ME414 - Estatística para Experimentalistas

Parte 19

Inferência para duas populações: intervalo de confiança para duas médias

Intervalo de Confiança para Duas Médias

População 1: Coletamos uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma população com média μ_1 e a variância σ_1^2 e usamos $\bar{X}$ para estimar μ_1.

População 2: Coletamos uma amostra aleatória $Y_1, Y_2, \ldots, Y_m$ de uma população com média μ_2 e a variância σ_2^2 e usamos $\bar{Y}$ para estimar μ_2.

A população 1 é independente da população 2.

Condições:

1. As populações 1 e 2 são aproximadamente normais ou
2. Os tamanhos amostrais n e m são suficientemente grandes.

Se pelo menos uma das condições acima é satisfeita, temos:

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \quad \text{e} \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Intervalo de Confiança para Duas Médias

Caso 1: Variâncias diferentes e conhecidas

Assumindo que as duas amostras $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ são independentes com $\sigma_1^2 \neq \sigma_2^2$ conhecidas, temos:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

E daí,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$$

Intervalo de Confiança para Duas Médias

Do resultando anterior, similar com o que fizemos para uma média, podemos construir um IC de $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ da seguinte forma:

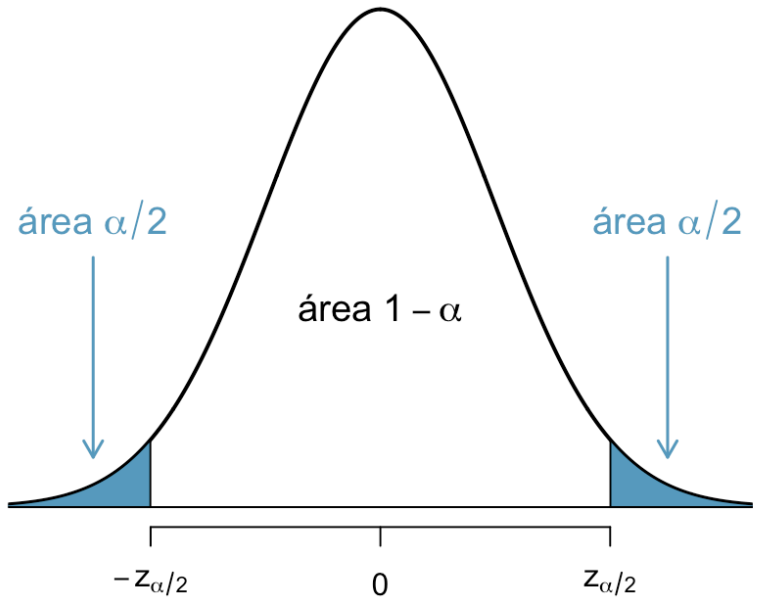
$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \leq z_{\alpha/2}\right) = 1-\alpha$$

Portanto, um IC de $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ é dado por:

$$IC(\mu_1-\mu_2, 1-\alpha) = (\bar{x}-\bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}$$

Relembrando: Como encontrar $z_{\alpha/2}$

$$P(|Z| \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$



Procure na tabela o valor de z tal que a probabilidade acumulada até o valor de z, isto é $P(Z \leq z) = \Phi(z)$, seja $1 - \alpha/2$.

Intervalo de Confiança para Duas Médias

Caso 2: Variâncias iguais e conhecidas

Considere o caso particular em que as variâncias são conhecidas e idênticas, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Então,

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0, 1)$$

E um IC de $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ é dado por:

$$IC(\mu_1 - \mu_2, 1 - \alpha) = (\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}$$

Intervalo de Confiança para Duas Médias

Caso 3: Variâncias iguais e desconhecidas

E se as variâncias das duas populações são idênticas porém desconhecidas, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, σ^2 desconhecida?

Assim como no caso de uma média com variância desconhecida, usamos uma estimativa de σ^2 e a distribuição normal é substituída pela distribuição t.

