MAE116 – Noções de Estatística

Grupos B e D – II Semestre de 2020

Lista de exercícios 7 – Estimação II – CASA (gabarito)

Exercício 1.

Um determinado estado deseja estudar o número de mortes infantis causadas por lesões, com o intuito de promover um programa educacional de redução desse número. As causas de morte por lesões para crianças com idades entre 5 e 9 anos são: 1 - acidente por veículo motorizado, 2 - afogamento, 3 - incêndio no lar, 4 - homicídio e 5 - outras causas, inclusive sufocamento, quedas e envenenamento. Deseja-se estimar a proporção p de crianças vítimas fatais por veículo motorizado ou afogamento.

Solução

(a) Qual deve ser o tamanho da amostra, para que o erro da estimativa de p seja de 0,05, com um nível de confiança de 0,96?

Sabemos que

$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p),$$

em que ε é o erro da estimativa de p, e z é o quantil da distribuição normal padrão tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$, sendo γ o coeficiente de confiança e $Z \sim \mathrm{N}(0,1)$.

Do enunciado temos $\varepsilon = 0,05$ e $\gamma = 0,96$. Para utilizar a tabela fornecida no e-disciplinas para encontrar o valor do quantil z, é necessário expressá-lo em termos de sua probabilidade acumulada à esquerda. Como

$$\mathrm{P}(-z \leq Z \leq z) = 0,96,$$

então z deixa uma probabilidade acumulada à esquerda igual a $0,98$. Verificando a tabela obtemos que $z = 2,05$.

Note que para calcular o tamanho amostral ainda é necessário fornecer um valor para a proporção p. Como não temos nenhuma informação sobre esta proporção, vamos usar o fato de que a função $p(1-p)$ é uma parábola simétrica em torno de $p = 0,5$, e que atinge o valor máximo em $p = 0,5$. Graficamente,

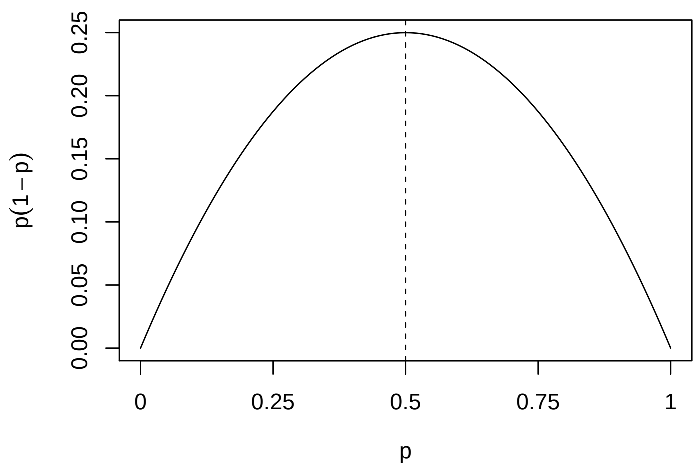


Figura 1: Gráfico de $p(1-p)$ em função de p.

Utilizar o valor $p = 0,5$ significa calcular o tamanho amostral (controlando o erro da estimativa e o coeficiente de confiança) no cenário com maior variabilidade possível. Assim, temos que

$$n = \left(\frac{2,05}{0,05}\right)^2 \times 0,5 \times 0,5 = 420,25.$$

Logo, o tamanho da amostra a ser utilizado deve ser de $n = 421$ crianças.

(b) Por meio de registros nos Juizados de menores do estado, sabe-se que essa proporção p é superior a 70%. Com essa informação seria possível considerar em (a) uma amostra de tamanho menor? Se sim, de quanto? Se não, por quê?

Note que apesar de sabermos que $p > 0,7$, ainda precisamos fornecer um valor para p para que seja possível realizar o cálculo do tamanho amostral. Mais uma vez iremos considerar o cenário de maior variabilidade possível, mas agora sabendo que $p > 0,7$. A figura abaixo apresenta o gráfico da função $p(1-p)$, em que destaca-se em vermelho os valores da função quando restringimos que $p > 0,7$. Note que o valor máximo ocorre em $p = 0,7$.

