IAA005 - ESTATÍSTICA APLICADA I
Parte 3

Prof. Arno P. Schmitz

UFPR – Universidade Federal do Paraná

Análise de Regressão Linear

A análise de regressão linear baseia-se em um modelo matemático teórico, como por exemplo a função consumo das famílias:

$$Y = \beta_1 + \beta_2 X$$

Em que:

Y = Despesas de consumo;

X = Renda disponível;

β_1 e β_2 = Parâmetros a serem estimados (intercepto e coeficiente angular, respectivamente).

➔β_2 mede a propensão marginal a consumir da renda disponível;

➔β_1 é interpretado como o consumo autônomo que independe da renda;

Análise de Regressão Linear

Para qualquer modelo matemático teórico exemplificado abaixo:

$$Y = \beta_1 + \beta_2 X$$

- A variável que aparece do lado esquerdo da igualdade é chamada de **“variável dependente”**;

- As variáveis do lado direito são chamadas de **“variáveis independentes”** ou **“variáveis explanatórias”**, bem como alguns outros nomes;

- Os β_s são parâmetros a serem estimados.

Análise de Regressão Linear

O modelo matemático exige uma relação exata ou determinística entre as variáveis. Mas as relações entre variáveis econômicas e sociais são em geral, inexatas.

- Portanto, se coletarmos dados sobre despesas de consumo e renda disponível (a renda depois de descontados os impostos) de uma amostragem de, digamos, 500 famílias e traçarmos um gráfico em que o eixo vertical representa as despesas de consumo e, o eixo horizontal a renda disponível, não devemos esperar que as 500 observações se situem exatamente sobre a reta dada pela Equação.
- Isto porque, além da renda, outras variáveis afetam o consumo tais como: tamanho da família, idade dos componentes da família, religião, localização geográfica, etc.

Análise de Regressão Linear

- Para considerar as relações inexatas entre as variáveis deve-se ter em mente um modelo estatístico ou econométrico, tal como:

$$Y = \beta_1 + \beta_2 X + u$$

- Este modelo econométrico acima é o caso de regressão simples, pois existe uma variável dependente e apenas uma variável explicativa. Mas é possível ter uma modelo de regressão múltipla, que considera duas ou mais variáveis explicativas:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + u$$

- Para resolver o modelo deve-se dispor dos valores das variáveis em um período ou espaço determinado.
- Em outras palavras deve-se ter em mãos uma matriz com os valores das variáveis.

Análise de Regressão Linear

Ano	DCP(*Y*)	PIB(*X*)
1960	1597,4	2501,8
1961	1630,3	2560,0
1962	1711,1	2715,2
1963	1781,6	2834,0
1964	1888,4	2998,6
1965	2007,7	3191,1
1966	2121,8	3399,1
1967	2185,0	3484,6
1968	2310,5	3652,7
1969	2396,4	3765,4
1970	2451,9	3771,9
1971	2545,5	3898,6
1972	2701,3	4105,0
1973	2833,8	4341,5
1974	2812,3	4319,6
1975	2876,9	4311,2
1976	3035,5	4540,9
1977	3164,1	4750,5
1978	3303,1	5015,0
1979	3383,4	5173,4
1980	3374,1	5161,7
1981	3422,2	5291,7
1982	3470,3	5189,3
1983	3668,6	5423,8
1984	3863,3	5813,6
1985	4064,0	6053,7
1986	4228,9	6263,6
1987	4369,8	6475,1
1988	4546,9	6742,7
1989	4675,0	6981,4
1990	4770,3	7112,5
1991	4778,4	7100,5
1992	4934,8	7336,6
1993	5099,8	7532,7
1994	5290,7	7835,5
1995	5433,5	8031,7
1996	5619,4	8328,9
1997	5831,8	8703,5
1998	6125,8	9066,9
1999	6438,6	9470,3
2000	6739,4	9817,0
2001	6910,4	9890,7
2002	7099,3	10048,8
2003	7295,3	10301,0
2004	7577,1	10703,5
2005	7841,2	11048,6

Análise de Regressão Linear

- A matriz de dados apresentada tem seus dados que podem ser classificados como dados temporais, ou uma série temporal. Isto porque apresentada dados de 1969 a 2005, para dados de consumo - DCP(Y) - e renda bruta da sociedade – PIB(X).

Portanto, o modelo estatístico deve ser corretamente apresentado como:

$$Y_t = \beta_1 + \beta_2 X_t$$

O subscrito "t" apresenta o modelo como sendo um modelo de série temporal.

Alternativamente, um modelo estatístico pode ser apresentado como do tipo cross-section, ou seja, cujos dados são apresentados todos em um mesmo período de tempo e desfrutam da mesma localização geográfica. Este modelo pode ser expresso por:

$$Y_i = \beta_1 + \beta_2 X_i$$

Ou seja, o subscrito "i" apresenta o modelo como sendo do tipo cross-section (corte temporal). Exemplos destes dados podem uma função que deseja saber a produtividade média das máquinas em uma linha de produção e se utiliza dos seguintes dados: produtividade da máquina e gastos com manutenção.

Análise de Regressão Linear

- Um outro tipo de modelo estatístico é aquele no qual os dados estão dispostos segundo sua disposição geográfica, ou seja os dados carregam consigo a informação da sua localização (latitude e longitude).

$$Y_{ui,vi} = \beta_1 + \beta_2 X_{ui,vi}$$

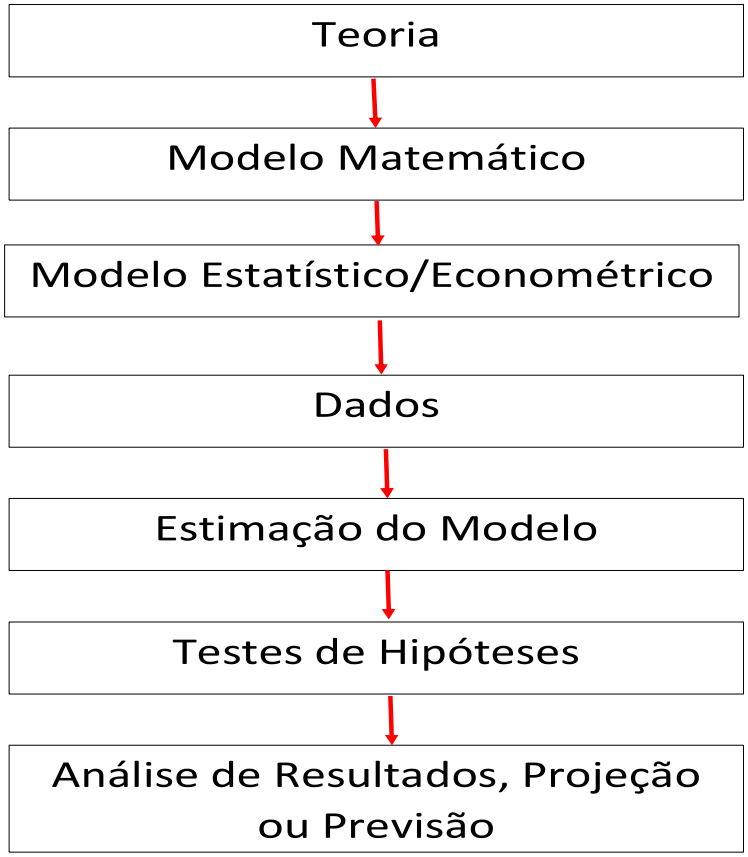
Neste caso, os subscritos ui e vi representam respectivamente a latitude e longitude da observação i, que vai de 1 a k. Sendo assim poderemos ter k observações geograficamente distribuídas.

Deve-se notar que, para o modelo acima, os parâmetros são genéricos para toda a distribuição espacial. Mas, também é possível ter modelos com parâmetros espacialmente distribuídos tal como no seguinte modelo:

$$Y_{ui,vi} = \beta_{1;ui,vi} + \beta_{2;ui,vi} X_{ui,vi}$$

Escolha da Estrutura Inicial do Modelo

- Parte-se de um problema a ser tratado, para tanto deve buscar primeiramente uma teoria que forneça uma base de conhecimento para elaborar o modelo e fazer as análises necessárias.



Escolha da Estrutura Inicial do Modelo

- Contudo, se não houver uma teoria plausível para o problema a ser tratado, pode-se utilizar de experimentação, intuição científica e evidências coletadas em bases de dados.
- Neste caso, elimina-se a necessidade de uma teoria subjacente ao problema de pesquisa.
- Os modelos econométricos tratam da dependência de uma variável em relação a outras, mas não existe necessariamente causação entre as variáveis explicativas e a variável dependente.
- Para ver se uma variável causa impacto em outra ou outras, deve-se utilizar metodologia específica, tal como os testes de causalidade de Granger e outros testes disponíveis na atualidade.

Regressão *versus* Correlação

- A análise de correlação tem como objetivo medir a “força” ou o “grau” de associação linear entre duas ou mais variáveis está estritamente relacionada à análise de regressão, mas conceitualmente é muito diferente.

