(Introdução às) Probabilidades e Estatística

— TODOS OS CURSOS —

2019/2020
08/09/2020 – **09:00**

Duração: 120 minutos **Exame de Época Especial**

Pergunta 1

Considere os acontecimentos A, B e C tais que: A e B são independentes condicionalmente a C; $P(A \mid C) = a$, $P(A \cap B) = b$ e $P(B \cap C) = c$. Obtenha $P[C \mid (A \cap B)]$.

Indique o resultado com pelo menos quatro casas decimais.

- **Prob. pedida**

  Uma vez que $P(A \mid C) = a$, $P(A \cap B) = b$, $P(B \cap C) = c$, e A e B são independentes condicionalmente a C, i.e.,

  $$P[(A \cap B) \mid C] = P(A \mid C) \times P(B \mid C),$$

  temos

  $$\begin{aligned} P[C \mid (A \cap B)] &= \frac{P(A \cap B \cap C)}{P(A \cap B)} \\ &= \frac{P[(A \cap B) \mid C] \times P(C)}{P(A \cap B)} \\ &= \frac{P(A \mid C) \times [P(B \mid C) \times P(C)]}{P(A \cap B)} \\ &= \frac{P(A \mid C) \times P(B \cap C)}{P(A \cap B)} \\ &= \frac{a\,c}{b}. \end{aligned}$$

Pergunta 2

Durante uma epidemia admite-se vir a aplicar um teste de diagnóstico da doença a pessoas sorteadas ao acaso numa grande população. Admitindo que c% da população está infetada e que há independência entre os estados de diferentes pessoas selecionadas, calcule a probabilidade de virem a ser necessárias mais de a e não mais de b aplicações do teste para que seja detetada a primeira pessoa infetada.

Indique o resultado com pelo menos quatro casas decimais.

- **V.a. de interesse**

  $X =$ número de aplicações do teste de diagnóstico até ser detetada a primeira pessoa infetada

- **Distribuição e f.p. de X**

  $X \sim$ **geométrica**(p), **onde** $p = \frac{c}{100}$

  $P(X = x) = (1-p)^{x-1}\,p, \quad x \in \mathbb{N}$

- **Prob. pedida**

  $$\begin{aligned} P(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \sum_{x=1}^{b} (1-p)^{x-1}\,p - \sum_{x=1}^{a} (1-p)^{x-1}\,p \\ &= p\,\frac{1-(1-p)^b}{1-(1-p)} - p\,\frac{1-(1-p)^a}{1-(1-p)} \\ &= (1-p)^a - (1-p)^b \end{aligned}$$

Pergunta 3

O tempo, em horas, que um algoritmo de força bruta leva para obter acesso a uma conta de utilizador de um sistema informático é uma variável aleatória X com distribuição exponencial tal que $P(X>1) = e^{-a/7}$. No caso de esse tempo exceder $c+b$ horas, o administrador do sistema é notificado.

Admitindo que o algoritmo está a ser executado há (mais de) c horas, calcule a probabilidade de o administrador do sistema não ser notificado.

Indique o resultado com pelo menos quatro casas decimais.

- **V.a. de interesse**

  X = tempo que algoritmo leva para obter acesso a conta de um utilizador de um sistema informático

- **Distribuição, f.d.p. e f.d. de** X

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

  $f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x>0 \\ 0, & \text{caso contrário} \end{cases}$

  $F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 1-e^{-\lambda x}, & x>0 \\ 0, & \text{caso contrário} \end{cases}$

  onde $\quad \lambda : P(X>1) = 1-F_X(1) = e^{-\lambda} = e^{-a/7} \Rightarrow \lambda = a/7.$

- **Prob. pedida**

  Ao invocarmos a propriedade da falta de memória da distribuição exponencial, obtemos

$$\begin{aligned} P(X \le c+b \mid X>c) &= 1-P(X>c+b \mid X>c) \\ &= 1-P(X>b) \\ &= F_X(b) \\ &= 1-e^{-ab/7} \end{aligned}$$

Pergunta 4

Seja (X,Y) um par aleatório com função de densidade de probabilidade conjunta dada por

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{12}{ab^2(3a+4b)}\,y(x+y), & 0<x<a,\ 0<y<b \\ 0, & \text{caso contrário.} \end{cases}$$

Determine $E[X \mid Y = cb]$.