No caso de duas populações, o estimador da variância σ^2 é a combinação das variâncias amostrais de cada população, ou seja,

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

sendo S_i^2 é a variância amostral da população i.

Intervalo de Confiança para Duas Médias

Então temos: $$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t_{n+m-2}$$

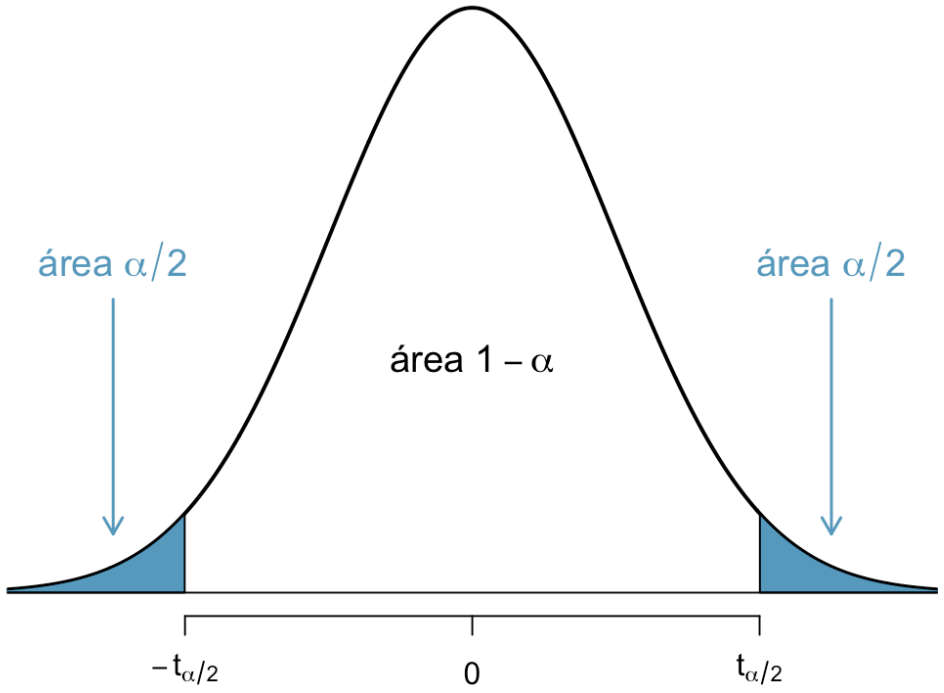
E um IC de $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ é dado por:

$$IC(\mu_1 - \mu_2, 1-\alpha) = (\bar{x} - \bar{y}) \pm t_{n+m-2,\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}$$

Observação: Se n e m são pequenos, as duas amostras devem vir de populações aproximadamente normais. Se n e m são grandes, então a distribuição t com $n + m - 2$ graus de liberdade aproxima-se de uma normal.

Relembrando: Como encontrar $t_{\nu,\alpha/2}$

$$P(-t_{\nu,\alpha/2} < T < t_{\nu,\alpha/2}) = 1 - \alpha$$



Nesse caso, $\nu = n + m - 2$ e os valores da distribuição t encontram-se tabelados.

Intervalo de Confiança para Duas Médias

Resumindo:

Variâncias	Margem de Erro	$IC(\mu_1 - \mu_2, 1-\alpha)$
Diferentes e conhecidas ($\sigma_1^2 \neq \sigma_2^2$)	$z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$	$(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$
Iguais e conhecidas ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)	$z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}$	$(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}$
Iguais e desconhecidas ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)	$t_{n+m-2,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$	$(\bar{x} - \bar{y}) \pm t_{n+m-2,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}$

Exemplo: Tempos de Incubação

Suspeita-se que o tempo de incubação do vírus 1 é maior que o do vírus 2.

Realizaram um estudo de controle e os tempos de incubação (em meses) desses dois vírus foram registrados.