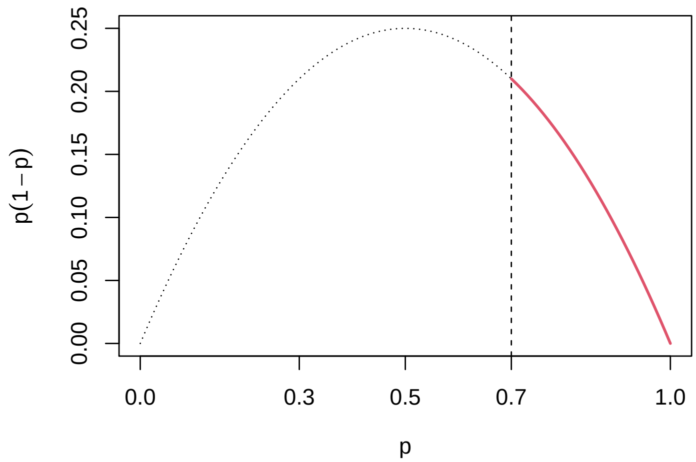


Figura 2: Gráfico de $p(1-p)$ em função de p com a restrição de que $p > 0,7$.

Portanto, o tamanho de amostra necessário para que o erro da estimativa de p seja de 0,05, com um coeficiente de confiança de 0,96, deve ser de

$$n = \left(\frac{2,05}{0,05}\right)^2 \times 0,7 \times 0,3 = 353,01.$$

Logo, o tamanho da amostra a ser utilizado deve ser de $n = 354$ crianças. Portando, é possível considerar uma amostra com uma redução de 67 crianças em relação ao item anterior.

(c) A seguir é apresentado um conjunto de dados que indica as causas de morte para uma amostra de 100 crianças entre as idades de 5 a 9 anos vítimas fatais de lesões.

1	5	3	1	2	4	1	3	1	5	2	1	1	1	2	1	2	1	4	1
4	1	3	1	1	1	2	1	1	2	5	1	1	1	1	2	3	1	2	1
2	3	1	1	2	1	5	1	5	1	1	2	5	1	1	1	3	4	1	1
1	1	2	1	1	2	1	1	2	1	1	3	1	5	2	3	2	1	3	4
1	1	2	4	5	2	1	5	1	5	2	1	1	5	1	1	1	1	1	5

Dê uma estimativa pontual para p e, com base nela, construa um intervalo de 96% de confiança para p. Qual é o erro amostral de sua estimativa?

A estimativa pontual para p é dada por

$$\widehat{p} = \frac{k}{n},$$

em que k é o número de crianças com idade entre 5 e 9 anos vitimas fatais por veículo motorizado ou afogamento observadas na amostra, e n é o tamanho amostral.

Dentre as 100 mortes selecionadas, 73 foram por veículo motorizado ou por afogamento. Logo

$$\widehat{p} = \frac{73}{100} = 0,73.$$

Utilizando o Teorema Limite Central para aproximar a distribuição amostral do estimador da proporção, um intervalo de confiança aproximado para p, com coeficiente de confiança γ é dado por

$$\text{IC}(p;\gamma) = \left[\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}};\;\; \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right],$$

em que z é o quantil da distribuição normal padrão tal que $\text{P}(-z \leq Z \leq z) = \gamma$, sendo γ o coeficiente de confiança e $Z \sim \text{N}(0,1)$.

Como já calculado no item (a), $z = 2,05$. Logo,

$$\begin{aligned}\text{IC}(p;96\%) &= \left[0,73 - 2,05\sqrt{\frac{0,73(1-0,73)}{100}};\;\; 0,73 + 2,05\sqrt{\frac{0,73(1-0,73)}{100}}\right] \\ &= [0,73 - 0,0910;\;\; 0,73 + 0,0910] \\ &= [0,639;\;\; 0,821]\end{aligned}$$

é a estimativa intervalar para p com um erro amostral igual a 0,0910.

Exercício 2

O tempo médio que um analgésico administrado via oral leva para penetrar na corrente sanguínea é de 43 min. A fim de acelerar esse tempo, um bioquímico acrescentou certo ingrediente à fórmula original. O medicamento com a nova fórmula foi aplicado em 70 pessoas, aleatoriamente selecionadas, obtendo-se um tempo médio de penetração do analgésico no sangue de 41 min, com desvio padrão de 6,7 min.