- O coeficiente de correlação mede a força de associação (linear) – tal como na matriz de correlação.

- Na análise de regressão, busca-se estimar ou prever o valor médio de uma variável com base nos valores fixos de outras variáveis.

- A regressão e a correlação têm algumas diferenças fundamentais. Na análise de regressão, existe uma assimetria na maneira como as variáveis dependente e explanatórias são tratadas. A variável dependente têm uma distribuição de probabilidade. Já para as variáveis explanatórias, considera-se que essas variáveis tem valores fixos em amostras repetidas.

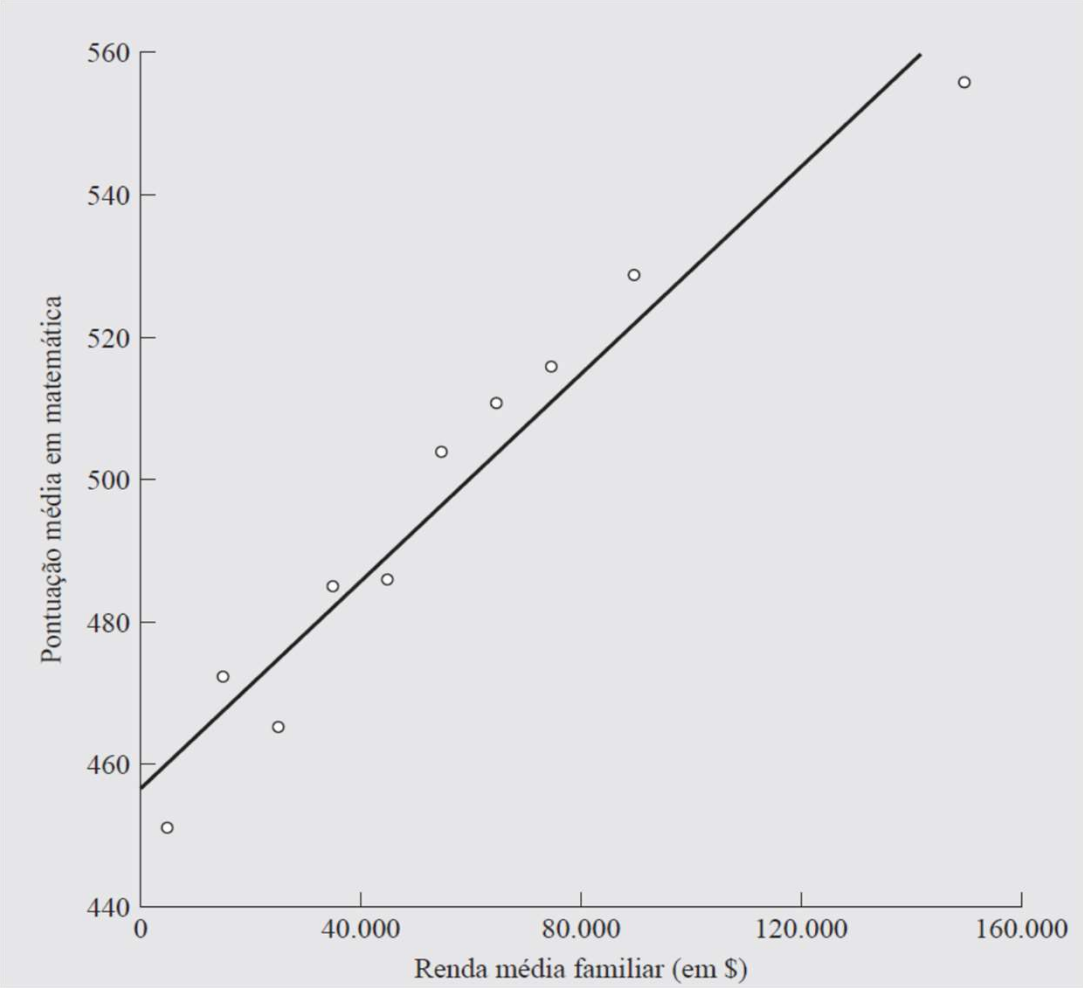
- Na análise de correlação trata-se as variáveis simetricamente, não existe distinção entre variáveis explicativas e dependentes.

Significado do Termo de Erro estocástico (u)

1. Caráter vago da teoria;
2. Indisponibilidade de dados;
3. Variáveis essenciais x Variáveis secundárias;
4. Caráter aleatório do comportamento humano;
5. Varáveis proxy pouco adequadas;
6. Princípio da parcimônia;
7. Forma Funcional errada.

Origem do Termo de Erro estocástico (u)

$Y_i = \beta_1 + \beta_2 X_i + u_i$; Y = Pontuação média matemática , X = Renda média familiar



O Método dos Mínimos Quadrados Ordinários - MQO

Função de Regressão Populacional (FRP):

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

A FRP não pode ser obtida diretamente, portanto deve ser estimada via uma Função de Regressão Amostral (FRA):

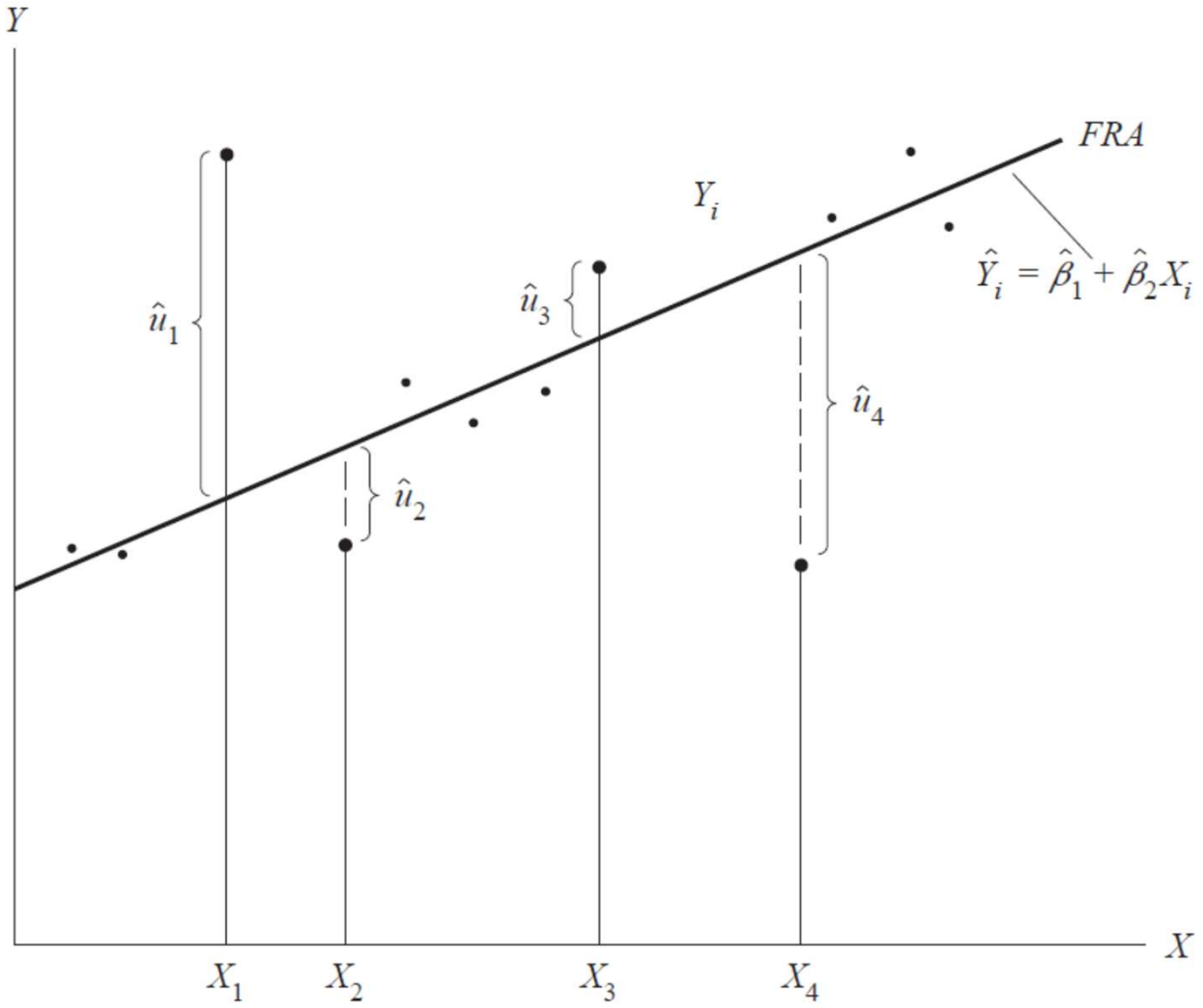
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$$

$$Y_i = \widehat{Y}_i + u_i$$

- As variáveis e parâmetros com “^”, significa que são “estimados”.

