Probabilidades e Estatística

TODOS OS CURSOS

2º semestre – 2020/2021
09/07/2021 – **15:00**

Duração: 60+15 minutos **Teste 2C**

Justifique convenientemente todas as respostas

Pergunta 1 4 valores

Considere a variável aleatória X, que representa o número de nascimentos por hora em determinado hospital e que se admite ter uma distribuição de Poisson de parâmetro $\lambda > 0$ (desconhecido).

Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de dimensão n proveniente de X e denote-se por $\bar{X}$ a média da amostra aleatória. Sabe-se que $T = \bar{X}(1+\bar{X})$ é um estimador enviesado de $E(X^2)$.

Calcule o valor do enviesamento de T na estimação de $E(X^2)$, quando $n = 19$ e $\lambda = 1.4$.

- **V.a. de interesse**

  X = número de nascimentos por hora durante um dia em determinado hospital

  $X \sim \text{Poisson}(\lambda)$

  $E(X) = V(X) = \lambda$ DESCONHECIDO

  Outro parâmetro desconhecido

  $E(X^2) = V(X) + E^2(X) = \lambda + \lambda^2 = \lambda(1+\lambda), \quad \lambda > 0$

- **Estimador de $E(X^2)$**

  $T = \bar{X}(1+\bar{X})$

- **Valor esperado de T**

  Ao notarmos que $E(\bar{X}) = E(X) = \lambda$, $V(\bar{X}) = \frac{V(X)}{n} = \frac{\lambda}{n}$, segue-se, para qualquer $\lambda > 0$,

$$\begin{aligned} E(T) &= E[\bar{X}(1+\bar{X})] \\ &= E(\bar{X}) + E(\bar{X}^2) \\ &= E(\bar{X}) + [V(\bar{X}) + E^2(\bar{X})] \\ &= \lambda + \left(\frac{\lambda}{n} + \lambda^2\right) \\ &= \lambda(1+\lambda) + \frac{\lambda}{n}. \end{aligned}$$

  [Uma vez que $E(T) \neq \lambda(1+\lambda)$, $\forall \lambda > 0$, T é, efectivamente, um estimador enviesado de λ.]

- **Enviesamento de T**

$$\begin{aligned} E(T) - E(X^2) &= \left[\lambda(1+\lambda) + \frac{\lambda}{n}\right] - \lambda(1+\lambda) \\ &= \frac{\lambda}{n} \\ &= \frac{1.4}{19} \\ &\simeq 0.073684. \end{aligned}$$

Pergunta 2 4 valores

Admita que X_1 (resp. X_2) representa a idade de um indivíduo que teve pelo menos um evento coronário em 2020 (resp. em 2019). Para estimar a diferença de idades esperadas entre doentes coronários em 2020 e 2019,

$\mu_1 - \mu_2$, foram recolhidas duas amostras independentes com dimensões $n_1 = 7$ e $n_2 = 14$ (respetivamente), tendo-se obtido os seguintes resultados: $\bar{x}_1 = 51$ e $s_1 = 0.8$; $\bar{x}_2 = 49$ e $s_2 = 0.6$.

Suponha que X_1 e X_2 são variáveis aleatórias normalmente distribuídas com variâncias desconhecidas mas iguais. Obtenha um intervalo de confiança a 95% para $\mu_1 - \mu_2$.

- **V.a. de interesse**

  X_1 = idade de um indivíduo que teve pelo menos um evento coronário em 2020

  X_2 = idade de um indivíduo que teve pelo menos um evento coronário em 2019

- **Situação**

  X_1 e X_1 v.a. independentes com distribuições normais

  $(\mu_1 - \mu_2)$ DESCONHECIDO

  σ_1^2 e σ_2^2 desconhecidas mas iguais

- **Obtenção do IC para $\mu_1 - \mu_2$**

  Passo 1 — Seleção da v.a. fulcral para $(\mu_1 - \mu_2)$

$$Z \quad = \quad \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

  Passo 2 — Obtenção dos quantis de probabilidade

  Uma vez que $(1-\alpha) \times 100\% = 95\%$ e $n_1 = 7$ e $n_2 = 14$, temos $\alpha = 0.05$ e

$$\begin{cases} a_\alpha = F^{-1}_{t_{(n_1+n_2-2)}}(\frac{\alpha}{2}) = -F^{-1}_{t_{(19)}}(0.975) \stackrel{tabelas, calc.}{=} -2.093 \\ b_\alpha = F^{-1}_{t_{(n_1+n_2-2)}}(1-\frac{\alpha}{2}) = F^{-1}_{t_{(19)}}(0.975) \stackrel{tabelas, calc.}{=} 2.093 \end{cases}$$