Solução

(a) Construa um intervalo de 90% de confiança para o tempo médio de penetração do analgésico com a nova fórmula na corrente sanguínea. Com base nesse intervalo, você diria que o tempo médio diminuiu?

Defina a seguinte variável aleatória:

X : "Tempo que um analgésico administrado via oral leva para penetrar na corrente sanguínea com a nova fórmula",

e denote por μ e σ o valor esperado e o desvio padrão de X, respectivamente.

Do enunciado, obtemos que o tamanho amostral foi de $n = 70$ pessoas, a média amostral foi igual a $\overline{x} = 41$ min e o desvio padrão amostral foi de $s = 6,7$ min. Note que o enunciado não nos fornece a distribuição da variável aleatória X.

Assumindo que $n = 70$ é um tamanho amostral suficientemente grande, podemos aproximar a distribuição da média amostral pela distribuição normal utilizando o Teorema Limite Central, de forma que

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{aproximadamente.}$$

Portanto, um intervalo de confiança aproximado para μ com coeficiente de confiança γ é dado por

$$\mathrm{IC}[\mu;\gamma] = \left[\overline{X} - z\frac{S}{\sqrt{n}};\ \ \overline{X} + z\frac{S}{\sqrt{n}}\right],$$

em que z é o quantil da distribuição normal padrão tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$, sendo $Z \sim \mathrm{N}(0,1)$.

Sendo $\gamma = 0,90$, segue que

$$\mathrm{P}(-z \leq Z \leq z) = 0,90,$$

e, portanto, o quantil z deixa uma probabilidade acumulada à esquerda igual a $0,95$. Verificando a tabela da distribuição normal padrão obtemos que $z = 1,65$.

Logo, uma estimativa intervalar aproximada para μ com coeficiente de confiança $\gamma = 0,90$ é dada por

$$\begin{aligned}
\mathrm{IC}[\mu; 0,90] &= \left[\overline{x} - z\frac{s}{\sqrt{n}};\ \ \overline{x} + z\frac{s}{\sqrt{n}}\right] \\
&= \left[41 - 1,65\frac{6,7}{\sqrt{70}};\ \ 41 + 1,65\frac{6,7}{\sqrt{70}}\right] \\
&= [41 - 1,3213;\ \ 41 + 1,3213] \\
&= [39,6787;\ \ 42,3213]
\end{aligned}$$

Note que o tempo médio antes do bioquímico acrescentar o igrediente à fórmula original era de 43 min. A estimativa intervalar calculada nos diz que o tempo médio com a nova fórmula está entre

39,6787 min e 42,3213 min aproximadamente, com confiança de 90%. Como o valor 43 está acima do limite superior calculado, então o tempo médio com a nova fórmula é estatisticamente menor do que 43, e portanto, existem evidências para afirmar que o tempo médio diminuiu ao acrescentar o igrediente à fórmula original.

(b) Que tamanho de amostra seria necessário para que o intervalo de 90% de confiança tivesse comprimento igual a 1?

Sabemos que um intervalo de confiança aproximado para μ com coeficiente de confiança $0,90$ é dado por

$$\text{IC}[\mu; 0,90] = \left[\overline{X} - 1,65\frac{S}{\sqrt{n}};\;\; \overline{X} + 1,65\frac{S}{\sqrt{n}}\right].$$

Deseja-se que este intervalo confiança tenha comprimento igual a 1, isto é,

$$\overline{X} + 1,65\frac{S}{\sqrt{n}} - \left(\overline{X} - 1,65\frac{S}{\sqrt{n}}\right) = 1.$$

Isolando n nesta equação, obtemos que

$$\begin{aligned} 2 \times 1,65\frac{S}{\sqrt{n}} &= 1 \\ \Rightarrow \sqrt{n} &= 3,3 \times S \\ \Rightarrow n &= 10.89 \times S^2. \end{aligned}$$

Utilizando a informação do estudo realizado no item (a), obtemos que

$$n = 10.89 \times 6,7^2 = 488,8521.$$

Logo, o tamanho da amostra necessário para que o intervalo de 90% de confiança tenha comprimento igual a 1 deve ser igual a 489 pessoas.