O Método dos Mínimos Quadrados Ordinários

.



O Método dos Mínimos Quadrados Ordinários

Propriedades da Reta de Regressão

1. Passa pelas médias de Y e X;
2. $\bar{\hat{Y}} = \bar{Y}$;
3. A soma dos valores dos resíduos $\hat{u}_i$ é igual a zero $(\sum \hat{u}_i = 0)$;
4. Os resíduos $\hat{u}_i$ não são correlacionados os valores de Y_i;
5. Os resíduos $\hat{u}_i$ não são correlacionados os valores de X_i;

O Método dos Mínimos Quadrados Ordinários

Estimativa do Parâmetro $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\beta}_2 = \frac{\sum x_i\, y_i}{\sum x_i^2}$$

O Método dos Mínimos Quadrados Ordinários

Estimativa do Parâmetro $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

O Método dos Mínimos Quadrados Ordinários

Hipóteses do Método

1. O modelo de regressão é linear – linearidade nos parâmetros;
2. Os valores de X são fixos em amostras repetidas ou são independentes do termo de erro: valores assumidos pelo regressor X podem ser fixos em amostras repetidas (caso do regressor fixo) ou seus valores podem mudar de acordo com a variável dependente Y (no caso do regressor estocástico). No segundo caso, supõe-se que as variáveis X e o termo de erro são independentes, isto é, $cov(X_i, u_i) = 0$.
3. O valor médio do termo de erro é zero;
4. Homocedasticidade ou variância constante de u_i;

O Método dos Mínimos Quadrados Ordinários

Hipóteses do Método

5. Não existe autocorrelação entre os termos de erro: $cov(u_i, u_j) = 0$;
6. O número de observações (n = tamanho da amostra) deve ser maior que o número de parâmetros;
7. Deve haver variabilidade dos valores de X.

O Método dos Mínimos Quadrados Ordinários

Principais Testes de Consistência do Modelo

- **Alguns Testes de heterocedasticidade, contrário à Homocedasticidade**:

a) Teste de Goldfeld-Quandt – para pequenas amostras;

b) Teste de Breusch-Pagan – para grandes amostras.

- **Teste de autocorrelação dos resíduos**:

a) Teste de Durbin-Watson.

O Método dos Mínimos Quadrados Ordinários

Testes e Estimativas Adicionais

- **Alguns Testes de Normalidade**:

a) Teste de Kolmogorov-Smirnov;
b) Teste de Shapiro-Wilk.

- **Testes de significância dos coeficientes (parâmetros) estimados**:

a) Teste t de Student;
b) Teste F de Snedecor-Fischer;
c) Teste Z (normal).

- **Estimativa dos intervalos de confiança para os parâmetros**

O Método dos Mínimos Quadrados Ordinários

Erros Padrão das Estimativas de (MQO)

$$ep(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

$$ep(\hat{\beta}_1) = \sqrt{\frac{\sum x_i^2}{n \sum x_i^2}}\,\hat{\sigma}$$

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-k}}$$

k = número de parâmetros estimados.

O Método dos Mínimos Quadrados Ordinários

Propriedades dos Estimadores de MQO: Teorema de Gauss-Markov

- Cada estimador de MQO é BLUE – Best Linear Unbiased Estimator, dado que:

1. É linear, ou seja, uma função linear de uma variável aleatória, tal como a variável dependente Y na função de regressão;
2. É não viesado, isto é, o seu valor médio ou esperado $E(\hat{\beta}_2)$é igual ao verdadeiro valor β_2;
3. Tem variância mínima na classe de todos os estimadores lineares não viesados; um estimador não viesado com a menor variância é conhecido como um estimador eficiente.

O Método dos Mínimos Quadrados Ordinários

ANOVA – Análise de Variância para uma Regressão por MQO

Fonte da Variação	SQ*	gl	MSQ†
Devido à regressão (SQE)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Devido aos resíduos (SQR)	$\sum \hat{u}_i^2$	$n-2$	$\frac{\sum u_i^2}{n-2} = \hat{\sigma}^2$
STQ	$\sum y_i^2$	$n-1$	

Fonte de variação	SQ	gl	MSQ	
Devido à regressão (SQE)	95,4255	1	95,4255	$F = \frac{95,4255}{0,8811}$
Devido aos resíduos (SQR)	9,6928	11	0,8811	$= 108,3026$
STQ	105,1183	12		

Em que:

SQ = Soma dos quadrados;

SQR = Soma dos quadrados dos resíduos;

gl = Graus de liberdade;

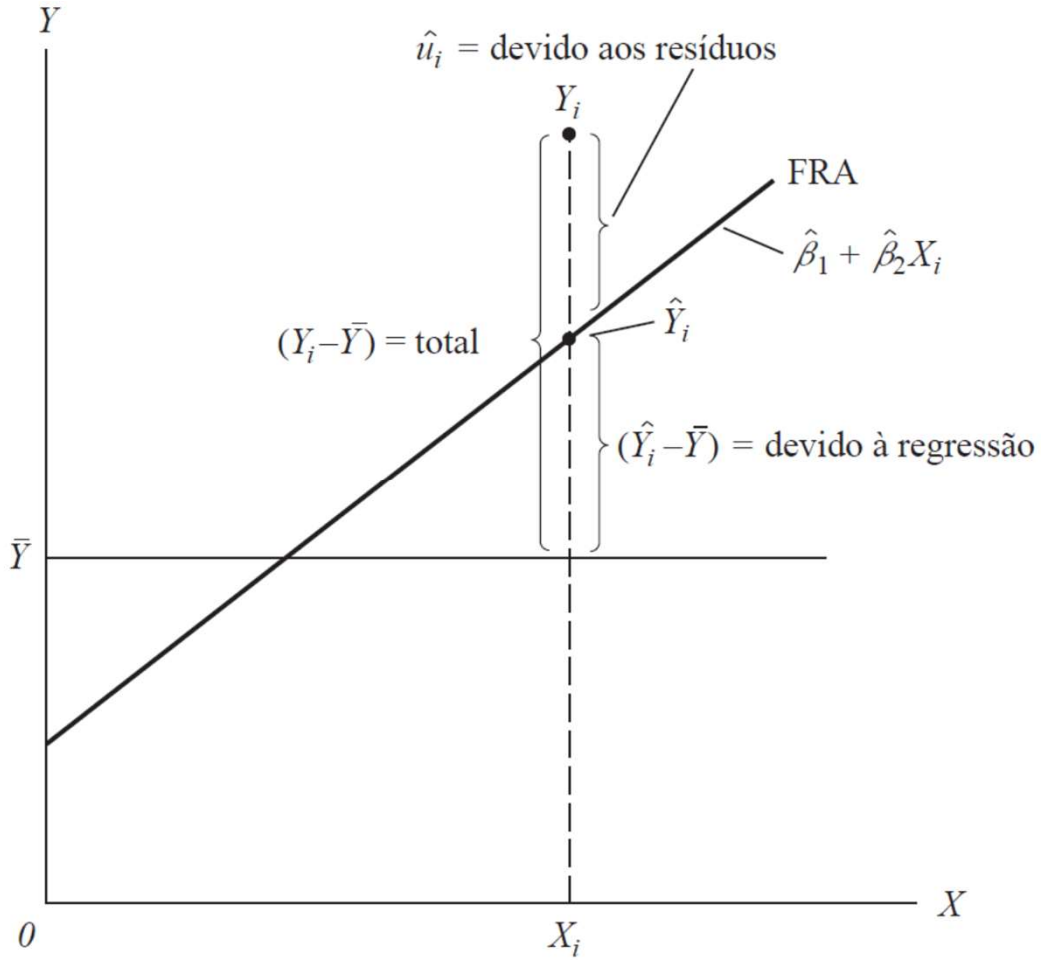
SQE = Soma dos quadrados explicados;

STQ = Soma total dos quadrados;

MSQ = Média da soma dos quadrados.

O Método dos Mínimos Quadrados Ordinários

ANOVA – Análise de Variância para uma Regressão por MQO



O Método dos Mínimos Quadrados Ordinários

Coeficiente de Determinação $(\boldsymbol{R^2})$

$$R^2 = \frac{SQE}{SQT} = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

➔Medida resumo que diz quanto a reta de regressão amostral se ajusta aos dados, portanto é uma medida da qualidade de ajustamento da reta.

➔Pode-se dizer que o coeficiente de determinação apresenta qual o percentual de explicação da variação total ocorrida na variável dependente, frente as variações das variáveis explicativas.

Por exemplo: R^2 = 0,85 significa que as variáveis explicativas conseguiram explicar 85% do comportamento (variações) da variável dependente.

O Método dos Mínimos Quadrados Ordinários

Coeficiente de Correlação Amostral (R) ou Coeficiente de Correlação Simples ou Coeficiente de Ordem Zero ou Coeficiente Produto Momento de Pearson

$$R = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \quad \text{ou} \quad r_{12} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$$

Em que:

r_{12}= correlação entre Y (1) e X (2) .