Indique o resultado com pelo menos quatro casas decimais.

- **Par aleatório e sua f.d.p. conjunta**

  (X,Y)

  $f_{X,Y}(x,y) = \begin{cases} \dfrac{12\,y(x+y)}{ab^2(3a+4b)}, & 0<x<a,\ 0<y<b \\ 0, & \text{caso contrário} \end{cases}$

- **F.d.p. marginal de** Y

  Para $0<y<b$,

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx \\
&= \int_0^{a} \frac{12\,y\,(x+y)}{ab^2\,(3a+4b)}\,dx \\
&= \frac{12\,y}{ab^2\,(3a+4b)} \times \int_0^{a} (x+y)\,dx \\
&= \frac{12\,y}{ab^2\,(3a+4b)} \times \left(\frac{x^2}{2}+x\,y\right)\Bigg|_0^{a} \\
&= \frac{12\,y}{ab^2\,(3a+4b)} \times \frac{a\,(a+2y)}{2} \\
&= \frac{6\,y(a+2y)}{b^2(3a+4b)}.
\end{aligned}
$$

- **F.d.p. de X condicional a $Y = c\,b$**

  Para $0 < x < a$,

$$
\begin{aligned}
f_{X|Y=c\,b}(x) &= \frac{f_{X,Y}(x,c\,b)}{f_Y(c\,b)} \\
&= \frac{\dfrac{12\,c\,b\,(x+c\,b)}{ab^2\,(3a+4b)}}{\dfrac{6c\,b(a+2c\,b)}{b^2(3a+4b)}} \\
&= \frac{2(x+c\,b)}{a\,(a+2c\,b)}.
\end{aligned}
$$

- **Valor esperado condicional pedido**

$$
\begin{aligned}
E[X \mid Y = c\,b] &= \int_{-\infty}^{+\infty} x\,f_{X|Y=c\,b}(x)\,dx \\
&= \int_0^{a} x\,\frac{2(x+c\,b)}{a\,(a+2c\,b)}\,dx \\
&= \frac{2}{a\,(a+2c\,b)} \int_0^{a} x\,(x+c\,b)\,dx \\
&= \frac{2}{a\,(a+2c\,b)} \left(\frac{x^3}{3}+\frac{c\,b\,x^2}{2}\right)\Bigg|_0^{a} \\
&= \frac{2}{a\,(a+2c\,b)}\,\frac{2\,a^3+3\,a^2\,c\,b}{6} \\
&= \frac{a(2a+3bc)}{3(a+2bc)}
\end{aligned}
$$

Pergunta 5

Um distribuidor de produtos alimentares frescos recebe diariamente pedidos de a clientes. Admita que as quantidades encomendadas pelos clientes, em kg, são variáveis aleatórias contínuas independentes e uniformemente distribuídas no intervalo $(0, b)$.

Considerando que o distribuidor só tem capacidade para expedir diariamente d kg de produtos, calcule a probabilidade de todos os pedidos serem satisfeitos num dia escolhido ao acaso.

Indique o resultado com pelo menos quatro casas decimais.

- **V.a.; distribuição, valor esperado e variância comuns**

  X_i = quantidade encomendada pelo cliente i, em kg, $\quad i = 1,\ldots,a$

$X_i \overset{i.i.d.}{\sim} X \sim \text{uniforme}(0, b), \quad i = 1, \ldots, a$

$\mu = E(X_i) = E(X) \overset{form.}{=} \frac{b}{2}$

$\sigma^2 = V(X_i) = V(X) = \frac{b^2}{12}$

- **V.a. de interesse**

  $S_a = \sum_{i=1}^{a} X_i$ = quantidade total encomendada pelos a clientes

  $E(S_a) = \cdots = a \times \mu$

  $V(S_a) = \cdots = a \times \sigma^2$

- **Distribuição aproximada de S_a**

  $\frac{S_a - E(S_a)}{\sqrt{V(S_a)}} = \frac{S_a - \frac{ab}{2}}{b\sqrt{\frac{a}{12}}} \overset{a}{\sim} \text{normal}(0,1)$

