# Estatística Aplicada a Recursos Hídricos

Docente: Rachid Muleia

(rachid.muleia@uem.mz)

Mestrado em Gestão de Recursos Hídricos - DGEO/UEM

Tema: Inferência Estatística: Comparação entre médias

Ano lectivo: 2023

# Comparações de Duas Médias

- Objectivo: Testar a significância estatística da diferença $\mu_1 - \mu_2$ entre as médias de duas distribuições de populações diferentes.
- Por exemplo: diferença entre as pontuações médias de duas turmas distintas submetidas a um mesmo teste;
- Possíveis situações na comparação de duas populações:

# Caso 1: Amostras emparelhadas (dependentes)

- Objectivo: comparar duas médias populacionais sendo que, para cada unidade amostral, realizamos duas medições da característica de interesse.
- De modo geral, essas observações correspondem a medidas tomadas (em um único indivíduo) antes e após uma dada intervenção. Essa técnica é conhecida como auto-emparelhamento.
- Para exemplificar, tomaremos um grupo de pessoas que fizeram determinada dieta por uma semana. Medimos o peso no início e no final da dieta, representado pelas v.a.'s $X_i$ e $Y_i$, respectivamente;
- O efeito produzido pela dieta pode ser representado, para o $i-$ésimo indivíduo, pela variável $D_i = Y_i - X_i$. Supondo, para $i = 1, \ldots, n$ ($n$ diferenças),

$$d_i \sim N\left(\mu_d, \sigma_d^2\right)$$

- queremos testar as hipóteses:

$H_0 : \mu_d = 0$ (a dieta não produziu efeito)

$H_1 : \mu_d \neq 0$ ou $\mu_d > 0$ ou $\mu_d < 0$ (a dieta produziu algum efeito)

- O método de análise apropriado é o teste t-pareado

# Caso 1: Amostras emparelhadas (dependentes)

- O estimador "natural" do parâmetro $\mu_d$ é a média amostral $\bar{d}$;
- O estimador da variância $\sigma_d^2$ é a variância amostral $s_d^2$, dado por,

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

- O teste de hipóteses é realizado utilizando-se a seguinte estatística do teste:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}, \quad \text{onde} \quad s_{\bar{d}} = \frac{s_d}{\sqrt{n}}.$$

que, sob $H_0$, segue uma distribuição t-Student com $n-1$ graus de liberdade.

$\bar{d}$ : média da amostra das diferenças;

$\mu_d$ : valor das diferenças entre médias das populações a ser testado;

$s_d$ : desvio padrão da amostra das diferenças;

$n$ : tamanho da amostra das diferenças

# Caso 1: Amostras emparelhadas (dependentes)

Exemplo 1: Um grupo de 10 pessoas é submetido a um tipo de dieta por 10 dias, estando o peso (em Kg) antes do início ($x_i$) e no final da dieta ($y_i$) marcados na tabela abaixo. Ao nível de 5%, podemos concluir que houve diminuição do peso médio pela aplicação da dieta?

# Caso 1: Amostras emparelhadas (dependentes)

Exemplo 1: Um grupo de 10 pessoas é submetido a um tipo de dieta por 10 dias, estando o peso (em Kg) antes do início ($x_i$) e no final da dieta ($y_i$) marcados na tabela abaixo. Ao nível de 5%, podemos concluir que houve diminuição do peso médio pela aplicação da dieta?

Resposta: Definição das hipóteses: Seja $\mu_d = \mu_y - \mu_x$, então

$$H_0 : \mu_d = 0 \quad \text{(a dieta não produziu efeito)}$$
$$H_1 : \mu_d < 0 \quad \text{(houve diminuição do peso médio)}$$