➔No caso de uma regressão simples com duas variáveis (uma variável Y – dependente; e uma variável X – explicativa) significa o grau de associação entre essas duas variáveis. Por exemplo: R = -0,75 significa que quando a vaiável Y cresce em uma unidade, a variável X decresce em média 0,75;

➔Portanto, R pode variar entre: $-1 \leq R \leq 1$;

➔Não é um indicador significantes para descrever relações não lineares;

➔Não implica ou apresenta qualquer relação de causa-efeito;

➔Para uma regressão múltipla (com várias variáveis X – explicativas), o valor de R pode representar a correlação conjunta dos valores de X frente a variável Y, mas é um indicador duvidoso.

O Método dos Mínimos Quadrados Ordinários

Matrizes de Correlação

- **Correlação Parcial ou de primeira ordem**: Indica a correlação entre duas variáveis, mantendo as demais variáveis do modelo constantes.

1. $r_{12,3}$= Coeficiente de correlação parcial entre Y e X_2, mantendo X_3 constante;
2. $r_{13,2}$= Coeficiente de correlação parcial entre Y e X_3, mantendo X_2 constante;
3. $r_{23,1}$= Coeficiente de correlação parcial entre X_2 e X_3, mantendo Y constante.

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \qquad r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

O Método dos Mínimos Quadrados Ordinários

A RAZÃO "F"

$$F = \frac{MSQ\ da\ SQE}{MSQ\ da\ SQR}$$

- A Razão F proporciona um teste da hipótese nula H_0: $\beta_s = 0$ ou alternativamente H_a: $\beta_s \neq 0$ estatisticamente.
- Como todas as quantidades que entram nessa equação podem ser obtidas por meio da amostra disponível, essa razão F oferece um teste estatístico para verificar a hipótese nula de que os verdadeiros β_s são estatisticamente iguais a zero.
- Calcula-se a razão F e compara-se com o valor crítico de F (distribuição F) apresentado nas tabelas F ao nível de significância escolhido ou obter o valor p da estatística F calculada.
- Com o teste de F, testa-se a "existência da reta de regressão", pois se todos os betas calculados forem estatisticamente iguais a zero, não existe reta de regressão. Em outras palavras, as variáveis explicativas não explicam nada do comportamento da variável dependente.

Testes de Hipótese com uma Amostra

Distribuição de Amostragem da Média

- Representa a distribuição de uma amostra de valores coletados de uma população.
- Uma distribuição de amostragem da média é uma distribuição de probabilidade para os possíveis valores da média da amostra $\bar{X}$, baseados em um tamanho da amostra particular.
- Para qualquer tamanho dado "n" de amostra de uma população com média μ, o valor da média da amostra $\bar{X}$ irá variar de amostra para amostra. Esta variação é a base da distribuição de amostragem.
- A distribuição de amostragem da média é descrita pela determinação do valor esperado $E(\bar{X})$, ou média, da distribuição e do desvio padrão da distribuição das médias $\sigma_{\bar{X}}$.
- O desvio padrão indica a acurácia da média da amostra como um estimador por ponto, $\sigma_{\bar{X}}$ é usualmente chamado de erro padrão da média.

Testes de Hipótese com uma Amostra

Distribuição de Amostragem da Média

Para uma população infinita (Amostra menor que 5% da população) :

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

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

Para população finita ou quando o erro padrão da população não é conhecido:

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

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

$\bar{X}$ = 180,00

σ = 14,00

n = 30

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

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

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

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

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

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;

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

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;

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

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;

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

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;

$\bar{X}_1$= 4000;

$\bar{X}_2$= 4600;

σ_1=200;

σ_2=250;

gl = 10 + 8 -2 = 16.

(4000 – 4600) $\pm$ 2,12 . 108,65

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

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

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;

$\bar{X}_1$= 4000;

$\bar{X}_2$= 4600;

σ_1=200;

σ_2=250;

gl = 10 + 8 -2 = 16.

(4000 – 4600) $\pm$ 2,12 . 108,65

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

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

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):

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

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 .

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

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

continua..

Testes de Hipótese com Duas Amostras

Teste de Diferença (Igualdade) entre duas médias

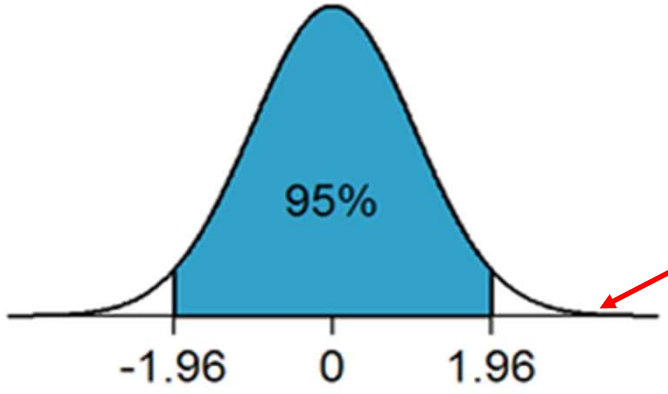
Exemplo:

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):

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

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

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



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;

σ_1= 2000;

σ_2= 2500;

σ_1^2 = 4000000;

σ_2^2 = 6250000.

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

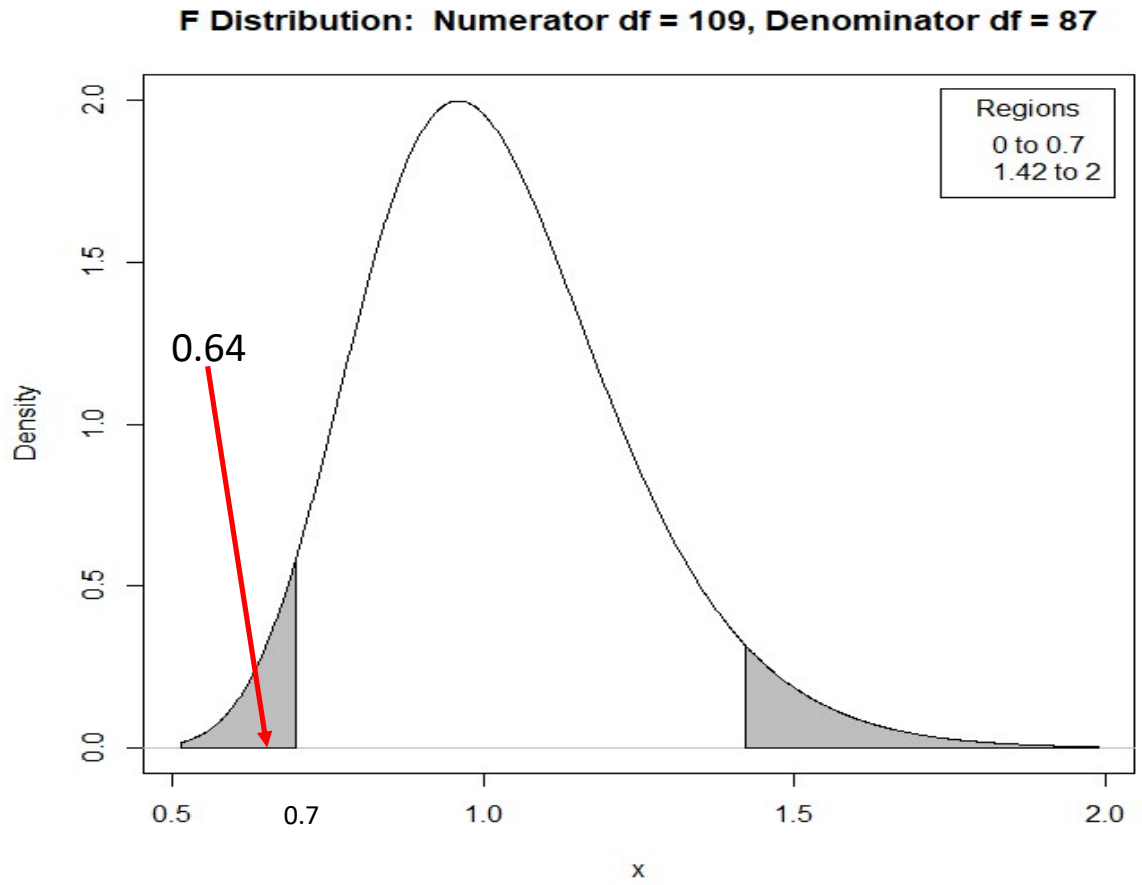
$$F_{gl_1,gl_2} = \frac{\sigma_1^2}{\sigma_2^2} = \frac{4000000}{6250000} = 0{,}64$$

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 $X_1, X_2, \dots, X_n$ e $Y_1, Y_2, \dots, Y_m$ 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

