UFRJ - CCMN - IM - Departamento de Métodos Estatísticos

Segunda Avaliação de Probabilidade e Estatística **13-11-2018**
Atenção: Não serão aceitas respostas sem justificativa.
Resolver as questões nos espaços apropriados.

Q1) Queremos desenvolver um software que será composto por 5 módulos. Admita que o número de comandos em cada módulo pode ser tratado como uma variável aleatória com distribuição Normal de média $\mu = 60$ comandos e desvio padrão $\sigma = 10$ comandos. Há independência entre os números de comandos dos diversos módulos. O objetivo é calcular a média e o desvio padrão do número total de comandos desse software.
Foi proposta a seguinte solução: Sejam X = número de comandos em um módulo e Y = número de comandos total no software. Então $X \sim N(60; 10^2)$. Além disso, $Y = 5X$.
Logo: $E(Y) = 5 \times 60 = 300$ comandos; $Var(Y) = 5^2 \times 10^2 = 2500$ comandos^2; $DP(Y) = \sqrt{2500} = 50$ comandos.
Pergunta-se:

(a) Está correta essa solução? Se sim, por que sim? Se não, por que não?

(b) Apresente outra resolução para o problema.

(c) A suposição de Normalidade é de fato essencial para se resolver o problema? E quanto à suposição de independência?

Q2) Os dados da tabela abaixo representam o número de acidentes relacionados a uma empresa com dez construções em andamento durante os anos de 2016 e 2017.

Ano	Número de acidentes									
2016 (x)	9	15	10	18	12	14	8	17	18	9
2017 (y)	11	0	10	11	9	8	5	10	15	7

$$\sum_{i=1}^{10} x_i = 130; \quad \sum_{i=1}^{10} x_i^2 = 1828; \quad \sum_{i=1}^{10} y_i = 86; \quad \sum_{i=1}^{10} y_i^2 = 886; \quad \sum_{i=1}^{10} x_i y_i = 1160$$

(a) Calcule o número médio de acidentes e o desvio padrão do número de acidentes para cada ano.

(b) Calcule os quartis do número de acidentes e construa um box-plot dos dados para cada ano. Os dois box-plots devem aparecer juntos na mesma figura, ou seja, utilize a mesma escala de forma que os gráficos fiquem comparáveis. Existe(m) valor(es) discrepante(s)?

(c) Calcule o coeficiente de correlação.

Q3) Uma determinada empresa de telefonia pretende estimar a proporção p de clientes satisfeitos com o plano básico de telefonia móvel oferecido. Responda às seguintes questões.

(a) Suponha que a empresa tenha realizado uma pesquisa de satisfação com 2000 clientes, dos quais 800 afirmam estar satisfeitos com o plano básico de telefonia. Obtenha um intervalo de confiança não conservativo com coeficiente de confiança de 96% para a proporção p.

(b) Apenas para raciocinar, admita que a empresa tenha realizado 80 pesquisas de satisfação sobre este mesmo assunto. Se para cada pesquisa um intervalo de confiança de 95% para a proporção p for construído, quantos desses intervalos se espera que não contenham a proporção p?

(c) Encontre o menor número de clientes que devem compor uma pesquisa de satisfação de modo que a proporção estimada $\hat{p}$ difira de p em menos de 2% com probabilidade de 95%?

Q4) O desempenho médio na estrada para um certo tipo de automóvel SUV movido a diesel é de 15 km/l segundo informações da montadora. A revista *Auto Rodas* especializada no assunto, verificou o desempenho na estrada em 27 desses automóveis e apurou um desempenho médio amostral de 14,3 km/l. Admitindo que o desempenho ao longo da estrada siga um modelo Normal com desvio padrão de 3 km/l:

(a) teste ao nível de significância de 5% a afirmação da montadora de que a média de consumo é igual a 15 km/l contra a alternativa de ser diferente de 15 km/l. O que você pode concluir?

(b) determine a probabilidade do erro do tipo II assumindo $\mu = 14$ km/l. O que esta probabilidade significa?

(c) calcule o p-valor desse teste.

Boa Prova!