Exercício 3

Um professor de educação física do Ensino Fundamental II, que leciona em várias escolas, deseja estimar a proporção p de alunos que conseguiram aprender a jogar Tênis de Mesa de forma satisfatória, depois de ensinar essa modalidade esportiva por um determinado período de tempo. Ele quer que essa proporção seja estimada com um erro de 0,02 e um nível de confiança de 0,90.

Solução

(a) Qual deve ser o tamanho de amostra necessário para atender às exigências do professor?

Sabemos que

$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p).$$

Do enunciado temos $\varepsilon = 0,02$ e $\gamma = 0,90$. Para utilizar a tabela fornecida no e-disciplinas para encontrar o valor do quantil z, é necessário expressá-lo em termos de sua probabilidade acumulada à esquerda. Como

$$\mathrm{P}(-z \leq Z \leq z) = 0,90,$$

então z deixa uma probabilidade acumulada à esquerda igual a $0,95$. Verificando a tabela obtemos que $z = 1,65$.

Como no Exercício 1 (a), não temos nenhuma informação sobre p e teremos que utilizar o fato de que a função $p(1-p)$ é uma parábola simétrica em torno de $p = 0,5$, que assume seu valor máximo em $p = 0,5$ como na Figura 1.

Logo,

$$n = \left(\frac{1,65}{0,02}\right)^2 0,5 \times 0,5 = 1701,562 \approx 1702,$$

e portanto o tamanho amostral deve ser igual a 1702 alunos.

(b) Que tamanho deveria ter a amostra supondo que p esteja entre 0,3 e 0,7? E supondo que p seja menor que 0,3?

Note que mesmo tendo informações sobre a proporção p, ainda precisamos fornecer um valor para ele para que seja possível realizar o cálculo do tamanho amostral. O procedimento que estamos adotando é considerar o cenário de maior variabilidade possível que corresponde ao valor máximo para a função $p(1-p)$.

Supondo que p esteja entre 0,3 e 0,7, o ponto máximo da função $p(1-p)$ ainda é atingido em $p = 0,5$, e portanto esta informação não provoca uma alteração no cálculo do tamanho amostral. A Figura 3 apresenta o gráfico da função $p(1-p)$, em que destaca-se em vermelho os valores da função quando restringimos que $0,3 < p < 0,7$.

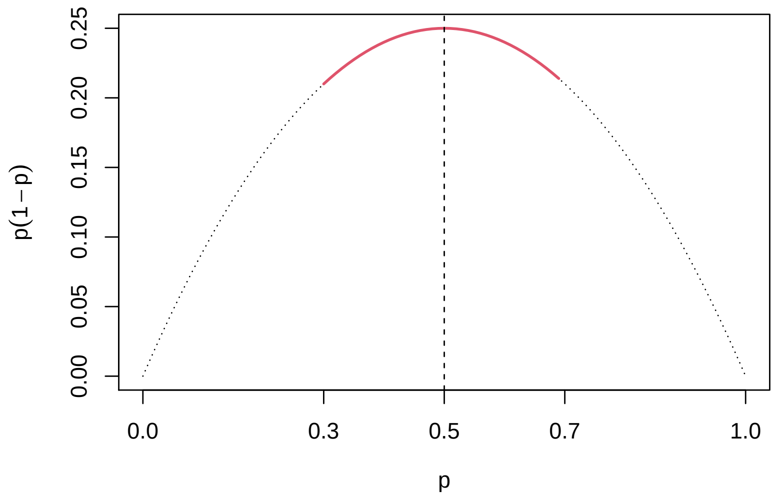


Figura 3: Gráfico de $p(1-p)$ em função de p com a restrição de que $0,3 < p < 0,7$.

Suponha agora que $p < 0,3$. Note que se $p < 0,5$, a função $p(1-p)$ é uma função crescente. Portanto, quando restringimos $p < 0,3$, o máximo desta função é atingido em $p = 0,3$. A Figura 4 ilustra este caso.