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

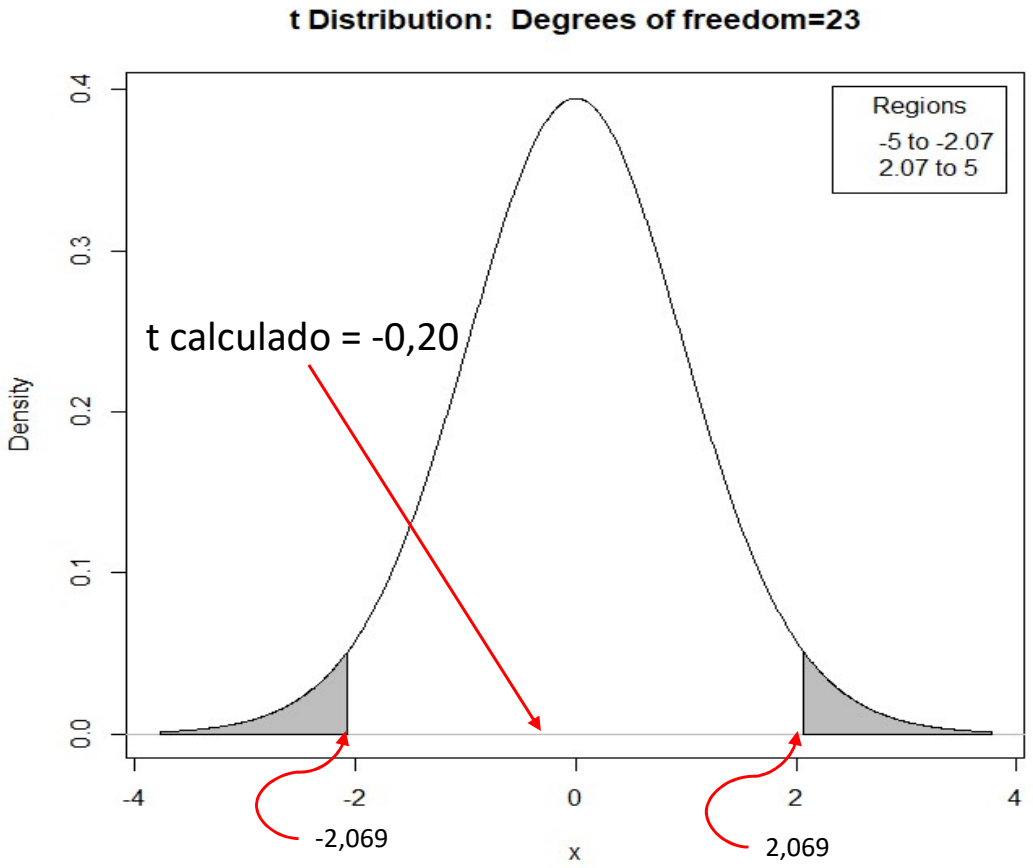
Testes de Hipótese com Duas Amostras

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) Teste de Wilcoxon-Mann-Whitney para amostras independentes

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) Teste de Wilcoxon-Mann-Whitney para amostras independentes

Seja $X_1, X_2, \dots, X_m$ os valores de uma amostra aleatória da população P_1; e $Y_1, Y_2, \dots, Y_n$ 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) Teste de Wilcoxon-Mann-Whitney para amostras independentes

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) Teste de Wilcoxon-Mann-Whitney para amostras independentes

Exemplo:

TESTE DE WILCOXON - INDEPENDENTES	
Resultados da Análise	

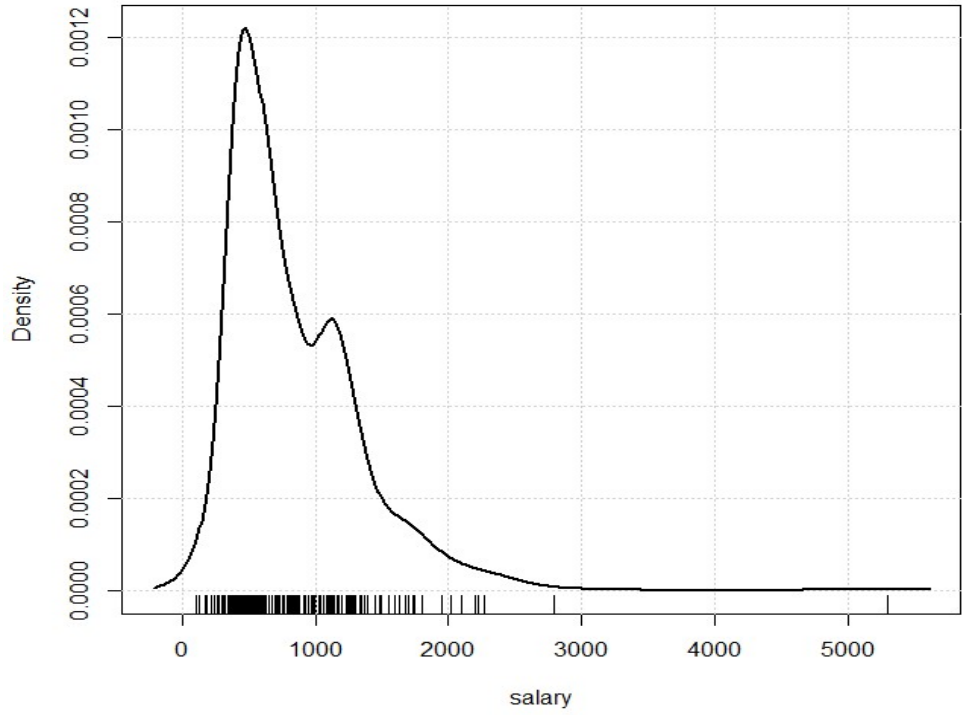
Tabela da Estatística do Teste (Wilcoxon)	
Informações	*Valores*
Estatística	141
P-valor	0,0373
Hipótese Nula	0
Limite Inferior	4
(Pseudo) Mediana	133,5
Limite Superior	240
Nível de Confiança	0,95

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

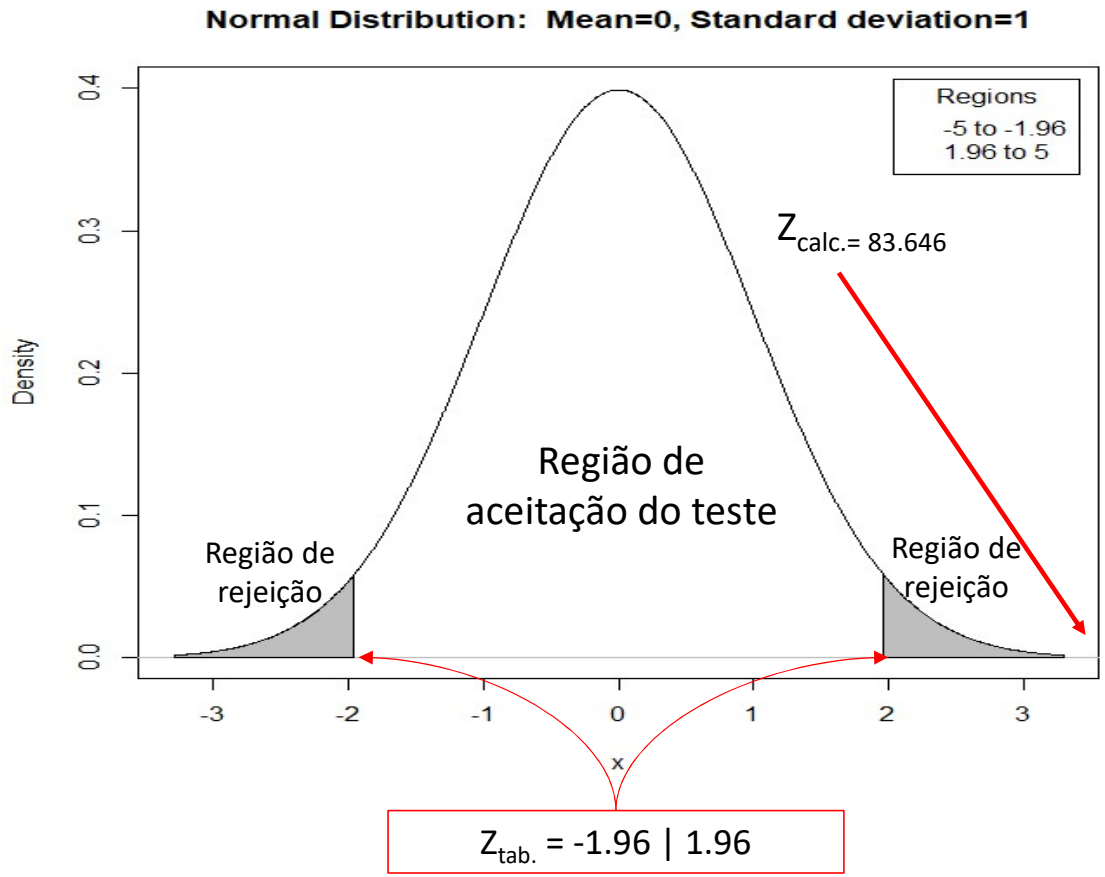
Teste de independência/equivalência de amostras

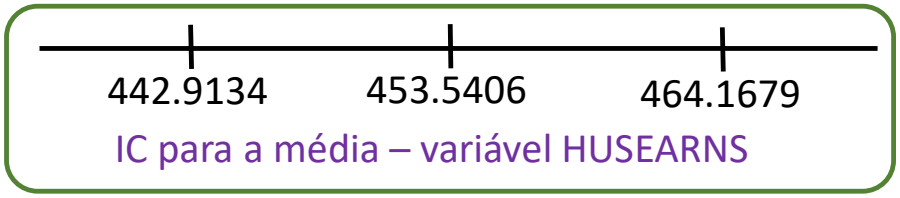
- Na prática de duas populações (e amostras) tem médias e variâncias estatisticamente iguais; pode-se dizer, a princípio que são populações (amostras) equivalentes.