Seja $d_i = y_i - x_i, \ i = 1, \ldots, 10.$

| Pessoa | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 120 | 104 | 93 | 87 | 85 | 98 | 102 | 106 | 88 | 90 |
| $y_i$ | 116 | 102 | 90 | 83 | 86 | 97 | 98 | 108 | 82 | 85 |

| Pessoa | A | B | C | D | E | F | G | H | I | J | $\sum$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $d_i$ | -4 | -2 | -3 | -4 | 1 | -1 | -4 | 2 | -6 | -5 | -26 |
| $d_i^2$ | 16 | 4 | 9 | 16 | 1 | 1 | 16 | 4 | 36 | 25 | 128 |

# Caso 1: Amostras emparelhadas (dependentes)

Exemplo 1 (continuação): $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i;\ \therefore \quad \bar{d} = -\frac{26}{10} = -2,6$

$$s_d^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right\}$$

$$s_d^2 = \frac{1}{9}\left\{128 - \frac{(26)^2}{10}\right\} = 6,71 \rightarrow s_d = \sqrt{6,71} = 2,59$$

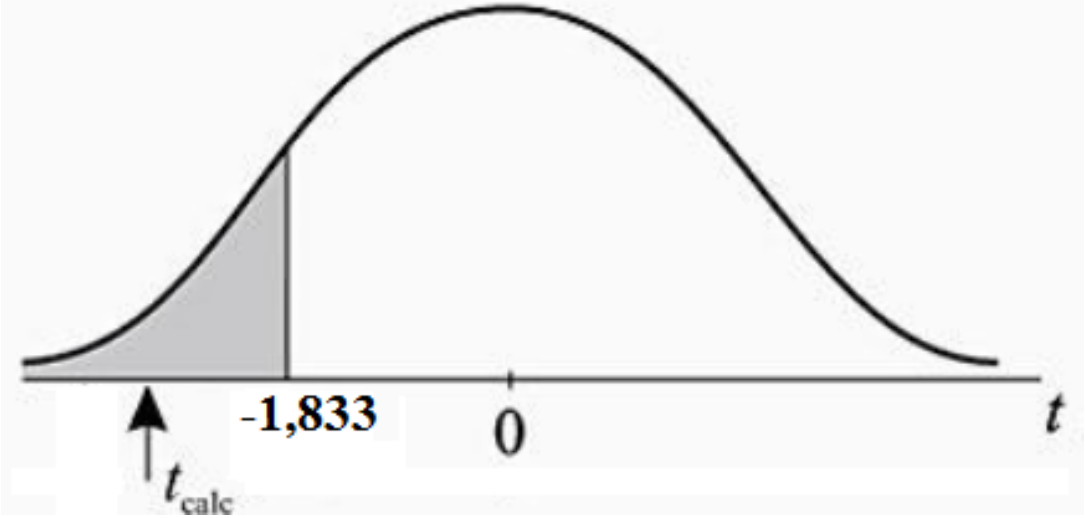
$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{2,59}{\sqrt{10}} = 0,82$$

Então

$$t_{calc} = \frac{\bar{d} - \mu_{d_{H0}}}{s_{\bar{d}}} = \frac{-2,6 - 0}{0,82} = -3,17 \quad \text{e} \quad t_{n-1,\alpha} = t_{9;0,05} = 1,833$$

# Caso 1: Amostras emparelhadas (dependentes)

Exemplo 1 (continuação): Aqui, consideramos $t_{9;0,05}$ com sinal negativo, por se tratar de um teste unilateral a esquerda.




Como $|t_{calc}| > t_{9;0,05}$, rejeita-se $H_0$, isto é, a 95%, concluímos que é a dieta teve um efeito significativo.

# Caso 2: Amostras independentes de populações normais com variâncias conhecidas

Teorema: Consideremos $X_1$ e $X_2$ duas amostras de populações independentes. Se

- $X_1 \sim N(\mu_1, \sigma_1^2) \rightarrow$ amostra de tamanho $n_1$;
- $X_2 \sim N(\mu_2, \sigma_2^2) \rightarrow$ amostra de tamanho $n_2$; Então

$$\bar{x}_d = \bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2) \text{ onde } \sigma_{\bar{x}_1}^2 = \frac{\sigma_1^2}{n_1} \text{ e } \sigma_{\bar{x}_2}^2 = \frac{\sigma_2^2}{n_2}$$