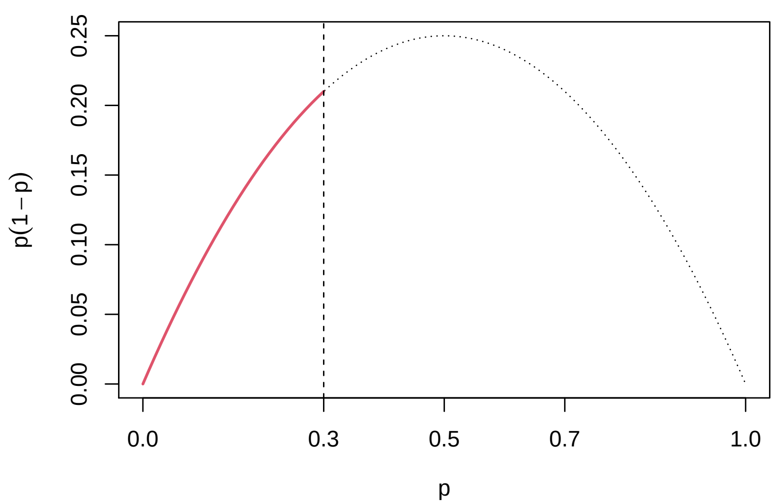


Figura 4: Gráfico de $p(1-p)$ em função de p com a restrição de que $p < 0,3$.

Assim

$$n = \left(\frac{1,65}{0,02}\right)^2 \times 0,3 \times 0,7 = 1429,312 \approx 1430.$$

Logo, sabendo que $p < 0.3$, o tamanho amostral deve ser de 1430 alunos. Uma redução em relação ao item anterior de aproximadamente 273 alunos.

(c) Coletada uma amostra de 150 estudantes, 30 apresentaram bom desempenho em uma competição em duplas. Com base nesses dados, determine um intervalo de confiança com coeficiente de confiança de 0,98 para p.

A estimativa pontual para p é dada por

$$\widehat{p} = \frac{k}{n},$$

em que k é o número de estudantes que apresentaram bom desempenho em uma competição em duplas observadas na amostra, e n é o tamanho amostral.

Como 30 em 150 alunos apresentaram bom desempenho na competição, a estimativa pontual para p é igual a

$$\widehat{p} = \frac{30}{150} = 0,20.$$

Utilizando o Teorema Limite Central para aproximar a distribuição amostral do estimador da proporção, um intervalo de confiança aproximado para p, com coeficiente de confiança γ é dado por

$$\mathrm{IC}(p;\gamma) = \left[\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}};\ \ \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right],$$

em que z é o quantil da distribuição normal padrão tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$, sendo γ o coeficiente de confiança e $Z \sim \mathrm{N}(0,1)$.

Como

$$\mathrm{P}(-z \leq Z \leq z) = 0,98,$$

então z deixa uma probabilidade acumulada à esquerda igual a $0,99$. Verificando a tabela obtemos que $z = 2,33$. Logo,

$$\begin{aligned} IC(p;98\%) &= \left[0,20 - 2,33\sqrt{\frac{0,20(1-0,20)}{150}};\ \ 0,20 + 2,33\sqrt{\frac{0,20(1-0,20)}{150}}\right] \\ &= [0,20 - 0,0761;\ \ 0,20 + 0,0761] \\ &= [0,1239; 0,2761]\,. \end{aligned}$$

Exercício 4

O gerente de uma grande academia deseja avaliar a eficácia de dois diferentes programas (A e B) de treinamento físico desenvolvidos para adultos jovens. Um grupo de 200 alunos da academia (todos adultos jovens) é aleatoriamente indicado para os dois programas, de modo que existam 100 sujeitos em cada programa. Ao final do período de treinamento, cuja duração foi um mês, é aplicado um teste de condicionamento físico padrão aos 200 sujeitos; as notas obtidas no teste (0 a 100) são apresentadas no arquivo Lista7_ex4_programas.xls.

Solução

(a) Descreva as variáveis envolvidas nesse experimento.