  Passo 3 — Inversão da desigualdade $a_\alpha \le Z \le b_\alpha$

$$P(a_\alpha \le Z \le b_\alpha) = 1-\alpha$$

$$P(a_\alpha \le \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \le b_\alpha) = 1-\alpha$$

$$P\left[(\bar{X}_1 - \bar{X}_2) - F^{-1}_{t_{(n_1+n_2-2)}}(1-\alpha/2) \times \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \le \mu_1 - \mu_2 \right.$$
$$\left. \le (\bar{X}_1 - \bar{X}_2) + F^{-1}_{t_{(n_1+n_2-2)}}(1-\alpha/2) \times \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right] = 1-\alpha$$

  Passo 4 — Concretização

  Tendo em conta os quantis acima, as dimensões amostrais n_1 e n_2, bem como as concretizações de $\bar{X}_1$, $\bar{X}_2$, S_1^2 e S_2^2 e a expressão geral do IC,

$$IC_{(1-\alpha)\times 100\%}(\mu_1 - \mu_2) \quad = \quad \left[(\bar{x}_1 - \bar{x}_2) \pm F^{-1}_{t_{(n_1+n_2-2)}}(1-\alpha/2) \times \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right],$$

  segue-se

$$
\begin{aligned}
IC_{95\%}(\mu_1-\mu_2) &= \left[(51-49) \pm 2.093 \times \sqrt{\frac{(7-1)\times 0.8^2+(14-1)\times 0.6^2}{7+14-2} \times \left(\frac{1}{7}+\frac{1}{14}\right)}\right] \\
&= [2 \pm 2.093 \times 0.309984] \\
&= [1.3512,\ 2.6488].
\end{aligned}
$$

Pergunta 3	4 valores

Uma engenheira biomédica conjetura que o número esperado de mutações em determinadas regiões de um cromossoma é igual a $\lambda_0 = 101$. Considere que o número X dessas mutações segue uma distribuição de Poisson.

Supondo que entre $n = 78$ cromossomas observados se identificaram 8255 mutações, teste a conjetura da engenheira biomédica contra a alternativa $H_1 : E(X) \neq \lambda_0$. Decida com base no valor-p aproximado.

- **V.a. de interesse**

  X = número de mutações em determinadas regiões de um cromossoma

- **Situação**

  $X \sim \text{Poisson}(\lambda)$

  λ DESCONHECIDO

- **Hipóteses**

  $H_0 : \lambda = \lambda_0 = 101$

  $H_1 : \lambda \neq \lambda_0$

- **Estatística de teste**

$$
T = \frac{\bar{X}-\lambda_0}{\sqrt{\frac{\lambda_0}{n}}} \overset{a}{\sim}_{H_0} \text{normal}(0,1)
$$

- **Região de rejeição de H_0**

  Teste bilateral ($H_1 : \lambda \neq \lambda_0$), logo a região de rejeição de H_0 é do tipo $W = (-\infty, -c) \cup (c, +\infty)$.

- **Decisão (com base no valor-p)**

  O valor observado da estatística de teste é igual a

$$
\begin{aligned}
t &= \frac{\frac{8255}{78}-101}{\sqrt{\frac{101}{78}}} \\
&\simeq 4.2475
\end{aligned}
$$

  e

$$
\begin{aligned}
valor-p &= 2 \times P(|T| > |t| \mid H_0) \\
&\simeq 2 \times [1-\Phi(|t|)] \\
&\simeq 2 \times [1-\Phi(4.2475)] \\
&\simeq 2 \times [1-\Phi(4.25)],
\end{aligned}
$$

onde $\Phi(4.25) > \Phi(4.09) \overset{tabelas, calc}{=} 0.999978$. Consequentemente,

$$\begin{aligned} valor-p &\simeq 2 \times [1-\Phi(4.25)] \\ &< 2 \times (1-0.999978) \\ &= 0.000044, \end{aligned}$$

pelo que devemos rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.0044\%$, designadamente a qualquer dos n.u.s. $(1\%, 5\%, 10\%)$.

[Altervativamente e usando uma calculadora, conclui-se que $valor-p \simeq 2 \times [1-\Phi(4.2475)] \simeq 0.0000216169$, pelo que:

- não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 0.00216169\%$;
- devemos rejeitar H_0 a qualquer n.s. $\alpha_0 > 0.00216169\%$, nomeadamente a qualquer dos n.u.s. $(1\%, 5\%, 10\%)$.]