- Observe que a independência entre as amostras foi necessária para obter a variância, uma vez que $Cov(\bar{x}_1, \bar{x}_2) = 0$. Temos, então, que $\sigma_{\bar{x}_d} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Se as populações não são normais e $n_1$, e $n_2$ são grandes ($> 30$), então (pelo TLC)

$$\bar{x}_d = \bar{x}_1 - \bar{x}_2 \cong N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- Genericamente, testaremos as hipóteses:

$$H_0 : \mu_1 - \mu_2 = \mu_0$$
$$H_1 : \mu_1 - \mu_2 \neq \mu_0 \text{ ou } \mu_1 - \mu_2 > \mu_0 \text{ ou } \mu_1 - \mu_2 < \mu_0$$

# Caso 2: Amostras independentes de populações normais

- Se $\mu_0 = 0$, testaremos $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2, \text{ ou } \mu_1 > \mu_2 \text{ ou } \mu_1 < \mu_2 \end{cases}$ .
- A estatística do teste é: $Z_{calc} = \frac{\bar{x}_d - \mu_{H_0}}{\sigma_{\bar{x}_d}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{H_0}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- Se há igualdade de variâncias, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, então $Z_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{H_0}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- Quando as variâncias forem desconhecidas e as amostras grandes usa-se (pelo TLC)

$$\sigma_{\bar{x}_d} = \sigma_(\bar{x}_1 - \bar{x}_2) \cong \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

onde $s_1^2$ e $s_2^2$ são estimativas de $\sigma_1^2$ e $\sigma_2^2$, feitas por meio de amostras de tamanhos $n_1$ e $n_2$. Dessa forma, a estatística do teste é:

$$Z_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_{H_0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

# Caso 2: Amostras independentes de populações normais

Exemplo 2: De duas populações normais $X_1$ e $X_2$ com variâncias $\sigma^2 = 25$, levantaram-se duas amostras de tamanhos $n_1 = 9$ e $n_2 = 16$, obtendo-se: $\sum_{i=1}^{9} = 27$ e $\sum_{j=1}^{16} = 32$. Ao nível de 10%, testar as hipóteses:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

# Caso 2: Amostras independentes de populações normais

Exemplo 2: De duas populações normais $X_1$ e $X_2$ com variâncias $\sigma^2 = 25$, levantaram-se duas amostras de tamanhos $n_1 = 9$ e $n_2 = 16$, obtendo-se: $\sum_{i=1}^{9} = 27$ e $\sum_{j=1}^{16} = 32$. Ao nível de 10%, testar as hipóteses:

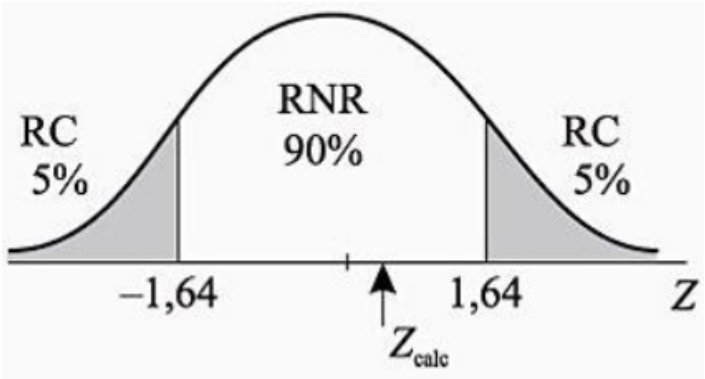
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Resolução:

- 1ª população: $X_1 \sim N(\mu_1, 25)$ $n_1 = 9$; $\bar{x}_1 = \frac{27}{9} = 3$;
- 1ª população: $X_2 \sim N(\mu_2, 25)$ $n_2 = 16$; $\bar{x}_1 = \frac{32}{16} = 2$;
- $\bar{x}_d = \bar{x}_1 - \bar{x}_2 = 3 - 2 = 1$; $\sigma_{\bar{x}_d} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 5\sqrt{\frac{1}{9} + \frac{1}{16}}$
  $\sigma_{\bar{x}_d} = 2,083$
- $Z_{calc} = \frac{\bar{x}_d - \mu_{H_0}}{\sigma_{\bar{x}_d}} = \frac{1-0}{2,083} = 0,48$;