- **Prob. pedida (valor aproximado)**

$$\begin{aligned} P(S_a \le d) &= P\left(\frac{S_a - \frac{ab}{2}}{b\sqrt{\frac{a}{12}}} \le \frac{d - \frac{ab}{2}}{b\sqrt{\frac{a}{12}}}\right) \\ &\overset{TLC}{\simeq} \Phi\left(\frac{d - \frac{ab}{2}}{b\sqrt{\frac{a}{12}}}\right) \end{aligned}$$

Pergunta 6

Considere uma população X que se reporta ao tempo de vida de um micro-organismo. Admita que X é uma variável aleatória contínua uniformemente distribuída no intervalo $[0, \theta]$, onde θ é uma constante positiva desconhecida.

Seja (X_1, X_2, X_3, X_4) uma amostra aleatória de dimensão 4 extraída da população X. Considere os estimadores $T = \frac{a X_1 + b X_2 + c X_3 + d X_4}{a+b+c+d}$ e a média da amostra aleatória, $\bar{X}$, na estimação do parâmetro $E(X)$.

Obtenha a razão entre os erros quadráticos médios de T e de $\bar{X}$.

Preencha a caixa abaixo com o resultado obtido com, pelo menos, quatro casas decimais.

- **V.a. de interesse, distribuição, valor esperado e variância**

  X = tempo de vida de um micro-organismo

  $X \sim \text{uniforme}(0, \theta)$

  $E(X) = \theta/2 \equiv \mu$

  $Var(X) = \theta^2/12$

- **Estimadores de $\mu = E(X)$**

  $T = \frac{a X_1 + b X_2 + c X_3 + d X_4}{a+b+c+d}$

  $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$

- **Erros quadráticos médios**

$$
\begin{aligned}
EQM_\mu(T) &= Var(T) + [E(T)-\mu]^2 \\
&\stackrel{X_i \stackrel{i.i.d.}{\sim} X}{=} \frac{(a^2+b^2+c^2+d^2)}{(a+b+c+d)^2}\, Var(X) + (\mu-\mu)^2 \\
&= \frac{(a^2+b^2+c^2+d^2)}{(a+b+c+d)^2}\, \frac{\theta^2}{12} \\
EQM_\mu(\bar{X}) &= Var(\bar{X}) + [E(\bar{X})-\mu]^2 \\
&\stackrel{X_i \stackrel{i.i.d.}{\sim} X}{=} \frac{Var(X)}{n} + (\mu-\mu)^2 \\
&= \frac{\frac{\theta^2}{12}}{4} \\
&= \frac{\theta^2}{48}
\end{aligned}
$$

- **Razão pedida**

$$
\frac{EQM_\mu(T)}{EQM_\mu(\bar{X})} = \frac{4\,(a^2+b^2+c^2+d^2)}{(a+b+c+d)^2}
$$

Pergunta 7

Um serviço de ambulâncias alega que a maioria das chamadas recebidas envolvem emergências devido ao covid-19. Uma amostra casual de n chamadas foi selecionada dos arquivos do serviço, tendo-se verificado que a dessas chamadas envolvem emergências devido ao covid-19.

Obtenha um intervalo de confiança aproximado a $100\,(1-\alpha)\,\%$ para a verdadeira proporção p de chamadas de emergências devido ao covid-19.

- **V.a. de interesse**

  $$X = \begin{cases} 1, & \text{se chamada de emergência devido ao covid-19,} \\ 0, & \text{no caso contrário.} \end{cases}$$

- **Situação**

  $X \sim \text{Bernoulli}(p = P(X=1))$, $\quad n$ suficientemente grande

- **Obtenção do IC aproximado para p**

  Passo 1 — Selecção da v.a. fulcral para p

  $$Z = \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \stackrel{a}{\sim} \text{normal}.(0,1)$$

  Passo 2 — Obtenção dos quantis de probabilidade

  $$\begin{cases} a = \Phi^{-1}(\alpha/2) \\ b = \Phi^{-1}(1-\alpha/2) \end{cases}$$