- 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)



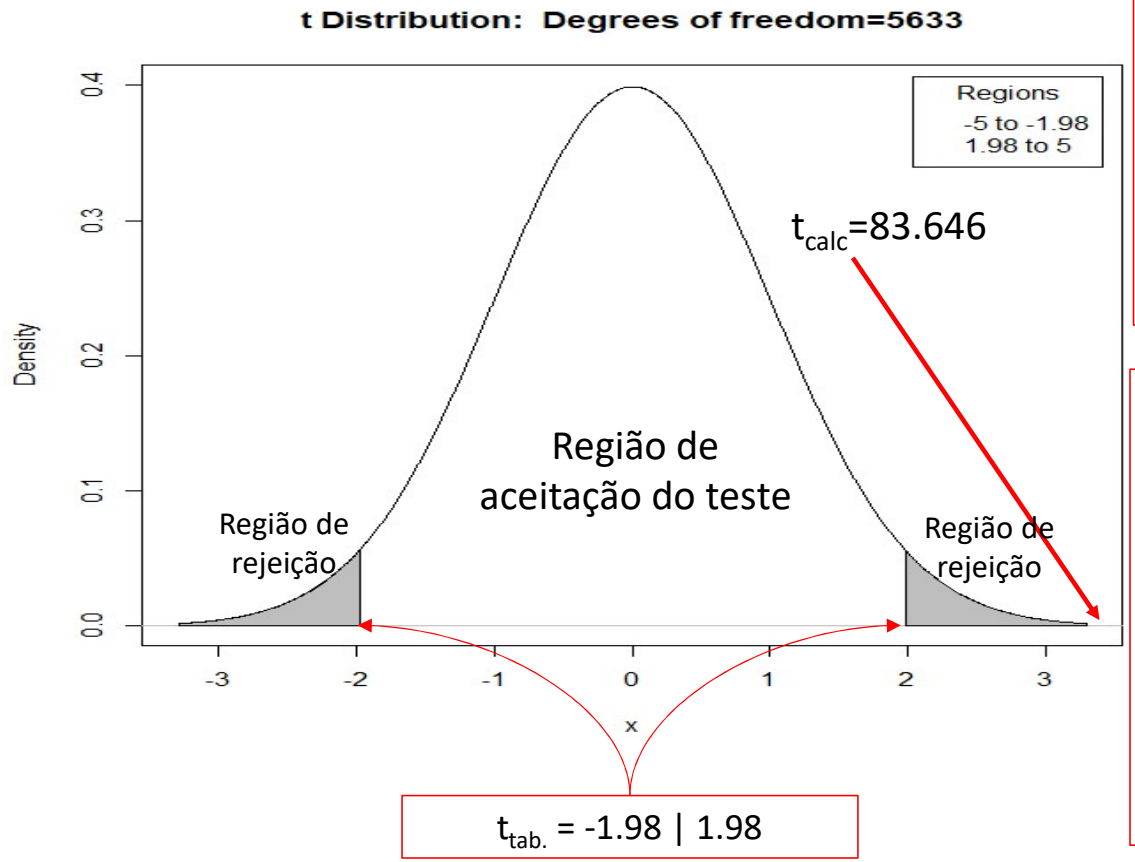


H0: valor verdadeiro da média de HUSEARNS = zero

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.

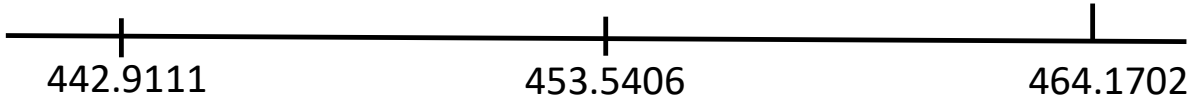
Intervalo de confiança para a média (t)



H0: valor verdadeiro da média de HUSEARNS = zero

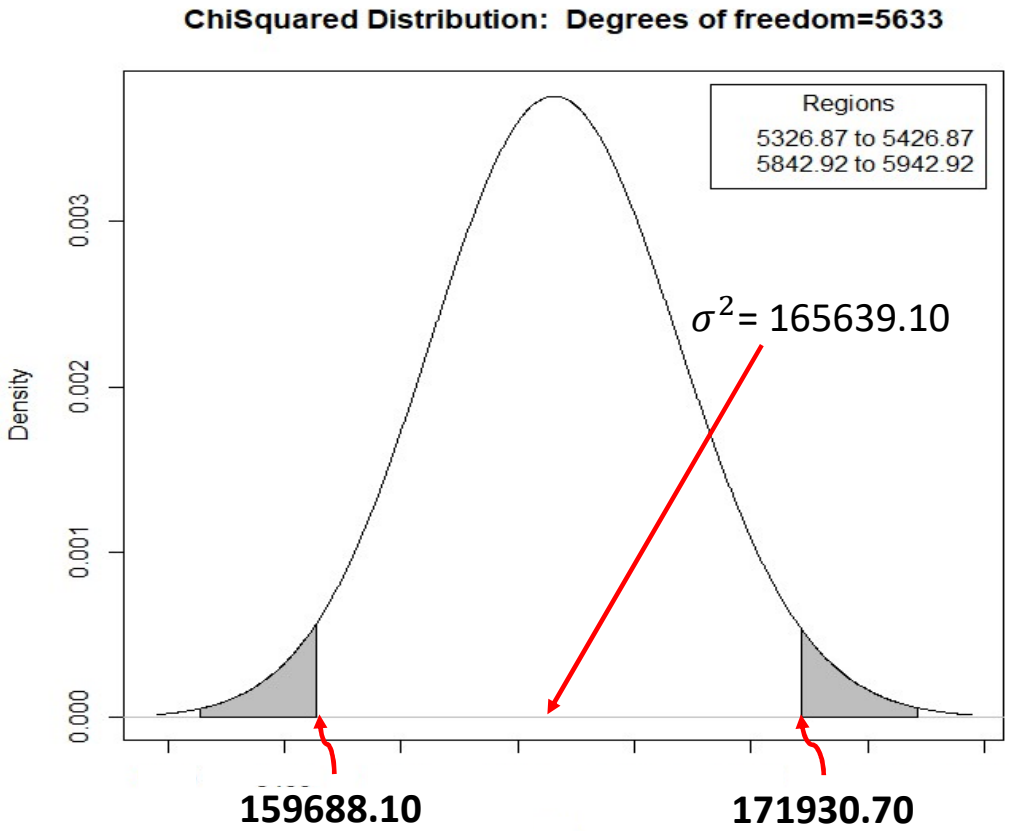
Ha: Valor verdadeiro da média de HUSEARNS ≠ zero

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.