# Caso 2: Amostras independentes de populações normais

- Se $\alpha = 10\% \rightarrow Z_{1-\frac{\alpha}{2}} = 1,64$




- Como $Z_{calc} \in$RNR, não se rejeita $H_0$, isto é, ao nível de 10% não é significativa a diferença entre as médias das duas populações.

- Outra forma de resolução:

$$RNR \rightarrow P(\mu_{H_0} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}_d} \leq \bar{x}_d \leq \mu_{H_0} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}_d}) = 1 - \alpha$$

$$P(0 - 1,64 \cdot 2,083 \leq \bar{x}_d \leq 0 + 1,64 \cdot 2,083) = 1 - \alpha$$

$$RNR = (-3,416; 3,416) \qquad RC = (-\infty; 3,416] \cup [3,416; +\infty)$$

Portanto, $\bar{x}_d = 1 \rightarrow \bar{x}_d \in RNR \rightarrow$ não se rejeita $H_0$.

# Caso 2: Amostras independentes de populações normais

Exemplo 3: Um supermercado não sabe se deve comprar lâmpadas da marca $A$ ou $B$, de mesmo preço. Testa uma amostra de 100 lâmpadas de cada uma das marcas, obtendo:

$$\bar{x}_A = 1.160 \text{ horas} \quad \text{e} \quad s_A = 90 \text{ horas}$$
$$\bar{x}_A = 1.140 \text{ horas} \quad \text{e} \quad s_B = 80 \text{ horas}$$

Ao nível de $2,5\%$, testar a hipótese de que as marcas são igualmente boas quanto contra a hipótese de que as da marca $A$ são melhores que as da marca $B$.

# Caso 2: Amostras independentes de populações normais

Exemplo 3: Um supermercado não sabe se deve comprar lâmpadas da marca $A$ ou $B$, de mesmo preço. Testa uma amostra de 100 lâmpadas de cada uma das marcas, obtendo:

$$\bar{x}_A = 1.160 \text{ horas} \quad \text{e} \quad s_A = 90 \text{ horas}$$
$$\bar{x}_A = 1.140 \text{ horas} \quad \text{e} \quad s_B = 80 \text{ horas}$$

Ao nível de $2,5\%$, testar a hipótese de que as marcas são igualmente boas quanto contra a hipótese de que as da marca $A$ são melhores que as da marca $B$.

Resolução: As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \mu_A - \mu_B = 0 \\ H_1 : \mu_A - \mu_B > 0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A > \mu_B \end{cases}$$

Como $n_1 = n_2 = 100$ lâmpadas, podemos usar $s_{A^2}$ e $s_{B^2}$ estimar $\sigma_A^2$ e $\sigma_B^2$

# Caso 2: Amostras independentes de populações normais

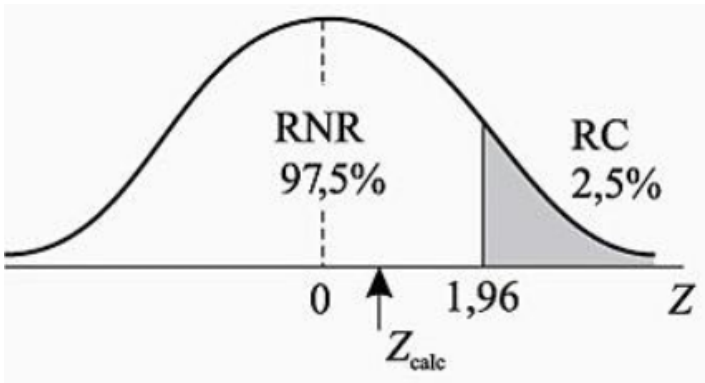
Exemplo 3 (cont.): Assim

$$s_{\bar{x}_d} = s_{(\bar{x}_A - \bar{x}_B)} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{8.100}{1000} + \frac{6400}{100}} = 12,0416$$

