UFRJ - CCMN - IM - Departamento de Métodos Estatísticos

Segunda Prova de Probabilidade e Estatística **02-07-2015**

Atenção: Não serão aceitas respostas sem justificativa. As soluções devem ser feitas nos espaços apropriados.

1. Túlio está vendendo cartelas de rifa com o objetivo de juntar dinheiro para comprar um vídeo-game que custa R$ 1.200,00. Cada cartela de rifa tem um preço aleatório, definido por raspadinhas no corpo da cartela. Em média, compradores pagam R$ 50,00 com desvio padrão de R$ 10,00.

   (a) Túlio vendeu 25 cartelas. Qual é a probabilidade de ele ter conseguido uma quantia suficiente para fazer a compra que deseja?

   (b) Quantas cartelas Túlio precisaria vender para conseguir recursos suficientes para essa compra com 50% de probabilidade?

   Obs.: Admita que, em ambos os casos, o tamanho amostral é suficientemente grande para que seja válida a aproximação fornecida pelo Teorema Central do Limite.

2. Cinco pessoas estão almoçando juntas. Os nomes dessas pessoas, ordenando-as em termos de idade (do mais novo para o mais velho), são: Paulo, Andrea, Jorge, Daniela e Vitor. Com base em suas idades, foram calculadas as seguintes medidas: Média = 21, Mediana = 25, Distância Interquartil = 10, Amplitude = 30 (todas elas expressas em anos) e Variância = 146 anos^2.

   (a) Quais são as idades de cada um deles?

   (b) Alguma dessas cinco observações pode ser considerada discrepante, pelo critério para identificação de outliers que se baseia nos quartis da variável? Qual?

   Obs.: Amplitude = (Maior valor) − (Menor valor).

3. Um administrador do governo quer avaliar a proporção p de aprovação da comunidade sobre determinado projeto. Para isso fez uma pesquisa de opinião selecionando ao acaso 40 pessoas da população de interesse. Considerando "1" = favorável e "0"= desfavorável, temos então 40 variáveis aleatórias iid $X_1, X_2, ..., X_{40}$, todas elas com distribuição de Bernoulli(p). Considere as quatro seguintes alternativas de estimador para p:

$$\hat{p}_A = \frac{X_{10} + X_{20} + X_{30} + X_{40}}{4}, \qquad \hat{p}_B = \frac{X_{10} + 2X_{20} + 3X_{30} + 4X_{40}}{10}, \qquad \hat{p}_C = \frac{\sum_{i=1}^{40} X_i}{40} \qquad e$$

$$\hat{p}_D = \frac{\sum_{i=1}^{10} X_i + 2.\sum_{i=11}^{20} X_i + 3.\sum_{i=21}^{30} X_i + 4.\sum_{i=31}^{40} X_i}{100}$$

   (a) Qual é o viés (ou vício) de cada um desses estimadores?

   (b) Expresse o erro quadrático médio (EQM) de cada um deles em função de p. Ordene esses estimadores em termos de seus EQM, e diga qual é o melhor de acordo com esse critério.

   (c) Se os dados coletados foram:

   10001 11111 10111 11110 11000 11111 11111 11100,

   qual é a estimativa de p para cada uma das quatro alternativas?

4. Um disparador de partículas realiza o lançamento de partículas idênticas e o tempo com que essas partículas percorrem uma certa distância é normalmente distribuido com média e variância desconhecidas. Com objetivo de estudar propriedades desse disparador, foram lançadas 25 partículas e foi computado o tempo, em segundos, que cada uma delas levou para percorrer uma distância D. Esses dados seguem abaixo:

   9, 12, 13, 15, 15, 17, 18, 18, 18, 19, 19, 20, 20, 21, 22, 24, 24, 24, 25, 26, 26, 27, 29, 31, 33

   Denotando por x_i as medições acima, temos: $\sum_{i=1}^{25} x_i = 525 \qquad \sum_{i=1}^{25} x_i^2 = 11877$

   (a) Construa um intervalo de 95% de confiança para o tempo médio com que as partículas lançadas percorrem essa distância D. Interprete o significado do intervalo obtido.

   (b) Um físico afirmou que o tempo médio necessário para as partículas disparadas percorrerem uma distância D, como essa, é de no máximo 18 segundos. Utilizando o conceito de p-valor, verifique se a afirmação H_0 do físico será constestada ao nível de significância de 2%.

