0 para níveis de significância ≥ 0.29 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 20 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:

$$\sum_{i=1}^{20} x_i = 134,\ \sum_{i=1}^{20} x_i^2 = 1424,\ \sum_{i=1}^{20} y_i = 167.7,\ \sum_{i=1}^{20} y_i^2 = 2220.21,\ \hat{\beta}_1 = 1.23, \hat{\sigma} = 1.08$$

(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 3 unidades da concentração conhecida. (1.0)

Pretende-se estimar $\delta = E[Y \mid x+3] - E[Y \mid x] = \beta_0 + \beta_1(x+3) - (\beta_0 + \beta_1 x) = 3\beta_1$. Como $\hat{\beta}_1 = 1.23$ tem-se $\hat{\delta} = 3.69$.

(b) Indicando as hipóteses de trabalho convenientes, teste a significância do modelo de regressão linear simples ao nível de significância de 1%. (3.0)

Admitindo que $Y_i \mid x_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$, $i = 1,\ldots,20$, quer-se testar $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$.
Seja $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 20\bar{x}^2}}} \sim t_{(18)}$. Sob H_0, tem-se a estatística de teste $T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 20\bar{x}^2}}} \sim t_{(18)}$.
Para o nível de significância de 0.01 deve-se rejeitar H_0 se $|T_0| > a$ em que $a = F^{-1}_{t_{(18)}}(0.995) \approx 2.8784$.
Como $t_0 \approx 26.12$ rejeita-se H_0 para o nível de significância de 0.01, 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)

$$R^2 = \frac{\left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right)^2}{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \times \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)} = \hat{\beta}_1^2 \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n y_i^2 - n\bar{y}^2} \approx 0.978.$$

Conclui-se que 97.8% da variabilidade observada nas medições efetuadas é explicada pelo MRLS o que evidencia o bom ajustamento desse modelo aos dados.