- A estatística do teste a ser considerada é $Z$ da normal padrão:

$$z_{calc} = \frac{(\bar{x}_A - \bar{x}_B) - \mu_{H_0}}{s_{\bar{x}_d}} = \frac{(1.160 - 1.140) - 0}{12,0416} = \frac{20 - 0}{12,0416} = 1,6609$$

- Temos que $\alpha = 2,5\% \rightarrow z_{1-\alpha} = 1,96$




- Como $|z_{calc}| < z_{1-\alpha}$, não se rejeita $H_0$, isto é, a diferença entre as vidas médias das lâmpadas não é estatisticamente significativa, ao nível de $2,5\%$

# Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

- Se $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (desconhecida) e $n_1 + n_2 \leq 30$, então usaremos a distribuição $t$ de Student.
- Temos que $s_1^2 = \dfrac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{n_1 - 1}$ e $s_2^2 = \dfrac{\sum_{j=1}^{n_2}(x_{1j} - \bar{x}_2)^2}{n_2 - 1}$ são ambos estimadores não viciados da variância;
- Determinamos $s^2$, uma estimativa de $\sigma^2$, como média ponderada entre $s_1^2$ e $s_2^2$:

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{1j} - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - n_2 - 2}$$

- Pelo facto de se usar estimador $s^2$, temos que a estatística do teste

$$t = \frac{\bar{x}_d - \mu_{H_0}}{s_{\bar{x}_d}}, \quad \text{onde} \quad s_{\bar{x}_d} = \sqrt{s^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

tem, sob $H_0$, distribuição $t$ de Student com $n_1 + n_2 - 2$ graus de liberdade.

# Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

Exemplo 3: Em uma prova de estatística, 12 alunos de uma classe conseguiram média $7,8$ e desvio padrão de $0,6$, ao passo que 15 alunos de outra turma, do mesmo curso, conseguiram média $7,4$ com desvio padrão de $0,8$. Considerando distribuições normais para as notas, verificar se o primeiro grupo é superior ao segundo, ao nível de 5%.

# Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

Exemplo 3: Em uma prova de estatística, 12 alunos de uma classe conseguiram média $7,8$ e desvio padrão de $0,6$, ao passo que 15 alunos de outra turma, do mesmo curso, conseguiram média $7,4$ com desvio padrão de $0,8$. Considerando distribuições normais para as notas, verificar se o primeiro grupo é superior ao segundo, ao nível de 5%.

Resolução: Definimos as seguintes hipóteses:

$$H_0 : \mu_1 - \mu_2 = 0 \rightarrow \mu_1 = \mu_2$$
$$H_1 : \mu_1 - \mu_2 > 0 \rightarrow \mu_1 > \mu_2$$

- Já que as turmas são do mesmo curso, as populações são normais, consideramos variâncias iguais, apesar de desconhecidas.

$$n_1 = 12; \quad \bar{x}_1 = 7,8; \quad s_1 = 0,6 \quad s_1^2 = 0,36;$$
$$n_2 = 15; \quad \bar{x}_2 = 7,4; \quad s_2 = 0,8 \quad s_2^2 = 0,64; \quad \bar{x}_d = \bar{x}_1 - \bar{x}_2 = 7,8 - 7,4 = 0,4$$
$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - n_2 - 2} = \frac{11 \cdot 0,36 + 14 \cdot 0,64}{25} = 0,5168$$

