Intervalo de Confiança para diferença de médias– Aula 03

Statistics for Business and Economics 7 edição, by Paul Newbold , William Carlson , Betty Thorne (cap. Estimation: Additional Topics)

Cap 12 Bussab e Morettin

Statistics for Economics, Accounting and Business Studies, capítulo 4, Barrow

Marislei Nishijima

Amostras Dependentes

Amostras dependentes

Teste de média de 2 Populações Relacionadas

- Amostras emparelhadas ou combinadas
- Medidas repetidas (antes / depois)
- Use a diferença entre os valores emparelhados:

$$d_i = x_i - y_i$$

- Elimina a variação entre objetos
- Hipótese:
  - Ambas populações são normalmente distribuídas

Diferença de média

Amostras dependentes

A $i^{ésima}$ diferença emparelhada é d_i , sendo

$$d_i = x_i - y_i$$

A estimative pontual para a diferença da média da população emparelhada é $\bar{d}$:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

O desvio padrão da diferença amostral é:

$$S_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

n é o número de pares combinados numa mesma amostra

Intervalo de Confiança para a diferença de Médias

Amostras dependentes

IC para a diferença de Médias populacionais, μ_d, é

$$\bar{d} - t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}$$

Sendo

n = tamanho da amostra

(número de pares combinados numa amostra emparelhada)

IC para a diferença de Médias

(cont.)

Amostras dependentes

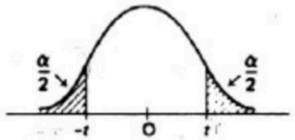
- A margem de erro é

$$ME = t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$$

- $t_{n-1,\alpha/2}$ é o valor da distribuição t de Student com (n – 1) graus de Liberdade para a probabilidade

$$P(t_{n-1} > t_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

TABELA DA DISTRIBUIÇÃO t de STUDENT



φ \ α	0,50	0,25	0,10	0,05	0,025	0,01	0,005
1	1,00000	2,4142	6,3138	12.706,	25,542	63,657	127,32
2	0,81650	1,6036	2,9200	4,3127	6,2053	9,9248	14,089
3	0,76489	1,4226	2,3534	3,1825	4,1765	5,8409	7,4533
4	0,74070	1,3444	2,1318	2,7764	3,4954	4,6041	5,5976
5	0,72669	1,3009	2,0150	2,5706	3,1634	4,0321	4,7733
6	0,71756	1,2733	1,9432	2,4469	2,9687	3,7074	4,3168
7	0,71114	1,2543	1,8946	2,3646	2,8412	3,4995	4,0293
8	0,70639	1,2403	1,8595	2,3060	2,7515	3,3554	3,8325
9	0,70272	1,2297	1,8331	2,2622	2,6850	3,2498	3,6897
10	0,69981	1,2213	1,8125	2,2281	2,6338	3,1693	3,5814
11	0,69745	1,2145	1,7959	2,2010	2,5931	3,1058	3,4966
12	0,69548	1,2089	1,7823	2,1788	2,5600	3,9545	3,4284
13	0,69384	1,2041	1,7709	2,1604	2,5326	3,0123	3,3725
14	0,692	1,2001	1,7613	2,1448	2,5096	2,9768	3,3257
15	0,69120	1,1967	1,7530	2,1315	2,4899	2,9467	3,2860
16	0,69013	1,1937	1,7459	2,1199	2,4729	2,9208	3,2520
17	0,68919	1,1910	1,7396	2,1098	2,4581	2,8982	3,2225
18	0,68837	1,1887	1,7341	2,1009	2,4450	2,8784	3,1966
19	0,68763	1,1866	1,7291	2,0930	2,4334	2,8609	3,1737
20	0,68696	1,1848	1,7247	2,0860	2,4231	2,8453	3,1534
21	0,68635	1,1831	1,7207	2,0796	2,4138	2,8314	3,1352
22	0,68580	1,1816	1,7171	2,0739	2,4055	2,8188	3,1188
23	0,68531	1,1802	1,7139	2,0687	2,3979	2,8073	3,1040
24	0,68485	1,1789	1,7109	2,0639	2,3910	2,7969	3,0905
25	0,68443	1,1777	1,7081	2,0595	2,3846	2,7874	3,0782
26	0,68405	1,1766	1,7056	2,0555	2,3788	2,7787	3,0669
27	0,68370	1,1757	1,7033	2,0518	2,3734	2,7707	3,0565
28	0,68335	1,1748	1,7011	2,0484	2,3685	2,7633	3,0469
29	0,68304	1,1739	1,6991	2,0452	2,3638	2,7564	3,0380
30	0,68276	1,1731	1,6973	2,0423	2,3596	2,7500	3,0298
40	0,68066	1,1673	1,6839	2,0211	2,3289	2,7045	2,9712
60	0,67862	1,1616	1,6707	2,0003	2,2991	2,6603	2,9146
120	0,67656	1,1559	1,6577	1,9799	2,2699	2,6174	2,8599
∞	0,67449	1,1503	1,6449	1,9600	2,2414	2,5758	2,8070

