UFRJ - CCMN - IM - Departamento de Métodos Estatísticos

Segunda Avaliação de Probabilidade e Estatística **17-06-2019**

Atenção: Não serão aceitas respostas sem justificativa.

Resolver as questões nos espaços apropriados.

Q1) Temos dois conjuntos de dados:

- $x_1, x_2, ..., x_m$, de tamanho m, cuja média é $\bar{x}$ e cuja variância é s_x^2; e
- $y_1, y_2, ..., y_n$, de tamanho n, cuja média é $\bar{y}$ e cuja variância é s_y^2;

e eles foram unidos para formar um novo conjunto de dados $x_1, x_2, ..., x_m, y_1, y_2, ..., y_n$, de tamanho m + n, cuja média é $\bar{z}$ e cuja variância é s_z^2.

(a) Calcule média e variância do conjunto de dados: 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1.

(b) Prove que $\bar{z} = \frac{m\bar{x}+n\bar{y}}{m+n}$

(c) Prove que $s_z^2 = \frac{(m-1)s_x^2+m(\bar{x}-\bar{z})^2+(n-1)s_y^2+n(\bar{y}-\bar{z})^2}{m+n-1}$
Sugestão: use as identidades: $\sum_{i=1}^{m}(x_i-\bar{z})^2 = \sum_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\bar{z})^2$ e $\sum_{j=1}^{n}(y_j-\bar{z})^2 = \sum_{j=1}^{n}(y_j-\bar{y}+\bar{y}-\bar{z})^2$
No 1^o somatório faça $a_i = x_i - \bar{x}$ e $b = \bar{x} - \bar{z}$, ficando com $\sum_{i=1}^{m}(a_i^2 + b^2 + 2a_ib)$.
Faça algo análogo com o 2^o somatório.

(d) Usando o que deve ser provado nos itens (b) e (c), recalcule a média e a variância do conjunto de dados do item (a). Desta vez, divida os dados em um primeiro conjunto com as 5 primeiras observações e um segundo conjunto com as 7 últimas observações.

Q2) Numa linha de produção se precisa estimar a proporção de peças defeituosas. Existem dois tipos de defeitos: A e B. Sejam p_A a proporção de peças que apresentam o defeito A e p_B a proporção de peças que apresentam o defeito B. Deseja-se que as estimativas não se desviem das proporções verdadeiras, tanto para p_A quanto para p_B, por mais de 0,08 com probabilidade de 98%.

(a) Sem nenhum conhecimento prévio, quantas peças se precisa examinar para satisfazer as exigências na estimativa da proporção p_A?

(b) Foi verificado que o tipo de defeito B afeta no máximo 20% das peças. Quantas peças se precisa examinar para satisfazer as exigências?

(c) Numa única amostra, quantas peças se precisa examinar para se satisfazer as duas exigências conjuntamente, tanto para p_A quanto para p_B?

(d) Repita o que foi feito no item (c), porém agora sendo um pouco mais tolerante com a precisão na estimação de p_A: o erro absoluto deve ser menor que 0,12 com 95% de probabilidade

Q3) Os seguintes dados foram coletados de uma população normal cuja média e variância são desconhecidas:

$$13, 10, 8, 11, 6, 7, 12, 5, 9$$

(a) Achar uma estimativa pontual da média e da variância da população

(b) Qual é o intervalo de confiança de 90% para a média da população?

(c) O intervalo de confiança de 95% para a média da população contem ou está contido no intervalo de confiança de 90%? Explique sua resposta.

(d) Se for testada a hipótese $H_0 : \mu = 11$ contra $H_1 : \mu \neq 11$ ao nível de significância $\alpha = 10\%$ que decisão será tomada?

Q4) Um grupo de arqueólogos propõe a teoria que o tamanho médio de crânios de seres humanos aumenta ao longo do tempo. Medições precisas garantem que atualmente tal quantidade comporta-se conforme uma distribuição normal, de média $\mu = 140$mm e desvio padrão de 26mm. Uma amostra de 30 crânios, datadas de 6.000 anos atrás, foi observada, com um tamanho médio de 131,37mm. Desejamos aferir a validade da teoria proposta, ou seja, verificar se no passado o tamanho médio dos crânios era de fato menor. Para isso, queremos testar $H_0 : \mu \geq 140$ contra $H_1 : \mu < 140$, admitindo que o desvio padrão não é afetado pela passagem do tempo.

(a) Diga, nesse contexto, o que significam os erros tipo I e II.

(b) Teste as hipóteses H_0 contra H_1, ao nível de significância de 5%.

(c) Calcule o p-valor desse teste, explicitando o seu significado.

Boa Prova!

Solução

Q1) (a) $\bar{z} = \frac{\sum_{i=1}^{(m+n)} z_i}{m+n} = \frac{7}{12} = 0,5833$

$s_z^2 = \frac{\sum_{i=1}^{(m+n)} z_i^2 - (m+n)\bar{z}^2}{m+n-1} = \frac{7-12\times 0,5833^2}{11} = 0,265$

(b) $\bar{z} = \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} y_j}{m+n} = \frac{m\bar{x}+n\bar{y}}{m+n}$

(c) $s_z^2 = \frac{\sum_{i=1}^{m}(x_i-\bar{z})^2 + \sum_{j=1}^{n}(y_j-\bar{z})^2}{m+n-1}$ (I)

$\sum_{i=1}^{m}(x_i-\bar{z})^2 = \sum_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\bar{z})^2 = \sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\bar{z})^2 + 2(\bar{x}-\bar{z})\sum_{i=1}^{m}(x_i-\bar{x})$

É claro que este último somatório é igual a zero, porque $\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$.