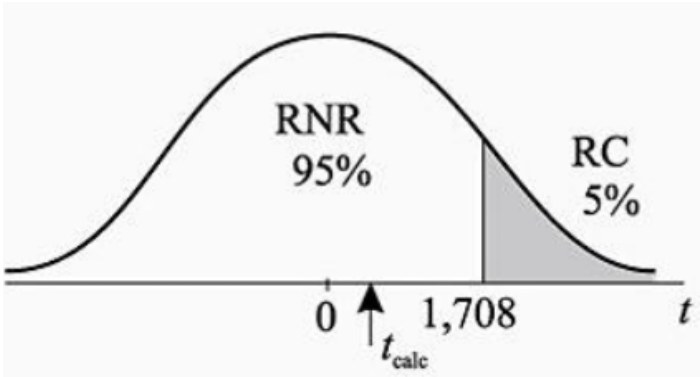
# Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

Exemplo 3 (Cont.):

$$s^2_{\bar{x}_d} = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0,5168 \cdot \left(\frac{1}{12} + \frac{1}{15}\right) = 0,0775 \rightarrow s_{\bar{x}_d} = \sqrt{0,0775} = 0,278$$

$$t_{calc} = \frac{\bar{x}_d - \mu_{H_0}}{s_{\bar{x}_d}} = \frac{0,4 - 0}{0,278} = 1,439$$

$$gl = n_1 + n_2 - 2 = 25 \rightarrow t_{gl,\alpha} = t_{25;0,05} = 1,708$$




Como $|t_{calc}| < t_{gl,\alpha} \rightarrow$ não se rejeita Ho. Concluímos que ao nível de 5%, não há motivos para considerar a primeira turma superior à segunda.

# Caso 3A: Amostras independentes com variâncias desconhecidas e iguais

## Exemplo 3: Resolução por intervalo de confiança

$$
\begin{aligned}
RNR \rightarrow & P(\bar{x} < \mu_0 + t_{gl,\alpha} \cdot \sigma_{\bar{x}_d}) = 0,95 \\
& P(\bar{x} < 0 + 1,708 \cdot 0,278) = 0,95 \\
RNR = & (-\infty; 0,478) \quad RC = [0,478; +\infty)
\end{aligned}
$$

Como $\bar{x}_d = 0,4 \rightarrow \bar{x}_d \in RNR$, o que nos leva a não rejeitar $H_0$

# Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

- Caso as populações sejam normais e $\sigma_1^2 \neq \sigma_2^2$ e desconhecidas, então para $n_1 + n_2 \leq 30$, teremos:

$$t_{calc} = \frac{\bar{x}_d - \mu_{H_0}}{s_{\bar{x}_d}}, \quad \text{onde} \quad s_{\bar{x}_d} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Portanto, a estatística do teste é

$$t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{H_0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

que tem distribuição $t$ de Student com com $\phi$ graus de liberdade, onde

$$\phi = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2.$$

- A sequência do teste é similar àquela apresentada nos casos anteriores.

# Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

Exemplo 4: O QI de 16 estudantes de uma zona pobre de certe cidade apresenta a média de 107 pontos com desvio padrão de 10 pontos, enquanto os 14 estudantes de outra região rica da cidade apresentam média de 112 pontos com desvio padrão de 8 pontos. O QI em ambas as regiões tem distribuição normal. Há uma diferença significativa entre os QI's médios dos dois grupos a 5%?.

# Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

Exemplo 4: O QI de 16 estudantes de uma zona pobre de certe cidade apresenta a média de 107 pontos com desvio padrão de 10 pontos, enquanto os 14 estudantes de outra região rica da cidade apresentam média de 112 pontos com desvio padrão de 8 pontos. O QI em ambas as regiões tem distribuição normal. Há uma diferença significativa entre os QI's médios dos dois grupos a 5%?.