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

  $P(a \le Z \le b) = 1-\alpha$

  $\ldots$

  $P\left(\bar{X} - b\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} < p < \bar{X} + b\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) = 1-\alpha$

  Passo 4 — Concretização

  $$IC_{(1-\alpha)\times 100\%}(p) \simeq \left[\frac{a}{n} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\frac{a}{n}\left(1-\frac{a}{n}\right)}{n}}\right].$$

Pergunta 8

Um médico conjetura que o valor esperado da temperatura basal de indivíduos saudáveis é de 37 graus Celsius. Nesse sentido, ele selecionou aleatoriamente n indivíduos, tendo observado uma média e variância corrigida da temperatura basal iguais a $\bar{x} = a$ e $s^2 = b$, respetivamente.

Admita que a temperatura basal de indivíduos saudáveis tem distribuição normal. Apoiarão os dados a conjetura do médico? Decida com base no valor-p.

- **V.a. de interesse**

  X = temperatura basal de indivíduo saudável (em graus Celsius)

- **Situação**

  $X \sim \text{normal}(\mu, \sigma^2)$

  $\mu = E(X)$ DESCONHECIDO

  $\sigma^2 = V(X)$ desconhecida

- **Hipóteses**

  $H_0 : \mu = \mu_0 = 37$

  $H_1 : \mu \neq \mu_0$

- **Estatística de teste**

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \stackrel{H_0}{\sim} t_{(n-1)}$$

- **Região de rejeição de** H_0 (para valores da estatística de teste)

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

- **Decisão (com base no p-value)**

$$\begin{aligned} valor - p &= P(|T| > |t| \mid H_0) \\ &= 2 \times \left[1 - F_{t_{(n-1)}}\left(\left|\frac{a - \mu_0}{\sqrt{b/n}}\right|\right)\right] \end{aligned}$$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[1] que se coaduna com:

  - a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor - p$;
  - a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor - p$.

Pergunta 9

Seja X a variável aleatória que representa o número de bombons com massa não conforme com a divulgada numa caixa de 10 bombons. O fabricante dos bombons defende a hipótese H_0 de que X possui uma distribuição binomial com parâmetros $n = 10$ e $p = 0.2$.

Os dados relativos a m destas caixas de 10 bombons são os seguintes:

Número de bombons com peso incorreto por caixa	0	1	2	3	>3
Frequência absoluta observada	o_1	o_2	o_3	o_4	o_5
Frequência absoluta esperada sob H_0	E_1	E_2	E_3	E_4	E_5

[1] 1. Rejeita-se H_0 para 1%, 5% e 10%. 2. Rejeita-se H_0 para 5% e 10% e não se rejeita H_0 para 1%. 3. Rejeita-se H_0 para 10% e não se rejeita H_0 para 1% e 5%. 4. Não se rejeita H_0 para 1%, 5% e 10%.

Após ter calculado as frequências absolutas esperadas E_1 e E_5 (aproximando-as às centésimas), averigúe se H_0 é consistente com este conjunto de dados. Decida com base no valor-p aproximado.

- **V.a. de interesse**

  X = número de bombons com massa não conforme, em caixa com 10 bombons

- **Hipóteses**

  $H_0 : X \sim \text{binomial}(n = 10, p = 0.2)$

  $H_1 : \neg H_0$

- **Estatística de teste**

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

  onde: k = no. de classes = 5; $\beta = 0$; o_i = freq. abs. observada da classe i; E_i = freq. abs. esperada, sob H_0, da classe i.

- **Região de rejeição de H_0**

  Ao lidarmos com um teste de ajustamento do qui-quadrado, a região de rejeição de H_0 é um intervalo à direita $W = (c, +\infty)$.

- **Frequências absolutas esperadas sob H_0**

$$\begin{aligned} E_1 &= m \times P(X = 0 \mid H_0) = m \times 0.1074 \\ E_5 &= m - \sum_{i=1}^{4} E_i \end{aligned}$$

- **Decisão (com base no valor-p)**

  $valor - p \simeq 1 - F_{\chi^2_{(k-1)}} \left(\sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i} \right)$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[2] que se coaduna com:

  - a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor - p$;
  - a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor - p$.