   (c) Construa um teste de hipóteses com nível de significância 8% para testar se o 3º quartil dos tempos com que as partículas disparadas percorrem essa distância D é igual a 23 contra a alternativa de que ele é diferente de 23. Qual a conclusão desse teste?

   Obs.: Note que é possível relacionar quartis com proporções.

Boa prova!

Soluções

1. (a) Sejam $X_1, \ldots, X_{25}$ variáveis aleatórias representando os valores pagos em cada cartela. Assim, a probabilidade do dinheiro ser suficiente é

$$P\left(\sum_{i=1}^{25} X_i \geq 1200\right) = P(\bar{X} \geq 1200/25), \quad \text{em que } \bar{X} \sim N(50, 100/25).$$

Logo,

$$P(\bar{X} \geq 1200/25) = P(\bar{X} \geq 48) = P\left(Z \geq \frac{48-50}{10/\sqrt{25}}\right) = P(Z \geq -1) = 0,8413.$$

(b) Deseja-se saber qual é o valor de n tal que

$$P\left(\sum_{i=1}^{n} X_i \geq 1200\right) = 0,50.$$

Temos que

$$P\left(\sum_{i=1}^{n} X_i \geq 1200\right) = P(\bar{X} \geq 1200/n), \quad \text{em que } \bar{X} \sim N(50, 100/n).$$

Logo,

$$\begin{aligned} P(\bar{X} \geq 1200/n) &= P\left(Z \geq \frac{1200/n - 50}{10/\sqrt{n}}\right) = 0,50 \\ &\Rightarrow \frac{1200/n - 50}{10/\sqrt{n}} = 0 \Rightarrow n = 1200/50 = 24. \end{aligned}$$

2. (a) Sejam $x_1 \leq x_2 \leq x_3 \leq x_4 \leq x_5$ as idades de Paulo, Andrea, Jorge, Daniela e Vitor, respectivamente. Então, temos:
Mediana $= x_3 = 25$, (I)
Distância Interquartil $= x_4 - x_2 = 10$, (II)
Amplitude $= x_5 - x_1 = 30$, (III)
Média $= 21 \rightarrow x_1 + x_2 + x_3 + x_4 + x_5 = 5 \times 21 = 105$ (IV)
Variância $= 146 \rightarrow x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 5 \times 21^2 = 4 \times 146$ (V)

Substituindo (I), (II) e (III) em (IV), obtemos:
$x_1 + x_2 + 25 + (x_2 + 10) + (x_1 + 30) = 105 \rightarrow 2x_1 + 2x_2 = 40 \rightarrow x_1 + x_2 = 20 \rightarrow x_2 = 20 - x_1$ (VI)
Além disso, de (II), $x_4 = x_2 + 10 = 20 - x_1 + 10 = 30 - x_1$ (VII)
Substituindo (VI), (I), (VII) e (III) em (V), chegamos a:
$x_1^2 + (20 - x_1)^2 + 25^2 + (30 - x_1)^2 + (30 + x_1)^2 = 4 \times 146 + 5 \times 441 = 2789.$
Isto nos leva a equação do 2^0 grau : $x_1^2 - 10x_1 + 9 = 0$, cujas raízes são 9 e 1.
Experimentemos inicialmente a solução $x_1 = 9$.
Ela nos leva a: $x_2 = 20 - 9 = 11$, $x_3 = 25$, $x_4 = 30 - 9 = 21$ e $x_5 = 30 + 9 = 39$.
Essa solução não serve, porque aqui temos $x_3 = 25 > x_4 = 21$, contrariando a premissa de que os dados estão ordenados.
Experimentemos agora a solução $x_1 = 1$. Ela nos leva a:
$x_2 = 20 - 1 = 19, x_3 = 25, x_4 = 30 - 1 = 29$ e $x_5 = 30 + 1 = 31$. Solução correta!
As idades são, portanto, as seguintes: Paulo: 1, Andrea: 19, Jorge: 25, Daniela: 29 e Vitor: 31.

(b) As cercas neste caso são: Cerca inferior $= x_2 - 1,5 \times DIQ = 19 - 1,5 \times 10 = 4$ e
Cerca superior $= x_4 + 1,5 \times DIQ = 29 + 1,5 \times 10 = 44$.
Sendo assim, temos um único outlier, que é a idade de Paulo = 1 ano.

