IAA005 - ESTATÍSTICA APLICADA I
Parte 2

Prof. Arno P. Schmitz

UFPR – Universidade Federal do Paraná

Medidas de Posição

MÉDIA ARITMÉTICA

$$\bar{X} = \frac{1}{n} . \sum_{i=1}^{n} X_i$$

Em que:

$\bar{X}$ = Média dos valores de X;

X_i = Valores de X da amostra;

n = Tamanho da amostra.

Medidas de Posição

MEDIANA: Valor que divide a amostra ao meio (valor central)

Forma de cálculo:

1) Ordenar a amostra com a variável que se deseja obter a mediana, em ordem crescente de valores;

2) **Tratamento:**

✓ **Para uma amostra de número ímpar de elementos:**

$$EM_d = \frac{n}{2} + 1/2$$

Exemplo: Se a amostra tem 5 elementos ➔ $\mathrm{E}M_d = \frac{5}{2} + 0,5 = 3$; a mediana será o 3º elemento da amostra ordenada;

✓ **Para uma amostra de número par de elementos:**

$$EM_d = \frac{n}{2} + 1/2$$

Exemplo: Se a amostra tem 4 elementos ➔ $EM_d = \frac{4}{2} + 0,5 = 2,5$; a mediana será o valor da média aritmética do 2º e 3º elementos da amostra ordenada;

Em que: EM_d = Elemento da mediana ; n = tamanho da amostra.

Medidas de Posição

MODA

Valor mais frequente em uma amostra, aquele que mais “aparece”;

Exemplo: Amostra = 20, 32, 10, 32, 20, 32, 41, 53

Moda = 32 (aparece 3 vezes)

✓ A moda é útil para variáveis qualitativas e em alguns casos para variáveis quantitativas;

Medidas de Dispersão

- São medidas de espalhamento em relação ao valor médio;

- Medidas quantitativas que expressam quanto os elementos da amostra se distanciam da média;

- As principais medidas de dispersão são: Variância, Desvio padrão e coeficiente de variação.

Medidas de Dispersão

Variância

Para uma população:

$$\sigma^2 = \frac{\sum(X - \mu)^2}{N}$$

Em que:

σ^2 = Variância da população;

X = Cada observação (dado) da variável X (que se deseja calcular a variância;

μ = Média da variável X (população);

N = Tamanho da população.

Medidas de Dispersão

Variância

Para uma amostra:

$$s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$

Em que:

s^2 = Variância da amostra;

X = Cada observação (dado) da variável X (que se deseja calcular a variância;

$\bar{X}$ = Média da variável *X (da amostra)*;

n = Tamanho da amostra.

Medidas de Dispersão

Desvio Padrão

Para uma população:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

Em que:

σ= Variância da população;

X = Cada observação (dado) da variável X (que se deseja calcular a variância;

μ = Média da variável X (população);

N = Tamanho da população.

Medidas de Dispersão

Desvio Padrão

Para uma amostra:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Em que:

s = Variância da amostra;

X = Cada observação (dado) da variável X (que se deseja calcular a variância;

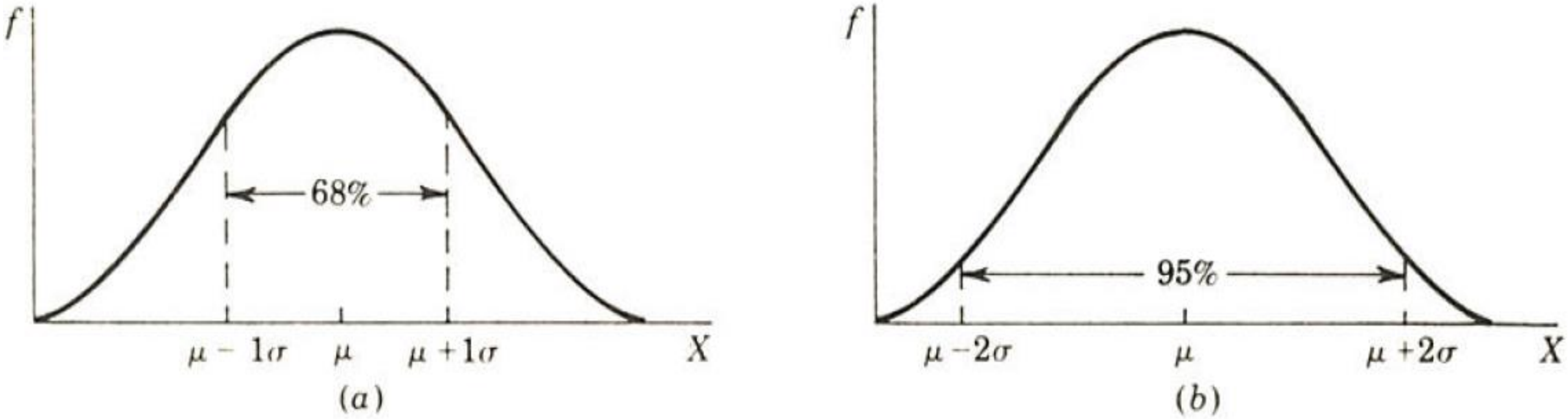
$\bar{X}$ = Média da variável *X (amostra)*;

n = Tamanho da amostra.

Medidas de Dispersão

- Uma das utilidades do desvio padrão

Distribuição de frequência, curva simétrica e mesocústica ➔ Curva normal



Medidas de Dispersão

- Exemplo de uso do Desvio Padrão

Foi observado que as contas de energia elétrica para uma área municipal, no mês de junho, são normalmente distribuídas. Se a média das contas for \$42,00 e o desvio padrão \$12,00, então aproximadamente 68% das contas estão entre \$30,00 e \$54,00. Também pode-se dizer que, aproximadamente 95% das contas estão entre \$18,00 e \$66,00.

Medidas de Dispersão

Coeficiente de Variação (CV)

- Indica a magnitude relativa do desvio padrão quando comparado com a média de uma variável.

Para a população:

$$V = \frac{\sigma}{\mu}$$

Para a amostra:

$$V = \frac{s}{\bar{X}}$$

Em que:

V = Coeficiente de variação;

σ , s = Desvio padrão; população e amostra, respectivamente;

μ , $\bar{X}$ = Média; população e amostra, respectivamente.

Medidas de Dispersão

Coeficiente de Variação (CV)

Exemplo de uso do Coeficiente de Variação:

Para duas emissões de ações ordinárias de duas empresas da indústria eletrônica, na bolsa de valores, o preço médio diário no fechamento dos negócios, durante o período de 1 mês, para as ações “A” foi de $150 com desvio padrão de $5. Para as ações “B”, o preço médio foi de $50 com um desvio padrão de $3.

$$V(A) = \frac{5}{150} = 0{,}03333 = 3{,}33\%$$

$$V(B) = \frac{3}{50} = 0{,}06000 = 6{,}00\%$$

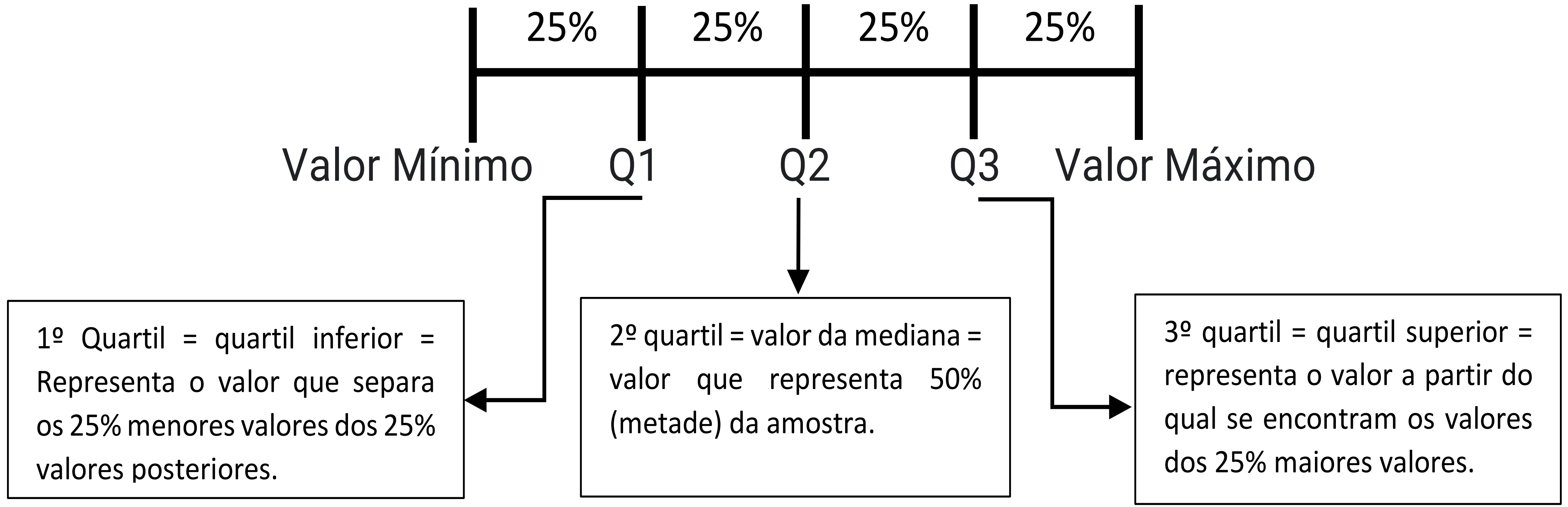
Baixa dispersão: cv ≤ 15%

Média dispersão: cv ➔ entre 16-29%

Alta dispersão: cv ≥ 30%

Medidas Separatrizes - Quartis

- Quartis são os valores que dividem o conjunto ordenado de dados em quatro partes iguais, e assim cada parte representa 1/4 da amostra ou população.



Medidas Separatrizes - Quartis

- Exemplo: amostra com dados ordenados (crescente) – amostra com número impar de elementos:

60, 65, 66, 67, 68, 68, 69, 70, 71, 72, 77

Valor Mínimo = 60

Valor Máximo = 77

n = tamanho da amostra = 11

- $Q_1 = \frac{(n+1)}{4} = \frac{(11+1)}{4} = 3^{\underline{o}}\ elemento\ da\ amostra = 66$

- Logo, 25% das observações são “iguais a” ou “abaixo” de 66 e 75% das observações estão acima de 66.

- $Q_2 = 2 \,.\left(\frac{(n+1)}{4}\right) = 2 \,.\left(\frac{(11+1)}{4}\right) = 6^{\underline{o}}\ elemento\ da\ amostra = 68$

- Logo, a observação que divide a amostra ao meio é o 6º elemento, cujo valor é 68.