Resolução: Definimos as seguintes hipóteses:

$$H_0 : \mu_1 - \mu_2 = 0 \rightarrow \mu_1 = \mu_2$$
$$H_1 : \mu_1 - \mu_2 \neq 0 \rightarrow \mu_1 \neq \mu_2$$

- Supomos que $\sigma_1^2$ e $\sigma_2^2$ desconhecidas e diferentes, já que se trata de QI's de estudantes de duas regiões distintas da mesma cidade;

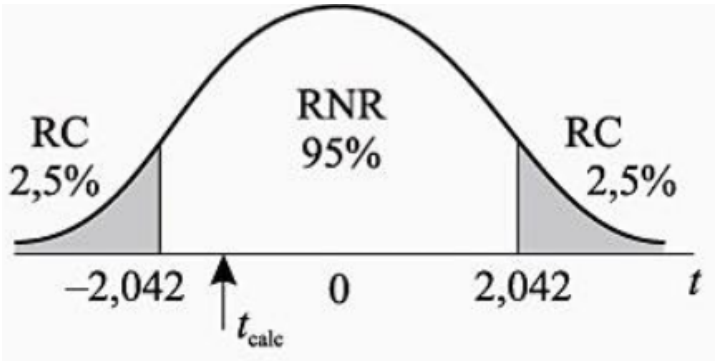
$$n_1 = 16; \quad \bar{x}_1 = 107; \quad s_1 = 10 \quad s_1^2 = 100;$$
$$n_2 = 14; \quad \bar{x}_2 = 112; \quad s_2 = 8 \quad s_2^2 = 64; \quad \bar{x}_d = \bar{x}_1 - \bar{x}_2 = 107 - 112 = -5$$

# Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

$$s_{\bar{x}_d} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{100}{16} + \frac{64}{14}} = \sqrt{10,8214} = 3,2896$$

$$t_{calc} = \frac{\bar{x}_d - \mu_{H_0}}{s_{\bar{x}_d}} = \frac{(107 - 112) - 0}{3,2896} = \frac{-5}{3,2896} = -1,52$$

$$\phi = \frac{\left(\frac{100}{16} + \frac{64}{14}\right)^2}{\frac{\left(\frac{100}{16}\right)^2}{17} + \frac{\left(\frac{64}{14}\right)^2}{15}} - 2 = 29,7425 \approx 30 \;\rightarrow\; t_{\phi,\frac{\alpha}{2}} = t_{30;0,025} = 2,042$$




Como $|t_{calc}| \in RNR \rightarrow$ não se rejeita $H_0$, isto é, ao nível de 5% não há evidências suficientes para afirmar a diferença entre os QI's das duas regiões é significativa.

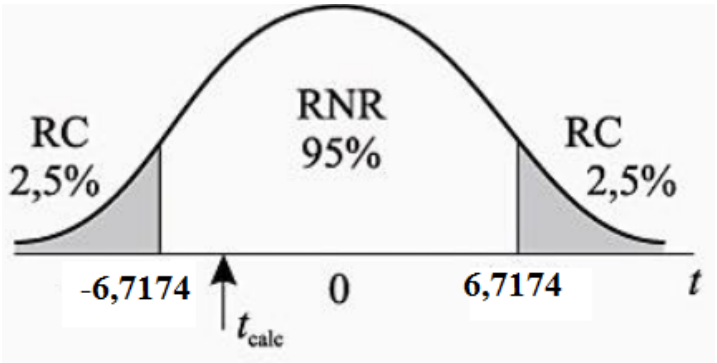
# Caso 3B: Amostras independentes com variâncias desconhecidas e diferentes

Resolução por intervalo de confiança:

$$RNR \rightarrow P(\mu_{H_0} - t_{\phi,\frac{\alpha}{2}} < \bar{x}_d < \mu_{H_0} + t_{\phi,\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(0 - 2,042 \cdot 3,2896 < \bar{x}_d < 0 + 2,042 \cdot 3,2896) = 0,95$$

$$RNR = (-6,7174; 6,7174) \qquad RC = (-\infty; -6,7174) \cup [6,7174; +\infty)$$




Como $\bar{x}_d = -5 \rightarrow \bar{x}_d \in RNR \rightarrow$ não se rejeita $H_0$.