Pergunta 10

Certa piscina é sujeita a uma limpeza. O tempo, em hora, decorrido desde a conclusão dessa limpeza é descrito pela variável x. O logaritmo dos valores para o resíduo de cloro é representado pela variável aleatória Y. Observaram-se os valores para o resíduo de cloro nessa piscina, em diversos momentos após a limpeza. Os resultados obtidos são apresentados na seguinte tabela

Tempo após a limpeza (x, em horas)	x_1	x_2	x_3	x_4	x_5	x_6
Logaritmo do resíduo de cloro (Y)	0.59	0.41	0.37	0.35	0.32	0.31

e verificam as seguintes igualdades

$$\sum_{i=1}^{6} x_i = \sum_{i=1}^{6} x_i, \quad \sum_{i=1}^{6} x_i^2 = \sum_{i=1}^{6} x_i^2, \quad \sum_{i=1}^{6} y_i = \sum_{i=1}^{6} y_i, \quad \sum_{i=1}^{6} y_i^2 = \sum_{i=1}^{6} y_i^2, \quad \sum_{i=1}^{6} x_i\, y_i = \sum_{i=1}^{6} x_i\, y_i$$

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

Acha que o logaritmo do resíduo de cloro é significativamente influenciado com o tempo após a limpeza? Decida com base no valor-p.

---

[2] 1. Rejeita-se H_0 para 1%, 5% e 10%. 2. Rejeita-se H_0 para 5% e 10% e não se rejeita H_0 para 1%. 3. Rejeita-se H_0 para 10% e não se rejeita H_0 para 1% e 5%. 4. Não se rejeita H_0 para 1%, 5% e 10%.

- **Modelo de RLS**

  Y = logaritmo do resíduo de cloro (v.a. resposta)

  x = tempo após limpeza, em horas (variável explicativa)

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

- **Hipóteses de trabalho**

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

- **Hipóteses**

  $H_0 : \beta_1 = \beta_{1,0} = 0$

  $H_1 : \beta_1 \neq \beta_{1,0}$

- **Estatística de teste**

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_i x_i^2 - n \times \bar{x}^2}}} \stackrel{H_0}{\sim} t_{(n-2)}$$

- **Região de rejeição de** H_0 (para valores da estatística de teste)

  Dado que o teste é bilateral, a região de rejeição de H_0 é a reunião de intervalos $W = (-\infty, -c) \cup (c, +\infty)$.

- **Valor observado da estatística de teste**

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_i x_i^2 - n \times \bar{x}^2}}}$$

  onde

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n \left(\frac{\sum_{i=1}^n x_i}{n}\right) \left(\frac{\sum_{i=1}^n y_i}{n}\right)}{\sum_{i=1}^n x_i^2 - n \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} \\
\hat{\sigma}^2 &= \frac{1}{n-2} \left[\left(\sum_{i=1}^n y_i^2 - n \bar{y}^2 \right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \right] \\
&= \frac{1}{n-2} \left\{ \left[\sum_{i=1}^n y_i^2 - n \left(\frac{\sum_{i=1}^n y_i}{n} \right)^2 \right] - (\hat{\beta}_1)^2 \left[\sum_{i=1}^n x_i^2 - n \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2 \right] \right\}.
\end{aligned}$$

  [Convinha que t fosse calculado com base nas fórmulas acima que tiram partido do valor de n e das somas que se encontram no enunciado do problema.]

- **Decisão (com base no valor-p)**

  $valor - p = 2 \times \left[1 - F_{t_{(n-2)}}(|t|)\right]$

  Devemos escolher a resposta, das quatro apresentadas em nota de rodapé,[3] que se coaduna com:

  - a não rejeição de H_0 a qualquer n.s. $\alpha_0 \leq valor - p$;
  - a rejeição de H_0 a qualquer n.s. $\alpha_0 > valor - p$.

[3] 1. Rejeita-se H_0 para 1%, 5% e 10%. 2. Rejeita-se H_0 para 5% e 10% e não se rejeita H_0 para 1%. 3. Rejeita-se H_0 para 10% e não se rejeita H_0 para 1% e 5%. 4. Não se rejeita H_0 para 1%, 5% e 10%.