Medidas Separatrizes - Quartis

- Exemplo: amostra com dados ordenados (crescente) – amostra com número impar de elementos:

60, 65, 66, 67, 68, 68, 69, 70, 71, 72, 77

Valor Mínimo = 60

Valor Máximo = 77

n = tamanho da amostra = 11

- $Q_3 = 3.\left(\frac{(n+1)}{4}\right) = 3\,.\left(\frac{(11+1)}{4}\right) = 9º\ elemento\ da\ amostra = 71$

- Portanto, 75% das observações são “iguais a” ou estão “abaixo” de 71 e 25% das observações estão acima de 71.

Medidas Separatrizes – Distância Interquartílica

- Exemplo: amostra com dados ordenados (crescente) – amostra com número impar de elementos:

60, 65, 66, 67, 68, 68, 69, 70, 71, 72, 77

Valor Mínimo = 60

Valor Máximo = 77

n = tamanho da amostra = 11

- Uma medida de dispersão alternativa ao desvio padrão é a distância interquartílica, definida como a diferença entre o terceiro e o primeiro quartis, ou seja:

$$d_q = Q_3 - Q_1 = 71 - 66 = 5$$

Medidas Separatrizes - Quartis

- Exemplo: amostra com dados ordenados (crescente) – amostra com número par de elementos:

60, 65, 66, 67, 68, 69, 70, 71, 72, 77

Valor Mínimo = 60

Valor Máximo = 77

n = tamanho da amostra = 10

- $Q_1 = \frac{(n)}{4} + 0{,}5 = \frac{(10)}{4} + 0{,}5 = 3º\ elemento\ da\ amostra = 66$

- Logo, 25% das observações são “iguais a” ou estão “abaixo” de 66 e 75% das observações estão acima de 66.

- $Q_2 = 2\,.\left(\frac{(n)}{4}\right) + 0{,}5 = 2\,.\left(\frac{(10)}{4}\right) + 0{,}5 = 5{,}5 =$ média aritmética entre o 5º e o 6º elemento =

$\frac{(68+69)}{2} = 68{,}5$

Medidas Separatrizes - Quartis

- Exemplo: amostra com dados ordenados (crescente) – amostra com número par de elementos:

60, 65, 66, 67, 68, 69, 70, 71, 72, 77

Valor Mínimo = 60

Valor Máximo = 77

n = tamanho da amostra = 10

- $Q_3 = 3.\frac{(n)}{4} + 0{,}5 = 3\,.\frac{(10)}{4} + 0{,}5 = 8^{\text{º}}\ elemento\ da\ amostra = 71$

- Portanto, 75% das observações são “iguais a” ou estão “abaixo” de 71 e 25% das observações estão acima de 71.

Medidas Separatrizes - Percentis

- O modo de calcular e o conceito são parecidos com os quartis.
- Ao contrário dos quartis, o objetivo não é subdividir a amostra em "quatro" partes, mas sim em "cem" partes.

Logo, para se calcular o valor do "décimo" percentil em uma amostra com número de observações ímpares, a fórmula é:

$$P_{10} = 10 \,.\left(\frac{(n+1)}{100}\right)$$

E, para se calcular o valor do "nonagésimo" percentil em uma amostra com número de observações pares, a fórmula é:

$$P_{90} = 90 \,.\left(\frac{(n)}{100}\right) + 0{,}5$$

Testes de Hipótese com uma Amostra

Distribuição de Amostragem da Média

<u>Para uma população infinita (Amostra menor que 5% da população)</u> :

$$E(\bar{X}) = \mu \qquad\qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Exemplo:

Suponha que a média do consumo de energia elétrica de uma população seja μ = 150 kWh/mês e um desvio padrão de σ = 36 kWh/mês. A amostra tem tamanho n = 36, em termos de valor esperado (média) e de erro-padrão da distribuição, tem-se:

$$E(\bar{X}) = \mu = 150 \text{ kWh/mês}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{36}{\sqrt{36}} = \frac{36}{6} = 6$$

Ou seja, a média do consumo de energia elétrica pode variar entre 156 e 144kWh/mês.

Testes de Hipótese com uma Amostra

Distribuição de Amostragem da Média

<u>Para população finita ou quando o erro padrão da população não é conhecido</u>:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Exemplo:

Um auditor utiliza uma amostra aleatória de n = 16 de uma população de 100 contas a receber de uma empresa. Não se conhece o desvio padrão dos valores das 100 contas a receber. Contudo, o desvio padrão da amostra é 57,00. Determinar o valor do erro padrão da amostra da média:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{57}{\sqrt{16}} \sqrt{\frac{100-16}{100-1}} = \frac{57}{4} \sqrt{\frac{84}{99}} \cong 13{,}13$$

Testes de Hipótese com uma Amostra

Intervalo de Confiança para a Média Utilizando a Distribuição Normal

- Um intervalo de confiança para a média é um intervalo estimado, construído com respeito à média da amostra, pelo qual pode ser especificada a probabilidade de o intervalo incluir o valor da média da população.
- O grau de confiança associado a um intervalo de confiança indica a percentagem de tais intervalos que incluiriam o parâmetro que se está estimando.
- Quando o uso da distribuição normal de probabilidade está garantido, o intervalo de confiança para a média amostral é determinado por:

$$\bar{X} \pm Z\,\sigma_{\bar{X}}$$

*A distribuição Z é utilizada para grandes amostras

Testes de Hipótese com uma Amostra

Intervalo de Confiança para a Média Utilizando a Distribuição Normal (Z)

- Exemplo:

Em uma dada semana, foi utilizada uma amostra aleatória de 30 empregados selecionados dentre um grande número de empregados de uma fábrica, a qual apresentou um salário médio de $\bar{X}$ = 180,00 com um desvio padrão da amostra de σ = 14,00. Estimar o salário médio para todos os empregados da fábrica de tal maneira que tenhamos uma confiança de 95% de que o intervalo estimado inclua a média da população:

Z = 1,96

Calculando o desvio padrão da média ➔ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{14}{\sqrt{30}} = 2,56$

$\bar{X}$ = 180,00

Para $\bar{X} \pm Z\, \sigma_{\bar{X}}$

σ = 14,00

$$180 - 1,96 \,.\, 2,56 \leq \bar{X} \leq 180 + 1,96 \,.\, 2,56$$

n = 30

$$180 - 5,02 \leq \bar{X} \leq 180 + 5,02$$

$$174,98 \leq \bar{X} \leq 185,02$$

Portanto, o salário médio da empresa como um todo deve se situar entre 174,98 e 185,02.

Testes de Hipótese com uma Amostra

A distribuição *t* de Student e o intervalo de confiança para a média

- Neste caso a amostra é pequena, a população normalmente distribuída e o desvio padrão é desconhecido.
- Para o caso da estimativa do intervalo para a média utiliza-se "*n*-1" graus de liberdade, pois temos apenas 1 parâmetro (a média).

$$\bar{X} \pm t_{gl} . \sigma_{\bar{X}}$$

Em que:

t = valor tabelado do valor da estatística t no nível de confiança escolhido (95%);

gl = graus de liberdade da estimativa.

Testes de Hipótese com uma Amostra

A distribuição *t* de Student e o intervalo de confiança para a média

Exemplo:

A vida média de funcionamento de lâmpadas produzidas é $\bar{X}$ = 4000 horas para uma amostra de n = 10, com desvio padrão de σ = 200 horas. Supõe-se que o tempo de operação das lâmpadas em geral tenha distribuição aproximadamente normal. Estimar a vida média de operação para a população de lâmpadas da qual foi extraída a amostra, usando o intervalo de confiança de 95%:

n = 10; Estimativa do desvio padrão da média ➔ $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{10}} = 63{,}30$

σ = 200; $\bar{X} \pm t_{gl}.\sigma_{\bar{X}}$

$\bar{X}$ = 4000; $4000 - 2{,}262 \,.\, 63{,}30 \leq \bar{X} \leq 4000 + 2{,}262 \,.\, 63{,}30$

gl = 9. $3856{,}81 \leq \bar{X} \leq 4143{,}18$

Portanto, a vida média de funcionamento da população de lâmpadas produzidas situa-se entre aproximadamente 3.857 e 4.143 horas.

Testes de Hipótese com uma Amostra

Intervalo de confiança para o Desvio Padrão e Variância

Exemplo:

$$\sqrt{\frac{(n-1)\sigma^2}{\chi^2_{gl;inf.}}} \leq \sigma \leq \sqrt{\frac{(n-1)\sigma^2}{\chi^2_{gl;\ sup.}}}$$

O salário médio de uma amostra de 100 empregados de uma grande empresa é $\bar{X}$=180,00, com um desvio padrão amostral de σ = 14,00. Sabe-se que os montantes de salários semanais da empresa estão normalmente distribuídos. O intervalo de confiança de 95% para estimar o desvio padrão dos salários é:

n = 100;

$$\sqrt{\frac{(n-1)\sigma^2}{\chi^2_{99;\ 0,025.}}} \leq \sigma \leq \sqrt{\frac{(n-1)\sigma^2}{\chi^2_{99;\ 0,975.}}}$$

$\bar{X}$=180,00;

$$\sqrt{\frac{(100-1)14^2}{129,6}} \leq \sigma \leq \sqrt{\frac{(100-1)14^2}{74,22}}$$

σ = 14,00;

$$12,24 \leq \sigma \leq 16,17$$

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "normal"

- Frequentemente existe a necessidade de se estimar a diferença entre duas médias, tal como a diferença entre os níveis de salários de duas empresas.

$$(\bar{X}_1 - \bar{X}_2) \pm Z\, \sigma_{\bar{X}_1 - \bar{X}_2}$$

- O erro padrão da diferença entre as médias é dado por:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}}$$

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "normal"

Exemplo:

A média de salários para uma amostra de *n* = 100 empregados de uma empresa é de $\bar{X}$ = 180,00 com um desvio padrão amostral de σ = 14,00. em uma outra empresa, uma amostra aleatória de *n* = 140 empregados apresentou um salário médio de $\bar{X}$ = 170,00 com um desvio padrão amostral de σ = 10,00. O intervalo de confiança de 95% para estimar a diferença entre as duas médias salariais é:

σ_1 = 14,00 ;

σ_2= 10,00 ; $\sigma_{\bar{X}_1} = \frac{\sigma_1}{\sqrt{n_1}} = \frac{14}{\sqrt{100}} = 1{,}40$ e $\sigma_{\bar{X}_2} = \frac{\sigma_2}{\sqrt{n_2}} = \frac{10}{\sqrt{140}} = 0{,}85$

n_1= 100;

n_2= 140;

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{(1{,}40)^2 + (0{,}85)^2} = 2{,}68$$

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "normal"

Exemplo:

$(\bar{X}_1 - \bar{X}_2)$ = 180 – 170 = 10

$$(\bar{X}_1 - \bar{X}_2) \pm Z\ \sigma_{\bar{X}_1 - \bar{X}_2}$$

$$10\ \pm 1{,}96\ .\ 2{,}68$$

$$10 - 1{,}96\ .\ 2{,}68\ \leq\ (\bar{X}_1 - \bar{X}_2) \leq 10 + 1{,}96\ .\ 2{,}68$$

$$4{,}75\ \leq\ (\bar{X}_1 - \bar{X}_2) \leq 15{,}25$$

Portanto, a diferença entre as médias das duas populações se encontra entre 4,75 e 15,25.