# Comparação de proporções para duas populações independentes (Amostras grandes)

- Objectivo: verificar o comportamento de uma certa característica em duas populações;
- Suponhamos amostras retiradas de duas populações independentes;
- Podemos obter então duas proporções amostrais independentes
- Se a amostra for suficientemente grande sabemos, pelo TLC, que $\hat{p} \approx$ Normal;
- Assim, se o interesse é comparar proporções de duas populações

$$H_0 : p_1 - p_2 = 0 \rightarrow p_1 = p_2$$
$$H_1 : p_1 - p_2 \neq 0 \rightarrow p_1 \neq p_2$$

então o estimador a ser utilizado será $\hat{p}_1 - \hat{p}_2$, cuja distribuição será aproximada pela Normal;

- É fácil demonstrar-se que $\hat{p}_1 - \hat{p}_2$ é estimador não viciado de $p_1 - p_2$;

# Comparação de proporções para duas populações independentes (Amostras grandes)

- $\hat{p}_1 - \hat{p}_2$ tem aproximadamente uma distribuição normal com parâmetros

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

- Note que, para calcular a variância, a independência entre as amostras garantiu a independência entre $\hat{p}_1$ e $\hat{p}_2$, $Cov(\hat{p}_1, \hat{p}_2) = 0$

- Sob $H_0$ verdadeira, denotamos $p_1 = p_2 = p$, e obtemos seu estimador através da ponderação dos estimadores $\hat{p}_1$ e $\hat{p}_2$:

$$\hat{p}_p = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

- Substituindo os valores de $p_1$ e $p_2$ por $\hat{p}_p$ na expressão da $Var(\hat{p}_1 - \hat{p}_2)$, obtemos a estatística do teste

$$z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)(1/n_1 + 1/n_2)}} \sim N(0,1)$$

- A sequência do teste é similar àquela apresentada nos casos anteriores.

# Comparação de proporções para duas populações independentes (Amostras grandes)

Exemplo 6: Num estudo sobre doenças infantis, desejamos investigar se a incidência de casos de contaminação por vermes é afectada pela idade. Dois grupos de crianças, um com idades de 2 a 4 anos (Grupo I) e outro, com idades de 7 a 9 anos (Grupo II) foram escolhidos para serem examinados quanto à ocorrência de vermes. Os dados são apresentados a seguir:

| Grupo | Amostra | Proporção com verminose |
|---|---|---|
| I | 120 | 0,083 |
| II | 260 | 0,104 |

Será que a faixa etária influencia na incidência dessa doença? Teste ao nível de 8%

Resposta: Vamos testar as seguintes hipóteses

$$H_0 : p_1 - p_2 = 0 \rightarrow p_1 = p_2$$
$$H_1 : p_1 - p_2 \neq 0 \rightarrow p_1 \neq p_2$$

- Temos que $n_1 = 120$, $n_2 = 260$, $\hat{p}_1 = 0,083$ e $\hat{p}_2 = 0,104$. Logo, Sob $H_0$

# Comparação de proporções para duas populações independentes (Amostras grandes)

$$\hat{p}_p = \frac{n_1 \cdot \hat{p}_1 + n_2 \cdot \hat{p}_2}{n_1 + n_2} = \frac{120 \times 0,083 + 260 \times 0,104}{120 + 260} = 0,097;$$

e também

$$Var(\hat{p}_1 - \hat{p}_2) = \hat{p}_p(1-\hat{p}_p)(1/n_1 + 1/n_2) = 0,097 \times 0,903 \times (1/120 + 1/260) = 0,0011$$

- Segue então que
$$z_{calc} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{0,0011}} \sim N(0,1).$$
- Como o teste é bilateral, $z_{1-\frac{\alpha}{2}} = z_{0,96} = 1,75$. Portanto $RNR = (-1,75; 1,75)$ e $RC = (-\infty; -1,75] \cup [1,75; +\infty)$
- Observe que $z_{calc} = -0,633 \in RNR$, então, não se rejeita $H_0$. Ao nível de significância de 8%, concluímos que a incidência de casos de contaminação por vermes não é afectada pela idade ou, dita de outra forma, a faixa etária não tem influência significativa na a incidência de casos de vermes em crianças;