Extraída do livro "Curso de Estatística"-Jairo Simon da Fonseca & Gilberto de Andrade Martins-Editora Atlas

Exemplo de amostras dependentes

Amostras dependentes

Seis pessoas se inscreveram em um programa de perda de peso. Você coleta os seguintes dados:

Peso: Pessoa	Antes(x)	Depois (y)	Diferença, d_i
1	136	125	11
2	205	195	10
3	157	150	7
4	138	140	- 2
5	175	165	10
6	166	160	6
			42

$$\bar{d} = \frac{\sum d_i}{n}$$

$= 7.0$

$$S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

$= 4.82$

Exemplo de amostras dependentes (emparelhadas)

(cont.)

Amostras dependentes

- Para um IC de 95%, o valor t apropriado t é $t_{n-1,\alpha/2} = t_{5,.025} = 2.571$
- O IC de 95% para a diferença de médias, μ_d, é

$$\bar{d} - t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2}\frac{S_d}{\sqrt{n}}$$

$$7 - (2.571)\frac{4.82}{\sqrt{6}} < \mu_d < 7 + (2.571)\frac{4.82}{\sqrt{6}}$$

$$-1.94 < \mu_d < 12.06$$

Como este intervalo contém zero, não podemos ter 95% de confiança, dados esses dados limitados, de que o programa de perda de peso ajuda as pessoas a perderem peso

Diferenças entre duas medias: Amostras Independentes

Médias população, amostras independentes

Objetivo: Construir um intervalo de confiança para a diferença entre duas médias populacionais, $\mu_x - \mu_y$

- Diferente fontes de dados
  - Não relacionados
  - Independente
    - A amostra selecionada de uma população não tem efeito sobre a amostra selecionada de outra população
    - A estimativa pontual é a diferença entre as duas médias amostrais:

$$\boxed{\bar{x} - \bar{y}}$$

Diferenças entre duas medias: Amostras Independentes

(cont.)

- Médias população, amostras independentes
  - σ_x^2 e σ_y^2 conhecidas → IC usar $\mathbf{z_{\alpha/2}}$
  - σ_x^2 e σ_y^2 desconhecidas
    - σ_x^2 e σ_y^2 assumido serem iguais → IC usar valores da distribuição t de Student ou $\mathbf{t_{(n-1),\alpha/2}}$
    - σ_x^2 e σ_y^2 assumido diferentes → IC usar valores da distribuição t de Student ou $\mathbf{t_{(n-1),\alpha/2}}$

σ_x^2 e σ_y^2 conhecidos

Médias população, amostras independentes

- σ_x^2 e σ_y^2 conhecidas *
- σ_x^2 e σ_y^2 desconhecidas

Hipóteses:

- As amostras são aleatórias e independentemente retiradas
- ambas as distribuições de população são normais
- As variâncias populacionais são conhecidas

σ_x^2 e σ_y^2 conhecidas

(cont.)

Médias população, amostras independentes

- σ_x^2 e σ_y^2 conhecidas *
- σ_x^2 e σ_y^2 desconhecidas

Quando σ_x e σ_y são conhecidos e ambas populações são normais, a variância de $\overline{X} - \overline{Y}$ é

$$\sigma^2_{\overline{X}-\overline{Y}} = \frac{\sigma_x^{\ 2}}{n_x} + \frac{\sigma_y^{\ 2}}{n_y}$$

…e a variável aleatória

$$Z = \frac{(\overline{x} - \overline{y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_x^2}{n_X} + \dfrac{\sigma_y^2}{n_Y}}}$$

tem distribuição normal padrão

IC, σ_x^2 e σ_y^2 conhecidas

Médias população, amostras independentes

- σ_x^2 e σ_y^2 conhecidas
- σ_x^2 e σ_y^2 desconhecidas

* O IC para $\mu_x - \mu_y$ é:

$$(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_y}} < \mu_X - \mu_Y < (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_y}}$$

σ_x^2 e σ_y^2 desconhecidas, Assumidas iguais

- Médias população, amostras independentes
  - σ_x^2 e σ_y^2 conhecidas
  - σ_x^2 e σ_y^2 desconhecidas
    - σ_x^2 e σ_y^2 supostas iguais *
    - σ_x^2 e σ_y^2 supostas diferentes

Hipóteses:

- As amostras são aleatórias e retiradas de forma independente
- Populações são normalmente distribuídas
- As variações populacionais são desconhecido, mas supostas iguais

σ_x^2 e σ_y^2 desconhecidas, Assumidas iguais

(cont.)