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "t de Student"

$$(\bar{X}_1 - \bar{X}_2) \pm t_{gl}\, \sigma_{\bar{X}_1 - \bar{X}_2}$$

Exemplo:

Para uma amostra aleatória de *n* = 10 lâmpadas, a vida média de funcionamento é de $\bar{X}$ = 4000 horas com σ = 200 horas. Supõe-se que a duração das lâmpadas tenha uma distribuição normal. Para uma outra marca de lâmpadas, cuja duração também é suposta normalmente distribuída, uma amostra de *n* = 8 apresentou uma média amostral de $\bar{X}$ = 4600 e um desvio padrão de σ = 250. Calcular o intervalo de confiança de 95% para a diferença entre as médias.

n_1= 10;

$$\sigma_{\bar{X}_1} = \frac{\sigma_1}{\sqrt{n_1}} = \frac{200}{\sqrt{10}} = 63{,}3$$

n_2= 8;

$\bar{X}_1$= 4000;

$$\sigma_{\bar{X}_2} = \frac{\sigma_2}{\sqrt{n_2}} = \frac{250}{\sqrt{8}} = 88{,}3$$

$\bar{X}_2$= 4600;

σ_1=200;

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{(63{,}3)^2 + (88{,}3)^2} = 108{,}65$$

σ_2=250.

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "t de Student"

Exemplo:

Para uma amostra aleatória de *n* = 10 lâmpadas, a vida média de funcionamento é de $\bar{X}$ = 4000 horas com σ = 200 horas. Supõe-se que a duração das lâmpadas tenha uma distribuição normal. Para uma outra marca de lâmpadas, cuja duração também é suposta normalmente distribuída, uma amostra de *n* = 8 apresentou uma média amostral de $\bar{X}$ = 4600 e um desvio padrão de σ = 250. Calcular o intervalo de confiança de 95% para a diferença entre as médias.

$$(\bar{X}_1 - \bar{X}_2) \pm t_{gl}\ \sigma_{\bar{X}_1 - \bar{X}_2}$$

n_1= 10;

n_2= 8;

$$(4000 - 4600) \pm 2{,}12\ .\ 108{,}65$$

$\bar{X}_1$= 4000;

$\bar{X}_2$= 4600;

$$-600 - 230{,}34 \leq (\bar{X}_1 - \bar{X}_2) \leq -600 + 230{,}34$$

σ_1=200;

σ_2=250;

$$-830{,}34 \leq (\bar{X}_1 - \bar{X}_2) \leq -369{,}66$$

gl = 10 + 8 -2 = 16.

Portanto, entende-se que a segunda marca tenha uma vida média maior do que a primeira marca entre aproximadamente 370 a 830 horas.

Testes de Hipótese com Duas Amostras

Intervalo de confiança para a diferença entre duas médias utilizando a distribuição "t de Student"

Exemplo:

Para uma amostra aleatória de n = 10 lâmpadas, a vida média de funcionamento é de $\bar{X}$ = 4000 horas com σ = 200 horas. Supõe-se que a duração das lâmpadas tenha uma distribuição normal. Para uma outra marca de lâmpadas, cuja duração também é suposta normalmente distribuída, uma amostra de n = 8 apresentou uma média amostral de $\bar{X}$ = 4600 e um desvio padrão de σ = 250. Calcular o intervalo de confiança de 95% para a diferença entre as médias.

$$(\bar{X}_1 - \bar{X}_2) \pm t_{gl}\, \sigma_{\bar{X}_1 - \bar{X}_2}$$

n_1= 10;

n_2= 8;

$$(4000 – 4600) \pm 2{,}12 \,.\, 108{,}65$$

$\bar{X}_1$= 4000;

$\bar{X}_2$= 4600;

$$-600 - 230{,}34 \leq (\bar{X}_1 - \bar{X}_2) \leq -600 + 230{,}34$$

σ_1=200;

σ_2=250;

$$- 830{,}34 \leq (\bar{X}_1 - \bar{X}_2) \leq -369{,}66$$

gl = 10 + 8 -2 = 16.

Portanto, entende-se que a segunda marca tenha uma vida média maior do que a primeira marca entre aproximadamente 370 a 830 horas.

Testes de Hipótese com Duas Amostras

Teste de Diferença (Igualdade) entre duas médias

Utilizando a distribuição normal (grandes amostras):

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

Utilizando a distribuição *t* de Student (pequenas amostras):

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sigma_{\bar{X}_1 - \bar{X}_2}} \qquad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2 + (n_2 - 1)}}$$

Para estimar o desvio padrão da diferença entre as médias:

$$\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}$$

Testes de Hipótese com Duas Amostras

Teste de Diferença (Igualdade) entre duas médias

Exemplo:

A média de salários de uma amostra de n_1 = 100 empregados em uma grande companhia industrial é de $\bar{X}_1$= 180,00 com desvio padrão amostral de σ_1=14,00. Para uma outra grande empresa, uma amostra aleatória de n_2 = 140 apresentou uma média de $\bar{X}_2$= 170,00 com um desvio padrão amostral de σ_2=10,00. Não é feita a suposição de que os desvios padrões das duas populações sejam iguais. Testar a hipótese de que não existe diferença entre os valores dos salários médios das duas empresas, utilizando um nível de significância de 5% (95% de confiança):

$\bar{X}_1$= 180,00 ;

$\bar{X}_2$= 170,00 ;

n_1 = 100 ;

n_2 = 140 ;

σ_1=14,00 ;

σ_2=10,00 .

$$\sigma_{\bar{X}_1} = \frac{\sigma_1}{\sqrt{n_1}} = \frac{14}{\sqrt{100}} = 1{,}40 \;\; ; \quad \sigma_{\bar{X}_2} = \frac{\sigma_2}{\sqrt{n_2}} = \frac{10}{\sqrt{140}} = 0{,}85$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{1{,}4^2 + 0{,}85^2} = 1{,}64$$

continua..

Testes de Hipótese com Duas Amostras

Teste de Diferença (Igualdade) entre duas médias

Exemplo:

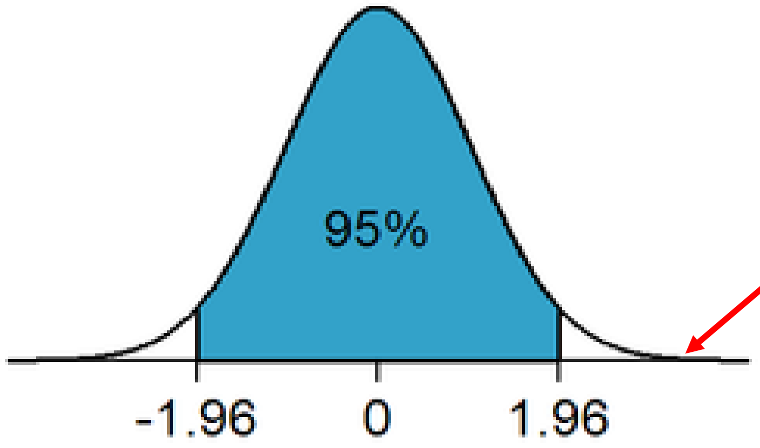
A média de salários de uma amostra de n_1 = 100 empregados em uma grande companhia industrial é de $\bar{X}_1$= 180,00 com desvio padrão amostral de σ_1=14,00. Para uma outra grande empresa, uma amostra aleatória de n_2 = 140 apresentou uma média de $\bar{X}_2$= 170,00 com um desvio padrão amostral de σ_2=10,00. Não é feita a suposição de que os desvios padrões das duas populações sejam iguais. Testar a hipótese de que não existe diferença entre os valores dos salários médios das duas empresas, utilizando um nível de significância de 5% (95% de confiança):

$H_0: \bar{X}_1 = \bar{X}_2$; $H_a: \bar{X}_1 \neq \bar{X}_2$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}} = \frac{180 - 170}{1,64} = 6,10$$

$Z_{tabelado} = \pm 1,96$

Decisão:



"Rejeita-se H_0, as médias são estat. diferentes"

Testes de Hipótese com Duas Amostras

Teste de diferença entre duas variâncias – Distribuição F

$$F_{gl_1,gl_2} = \frac{\sigma_1^2}{\sigma_2^2}$$

$$F_{gl_1,gl_2\,inferior} = \frac{1}{F_{gl_2,gl_1}}$$

Testes de Hipótese com Duas Amostras

Teste de diferença entre duas variâncias – Distribuição F

Exemplo:

Para uma amostra aleatória de n_1= 110 pneus, a vida útil média foi de $\bar{X}_1$= 40000 quilômetros, com σ_1= 2000. Para outra marca de pneus, cuja vida útil também supõe-se ser normalmente distribuída, uma amostra aleatória de n_2= 88 apresentou uma média amostral de $\bar{X}_2$= 43000 e um desvio padrão amostral de σ_2= 2500. Testar a hipótese de que as amostras foram obtidas de populações com variâncias iguais, usando o nível de significância de 10% (90% de confiança).

Hipóteses ➔ H_0: $\sigma_1^2 = \sigma_2^2$, H_a: $\sigma_1^2 \neq \sigma_2^2$

n_1= 110;

n_2= 88;

$\bar{X}_1$= 40000;

$\bar{X}_2$= 43000;

$$F_{109,\,87}\ crítico\ (5\%\ inferior) = \frac{1}{F_{87,\,109}(10\%\ superior)} = \frac{1}{1{,}27} = 0{,}79$$

σ_1= 2000;

σ_2= 2500;

σ_1^2 = 4000000;

σ_2^2 = 6250000.