3. (a) Usando as propriedades da esperança de uma combinação linear de variáveis aleatórias iid:

$$E(\hat{p}_A) = E(\frac{X_{10} + X_{20} + X_{30} + X_{40}}{4}) = p, \qquad E(\hat{p}_B) = E(\frac{X_{10} + 2.X_{20} + 3.X_{30} + 4.X_{40}}{10}) = p,$$

$$E(\hat{p}_C) = E(\frac{\sum_{i=1}^{40} X_i}{40}) = p, \quad E(\hat{p}_D) = E(\frac{\sum_{i=1}^{10} X_i + 2.\sum_{i=11}^{20} X_i + 3.\sum_{i=21}^{30} X_i + 4.\sum_{i=31}^{40} X_i}{100}) = p$$

O viés é $B(\hat{\theta}) = E(\hat{\theta}) - \theta$. Assim, $B(\hat{p}_A) = B(\hat{p}_B) = B(\hat{p}_C) = B(\hat{p}_D) = 0$

(b) Usando as propriedades da variância de uma combinação linear de variáveis aleatórias iid:

$$Var(\hat{p}_A) = Var(\frac{X_{10} + X_{20} + X_{30} + X_{40}}{4}) = \frac{1}{16}(Var(X_{10}) + Var(X_{20}) + Var(X_{30}) + Var(X_{40})) = \frac{p(1-p)}{4},$$

$$Var(\hat{p}_B) = \frac{Var(X_{10}) + 4.Var(X_{20}) + 9.Var(X_{30}) + 16.Var(X_{40})}{100} = \frac{3}{10}p(1-p)$$

$$Var(\hat{p}_C) = Var(\frac{\sum_{i=1}^{40} X_i}{40}) = \frac{\sum_{i=1}^{40} Var(X_i)}{40^2} = \frac{p(1-p)}{40}$$

$$Var(\hat{p}_D) = \frac{Var(\sum_{i=1}^{10} X_i) + 4.Var(\sum_{i=11}^{20} X_i) + 9.Var(\sum_{i=21}^{30} X_i) + 16.Var(\sum_{i=31}^{40} X_i)}{100^2} = \frac{3}{100}p(1-p)$$

Colocando as frações: $\frac{1}{4}$ $\frac{3}{10}$ $\frac{1}{40}$ $\frac{3}{100}$ com o mesmo denominador: $\frac{100}{400}$ $\frac{120}{400}$ $\frac{10}{400}$ $\frac{12}{400}$. Assim, $Var(\hat{p}_C) \leq Var(\hat{p}_D) \leq Var(\hat{p}_A) \leq Var(\hat{p}_B)$. Como os estimadores são não viciados, temos que:

$$EQM(\hat{p}_C) \leq EQM(\hat{p}_D) \leq EQM(\hat{p}_A) \leq EQM(\hat{p}_B)$$

O melhor estimador é o $\hat{p}_C$.

(c) Para esse conjunto de dados, temos as seguintes estimativas:

$$\hat{p}_A = \frac{2}{4} = 0,5$$

$$\hat{p}_B = \frac{1 + 2 \times 0 + 3 \times 1 + 4 \times 0}{10} = \frac{4}{10} = 0,4$$

$$\hat{p}_C = \frac{30}{40} = 0,75$$

$$\hat{p}_D = \frac{7 + 2 \times 8 + 3 \times 7 + 4 \times 8}{100} = \frac{76}{100} = 0,76$$

4. **(a)** Estatísticas necessárias: $\bar{x} = \sum_{i=1}^{25} \frac{x_i}{25} = \frac{525}{25} = 21$ e $s_x^2 = \frac{\sum_{i=1}^{25} x_i^2 - 25\bar{x}^2}{25-1} = \frac{11877 - 25.(21)^2}{24} = 35,5$

Seja μ o tempo médio. Como temos variância desconhecida e tamanho amostral pequeno ($n < 30$), é preciso construir o intervalo através da distribuição t-student. Temos que:

$$P\left(t_{n-1}\left(\frac{\alpha}{2}\right) < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$P\left(-t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\bar{X} - t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Os extremos do Intervalo de Confiança a 95% para μ são, portanto:

$$\bar{x} \pm t_{24}(0,975)\frac{s_x}{\sqrt{n}}, \quad \text{ou seja,} \quad 21 \pm 2,064\sqrt{\frac{35,5}{25}}$$

Logo, $IC(\mu, 95\%) = [\ 18,540\ ;\ 23,460\ ]$ em segundos.