Sabe-se que:

- O tempo de incubação do vírus 1 segue uma distribuição normal com média μ_1 e desvio padrão $\sigma_1 = \sqrt{2}$.
- O tempo de incubação do vírus 2 segue uma distribuição normal com média μ_2 e desvio padrão $\sigma_2 = 1$.
- Os tempos de incubação de ambos os vírus são considerados independentes.

Construa um IC de 95% para a diferença do tempo médio de incubação entre os vírus, isto é, $\mu_1 - \mu_2$.

Exemplo: Tempos de Incubação

Os tempos de incubação (em meses) registrados foram:

X: tempo de incubação do vírus 1 (20 observações)

```
##  [1] 4.56 3.72 3.45 2.86 4.03 4.08 6.56 4.31 0.42 5.56 5.92 2.65 4.54 4.04 4.23
## [16] 6.24 6.16 5.46 3.22 2.28
```

Y: tempo de incubação do vírus 2 (22 observações)

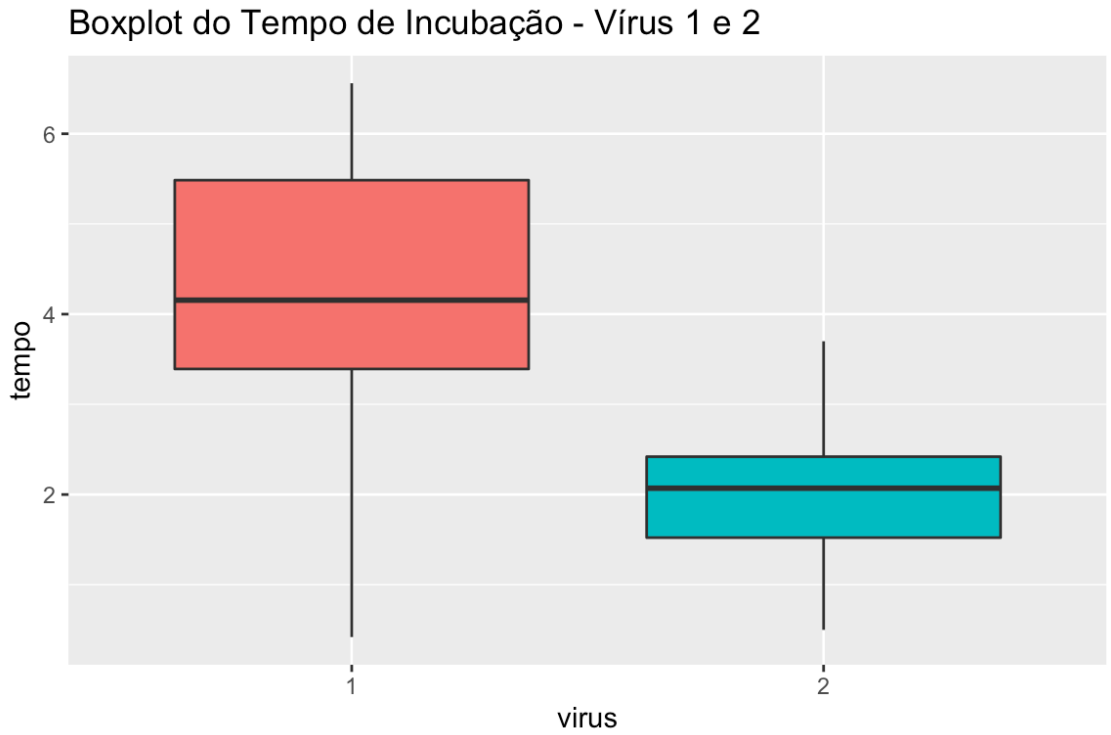
```
##  [1] 2.44 1.49 2.68 2.60 1.51 1.60 1.47 3.70 2.22 1.78 2.36 1.56 2.98 3.33 2.22
## [16] 0.58 2.26 2.26 1.92 0.50 1.17 1.70
```

Pelo enunciado, as duas populações são normais e as variâncias são diferentes mas conhecidas: $\sigma_1^2 = 2$ e $\sigma_2^2 = 1$.