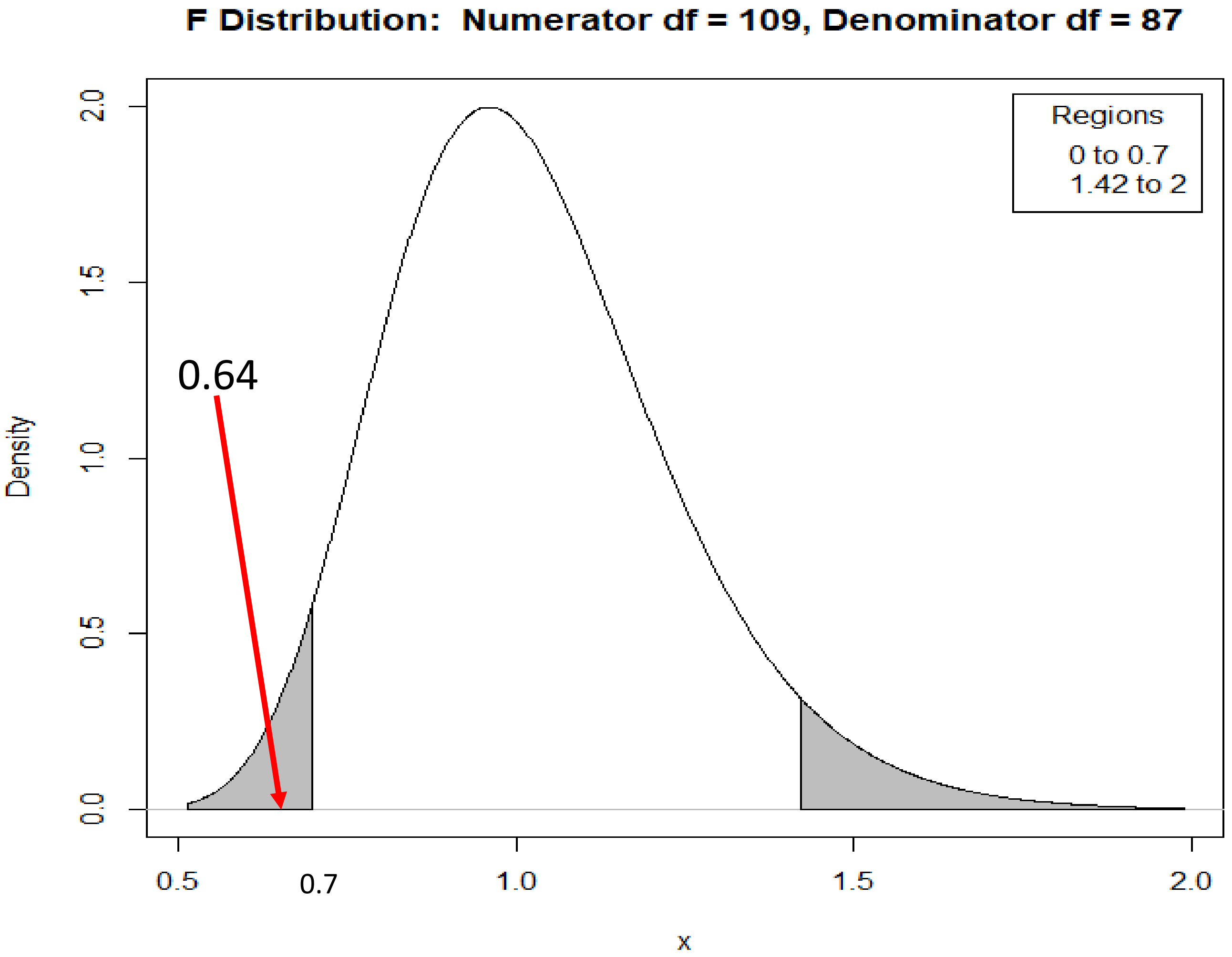
$$F_{gl_1,gl_2} = \frac{\sigma_1^2}{\sigma_2^2} = \frac{4000000}{6250000} = 0{,}64$$

Continua...

Testes de Hipótese com Duas Amostras

Teste de diferença entre duas variâncias – Distribuição F

Exemplo:



O valor de 0.64 se situa na região de rejeição da hipótese nula, ou seja, as variâncias não são iguais estatisticamente.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

- São testes que servem para identificar se as amostras são estatisticamente parecidas;

Os testes mais populares são:

1. Teste de t para amostras – necessita que a amostra seja normalmente distribuída e que as variâncias sejam iguais;
2. Teste de Wilcoxon-Mann-Whitney para amostras independentes– não tem restrições.

Obs: para diferentes amostras com mais de uma variável, cuja intensão é utilizar um modelo estatístico/econométrico, pode-se empregar os valores dos resíduos da regressão para fazer os testes.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

Sejam $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ duas amostras independentes aleatórias de duas populações normais $N(\bar{X}_1, \sigma^2)$ e $N(\bar{X}_2, \sigma^2)$. As amostras podem ter tamanhos diferentes $(n = ou \neq m)$. Mas as amostras devem ter origem em populações normais com variâncias iguais.

Nesse contexto, tem-se como hipótese nula H_0: $\bar{X}_1 = \bar{X}_2$. Como hipóteses alternativas tem-se:

H_a: $\bar{X}_1 \neq \bar{X}_2$ $(testebilateral)$;

$\bar{X}_1 < \bar{X}_2 (teste\ unilateral\ a\ esquerda)$;

$ou\ \ \bar{X}_1 > \bar{X}_2\ (teste\ unilateral\ a\ direita)$.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$\mu_1 e\ \mu_2$=médias das populações 1 e 2;

S = Estimativa combinada do desvio padrão populacional das duas amostras.

- Como testamos que as populações (e amostras) são equivalentes, então $\mu_1 = \mu_2$, ou seja, $\mu_1 - \mu_2 = 0$. Portanto, a igualdade acima pode ser resumida a:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

- Para estimar o desvio padrão combinado das duas populações, tem-se:

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

Em que:

S_1^2 e S_2^2= Variâncias amostrais das amostras das populações 1 e 2.

- A estatística de teste para “t” tem gl = $n+m-2$.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

Exemplo:

Suspeita-se que a maconha afeta a memória. Para averiguar esta afirmação, um experimento foi conduzido da seguinte forma:

a) Duas amostras foram construídas, uma com 13 pessoas usuários de maconha e outra com 12 pessoas não usuárias.

b) Cada grupo recebeu uma lista que continha 15 palavras para memorizar em 5 minutos.

- Utilizar 5 % de significância (95% de confiança).
- Testar se as duas populações (amostras) são equivalentes.

O número de palavras memorizadas por cada pessoa foi registrado na tabela a seguir:

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

número da observação	Amostra de Usuários (1 ou n)	Amostra de Não-Usuários (2 ou m)
1	6	10
2	11	7
3	7	5
4	4	6
5	6	5
6	4	5
7	5	9
8	10	6
9	6	7
10	9	8
11	10	12
12	9	10
13	8	
Média	7.31	7.50
Variância	6.00	5.38
Desvio Padrão	2.45	2.32

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

$$S^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{(13-1).\,6{,}00 + (12-1).\,5{,}38}{13+12-2} = 5{,}70$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{7{,}31 - 7{,}50}{2{,}38\sqrt{\frac{1}{13}+\frac{1}{12}}} = \frac{-0{,}19}{0{,}95} = -0{,}20$$

gl = 13 + 12 - 2 = 23

$$t_{0{,}05;\ 23gl(tabelado)} = 2{,}069$$

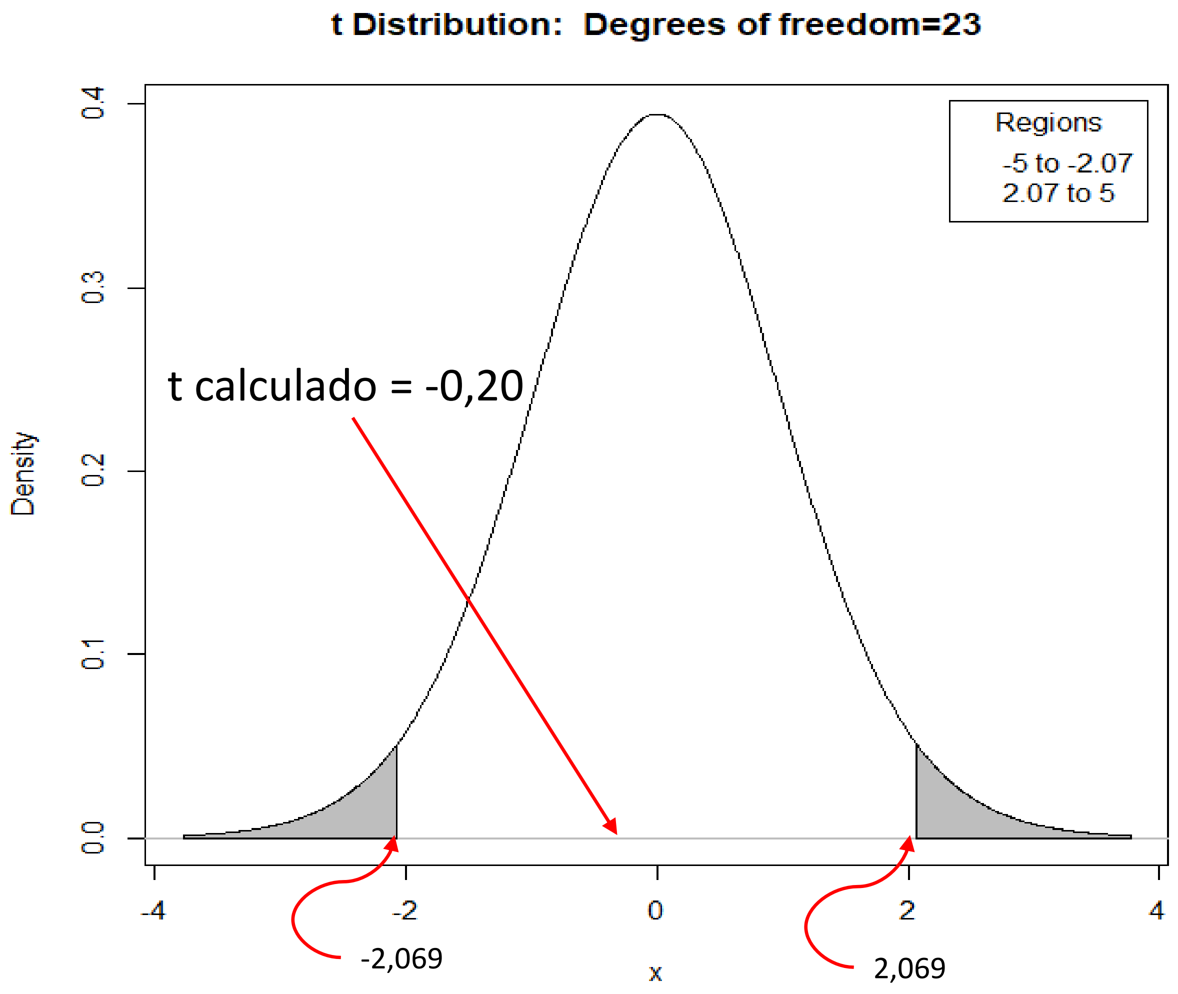
Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) Teste de t para equivalência de duas amostras

$H_0\!: \bar{X}_1 = \bar{X}_2$
$H_a\!: \bar{X}_1 \neq \bar{X}_2 \; (testebilateral)$

Como o valor "t" calculado situa-se na região de aceitação, aceita-se H_0, ou seja, as duas populações (amostras) são equivalentes.



Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) <u>Teste de Wilcoxon-Mann-Whitney para amostras independentes</u>

Considere duas populações, P1 e P2, das quais não se dispõe de informações a respeito de suas distribuições. Pode-se abordar o teste a partir de variáveis aleatórias qualitativas ordinais ou quantitativas. Considere também duas amostras independentes destas duas populações. Deseja-se testar se as distribuições são iguais em localização, ou seja, busca-se saber se uma população tende a possuir valores maiores do que a outra, ou se têm a mesma mediana.

Este teste é baseado nos “postos” dos valores obtidos combinando-se as duas amostras. Primeiro ordena-se os valores, em ordem crescente, independentemente de qual população cada valor provém.

No caso de haver uma variável aleatória qualitativa ordinal, comumente associa-se os números às diversas categorias (ou classes, ou atributos), segundo as quais a variável é classificada. P. ex., a amostra pode ter 1 reprovado, 2 em exame final e 3 aprovado. Logo, esses valores são “postos”.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) <u>Teste de Wilcoxon-Mann-Whitney para amostras independentes</u>

Seja $X_1, X_2, \ldots, X_m$ os valores de uma amostra aleatória da população P_1; e $Y_1, Y_2, \ldots, Y_n$ os valores de uma amostra aleatória da população P_2; de modo que os X_i´$s$são independentes e identicamente distribuídos (iid) e os Y_i´$s$são iid. Além disso, supõe-se que os X_i´s e os Y_i´s são mutuamente independentes e tome-se como amostra Y aquela amostra que detenha o menor tamanho amostral, ou seja, $n \leq m$.

Na aplicação do teste, supõe-se que F e G sejam as funções de distribuição das populações P_1 e P_2, respectivamente e, neste caso, consideramos como hipótese nula:
Hipóteses do teste ➔ $H_0\!: F(t\ -\ \Delta) = G(t)\ \forall\ "t"\ e\ "\Delta = 0"$
$H_a\!: F(t\ - \Delta\ ) \neq G(t)\ \forall\ "t"e\ "0 < \Delta < 0"$

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) <u>Teste de Wilcoxon-Mann-Whitney para amostras independentes</u>

No teste, ordena-se todos os valores (das duas amostras) em ordem crescente e calcula-se os postos associados. Considera-se S_m e S_n as somas dos postos relacionados aos elementos das amostras X e Y, respectivamente. Com os valores de S_m e S_n, calcula-se os valores:

$$U_m = S_m - \frac{1}{2}m(m+1) \qquad e \qquad U_n = S_n - \frac{1}{2}n(n+1)$$

Como S_m + S_n é igual a soma de todos os postos (das duas amostras), tem-se a seguinte relação:

$$U_m = m\,n \, - \, U_n$$

Logo, apenas U_m ou U_n precisa ser calculado e, na equação acima encontra-se o valor do outro.

No teste de Wilcoxon-Mann-Whitney, a estatística do teste "W" é dada por U_n.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

1) <u>Teste de Wilcoxon-Mann-Whitney para amostras independentes</u>

<u>Exemplo:</u>

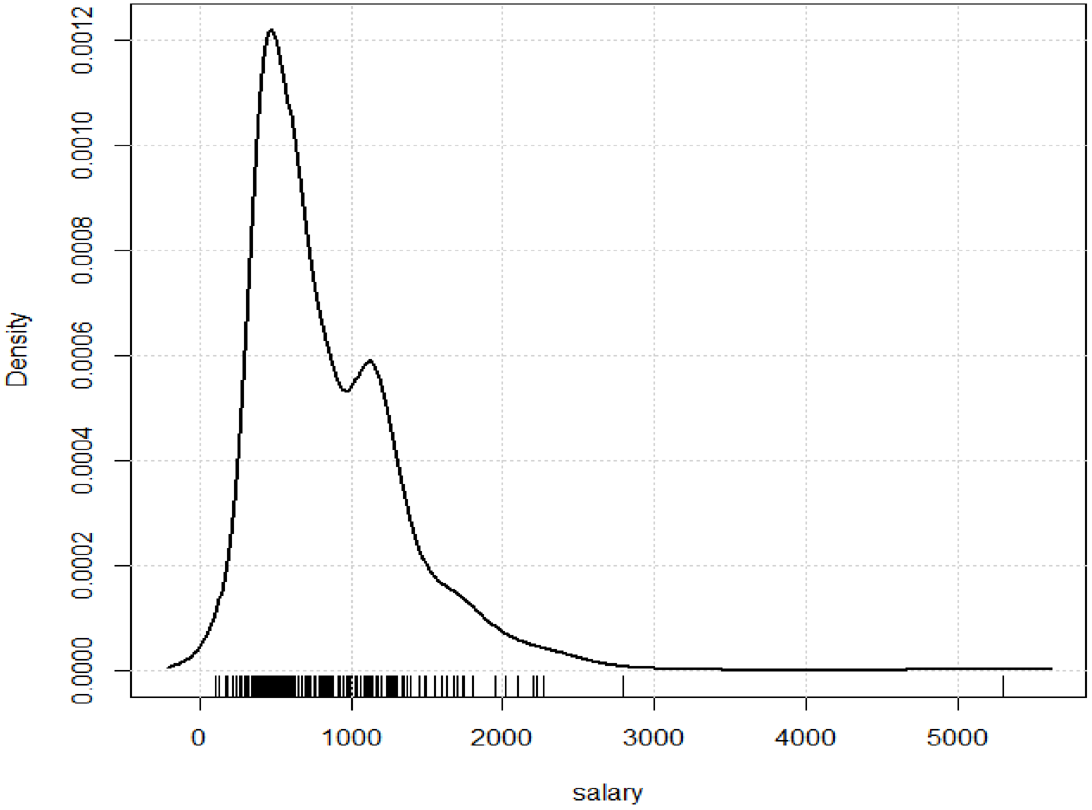
TESTE DE WILCOXON - INDEPENDENTES	
Resultados da Análise	
Tabela da Estatística do Teste (Wilcoxon)	
Informações	***Valores***
Estatística	141
P-valor	0,0373
Hipótese Nula	0
Limite Inferior	4
(Pseudo) Mediana	133,5
Limite Superior	240
Nível de Confiança	0,95

A estatística do teste é W = 141, o p-valor é igual a 0,0373 = 3,73% (portanto, menor que 5%), logo rejeita-se a hipótese nula. Em outras palavras, tem-se evidências de que as amostras vem de populações que possuem medianas diferentes.

Testes de Hipótese com Duas Amostras

Teste de independência/equivalência de amostras

- Adicionalmente, para comparar duas amostras, pode-se obter a distribuição de densidade dos dados (ou resíduos da regressão) de cada amostra para ver se elas se parecem.



Testes de Hipótese com Duas Amostras

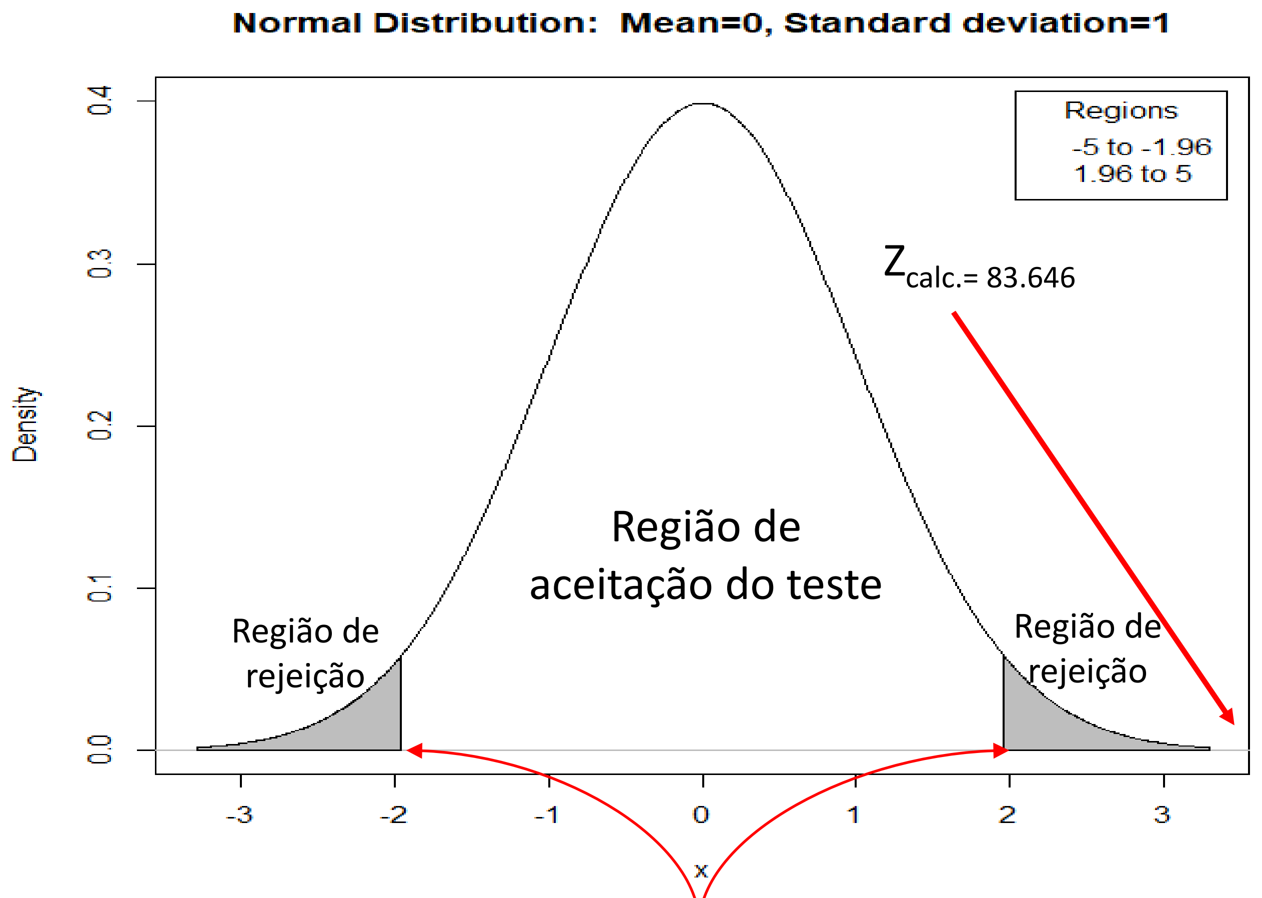
Teste de independência/equivalência de amostras

- Na prática de duas populações (e amostras) tem médias e variâncias estatisticamente iguais; pode-se dizer, a princípio que são populações (amostras) equivalentes.