Solução

Q1) (a) A solução proposta acima não está correta. Ao afirmar que Y = 5X, ela pressupõe que o número de comandos será o mesmo em todos os 5 módulos, o que quase certamente não acontecerá.

(b) Solução correta: Sejam X_i = número de comandos no módulo i, i = 1,...,5. Então X_i é $N(60;10^2)$ para todo i; e os $X_i's$ são v.a.'s independentes. Ou seja, são iid. Além disso, Y é igual à soma dos $X_i's$.
Logo, E(Y) = 5 × 60 = 300 comandos; Var(Y) = 5 × 10^2 = 500 comandos^2; DP(Y) = $\sqrt{500}$ = 22,36 comandos.

(c) A suposição de Normalidade é desnecessária. Observe que ela não foi usada em nenhum ponto da solução. Já a suposição de independência é essencial. Foi o que nos permitiu afirmar que Var(Y) é igual à soma das variâncias dos $X_i's$, uma vez que independência implica em covariâncias nulas.

Q2) Sejam x= Número de acidentes em 2016 e y= Número de acidentes em 2017.

(a) $\bar{x} = \frac{\sum x}{n} = \frac{130}{10} = 13; \quad s_x^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} = \frac{1828-130^2/10}{9} = 15,33 = 3,91^2; \quad s_x = 3,91$ e
$\bar{y} = \frac{\sum y}{n} = \frac{86}{10} = 8,6; \quad s_y^2 = \frac{\sum y^2 - (\sum y)^2/n}{n-1} = \frac{886-86^2/10}{9} = 16.27 = 4,03^2; \quad s_y = 4,03$

(b) Pos(Q1)= (3+n)/4=13/4=3,25; Pos (Q2)= (11)/2 = 5,5; Pos(Q3)=(3n+1)/4=31/4=7,75; Os valores de X e de Y em ordem crescente são:
x $\Longrightarrow$ 8 9 9 10 12 14 15 17 18 18
y $\Longrightarrow$ 0 5 7 8 9 10 10 11 11 15

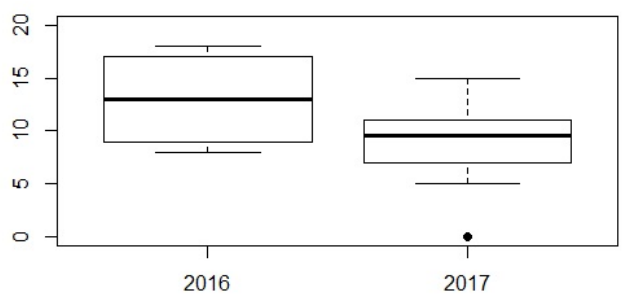
$Q1_x = 0,75 \times x_{(3)} + 0,25 \times x_{(4)} = 0,75 \times 9 + 0,25 \times 10 = 9,25;$
$Q1_y = 0,75 \times y_{(3)} + 0,25 \times y_{(4)} = 0,75 \times 7 + 0,25 \times 8 = 7,25;$

$Q2_x = 0,5 \times x_{(5)} + 0,5 \times x_{(6)} = 0,5 \times 12 + 0,5 \times 14 = 13;$
$Q2_y = 0,5 \times y_{(5)} + 0,5 \times y_{(6)} = 0,5 \times 9 + 0,5 \times 10 = 9,5;$

$Q3_x = 0,25 \times x_{(7)} + 0,75 \times x_{(8)} = 0,25 \times 15 + 0,75 \times 17 = 16,5;$
$Q3_y = 0,25 \times y_{(7)} + 0,75 \times y_{(8)} = 0,25 \times 10 + 0,75 \times 11 = 10,75;$

$DIQ_x = Q3_x - Q1_x = 16,5 - 9,25 = 7,25; \quad CI_x = 9,25 - 1,5 \times 7,25 = -1,625; \quad CS_x = 16,5 + 1,5 \times 7,25 = 27,375$
$DIQ_y = Q3_y - Q1_y = 10,75 - 7,25 = 3,5; \quad CI_y = 7,25 - 1,5 \times 3,5 = 2; \quad CS_y = 10,75 + 1,5 \times 3.5 = 16$

Assim, só o valor zero, no ano de 2017, é um valor discrepante.
O boxplot conjunto:



(c) $r_{xy} = cor(x,y) = \frac{s_{xy}}{s_x \times s_y} = \frac{(\sum xy - \sum x \times \sum y/n)/(n-1)}{s_x \times s_y} = \frac{(1160-130\times 86/10)/9}{3,91\times 4,03} \approx 0,30$

Q3) (a) Como $\hat{p} = 0,4$, o intervalo de confiança não conservativo com coeficiente de confiança de 96% para p é dado por:

$$IC(p) = \left(0,4 - 2,05\sqrt{\frac{(0,4)(0,6)}{2000}};\ 0,4 + 2,05\sqrt{\frac{(0,4)(0,6)}{2000}}\right) = (0,3775; 0,4224).$$

(b) Como os 80 intervalos de confiança foram construídos ao nível de 95%, se espera que cerca de 80 × (5/100) = 4 desses intervalos não contenham a proporção p desconhecida.

(c) Como $d = 0,02$, $\alpha = 0,05$ e não há informação a priori sobre a proporção p, o tamanho amostral n deve satisfazer

$$n \geq (1/4)(1,96/0,02)^2 = 2401.$$

Q4) (a) As hipóteses envolvem o parâmetro μ e podem ser escritas como:

$$\begin{cases} H_0: & \mu = 15 \\ H_1: & \mu \neq 15 \end{cases}$$

Como decidir pela a aceitação ou rejeição de H_0 sob o nível de significância de 5% já previamente estipulado? Uma vez que o teste envolve a média populacional, consideramos a média amostral $\bar{X}$ para construir a estatística de teste e usamos o fato de que $\bar{X} \sim N(\mu, \sigma^2/n = 3^2/27)$. A estatística de teste sob H_0 é

$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X}-15}{3/\sqrt{27}} \sim N(0,1)$.

Note que a regra de decisão pode ser formulada em termos da estatística Z ou em termos da própria média amostral $\bar{X}$. Rejeitamos H_0 se

$$z_{obs} < -1,96 \quad \text{ou} \quad z_{obs} > 1,96,$$

ou equivalentemente pela região crítica, $\bar{x} < 15 - 1,96\frac{3}{\sqrt{27}} = 13,86 \quad \text{ou} \quad \bar{x} > 15 + 1,96\frac{3}{\sqrt{27}} = 16,13$

Dada a amostra de tamanho $n = 27$, o valor observado da estatística de teste é :

$$z_{obs} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{14,3-15}{3/\sqrt{27}} = -1,212.$$

Como $-1,96 < z_{obs} = -1,212 < 1,96$, estamos na região de aceitação da hipótese e a decisão é não rejeitar a hipótese de que o consumo médio na estrada é de 15 *km/l*. Da mesma forma, como $\bar{x} = 14,3$ não pertence a região crítica, não rejeitamos a hipótese H_0 ao nível de significância de 5%. Concluímos que a média de desempenho na estrada é compatível com $H_0 : \mu = 15$ *km/l*, como divulgado pela montadora.

(b) Assumindo $\mu = 14$ *km/l*, a probabilidade do erro do tipo II é dada por

$$\begin{aligned} \beta(14) &= P(\text{não rejeitar}\, H_0 \mid H_0\, \text{falsa}) \\ &= P(13,86 \leq \bar{X} \leq 16,13 \mid \mu = 14) \\ &= P\left(\frac{13,86-14}{3/\sqrt{27}} \leq \frac{\bar{X}-14}{3/\sqrt{27}} \leq \frac{16,13-14}{3/\sqrt{27}}\right) \\ &= P(-0,24 \leq Z \leq 3,68) \\ &= 0,5938. \end{aligned}$$

Assim, com $\mu = 14$ *km/l* concluiríamos de forma equivocada que H_0 não deveria ser rejeitada com probabilidade de 0,5938.

(c) Como o teste é bilateral e $z_{obs} = -1,212$,
Temos que o p-valor=$2(1-\Phi(1,212)) = 2(1-0,8872) = 0,2256$
Como o p-valor é grande, maior do que 5%, confirmamos a conclusão do item (a), ou seja, não temos evidência para a rejeição de H_0.