Então, $\sum_{i=1}^{m}(x_i-\bar{z})^2 = (m-1)s_x^2 + m(\bar{x}-\bar{z})^2$ (II)

Analogamente: $\sum_{j=1}^{n}(y_j-\bar{z})^2 = (n-1)s_y^2 + n(\bar{y}-\bar{z})^2$ (III)

Substituindo (II) e (III) em (I), decorre imediatamente a expressão desejada.

Obs.: Na demonstração foi usada a propriedade $(a+b)^2 = a^2+b^2+2ab$.

(d) No conjunto de dados há 5 valores iguais a 0 e 7 valores iguais a 1. Portanto:

m=5, $\bar{x}=0$, $s_x^2 = 0$ n=7, $\bar{y}=1$, $s_y^2=0$

Logo, m+n=12; $\bar{z} = \frac{5\times 0+7\times 1}{12} = \frac{7}{12} = 0,583$

$s_z^2 = \frac{4\times 0+5\times(0-7/12)^2+6\times 0+7\times(1-7/12)^2}{5+7-1} = 0,265$

Q2) (a) O erro absoluto máximo tolerável da estimativa é d=0,08. $P(|\hat{p}-p|<d) = 0,98$, $1-\alpha = 0,98$, assim $1-\frac{\alpha}{2} = 0,99$. Como $\Phi(z) = 0,99$, temos z=2,33. Como não temos informação alguma sobre p, utilizaremos p = 0,5 que maximiza p(1-p) e calculamos o tamanho amostral:

$$n_A = \left(\frac{z}{d}\right)^2 p(1-p) = \left(\frac{2,33}{0,08}\right)^2 0,25 = 212,07 \approx 213$$

(b) Se p_2 for no máximo 0,2, o tamanho amostral será:

$$n_B = \left(\frac{z}{d}\right)^2 p(1-p) = \left(\frac{2,33}{0,08}\right)^2 0,2\times 0,8 = 135,72 \approx 136$$

(c) n = max(n_A, n_B) = max(213; 136)=213

(d) $n_A^d = \left(\frac{z}{d}\right)^2 p(1-p) = \left(\frac{1,96}{0,12}\right)^2 0,25 = 96,04 \approx 97$. Assim, max($n_A^d, n_B$) = max(97, 136) = 136.

Q3) (a) $\bar{x} = \frac{1}{n}\sum_{i=1}^{9} x_i = \frac{81}{9} = 9$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{9}(x_i-\bar{x})^2 = \frac{60}{8} = 7,5$ e $s = 2,74$

(b) $[\bar{x} - t_{8;0,95}\frac{s}{\sqrt{n}}, \bar{x} + t_{8;0,95}\frac{s}{\sqrt{n}}] = [9 - \frac{2,74}{3}1,86; 9 + \frac{2,74}{3}1,86] = [7,3; 10,7]$, já que $t_{0,95;8} = 1,86$

(c) Sendo o nível de confiança maior, isso implica que o intervalo contem o intervalo anterior. Isso pode ser também deduzido observando que $t_{0,95;8} < t_{0,975;8}$.

(d) Como 11 não pertence ao intervalo [7,3; 10,7], H_0 deve ser rejeitada, já que a distância entre $\bar{x} = 9$ e $\mu_0 = 11$ é maior que a metade da amplitude do intervalo de confiança obtido no item (b).

Q4) (a) O erro tipo I representa rejeitar a hipótese nula quando ela é verdadeira. Nesse contexto, representa aceitar H_1, uma nova teoria científica, quando ela, de fato, não é verdadeira. Por outro lado, o erro tipo II representa não rejeitar a hipótese nula quando ela é falsa, ou seja, não aceitar uma nova teoria científica quando ela é verdadeira. O erro tipo I pode ser considerado mais grave pois aceitar uma nova teoria em detrimento de uma antiga, já bem consolidada, potencialmente altera mais a perspectiva da realidade do que não aceitar uma nova teoria.

(b) Denote por $\overline{X}$ a média amostral. Parece razoável rejeitarmos H_0 se observarmos um valor suficientemente baixo x_c para $\overline{X}$. De fato, usemos o nível de significância para encontrar tal valor:

$$
\begin{aligned}
0,05 &= P(\text{rejeitar } H_0 \mid H_0 \text{ é verdade}) \\
&= P(\overline{X} < x_c \mid \mu = 140) \\
&= P\left(\frac{\overline{X} - 140}{26/\sqrt{30}} < \frac{x_c - 140}{26/\sqrt{30}}\right) \\
&= P\left(Z < \frac{x_c - 140}{26/\sqrt{30}}\right),
\end{aligned}
$$

onde $Z \sim \mathcal{N}(0,1)$. Consultando a tabela, temos que $\dfrac{x_c - 140}{26/\sqrt{30}} = z_{0,05} = -z_{0,95} = -1,65$, de modo que $x_c = 132,17$mm. Como $\overline{x} = 131,37 < x_c$, rejeitamos H_0 ao nível de significância de 5%.

(c) O p-valor é o menor nível de significância para o qual a hipótese nula seria rejeitada, para aqueles dados observados. Mais precisamente:

$$
\begin{aligned}
\tilde{\alpha} &= P(\overline{X} < \overline{x} \mid \mu = 140) \\
&= P\left(\frac{\overline{X} - 140}{26/\sqrt{30}} < \frac{\overline{x} - 140}{26/\sqrt{30}} \;\middle|\; \mu = 140\right) \\
&= P(Z < -1,82) \\
&= 1 - P(Z < 1,82) \\
&= 1 - 0,9656 \\
&= 0,0344,
\end{aligned}
$$

onde $Z \sim \mathcal{N}(0,1)$. Ou seja, o menor nível de significância para o qual ainda rejeitaríamos a hipótese nula é de 3,44%.