- Também, não há como dizer se uma amostra é melhor comparativamente a outra. O que se faz é testar a ocorrência de “outliers” e retirá-los da amostra. No mais é observar se o processo de amostragem foi bem feito, de acordo com o plano amostral.

- Se o plano amostral foi obedecido de maneira rigorosa diminui a ocorrência do “erro amostral”. Caso contrário, existe uma grande chance de ter um erro amostral grande e assim a amostra pode ser considerada ruim. Nestes casos, recomenda-se a **“re-amostragem”.**

Intervalo de confiança para a média (Z)



$Z_{tab.}$ = -1.96 | 1.96

442.9134 453.5406 464.1679

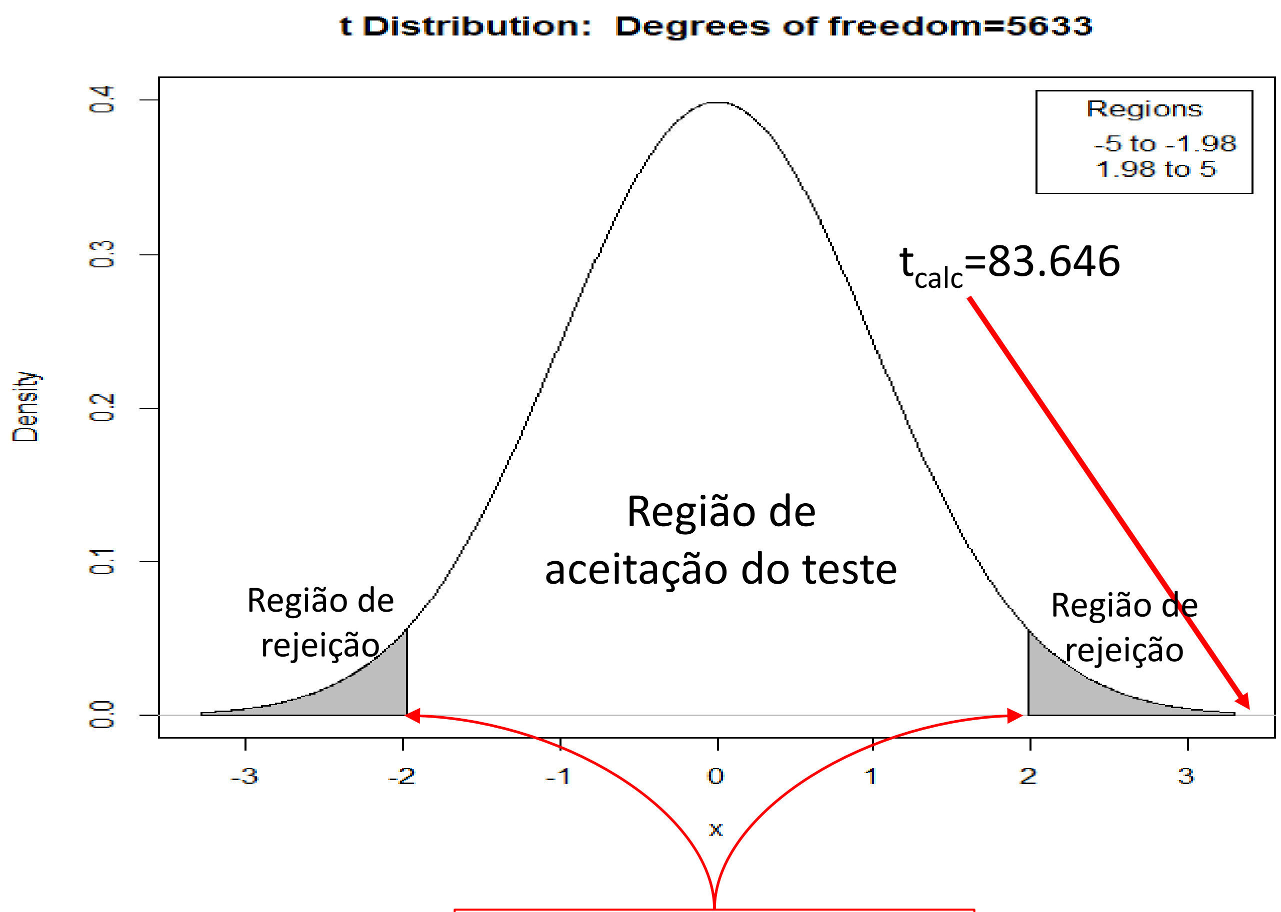
IC para a média – variável HUSEARNS

H0: valor verdadeiro da média de HUSEARNS = zero

Ha: Valor verdadeiro da média de HUSEARNS ≠ zero

Resultado do teste:
O valor calculado de Z se situa na região de rejeição de “H0”. Portanto, rejeita-se “H0” em favor da “Ha” de que a média calculada da variável HUSEARNS não é estatisticamente igual a zero. Logo, o valor calculado de 453.5406 é estatisticamente significativo com 95% de confiança.

Intervalo de confiança para a média (t)



$t_{tab.}$ = -1.96 | 1.96

H0: valor verdadeiro da média de HUSEARNS = zero

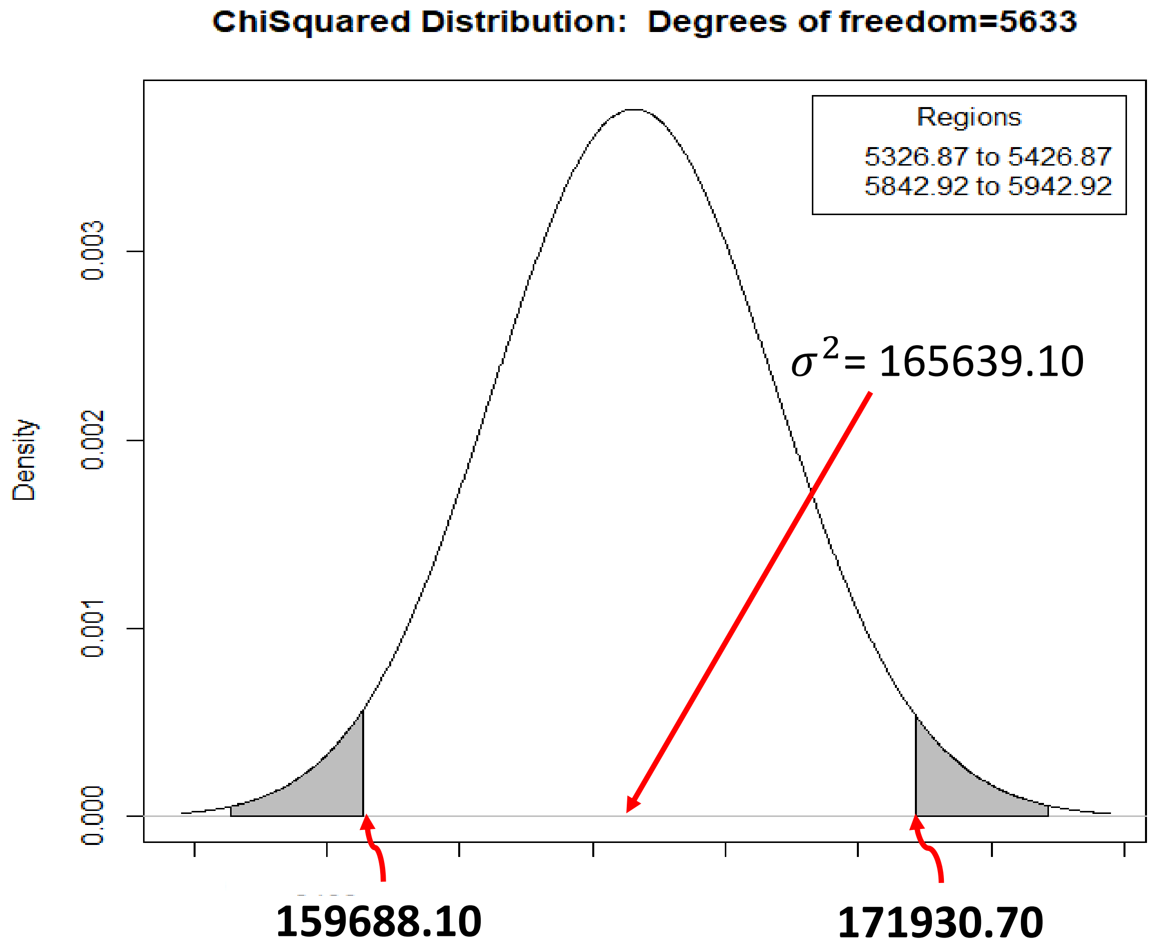
Ha: Valor verdadeiro da média de HUSEARNS ≠ zero

Resultado do teste:
O valor calculado de t se situa na região de rejeição de "H0". Portanto, rejeita-se "H0" em favor da "Ha" de que a média calculada da variável HUSEARNS não é estatisticamente igual a zero. Logo, o valor calculado de 453.5406 é estatisticamente significativo com 95% de confiança.