- Médias população, amostras independentes
  - σ_x^2 e σ_y^2 conhecidas
  - σ_x^2 e σ_y^2 desconhecidas
    - σ_x^2 e σ_y^2 supostas iguais *
    - σ_x^2 e σ_y^2 supostas diferentes

Criando estimativas por intervalo:

- as variâncias populacionais são assumidas iguais, usar então os dois desvio-padrão amostrais e juntá-los para estimar σ

- use o t valor com $(n_x + n_y - 2)$ graus de liberdade

σ_x^2 e σ_y^2 desconhecidas, Assumidas iguais

(cont.)

Médias população, amostras independentes

σ_x^2 e σ_y^2 conhecidas

σ_x^2 e σ_y^2 desconhecidas

σ_x^2 e σ_y^2 supostas iguais *

σ_x^2 e σ_y^2 supostas diferentes

A variância conjunta é

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

σ_x^2 e σ_y^2 desconhecidas, Assumidas iguais

σ_x^2 e σ_y^2 conhecidas

σ_x^2 e σ_y^2 supostas iguais *

σ_x^2 e σ_y^2 supostas diferentes

O IC para $\mu_1 - \mu_2$ é:

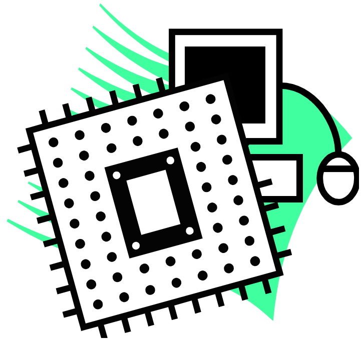
$$(\bar{x}-\bar{y}) - t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}} < \mu_X - \mu_Y < (\bar{x}-\bar{y}) + t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

Where $$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}$$

Exemplo: Variância conjunta

Você está testando a velocidade de dois processadores de computador. Forme um intervalo de confiança para a diferença na velocidade da CPU. Você coleta os seguintes dados de velocidade (em Mhz):

	CPU_x	CPU_y
Quant. Testada	**17**	**14**
Média amostra	**3004**	**2538**
Des-Pad amostra	**74**	**56**



Assuma que ambas as populações são normalmente distribuídas e com variâncias iguais. Use 95% de confiânça.

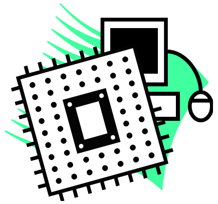
Calculando a Variância conjunta

a Variância conjunta é:

$$S_p^2 = \frac{(n_x - 1)S_x^{\,2} + (n_y - 1)S_y^{\,2}}{(n_x - 1) + (n_y - 1)} = \frac{(17-1)74^2 + (14-1)56^2}{(17-1) + (14-1)} = \boxed{4427.03}$$

O t valor para um IC de 95% é:

$$\boxed{t_{n_x + n_y - 2,\, \alpha/2} = t_{29,\, 0.025} = 2.045}$$



Calculando os limites do Intervalo

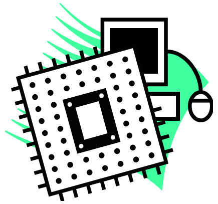
- O IC de 95% é:

$$(\bar{x}-\bar{y})-t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}<\mu_X-\mu_Y<(\bar{x}-\bar{y})+t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$$

$$(3004-2538)-(2.054)\sqrt{\frac{4427.03}{17}+\frac{4427.03}{14}}<\mu_X-\mu_Y<(3004-2538)+(2.054)\sqrt{\frac{4427.03}{17}+\frac{4427.03}{14}}$$

$$\boxed{416.69<\mu_X-\mu_Y<515.31}$$

Estamos 95% confiantes que a diferença de media em velocidade de CPU recai entre 416.69 e 515.31 Mhz.