Pergunta 4 4 valores

Seja X a massa (em gramas) de um carapau médio. Uma bióloga defende a hipótese H_0 de esta variável aleatória ter distribuição normal com valor esperado e desvio padrão iguais a 100 e 31 gramas (respetivamente). Numa amostra casual de 200 desses peixes, foram registadas as seguintes frequências por intervalos de massa:

Intervalo de massa	≤ 60	$]60, 80]$	$]80, 120]$	$]120, 140]$	> 140
Frequência absoluta observada	20	26	96	33	25
Frequência absoluta esperada sob H_0	E_1	32.2	96.2	32.2	E_5

Calcule as frequências absolutas esperadas sob H_0 omissas E_1 e E_5 (aproximando-as às décimas). Serão os dados consistentes com H_0? Decida com base no valor-p aproximado.

- **V.a. de interesse**

  X = massa (em gramas) de um carapau médio

- **Hipóteses**

  $H_0 : X \sim \text{normal}(\mu = 100, \sigma^2 = 31^2)$

  $H_1 : X \not\sim \text{normal}(\mu = 100, \sigma^2 = 31^2)$

- **Estatística de teste**

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

  onde:

  - k = no. de classes = 5;
  - O_i = freq. abs. observável da classe i;
  - E_i = freq. abs. esperada sob H_0 da classe i;
  - $\beta = 0$.

- **Frequência esperadas sob H_0 omissas**

$$\begin{aligned} E_1 &= n \times P(X \le 60 \mid H_0) \\ &= 200 \times \Phi\left(\frac{60-100}{31}\right) \\ &\simeq 200 \times \Phi(-1.29) \\ &\overset{tabelas, calc.}{=} 200 \times (1-0.9015) \\ &= 19.7 \\ E_5 &= n - \sum_{i=1}^{4} E_i \\ &\simeq 200 - (19.7 + 32.2 + 96.2 + 32.2) \\ &= 19.7. \end{aligned}$$

- **Região de rejeição de H_0** (para valores de T)

  Tratando-se de um teste de ajustamento do qui-quadrado, a região de rejeição de H_0 escrita para valores observados de T é o intervalo à direita $W = (c, +\infty)$.

- **Decisão (com base no valor-p)**

i	Classe i	Freq. abs. obs. o_i	Freq. abs. esp. sob H_0 E_i	Parcelas valor obs. estat. teste $\frac{(o_i-e_i)^2}{e_i}$
1	≤ 60	20	19.7	$\frac{(20-19.7)^2}{19.7} = 0.0046$
2	$]60, 80]$	26	32.2	1.1938
3	$]80, 120]$	96	96.2	0.0004
4	$]120, 140]$	33	32.2	0.0199
5	> 140	25	19.7	1.4259
		$\sum_{i=1}^{k} o_i = n$ $= 200$	$\sum_{i=1}^{k} e_i = n$ $= 200$	$t = \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{e_i}$ $\simeq 2.6445$

Dado que o valor observado da estatística de teste é $t = \sum_{i=1}^{k} \frac{(o_i-E_i)^2}{E_i} \simeq 2.6445$ e $W = (c, +\infty)$, obtemos

$$\begin{aligned} valor-p &= P(T > t \mid H_0) \\ &\simeq 1 - F_{\chi^2_{(k-1)}}(t) \\ &= 1 - F_{\chi^2_{(4)}}(2.6445) \\ &\simeq 0.618954 \end{aligned}$$

e devemos

- não rejeitar de H_0 a qualquer n.s. $\alpha_0 \le valor - p = 61.8954\%$, designadamente aos n.u.s. $(1\%, 5\%, 10\%)$;
- rejeitar de H_0 a qualquer n.s. $\alpha_0 > valor - p = 61.8954\%$.

Alternativamente e recorrendo às tabelas de quantis da distribuição do qui-quadrado, podemos obter um intervalo para o valor-p deste teste:

$$\begin{aligned} F^{-1}_{\chi^2_{(4)}}(0.30) = 2.195 &< t = 2.6445 < 2.753 = F^{-1}_{\chi^2_{(4)}}(0.40) \\ 0.30 &< F_{\chi^2_{(4)}}(2.6445) < 0.40 \\ 0.60 = 1 - 0.40 &< valor - p \simeq 1 - F_{\chi^2_{(4)}}(2.6445) < 1 - 0.30 = 0.70. \end{aligned}$$

Logo:

– não devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \leq 60\%$, nomeadamente aos n.u.s. $(1\%, 5\%, 10\%)$;
– devemos rejeitar H_0 a qualquer n.s. $\alpha_0 \geq 70\%$.

