UFRJ - CCMN - IM - Departamento de Métodos Estatísticos

Probabilidade e Estatística

Prova #1 **12-07-2016**

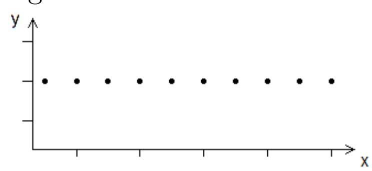
Atenção: Não serão aceitas respostas sem justificativa.

1. Marque verdadeiro ou falso para cada uma das afirmações abaixo, justificando a sua resposta.

   (a) Considere n pares (x_i, y_i) de modo que todas as observações da amostra Y tenham sido obtidas tomando as observações da variável X, multiplicando por 4 e depois somando 3 unidades ao resultado; ou seja, $y_i = 4x_i + 3$, para $i = 1, 2, \ldots n$. Nesse caso podemos afirmar que o coeficiente de variação da variável Y é inferior ao da variável X.

   (b) Considere uma amostra com amplitude (diferença entre o máximo e o mínimo) igual a 80. Se a diferença entre o terceiro quartil e o primeiro quartil for igual a 25, podemos afirmar que haverá pelo menos um outlier nessa amostra.

   (c) Se o gráfico de dispersão entre duas váriaveis X e Y for igual ao gráfico apresentado abaixo, então podemos afirmar que a correlação entre X e Y é igual a 1.



2. Deseja-se estimar a média do índice geral de inflação X, a partir de uma amostra de tamanho n. Sabe-se de estudos anteriores que X tem distribuição Normal com desvio padrão de 30 pontos. Propõe-se o uso de dois estimadores neste caso: a média amostral $\bar{X}$ e o estimador $T = \frac{X_1 + X_n}{2}$.

   (a) Compare ambos os estimadores com relação ao viés e o erro quadrático médio.

   (b) Suponha que os seguintes valores foram observados ao longo de 4 anos:

ano	1961	1962	1963	1964
índice (X)	108	120	128	192

   Dentre as suposições necessárias para os cálculos do item (a), existe uma que parece não ser satisfeita neste caso. Indique qual seria esta suposição e, em que parte do item (a), o fato de ela não ser atendida pode tornar questionáveis as conclusões obtidas.

   (c) Calcule as estimativas para $\bar{X}$ e T com base na amostra observada.

3. Os dados a seguir se referem ao tempo de tramitação (em anos) de 12 processos judiciais da mesma natureza: 7,2; 5,8; 0,8; 5,2; 7,7; 7,5; 6,5 ; 4,1 ; 6,2 ; 4,6 ; 5,5 ; 9,0. Admitindo que essa variável siga uma distribuição Normal:

   (a) Obtenha um intervalo de confiança a 98% para o tempo médio de tramitação de tais processos.

   (b) Qual deve ser o tamanho de uma nova amostra de processos a partir da qual se possa obter um intervalo de confiança a 98% para o tempo médio com amplitude igual a 1/4 da amplitude do intervalo obtido em (a)?

   Obs 1. Note que estimativas do mesmo parâmetro σ obtidas com base nessas duas amostras provavelmente devem estar próximas entre si. E lembre-se de que, se ν é suficientemente grande, a t de Student com ν graus de liberdade praticamente se confunde com a Normal(0,1).

   Obs 2. Para facilitar os cálculos, são fornecidos: $\sum_{i=1}^{12} x_i = 70,1$ anos e $\sum_{i=1}^{12} x_i^2 = 458,61$ anos^2.

4. O nível de aprovação da qualidade das refeições servidas em um restaurante universitário era de 25%, quando houve uma movimentação geral dos estudantes que forçou a direção do restaurante a fazer mudanças. Deseja-se verificar se as mudanças de fato surtiram efeito sobre a satisfação dos estudantes. Para isso sorteou-se uma amostra de 64 estudantes usuários do restaurante, dos quais 25 revelaram aprovar a qualidade da comida.

   (a) Verifique se as mudanças contribuiram para aumentar o nível de aprovação dos estudantes para com o restaurante universitário, sob um nível de significância de 5%.

   (b) Obtenha o menor nível de significância para o qual a hipótese nula do teste é rejeitada à luz do experimento realizado.

   (c) Calcule o valor da probabilidade do erro tipo II do teste do item (a) quando $p = 0,35$.

Boa prova!

SOLUÇÕES

1. **(a)** Verdadeiro.

   Como $y_i = 4x_i + 3$, pelas propriedades de esperança e variância, que também funcionam da mesma forma para média e variância amostral: $\bar{y} = 4\bar{x} + 3$ e $S_y^2 = 4^2 S_x^2 = 16S_x^2$. Daí:

$$C.V(Y) = \frac{S_y}{\bar{y}} = \frac{\sqrt{16S_x^2}}{4\bar{x}+3} = \frac{4S_x}{4\bar{x}+3} < \frac{4S_x}{4\bar{x}} = C.V(X)$$

   Logo, $C.V(Y) < C.V(X)$.

   (b) Falso.

   Para garantirmos a existência de pelo menos 1 outlier seria preciso que a amplitude dos dados fosse superior a diferença entre as cercas superior e inferior, que é dada por:

$$CS - CI = Q_3 + \frac{3}{2}(Q_3 - Q1) - \left(Q_1 - \frac{3}{2}(Q_3 - Q1)\right) = (Q_3 - Q_1) + \frac{3}{2}(Q_3 - Q_1) + \frac{3}{2}(Q_3 - Q_1) = 4(Q_3 - Q_1)$$

   Nesse caso temos $4(Q_3 - Q_1) = 4 \times 25 = 100$, que é superior a amplitude de 80. Logo, não há garantia de haver outliers.