Além disso, $n = 20$ e $m = 22$.

Exemplo: Tempos de Incubação

Calculamos as médias amostrais das duas populações: $\bar{x} = 4.21$ e $\bar{y} = 2.02$.



Exemplo: Tempos de Incubação

Portanto, um Intervalo de 95% de confiança para $\mu_1 - \mu_2$ é dado por:

$$
\begin{aligned}
IC(\mu_1 - \mu_2, 0.95) &= (\bar{x} - \bar{y}) \pm z_{0.025}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \\
&= (4.21 - 2.02) \pm 1.96\sqrt{\frac{2}{20} + \frac{1}{22}} \\
&= 2.19 \pm 1.96 \times 0.38 \\
&= 2.19 \pm 0.75 \\
&= [1.44; 2.94]
\end{aligned}
$$

Interpretação: Com grau de confiança igual a 95%, estimamos que a diferença entre o tempo médio de incubação do vírus 1 para o vírus 2 está entre 1.44 e 2.94 meses.

Exemplo: Tecidos

Dois tipos diferentes de tecido devem ser comparados. Uma máquina de testes *Martindale* pode comparar duas amostras ao mesmo tempo. Os pesos (em miligramas) para sete experimentos foram:

Tecido	1	2	3	4	5	6	7
A	36	26	31	38	28	20	37
B	39	27	35	42	31	39	22

Construa um IC de 95% para a diferença entre os pesos médios dos tecidos. Admita que a variância é a mesma, e igual a 49.

Quais outras suposições são necessárias para que o IC seja válido?

Adaptado de: Profa. Nancy Garcia, Notas de aula.

Exemplo: Tecidos

Os tecidos do tipo A tem uma média amostral igual a $\bar{x}_A = 30.86$. Já os tecidos do tipo B têm média amostral de $\bar{x}_B = 33.57$.

A variância populacional é igual a 49, enquanto as variâncias amostrais são 44.14 e 52.62, respectivamente.

Suposições: Como os tamanhos amostrais $n = m = 7$ são pequenos, devemos assumir os pesos dos tecidos dos dois tipos são normalmente distribuídos ou seja, $X_A \sim N(\mu_A, \sigma^2)$ e $X_B \sim N(\mu_B, \sigma^2)$. Além disso são independentes e com variâncias iguais.

Exemplo: Tecidos

Assumindo que as variâncias são iguais e **conhecidas** ($\sigma_1^2 = \sigma_2^2 = 49$), um IC de 95% para $\mu_A - \mu_B$ é dado por:

$$
\begin{aligned}
IC(\mu_A - \mu_B, 0.95) &= (\bar{x}_A - \bar{x}_B) \pm z_{0.025}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= (30.86 - 33.57) \pm 1.96\sqrt{49\left(\frac{1}{7} + \frac{1}{7}\right)} \\
&= -2.71 \pm 1.96 \times 3.74 \\
&= -2.71 \pm 7.33 \\
&= [-10.04; 4.62]
\end{aligned}
$$

Portanto, com um grau de confiança de 95%, estimamos que a diferença entre os pesos médios dos tecidos do tipo A e tipo B está entre -10.04 e 4.62mg.

Exemplo: Tecidos

Vamos assumir agora que a variância populacional não fosse conhecida.

Assumindo ainda que as variâncias são iguais mas **desconhecidas**, vamos então estimar a variância amostral combinada.