São duas as variáveis do experimento:

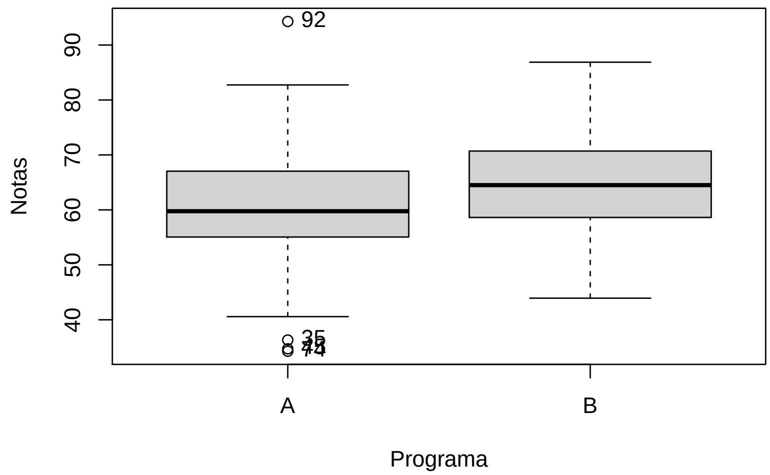
- `Notas`: notas obtidas no teste, que pode variar entre 0 e 100, obtida por cada jovem adulto após a realização do teste de condicionamento físico;
- `Programa`: programas (A ou B) de treinamento físico desenvolvidos pela academia.

(b) Construa um boxplot para as notas obtidas no teste, para cada um dos programas de treinamento físico.

A seguir apresentamos os passos utilizados no `Rcmdr` para a construção dos boxplots.

1) Ler o conjunto de dados: Clique no menu “Dados”→ “Importar arquivos de dados”→ “do Excel”. Defina um nome para o seu conjunto de dados (por exemplo, dados) e clique em “OK”. Navegue até o diretório onde o seu conjunto de dados está localizado e clique em “Abrir”. Selecione a aba do Excel onde está localizado o conjunto de dados (no nosso caso, “Plan1”) e clique em “OK”.

2) Construção do boxplot: Clique no menu “Gráficos” → “Boxplot”. Na aba “Dados” selecione a variável “Notas” e em “Gráficos por grupos” selecione a variável “Programa” e clique em “OK”. Finalize clicando em “OK”. O gráfico será exibido em uma nova janela. Para copiá-lo, clique com o botão direito sobre o gráfico e escolha a opção “Copiar como Bitmap”. Em seguida cole no documento da sua lista de exercícios. O gráfico construído é apresentado a seguir.



(c) Calcule a média, a mediana, o desvio padrão, o máximo e o mínimo das notas obtidas no teste, para cada um dos programas de treinamento físico. Construa uma tabela para apresentar os resultados encontrados.

Cálculo das medidas resumo: Clique no menu “Estatísticas” → “Resumos” → “Resumos numéricos”. Na aba “Dados” selecione a variável “Notas”, em “Resuma por grupos” selecione a variável “Programa” e clique em “OK”. Na aba “Estatísticas” selecione a média, o desvio padrão e os quartis. No campo correspondente aos quartis preencha com “0, 0.5, 1” (Lembre que no R o separador decimal é o “.”, e aqui estamos pedindo os quantis 0, 0.5 e 1, ou seja, o mínimo, a mediana e o máximo). Clique em “OK” e finalize clicando em “OK”.

O resultado é apresentado a seguir.

Programa	média	desvio padrão	mínimo	mediana	máximo	n
A	60.37408	10.13284	34.28397	59.75881	94.28377	100
B	65.03736	8.844297	43.93737	64.50737	86.87783	100

(d) Com base nos resultados dos itens (b) e (c), compare, descritivamente, as notas obtidas no teste, entre os dois programas de condicionamento físico.

Analisando os boxplots construídos no item (b) observamos que as distribuições das notas dos dois programas são simétricas e que a mediana das notas dos alunos do programa B é um pouco maior do que a dos alunos do programa A. No programa A observamos quatro adultos jovens (nas linhas #74, #45, #35 e #92) com notas discrepantes, sendo que três apresentaram notas abaixo de 40 e um apresentou nota acima de 90. No programa B não observamos nenhum adulto jovem com nota discrepante.

Analisando as medidas resumo observamos que a média das notas do programa A é menor do que a do programa B, porém a variabilidade das notas dos alunos do programa A é maior do que a dos alunos do programa B. A menor nota e a maior nota foram observadas no programa A.