   Nota: Também poderia ser feito esboçando um exemplo de base de dados com essas características (amplitude igual a 80 e distância interquartil igual a 25), mas sem nenhum outlier.

   (c) Falso.

   Nesse caso a correlação entre X e Y terá que ser igual a zero, pois Y assume um valor constante independentemente das variações de X. Logo, não há nenhuma dependencia entre X e Y, em particular, não há dependência linear; e consequentemente, a correlação entre eles é igual a zero.

2. **(a)** Seja X o índice geral de inflação. Deseja-se estimar o índice médio de inflação, dado por $E(X) = \mu$. Como $\mathrm{E}(\bar{X}) = \mu$ e $E[T] = \mu$, então ambos $\bar{X}$ e T são estimadores não viciados para μ.

   Por outro lado, $EQM(\bar{X}) = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$ e $EQM(T) = \mathrm{Var}(T) = \frac{\sigma^2}{2}$. O $EQM(\bar{X})$ diminui com o tamanho da amostra, portanto para $n > 2$ temos que $\bar{X}$ é mais eficiente que T.

   (b) A hipótese de independência entre os valores amostrados não parece ser satisfeita, devido a estrutura temporal das observações. Isto pode afetar principalmente as conclusões referentes ao EQM, porque podem haver covariâncias não nulas.

   (c) Estimativas são: $\bar{x} = 137$ e $t = 150$.

3. **(a)** Os extremos do intervalo de confiança pedido são obtidos pela expressão: $\bar{x} \pm t_{1-\alpha/2;n-1}\frac{s}{\sqrt{n}}$, onde

   $n = 12$; $\bar{x} = \frac{70{,}1}{12} = 5,84$ anos; $s = \sqrt{\frac{458{,}61 - \frac{70{,}1^2}{12}}{12-1}} = 2,11$ anos; $\alpha = 0,02$; $t_{0{,}99;11} = 2,72$.

   Fazendo os cálculos, obtemos (4,18; 7,50), em anos.

   (b) Seja m o tamanho da nova amostra a ser extraída. Para que a amplitude do novo intervalo seja igual a $1/4$ da amplitude do intervalo anterior, devemos ter m cerca de $4^2 = 16$ vezes maior que $n = 12$. Para esse tamanho de amostra o número de graus de liberdade $m - 1$ já será grande suficiente para que se considere que a t e a Normal(0;1) são praticamente iguais. Então, temos:

$$z_{0,99}\frac{s_m}{m} = \frac{1}{4}2,72\frac{s_n}{n},$$

onde s_m é o novo desvio padrão amostral e s_n é o antigo desvio padrão amostral. Como ambos s_m e s_n são estimativas do desvio padrão populacional σ, eles devem estar próximos entre si, mesmo se tratando de estimativas obtidas a partir de amostras de tamanhos diferentes. Admitiremos que eles são iguais. Chegamos então à expressão:

$$m = \left(4 \times \frac{2,33}{2,72}\right)^2 \times 12 \approx 141 \text{processos}$$

.

4. **(a)** Seja p: proporção de alunos satisfeitos com o restaurante universitário. Deseja-se testar:

   $H_0 : p \leq 0,25$ vs $H_1 : p > 0,25$

   (i) Estatística de teste: $Z = \frac{\hat{p}-0,25}{\sqrt{0,25(1-0,25)/n}}$, a qual sob H_0 apresenta distribuição $N(0,1)$.

   (ii) Pela tabela da $N(0,1)$, rejeita-se H_0 se $Z > 1,64$, ou equivalentemente, se $\hat{p} > 0,339$.

   (iii) Estatística de teste calculada para a amostra: $z = \frac{0,39-0,25}{\sqrt{0,25(1-0,25)/64}} = 2,59$.

   (iv) Decisão: como $z = 2,59$ pertence a região de rejeição, então rejeita-se H_0.

   Portanto, tudo indica que as mudanças implementadas contribuíram para que o nível de satisfação dos estudantes para com o restaurante universitário aumentasse.

   Outra maneira de resolver é substituindo o passo (iii) pelo cálculo do p-valor, da seguinte forma: a estatística de teste observada para a amostra é $z = 2,59$. Logo, p-valor $= P(Z > 2,59) = 0,0048$, o qual é menor que o nível de significância de 5%, portanto obtemos a mesma conclusão anterior.

   (b) Como p-valor $= 0,0048$, o menor nível de significância de 5% que levaria já à rejeição de H_0 é igual a $0,48\%$.

   (c) $\beta(p) = \mathrm{P}(\text{erro tipo II}) = \mathrm{P}(\text{não rejeitar } H_0 \mid H_1 \text{ verdadeira}) = \mathrm{P}(\hat{p} < 0,339)$, se $p > 0.25$.

   Logo, se $p > 0,25$

   $\beta(p) = \mathrm{P}\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < \frac{0,339-p}{\sqrt{p(1-p)/n}}\right) = \mathrm{P}\left(Z < \frac{0,339-p}{\sqrt{p(1-p)/64}}\right)$, em que $Z \sim N(0,1)$.

   Para $p = 0,35$, temos $\beta(0,35) = P\left(Z < \frac{0,339-0,35}{\sqrt{0,35(1-0,35)/64}}\right) = P(Z < -0,19) = 1 - P(Z < 0,19) = 1 - 0,5753 = 0,4247$