442.9111 | 453.5406 | 464.1702

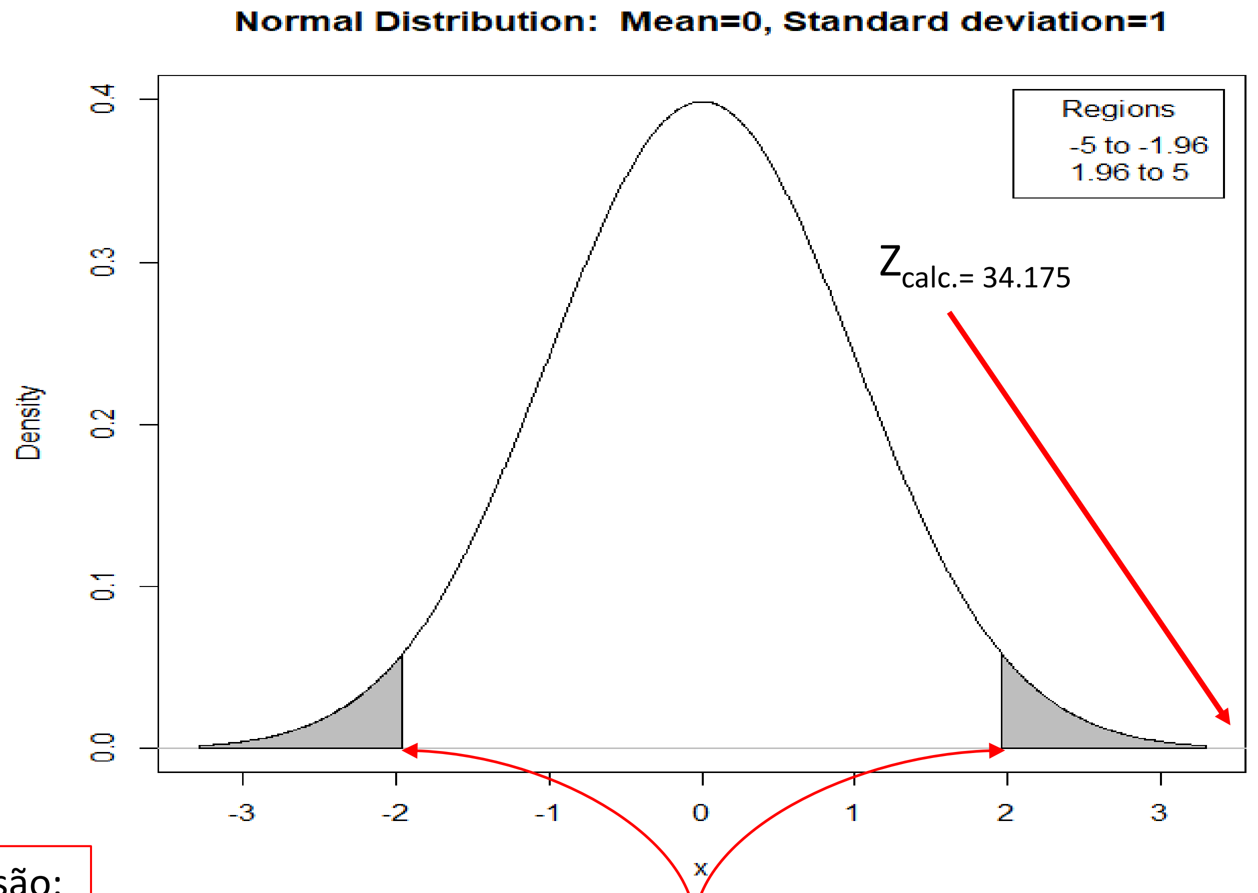
IC para a média – variável HUSEARNS

Intervalo de confiança para a variância



Intervalo de confiança construído com 95% de confiança

Teste da diferença entre duas médias (z)



$Z_{tab.}$ = -1.96 | 1.96

Os valores das médias são:

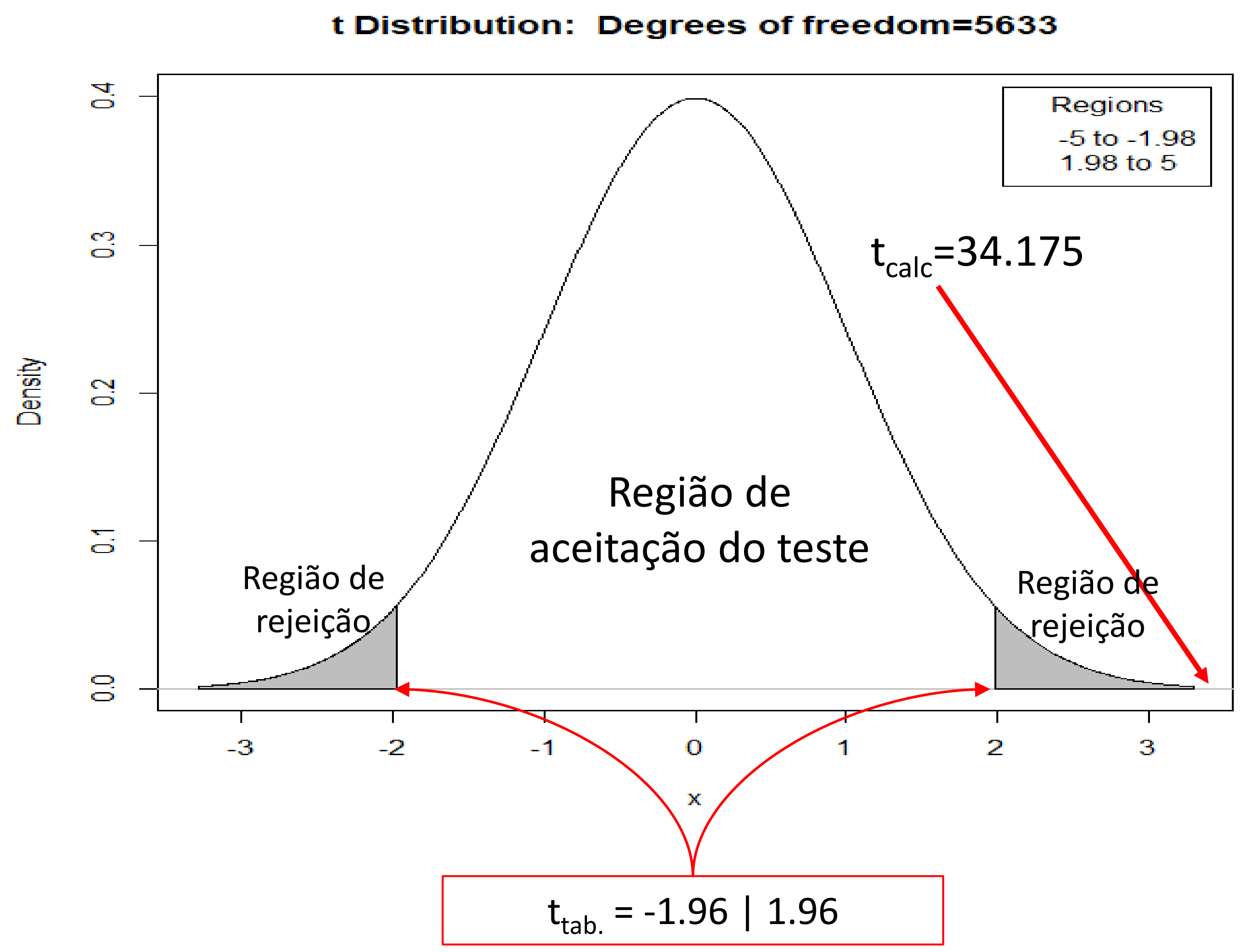
- HUSEARNS = 453.5406
- EARNS = 232.8330

H0: A diferença verdadeira entre as médias é igual a zero

Ha: A diferença verdadeira entre as médias não é igual a zero

Resultado do teste:
O valor calculado de Z se situa na região de rejeição de “H0”. Portanto, rejeita-se “H0” em favor da “Ha” de que a diferença verdadeira entre as médias não é igual a zero. Logo, pode-se dizer as médias são estatisticamente diferentes, com 95% de confiança.

Teste da diferença entre duas médias (t)

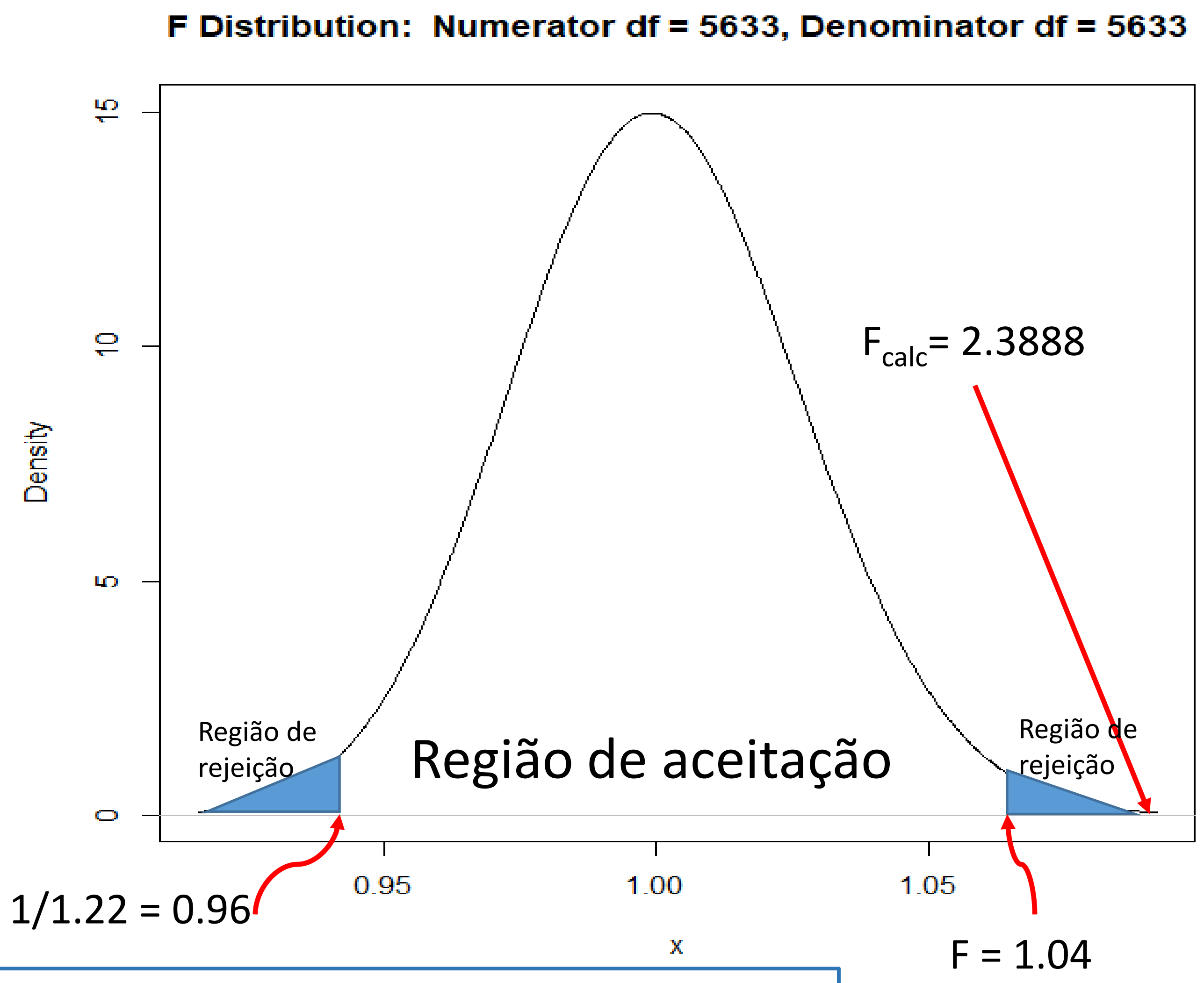


H0: A diferença verdadeira entre as médias é igual a zero

Ha: A diferença verdadeira entre as médias não é igual a zero

Resultado do teste:
O valor calculado de t se situa na região de rejeição de “H0”. Portanto, rejeita-se “H0” em favor da “Ha” de que a diferença verdadeira entre as médias não é igual a zero. Logo, pode-se dizer as médias são estatisticamente diferentes, com 95% de confiança.