Pergunta 5	4 valores

Os dados relativos à produção de trigo (x, em toneladas) e ao preço do quilo de farinha de trigo (Y, em cêntimos de euro) em 7 anos consecutivos conduziram a

$$\sum_{i=1}^{7} x_i = 34, \quad \sum_{i=1}^{7} x_i^2 = 190, \quad \sum_{i=1}^{7} y_i = 440, \quad \sum_{i=1}^{7} y_i^2 = 30\,150, \quad \sum_{i=1}^{7} x_i\, y_i = 2\,365.$$

Admita que as variáveis x e Y estão relacionadas de acordo com o modelo de regressão linear simples: $Y = \beta_0 + \beta_1\, x + \varepsilon$.

Após ter enunciado as hipóteses de trabalho que entender convenientes, obtenha um intervalo de confiança a 95% para β_1 e a amplitude deste intervalo de confiança, tirando partido dos resultados acima.

- **Modelo de RLS**

  Y = preço do quilo de farinha (v.a. resposta)

  x = produção de trigo (variável explicativa)

  $Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i, \quad i = 1, \ldots, n$

- **Hipóteses de trabalho**

  $\varepsilon_i \stackrel{i.i.d.}{\sim} \text{normal}(0, \sigma^2), \quad i = 1, \ldots, n$

- **Estimativas de MV de β_0 e β_1; estimativa de σ^2**

  Importa notar que

  - $n = 7$
  - $\sum_{i=1}^{n} x_i = 34$

    $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{34}{7} \simeq 4.857143$

    $\sum_{i=1}^{n} x_i^2 = 190$

    $\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2 \simeq 190 - 7 \times 4.857143^2 \simeq 24.857133$
  - $\sum_{i=1}^{n} y_i = 440$

    $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{440}{7} \simeq 62.857143$

    $\sum_{i=1}^{n} y_i^2 = 30\,150$

    $\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2 = 30\,150 - 7 \times 62.857143^2 \simeq 2492.857143$
  - $\sum_{i=1}^{n} x_i\, y_i = 2\,365$

    $\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y} \simeq 2\,365 - 7 \times 4.857143 \times 62.857143 \simeq 227.857075$

  Logo,

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \\ &\simeq \frac{227.857075}{24.857133} \\ &\simeq 9.166668 \end{aligned}$$

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&\simeq 62.857143 - 9.166668 \times 4.857143 \\
&\simeq 18.333326 \\
\hat{\sigma}^2 &= \frac{1}{n-2}\left[\left(\sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2\right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)\right] \\
&\simeq \frac{1}{7-2}(2492.857017 - 9.166668^2 \times 24.857133) \\
&\simeq 80.833352.
\end{aligned}
$$

- **Obtenção do IC para β_1**

Passo 1 — Seleção da v.a. fulcral

$$
Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \sim t_{(n-2)}
$$

Passo 2 — Obtenção dos quantis de probabilidade

$$
\begin{cases}
a_\alpha = F_{t_{(n-2)}}(\alpha/2) = -F_{t_{(7-2)}}(1-0.05/2) = -F_{t_{(5)}}(0.975) \stackrel{tabelas,calc.}{=} -2.571 \\
b_\alpha = F_{t_{(n-2)}}(1-\alpha/2) = F_{t_{(7-2)}}(1-0.05/2) = F_{t_{(5)}}(0.975) \stackrel{tabelas,calc.}{=} 2.571
\end{cases}
$$

Passo 3 — Inversão da desigualdade $a_\alpha \leq T \leq b_\alpha$

$$
P(a_\alpha \leq Z \leq b_\alpha) = 1-\alpha
$$

$$
P\left[a_\alpha \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}} \leq b_\alpha\right] = 1-\alpha
$$

$$
P\left[\hat{\beta}_1 - b_\alpha \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}} \leq \beta_1 \leq \hat{\beta}_1 - a_\alpha \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}\right] = 1-\alpha
$$

Passo 4 — Concretização

Tendo em conta a expressão geral do IC para β_1,

$$
IC_{(1-\alpha)\times 100\%}(\beta_1) = \left[\hat{\beta}_1 \pm F^{-1}_{t_{(n-2)}}(1-\alpha/2) \times \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}}\right],
$$

e os resultados anteriores, o IC pretendido e a sua amplitude são iguais a

$$
\begin{aligned}
IC_{95\%}(\beta_1) &\simeq \left[9.166668 \pm 2.571 \times \sqrt{\frac{80.833352}{24.857133}}\right] \\
&\simeq [9.166668 \pm 2.571 \times 1.803307] \\
&\simeq [4.530365, 13.802971] \\
13.802971 - 4.530365 &\simeq 9.27261.
\end{aligned}
$$