Sabendo que $s_1^2 = 44.14$, $s_2^2 = 52.62$ e $n = m = 7$ temos:

$$\begin{aligned} s_p^2 &= \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} \\ &= \frac{(7-1)44.14 + (7-1)52.62}{7+7-2} \\ &= 48.38 \end{aligned}$$

Exemplo: Tecidos

Nesse caso, um IC de 95% para $\mu_A - \mu_B$ é dado por:

$$
\begin{aligned}
IC(\mu_A - \mu_B, 0.95) &= (\bar{x}_A - \bar{x}_B) \pm t_{n+m-2,0.025}\sqrt{s_p^2\left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= (30.86 - 33.57) \pm 2.18\sqrt{48.38\left(\frac{1}{7} + \frac{1}{7}\right)} \\
&= -2.71 \pm 2.18 \times 3.72 \\
&= -2.71 \pm 8.11 \\
&= [-10.82; 5.4]
\end{aligned}
$$

Portanto, com um grau de confiança de 95%, estimamos que a diferença entre os pesos médios dos tecidos do tipo A e tipo B está entre -10.82 e 5.4mg.

Note que a margem de erro desse IC é maior que o caso anterior.

Exemplo: Tempo de Adaptação

Num estudo comparativo do tempo médio de adaptação (em anos), uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatística	Homens	Mulheres
Média	3.2	3.7
Desvio Padrão	0.8	0.9

Construa um IC de 95% para a diferença entre o tempo médio de adaptação para homens e mulheres.

Fonte: Adaptado de Morettin & Bussab, Estatística Básica 5^a edição, pág 365.

Exemplo: Tempo de Adaptação

Veja que não sabemos a variância populacional, mas temos os desvios padrão amostrais e estes são bem próximos. Então iremos assumir que as variâncias são iguais porém desconhecidas.

Nesse caso, vamos então estimar a variância amostral combinada.

Sabendo que $s_H = 0.8$, $s_M = 0.9$ e $n = m = 50$ temos:

$$\begin{aligned} s_p^2 &= \frac{(n-1)s_H^2 + (m-1)s_M^2}{n+m-2} \\ &= \frac{(50-1)(0.8)^2 + (50-1)(0.9)^2}{50+50-2} \\ &= 0.73 \end{aligned}$$

Exemplo: Tempo de Adaptação

Nesse caso, um IC de 95% para $\mu_H - \mu_M$ é dado por:

$$
\begin{aligned}
IC(\mu_H - \mu_M, 0.95) &= (\bar{x}_H - \bar{x}_M) \pm t_{n+m-2,0.025} \sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)} \\
&= (3.2 - 3.7) \pm 1.98 \sqrt{0.73 \left(\frac{1}{50} + \frac{1}{50}\right)} \\
&= -0.5 \pm 1.98 \times 0.17 \\
&= -0.5 \pm 0.34 \\
&= [-0.84; -0.16]
\end{aligned}
$$

Com um grau de confiança de 95%, estimamos que a diferença entre os tempos médios de adaptação entre homens e mulheres está entre -0.84 e -0.16 anos, ou seja, aproximadamente entre 2 e 10 meses a mais para as mulheres.

Intervalo de Confiança para Duas Proporções

Considere $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ duas amostras independentes de ensaios de Bernoulli tal que $X \sim b(p_1)$ e $Y \sim b(p_2)$, com probabilidade p_1 e p_2 de apresentarem uma certa característica.

Queremos encontrar um IC de confiança para a diferença entre as proporções p_1 e p_2, ou seja, um IC para $p_1 - p_2$.

Em aulas anteriores vimos que:

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{e} \quad \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

Como as variâncias de $\hat{p}_1$ e $\hat{p}_2$ dependem de p_1 e p_2 e, portanto, não são conhecidas, iremos usar uma estimativa dessas variâncias.

Intervalo de Confiança para Duas Proporções

Ou seja,

$$\hat{p}_1 \sim N\left(p_1, \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}\right) \quad \text{e} \quad \hat{p}_2 \sim N\left(p_2, \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)$$

Condições: Todas as quantidades $n_1\hat{p}_1,\ n_1(1-\hat{p}_1),\ n_2\hat{p}_2$ e $n_2(1-\hat{p}_2)$ devem ser pelo menos igual a 10 para que a aproximação pela normal seja válida.