(e) Construa um intervalo de 98% de confiança para a nota média populacional de adultos jovens submetidos a cada programa. Se não houver sobreposição (intersecção) entre os intervalos construídos, conclua que há diferença entre as notas médias de adultos jovens submetidos dos dois programas de treinamento físico. Baseado nesse critério estabeleça sua conclusão.

Defina as seguintes variáveis aleatórias:

$$X : \text{Nota obtida por um aluno do programa A}$$
$$Y : \text{Nota obtida por um aluno do programa B.}$$

Sejam μ_X e σ_X o valor esperado e o desvio padrão de X, e μ_Y e σ_Y o valor esperado e o desvio padrão de Y. Denote por $\overline{X}$ e S_X, e $\overline{Y}$ e S_Y a média amostral e o desvio padrão amostral de X e Y, respectivamente.

Do item (c), obtemos que $\overline{x} = 60,3741$, $\overline{y} = 65,0374$, $s_x = 10,1328$ e $s_y = 8,8443$. Note que o enunciado não nos fornece informação sobre as distribuições de X e Y. Assumindo que $n = 100$ (ambos os grupos possuem a mesma quantidades de unidades observacionais) é um tamanho amostral suficientemente grande, podemos aproximar as distribuições amostrais de $\overline{X}$ e $\overline{Y}$ pela distribuição normal utilizando o Teorema Limite Central de forma que

$$\overline{X} \sim \text{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right) \quad \text{e} \quad \overline{Y} \sim \text{N}\left(\mu_Y, \frac{\sigma_Y^2}{n}\right), \quad \text{aproximadamente.}$$

Logo, um intervalo de confiança aproximado para μ_X com coeficiente de confiança γ é dado por

$$\text{IC}[\mu_X; \gamma] = \left[\overline{X} - z\frac{S_X}{\sqrt{n}};\ \ \overline{X} + z\frac{S_X}{\sqrt{n}}\right],$$

e para μ_Y é dado por

$$\text{IC}[\mu_Y; \gamma] = \left[\overline{Y} - z\frac{S_Y}{\sqrt{n}};\ \ \overline{Y} + z\frac{S_Y}{\sqrt{n}}\right],$$

em que z é o quantil da distribuição normal padrão tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$, sendo $Z \sim \mathrm{N}(0, 1)$.

Sendo $\gamma = 0,98$, segue que

$$\mathrm{P}(-z \leq Z \leq z) = 0,98,$$

e, portanto, o quantil z deixa uma probabilidade acumulada à esquerda igual a $0,99$. Verificando a tabela da distribuição normal padrão obtemos que $z = 2,33$.

Logo, uma estimativa intervalar aproximada para μ_X com coeficiente de confiança $\gamma = 0,98$ é dada por

$$\begin{aligned}
\mathrm{IC}[\mu_X; 0,98] &= \left[\overline{x} - z\frac{s_x}{\sqrt{n}};\;\; \overline{x} + z\frac{s_x}{\sqrt{n}}\right] \\
&= \left[60,3741 - 2,33 \times \frac{10,1328}{\sqrt{100}};\;\; 60,3741 + 2,33 \times \frac{10,1328}{\sqrt{100}}\right] \\
&= [60,3741 - 2,3609;\;\; 60,3741 + 2,3609] \\
&= [58,0132;\;\; 62,7350]
\end{aligned}$$

e uma estimativa intervalar aproximada para μ_Y com coeficiente de confiança $\gamma = 0,98$ é dada por

$$\begin{aligned}
\mathrm{IC}[\mu_Y; 0,98] &= \left[\overline{y} - z\frac{s_y}{\sqrt{n}};\;\; \overline{y} + z\frac{s_y}{\sqrt{n}}\right] \\
&= \left[65,0374 - 2,33 \times \frac{8,8443}{\sqrt{100}};\;\; 65,0374 + 2,33 \times \frac{8,8443}{\sqrt{100}}\right] \\
&= [65,0374 - 2,0607;\;\; 65,0374 + 2,0607] \\
&= [62,9767;\;\; 67,0981]
\end{aligned}$$

Como os intervalos não se interceptam concluímos que, estatisticamente, existe diferença entre as notas médias dos programas A e B com coeficiente de confiança de 98%.