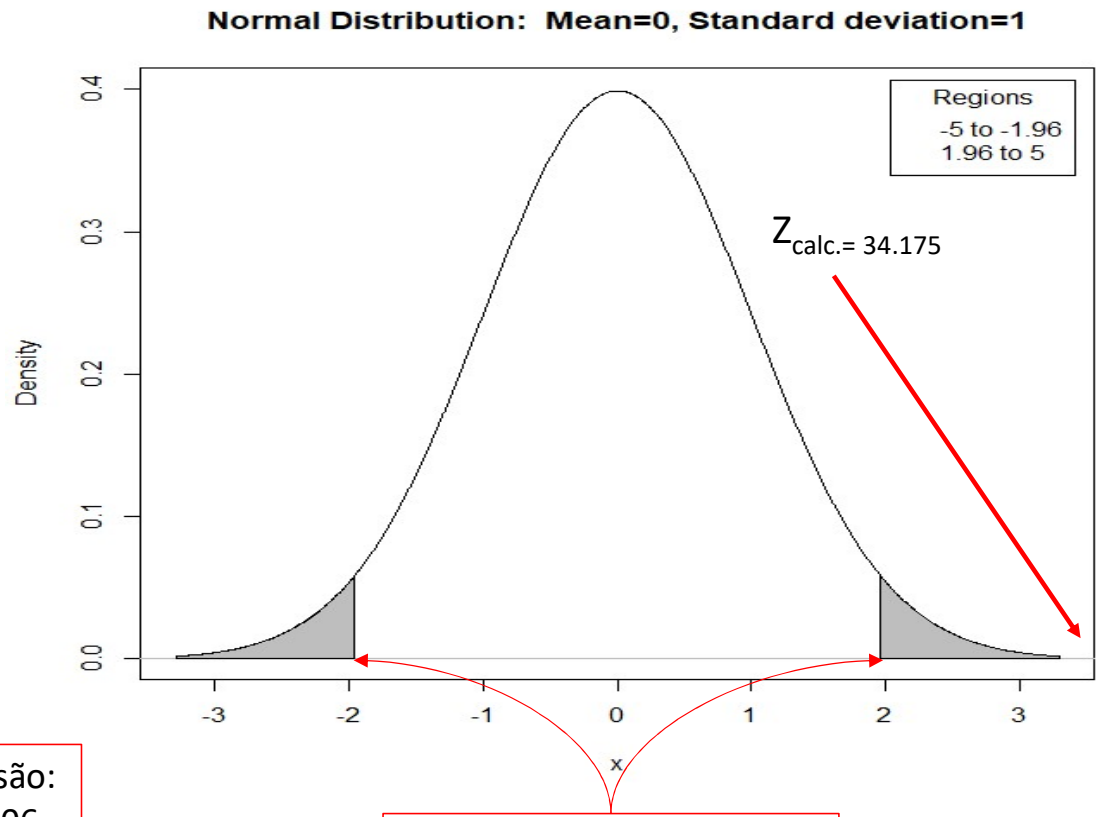
IC para a média – variável HUSEARNS

Intervalo de confiança para o desvio padrão



Intervalo de confiança construído com 95% de confiança

Teste da diferença entre duas médias (z)



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

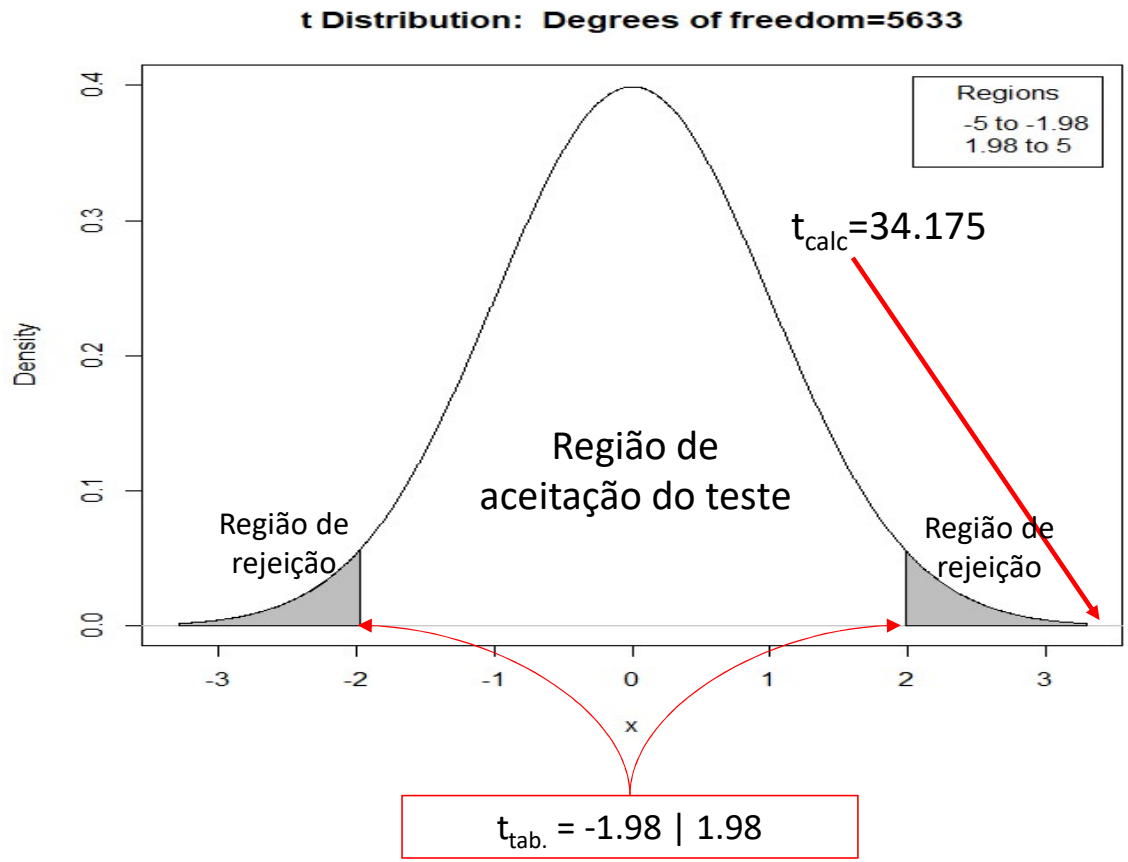
Os valores das médias são:
- HUSEARNS = 453.5406
- EARNS = 232.8330

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 a média (t)

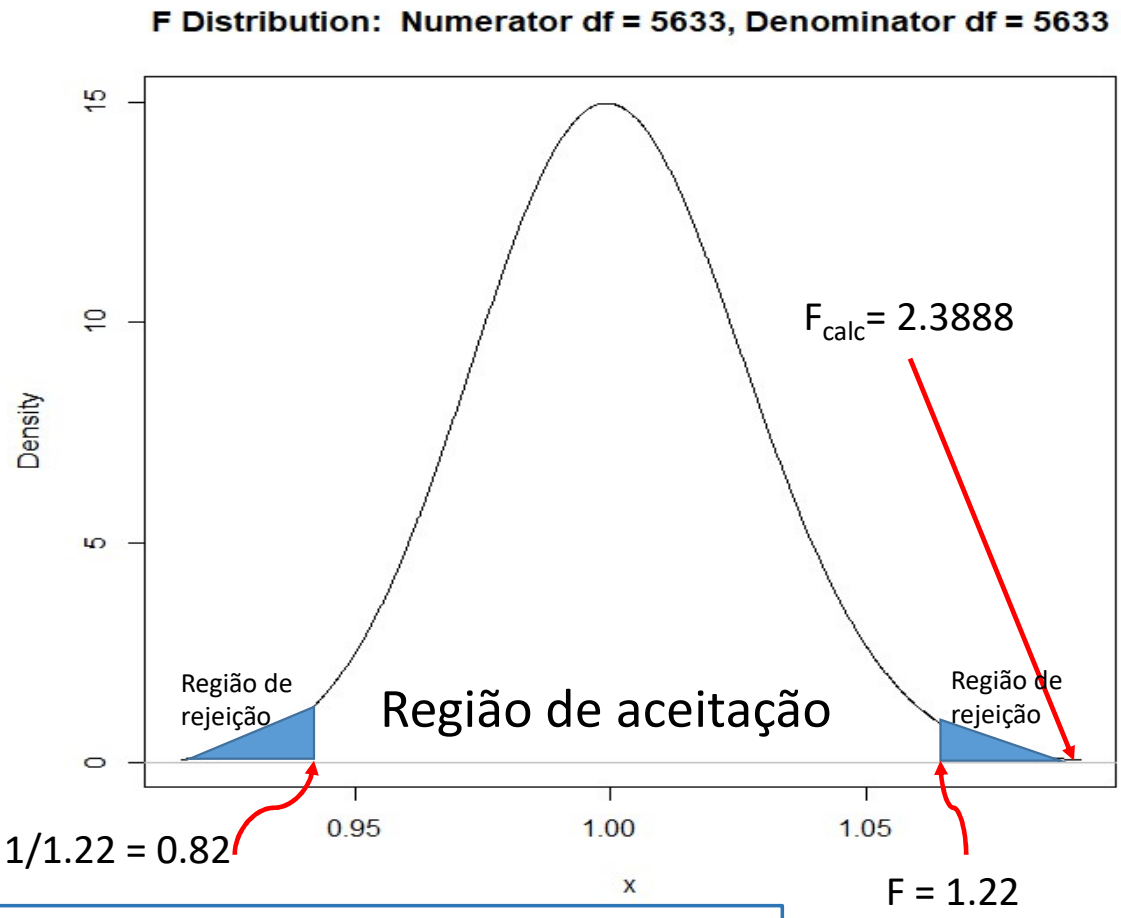


H0: A diferença verdadeira entre as médias é igual a zero

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)