Teste da diferença entre variâncias (F)



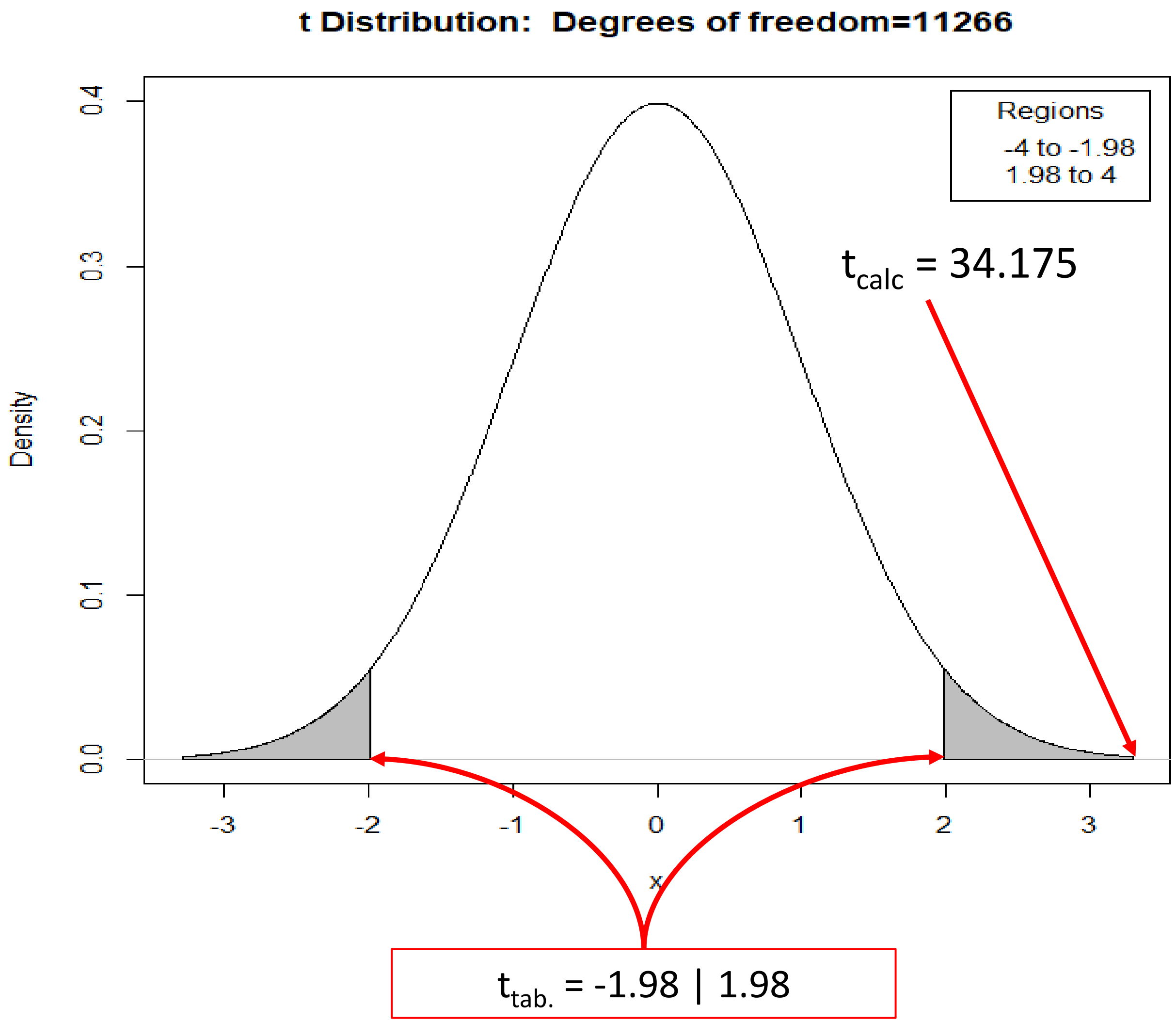
H0: A verdadeira razão entre as variâncias é igual a "um" (1)

Ha: A verdadeira razão entre as variâncias não é igual a "um" (1)

Resultado do teste:
O valor calculado de F se situa na região de rejeição de "H0". Portanto, rejeita-se "H0" em favor da "Ha" de que a verdadeira razão entre as variâncias não é igual a "um" (1). Logo, pode-se dizer as variâncias são estatisticamente diferentes, com 95% de confiança.

Obs: a razão das variâncias só é igual a "um" (1) quando as variâncias são iguais ou seja: $F = \frac{S_1^2}{S_2^2} = 1$; se $S_1^2 = S_2^2$

Teste de independência/equivalência entre duas amostras (t)



H0: As amostras são similares (equivalentes)

Ha: As amostras não são similares (equivalentes)

Resultado do teste:

O valor calculado de t se situa na região de rejeição de "H0". Portanto, rejeita-se "H0" em favor da "Ha" de que as amostras não são estatisticamente equivalentes, com 95% de confiança.

Se duas amostras são normalmente distribuídas e tem variâncias e médias estatisticamente iguais, pode-se dizer que essas amostras são similares ou equivalentes. Caso contrário, essas amostras são independentes.

Teste de Wilcoxon-Mann-Whitney para amostras independentes

H0: As amostras são similares (equivalentes)
Ha: As amostras são independentes

Wilcoxon rank sum test with continuity correction

data: x and y

W = 21091599, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

Obs: Como o p-value é ≤ 0.05 então rejeita-se "H0" em favor de "Ha" de que as amostras são independentes

Teste de Normalidade Kolmogorov-Smirnov

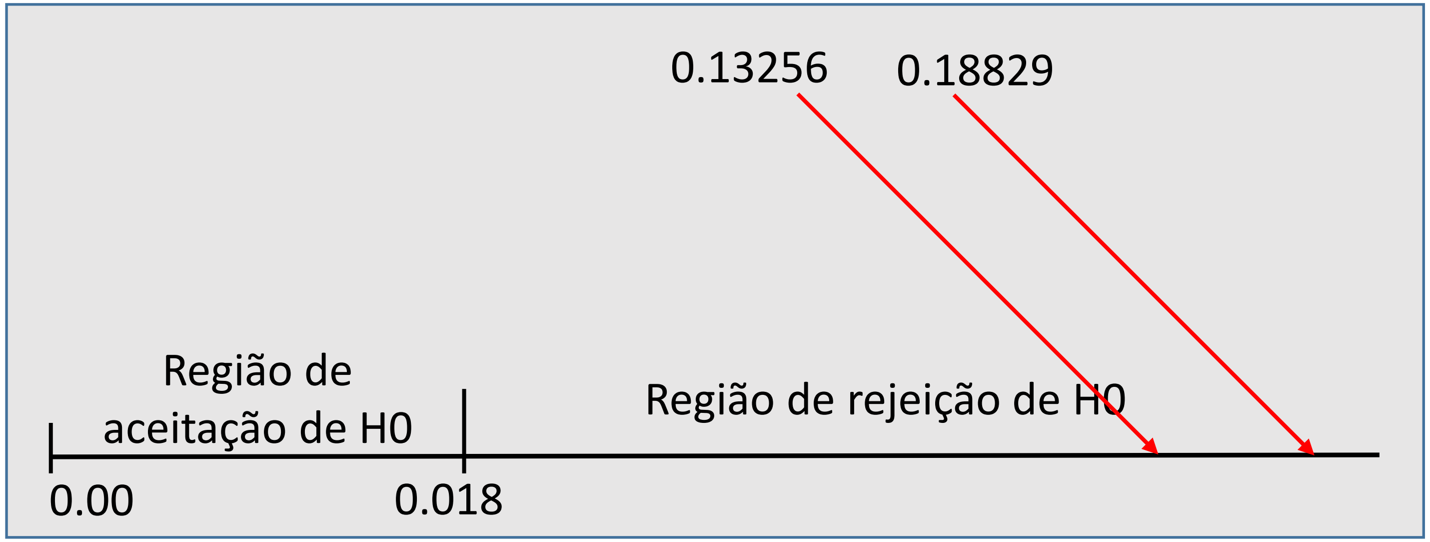
H0: A amostra provem de uma população normalmente distribuída

Ha: A amostra não provem de uma população normalmente distribuída

Ver tabela de valores críticos de “D” de Kolmogorov-Smirinov em: https://edisciplinas.usp.br/mod/resource/view.php?id=2637981&forceview=1

Pode ser aplicado a amostras com qualquer tamanho

$D_{crit} = \frac{1.36}{\sqrt{n}} = \frac{1.36}{75.06} = 0.018$



Resultados:

data: salarios$earns

D = 0.18829, p-value < 2.2e-16

data: salários$husearns

D = 0.13256, p-value < 2.2e-16

Regra de bolso: Ambos valores de p-value são inferiores a 0.05 (5%), logo rejeita-se H0.

Resultado do teste:

Ambos valores de “D “ são maiores que o valor crítico (tabelado), logo rejeita-se H0 em favor de Ha. Em outras palavras não se pode rejeitar a não normalidade das amostras.

Outros Testes de Normalidade

- Teste de normalidade de Shapiro-Wilk ➔ Para amostras com até 5000 observações

p-value > 0.05 indica que a variável possui distribuição normal

- Teste de normalidade de Anderson-Darling ➔ Para amostras de qualquer tamanho

P-value > 0.05 indica que a variável possui distribuição normal

- Teste de normalidade de Cramer- von Mises ➔ Para amostras de qualquer tamanho

P-value > 0.05 indica que a variável possui distribuição normal

Transformação de Box-Cox – Variáveis “Não-Normais”

- As vezes nos defrontamos com variáveis cuja distribuição é não-normal e desejamos que a mesma tenha uma distribuição “normal”.

Dado uma variável Y qualquer com distribuição “não-normal”, pode-se transformá-la em uma variável normal, sem perder as demais propriedades, por meio da seguinte expressão:

$$Y^*_{Box-Cox} = \frac{Y^\lambda - 1}{\lambda}$$

Qual o valor de λ (λ varia entre –∞ e +∞) que maximiza a aderência da distribuição da nova variável Y* à normalidade?

** Estimação por experimentação

Fim do Tema 3 !!!