Interpretação: Esse procedimento nos permite obter intervalos que conterão o verdadeiro valor do parâmetro μ, tempo médio, com 95% de chance. Vale observar que μ é um valor fixo, embora desconhecido, enquanto que os limites do intervalo são aleatórios.

(b) Queremos testar $\begin{cases} H_0 : \mu \leq 18 \\ H_1 : \mu > 18 \end{cases}$

Como σ^2 é desconhecido e $n < 30$, vamos utilizar o fato de que, sob H_0, $\dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Denotando p-valor por $\tilde{\alpha}$, temos que sendo H_0 verdadeira, ou seja, sendo $\mu = 18$.

$$\tilde{\alpha} = P(\bar{X} > \bar{x}) = P\left(T_{n-1} > \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}\right) = P\left(T_{24} > \frac{21 - 18}{\frac{\sqrt{35{,}5}}{\sqrt{25}}}\right) = P\left(T_{24} > \frac{3}{1,192}\right) =$$

$$= P(T_{24} > 2,527) = 1 - P(T_{24} \leq 2,527)$$

Pela tabela da t-student, podemos ver (olhando na linha associada a 24 graus de liberdade) que esse $P(T_{24} \leq 2,527)$ está entre $0,99$ e $0,995$. Logo, $\tilde{\alpha} \in [\ 0,005\ ;\ 0,01\ ]$.

Então, $\tilde{\alpha} < 0,02 = \alpha$, indicando que H_0 deve ser rejeitado ao nível de 2%, ou seja, a afirmação do físico deve sim ser contestada.

(c) Queremos testar $\begin{cases} H_0 : Q_3 = 23 \\ H_1 : Q_3 \neq 23 \end{cases}$

Defina p como sendo a proporção teórica de valores do tempo abaixo de 23 e $\hat{p}$ o estimador para p, dado pela proporção amostral de valores abaixo de 23. Como a probabilidade de um ensaio da variável "tempo de deslocamento da partícula" retornar um resultado abaixo de Q_3 é 75%, temos que o teste desejado equivale a testar, ao nível $\alpha = 0.08$:

$$\begin{cases} H_0 : p = 0,75 \\ H_1 : p \neq 0,75 \end{cases}$$

Como $np_0(1 - p_0) = 25 \times 0,75 \times 0,25 = 4,59 \geq 3$, podemos aproximar Binomial por Normal. Então:

$$\alpha = P(\text{ Rej } H_0), \text{ se } \mathrm{H}_0 \text{ é verdadeira} \qquad \Rightarrow$$

A região de aceitação é um intervalo da forma: $0,75 \pm c$, onde $P(0,75 - c < \hat{p} < 0,75 + c) = 0,92$, sendo $\hat{p} \sim N(0,75; \frac{0{,}75 \times 0{,}25}{25})$

$$P(\hat{p} < 0,75 + c) = 0,96 \quad \Rightarrow \quad P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0,75 + c - 0,75}{\sqrt{\frac{0{,}75(0{,}25)}{25}}}\right) = 0,96$$

$$P\left(Z < \frac{c}{0,0866}\right) = 0,96 \quad \Rightarrow \quad \frac{c}{0,0866} = Z_{0{,}96} \quad \Rightarrow \quad \frac{c}{0,0866} = 1,75 \quad \Rightarrow \quad c = 0.15155$$

Logo, a região de aceitação desse teste é dada por RA: $[0,75 - 0,15155\ ;\ 0,75 + 0,15155] = [0,59845\ ;\ 0,90155]$.

Como na amostra se observou $\hat{p} = \frac{15}{25} = 0,6$ e esse valor pertence a região de aceitação encontrada, temos que H_0 não foi rejeitado ao nível 8%, ou seja, não foi rejeitado que $Q_3 = 23$.

obs: Esse teste também poderia ser aplicado definindo por p a proporção de valores acima de 23. Nesse caso, testariamos se $p = 0,25$. Com cálculos similares chegariamos à região de aceitação RA:$[0,09845\ ;\ 0,40155]$ e concluiríamos pela não rejeição de H_0 pois $\hat{p} = 0,4 \in RA$.