Então,

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

Intervalo de Confiança para Duas Proporções

Similar com o que fizemos para uma proporção, podemos então construir um IC de $100(1-\alpha)\%$ para $p_1 - p_2$ da seguinte forma:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Então, um IC de $100(1-\alpha)\%$ para $p_1 - p_2$ é dado por:

$$IC(p_1 - p_2, 1-\alpha) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Exemplo: Embalagens de Sabonetes

Para o lançamento da nova embalagem de um sabonete a divisão de criação estuda duas propostas:

- A: amarela com letras vermelhas ou
- B: preta com letras douradas

Eles acreditam que a proposta A chama mais a atenção que B.

Realizaram uma pesquisa em dois supermercados “semelhantes” e perguntaram para um total de 1000 clientes se eles haviam notado a embalagem e então pediram para descrevê-la. Os resultados estão na tabela seguir.

Exemplo: Embalagens de Sabonetes

Proposta	Notaram	Não Notaram	Total
A	168	232	400
B	180	420	600
Total	348	652	1000

- Seja p_A a proporção de pessoas que notaram a proposta A e p_B a proporção de pessoas que notaram a proposta B.
- Encontre um IC de 95% para $p_A - p_B$.
- Usando os dados da tabela:

$$\hat{p}_A = \frac{168}{400} = 0.42 \quad \text{e} \quad \hat{p}_B = \frac{180}{600} = 0.3$$

Exemplo: Embalagens de Sabonetes

Veja que as condições são satisfeitas.

Então um IC de 95% para $p_A - p_B$ é dado por:

$$
\begin{aligned}
IC(p_A - p_B, 0.95) &= (\hat{p}_A - \hat{p}_B) \pm z_{0.025}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}} \\
&= (0.42 - 0.3) \pm 1.96\sqrt{\frac{0.42(0.58)}{400} + \frac{0.3(0.7)}{600}} \\
&= 0.12 \pm 1.96 \times 0.031 \\
&= 0.12 \pm 0.061 \\
&= [0.059; 0.181]
\end{aligned}
$$

Portanto, com um grau de confiança de 95%, estimamos que a diferença entre as proporções p_A e p_B está entre 0.059 e 0.181.

Exemplo: Ensaio Clínico

Um ensaio clínico é realizado para avaliar um novo tipo de tratamento contra uma doença e comparar os resultados com aqueles obtidos usando o tratamento tradicional.

- Dos 50 pacientes tratados com o tratamento novo, 36 se curaram.
- Dos 45 pacientes tratados com o tratamento antigo, 29 se curaram.

Seja p_1 a proporção de curados com o tratamento novo e p_2 a proporção de curados com o tratamento antigo.

Encontre um IC de 99% para $p_1 - p_2$.

Exemplo: Ensaio Clínico

A proporção de curados pelos tratamentos novo e antigo são, respectivamente:

$$\hat{p}_1 = \frac{36}{50} = 0.72 \quad \text{e} \quad \hat{p}_2 = \frac{29}{45} = 0.64$$

Então um IC de 99% para $p_1 - p_2$ é dado por:

$$\begin{aligned} IC(p_1 - p_2, 0.99) &= (0.72 - 0.64) \pm z_{0.005}\sqrt{\frac{0.72(0.28)}{50} + \frac{0.64(0.36)}{45}} \\ &= 0.08 \pm 2.58 \times 0.096 \\ &= 0.08 \pm 0.25 \\ &= [-0.17; 0.33] \end{aligned}$$

Portanto, com um grau de confiança de 99%, estimamos que a diferença entre as proporções de curados pelos tratamentos novo e antigo está entre -0.17 e 0.33.

Leituras

- Ross: capítulo 10.
- OpenIntro: seção 5.3.1. e 6.2.2
- Magalhães: capítulo 9.

Slides produzidos pelos professores:

- Samara Kiihl
- Tatiana Benaglia
- Larissa Matos
- Benilton Carvalho