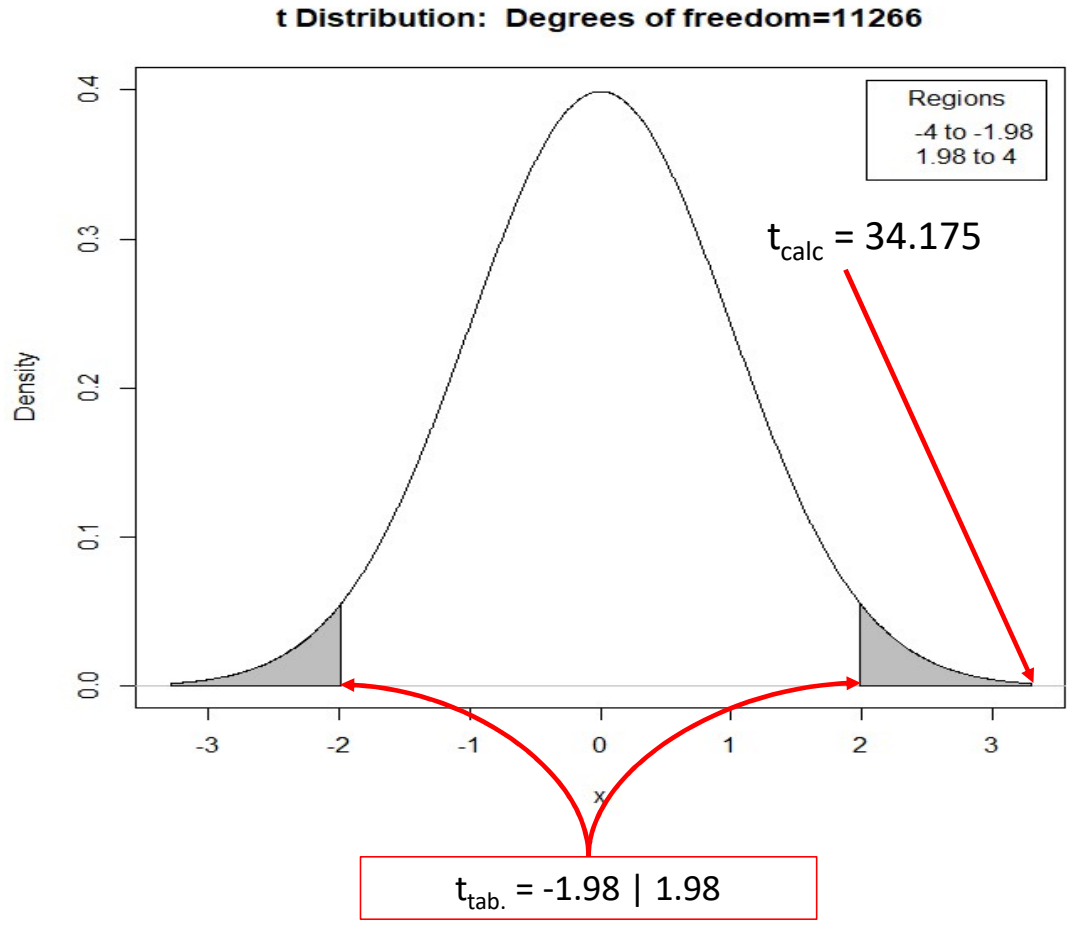
H0: A verdadeira razão entre as variâncias é igual a "um" (1)

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 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: http://www.portalaction.com.br/inferencia/62-teste-de-kolmogorov-smirnov

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

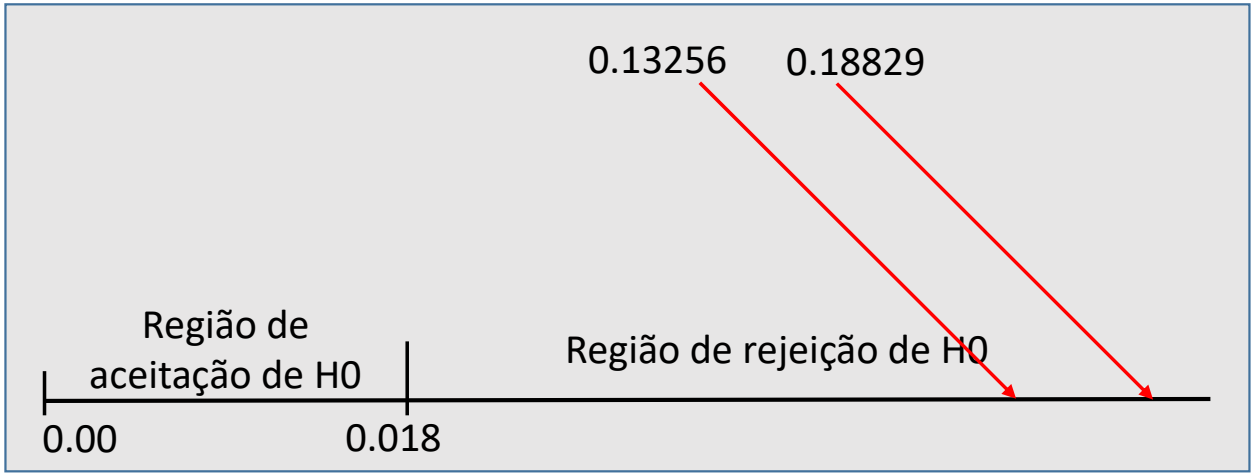
Resultados:

data: earns

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

data: 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.

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

Análise de regressão – Introd.

Coefficients:

		Estimate	Std. Error	t	value Pr(>|t|)	
(Intercept)	β_0	6.349e+02	4.027e+02	1.577	0.11673	
age	β_1	6.125e+00	5.552e+00	1.103	0.27146	
college	β_2	-1.527e+02	2.520e+02	-0.606	0.54541	
comten	β_3	-3.783e+00	3.872e+00	-0.977	0.33000	
grad	β_4	-5.935e+01	8.540e+01	-0.695	0.48805	
mktval	β_5	2.664e-02	9.717e-03	2.741	0.00677	**
sales	β_6	1.543e-02	1.019e-02	1.514	0.13195	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

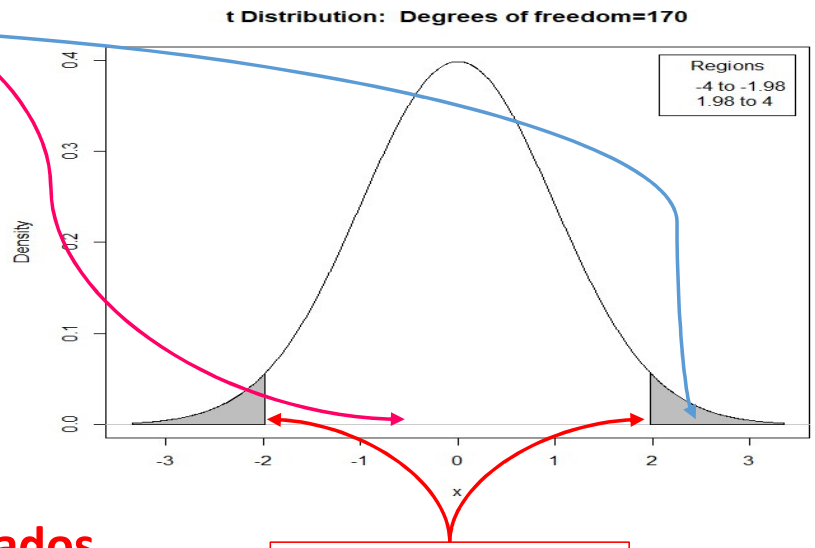
F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

Obs: deve-se fazer teste de t para todos os coeficientes calculados

Exemplos de testes de hipótese

H0: $\beta_4 = 0$
Ha: $\beta_4 \neq 0$

H0: $\beta_5 = 0$
Ha: $\beta_5 \neq 0$



t_{tab} = -1.98 | 1.98

Análise de regressão – Introd.

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

R-quadrado = 0.1896 ➔ quer dizer que as variáveis explicativas conseguem explicar 18,96% das variações da variável dependente. Ou seja, essas variáveis explicativas incluídas no modelo conseguem explicar somente 18,96% do salário dos CEOs. O R-quadrado considera apenas as variações ocorridas nas variáveis.

R-quadrado ajustado = 0.161 ➔ quer dizer que as variáveis explicativas conseguem explicar 16,10% das variações ocorridas nos salários dos CEOs. Este indicador considera as variações ocorridas nas variáveis, bem como o tamanho da amostra e o número de variáveis do modelo. Em outras palavras indica com maior precisão se o modelo é parcimonioso.

➔ Esse dois indicadores acima representam são usualmente chamados de coeficiente de determinação e apresentam a qualidade de ajustamento da reta de regressão aos dados das variáveis utilizadas no modelo.

Análise de regressão – Introd.

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

A estatística F testa se todos os parâmetros são estatisticamente diferentes de "zero", ou seja se existe reta de regressão.

$H0: \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 = 0$

$Ha: \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \neq 0$

Resultado do teste:

Como a estatística calculada situa-se na área de rejeição, rejeita-se H0 de que todos os coeficientes conjuntamente são iguais a "zero", em favor da hipótese alternativa de que pelo menos um dos parâmetros calculados é diferente de "zero", ou seja, existe reta de regressão.