σ_x^2 e σ_y^2 desconhecidas, supostas diferentes

- Médias população, amostras independentes
  - σ_x^2 e σ_y^2 conhecidas
  - σ_x^2 e σ_y^2 desconhecidas
    - σ_x^2 e σ_y^2 supostas iguais
    - σ_x^2 e σ_y^2 supostas diferentes *

Hipóteses:

- As amostras são aleatórias e retiradas de forma independente
- Populações são normalmente distribuídas
- As variações populacionais são desconhecido, mas supostas diferentes

σ_x^2 e σ_y^2 desconhecidas, supostas diferentes

(cont.)

Médias população, amostras independentes

- σ_x^2 e σ_y^2 conhecidas
- σ_x^2 e σ_y^2 desconhecidas
  - σ_x^2 e σ_y^2 supostas iguais
  - σ_x^2 e σ_y^2 supostas diferentes *

Forming interval estimates:

Criando estimativas por intervalo:

- as variâncias populacionais não são supostas iguais, usar então usar a variância conjunta não é adequado.
- use o t valor com ν graus de Liberdade, sendo

$$\nu = \frac{\left[\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2 / (n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2 / (n_y - 1)}$$

Intervalo de Confiança,
σ_x^2 e σ_y^2 desconhecidas, supostas diferentes

σ_x^2 e σ_y^2 desconhecidas

- σ_x^2 e σ_y^2 supostas iguais
- σ_x^2 e σ_y^2 supostas diferentes *

O IC para $\mu_1 - \mu_2$ é:

$$(\bar{x}-\bar{y}) - t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}} \; < \; \mu_X - \mu_Y \; < \; (\bar{x}-\bar{y}) + t_{\nu,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$$

Sendo

$$\nu = \frac{\left[(\frac{s_x^2}{n_x})+(\frac{s_y^2}{n_y})\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$$

Duas Proporções Populacionais

Proporções Populacionais

Objetivo: Construir um Intervalo de Confiança para a diferença de duas proporções populacionais, $P_x - P_y$

Hipóteses:

Ambas amostras são grandes (geralmente pelo menos 40 observações em cada amostra)

A estimativa pontual para a diferença é $\hat{p}_x - \hat{p}_y$

Duas Proporções Populacionais

(cont.)

Proporções Populacionais

- A variável aleatória

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\dfrac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \dfrac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}$$

É aproximadamente normalmente distribuída

Intervalo de Confiança for duas proporções populacionais

Proporções Populacionais

Os limites de confiança para

$P_x - P_y$ são:

$$(\hat{p}_x - \hat{p}_y) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

Exemplo:
Duas proporções populacionais

Construa um IC de 90% para a diferença entre as proporções de homens e mulheres que tem Ensino Superior completo

- Numa amostra aleatória, 26 de 50 homens e 28 de 40 mulheres obtiveram curso superior.

Exemplo:
Duas proporções populacionais

(cont.)

Homens: $\hat{p}_x = \frac{26}{50} = 0.52$

Mulheres: $\hat{p}_y = \frac{28}{40} = 0.70$

$$\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = \sqrt{\frac{0.52(0.48)}{50} + \frac{0.70(0.30)}{40}} = 0.1012$$

Para 90% confiança, $Z_{\alpha/2} = 1.645$

Exemplo:
Duas proporções populacionais

(cont.)

Os limites de confiança são:

$$(\hat{p}_x - \hat{p}_y) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$$

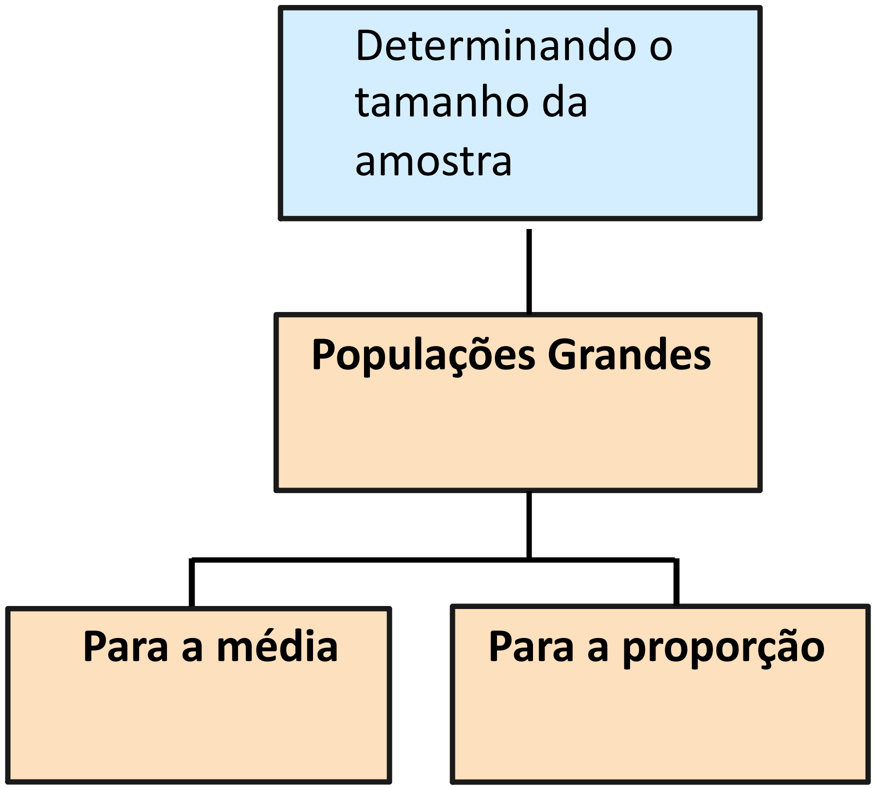
$$= (.52 - .70) \pm 1.645 (0.1012)$$

então o IC é

$$-0.3465 < P_x - P_y < -0.0135$$

Como o interval não inclui zero de estamos 90% confiantes que as duas proporções não são iguais.

Determinando o tamanho da amostra



Margem de Erro

- O tamanho requerido da amostra pode ser calculado para alcançar a margem de erro (ME) desejada com o nível de confiança de $(1 - \alpha)$

- A margem de erro também é conhecida por erro amostral
  - a quantidade de imprecisão na estimativa do parâmetro populacional
  - a quantidade adicionada e subtraída à estimativa pontual para formar o intervalo de confiança

Determinando o tamanho da amostra

Populações Grandes

Para a média

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Margem de Erro (Erro amostral)

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Determinando o tamanho da amostra

(cont.)

Populações Grandes

Para a média

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Resolvendo para n tem-se:

$$\boxed{n = \frac{z_{\alpha/2}^2 \sigma^2}{ME^2}}$$

Determinando o tamanho da amostra

(cont.)

- Para determinar o tamanho da amostra para a média, é preciso conhecer:

  - O nível de confiança desejado ($1 - \alpha$), que determina o valor $z_{\alpha/2}$
  - A margem de erro aceitável (erro de amostragem), ME
  - O desvio-padrão populaciona, σ

Tamanho da amostra requerido

Se σ = 45, qual o tamanho da amostra necessário para estimar a média dentro de ± 5 com 90% de confiança?

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{ME^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19$$

Então o tamanho requerido é **n = 220**

(Sempre arredondado)

Determinando o tamanho da amostra: Proporção da população

População grande

Para a proporção

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margem de Erro (erro amostral)

Determinando o tamanho da amostra: Proporção da população

(cont.)

População grande

Para a proporção

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\hat{p}(1-\hat{p})$ não pode ser maior que 0.25, quando $\hat{p}$= 0.5

Substituir por 0.25 e resolver para n para obter $\hat{p}(1-\hat{p})$

$$\boxed{n = \frac{0.25\, z_{\alpha/2}^2}{ME^2}}$$

Determinando o tamanho da amostra: Proporção da população

(cont.)

- As proporções amostral e populacional, $\hat{p}$ e P, são em geral desconhecidas (nenhuma amostra havia sido retirada)
- P(1 – P) = 0.25 gera uma margem de erro grande (então, garante que o resultado da amostra atenderá o nível de confiança desejado)
- Para determinar o tamanho da amostra necessário para a proporção, deve-se conhecer:
  - O tamanho desejado do nível de confiança $(1 - \alpha)$, que determina o valor crítico $z_{\alpha/2}$
  - A margem de erro de aceitação (margem de erro), ME
  - Estimativa P(1 – P) = 0.25

Tamanho necessário da amostra para a Proporção. Exemplo: Proporção Populacional

Qual o tamanho da amostra seria necessário para estimar a verdadeira proporção de defeitos em uma grande população dentro de ± 3%, com 95% de confiança?

Tamanho necessário da amostra para a Proporção

(cont.)

Solução:

Para 95% de confiança, use $z_{0.025} = 1.96$

ME = 0.03

Estimativa P(1 – P) = 0.25

$$n = \frac{0.25\, z_{\alpha/2}^2}{ME^2} = \frac{(0.25)(1.96)^2}{(0.03)^2} = \boxed{1067.11}$$

Usar n = 1068