Núcleos de Desenvolvimento

Correlação e Regressão Linear

ESEG
FACULDADE DO GRUPO ETAPA

Covariância – Definição

A covariância determina se dois conjuntos de dados, de duas variáveis, possuem movimentos sincronizados, ou seja, se os maiores valores de uma variável estiverem associados aos maiores valores da outra variável, poderá existir uma covariância positiva (representada pelo sinal +) entre os dados das duas variáveis.

Se os maiores valores de uma variável estiverem associados aos menores valores da outra variável, poderá existir uma covariância negativa (representada pelo sinal -) entre os dados das duas variáveis.

Se os valores das duas variáveis não estiverem associados, ou relacionados, a covariância será próxima de zero.

Covariância em população – Definição

$Cov(X,Y) = E[(X-E(X)).(Y-E(Y))]$

Interpretação: Covariância é o valor médio dos produtos dos desvios de cada dado das variáveis aleatórias X e Y em relação às suas respectivas médias.

Simplificando a fórmula: $Cov(X,Y) = E(X.Y)-E(X).E(Y)$

$$-\infty < Cov(X, Y) < +\infty$$

Covariância em amostra – Definição

A covariância medida em amostras de duas variáveis aleatórias X e Y:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X}).(Y_i - \bar{Y})}{n-1}$$

$$-\infty < \text{Cov}(X, Y) < +\infty$$

ESEG
FACULDADE DO GRUPO ETAPA

Coeficiente Linear de Correlação em população – Definição

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

Onde σ_X é o desvio padrão da variável X e σ_Y é o desvio padrão da variável Y.

$$-1 < \rho(X, Y) < +1$$

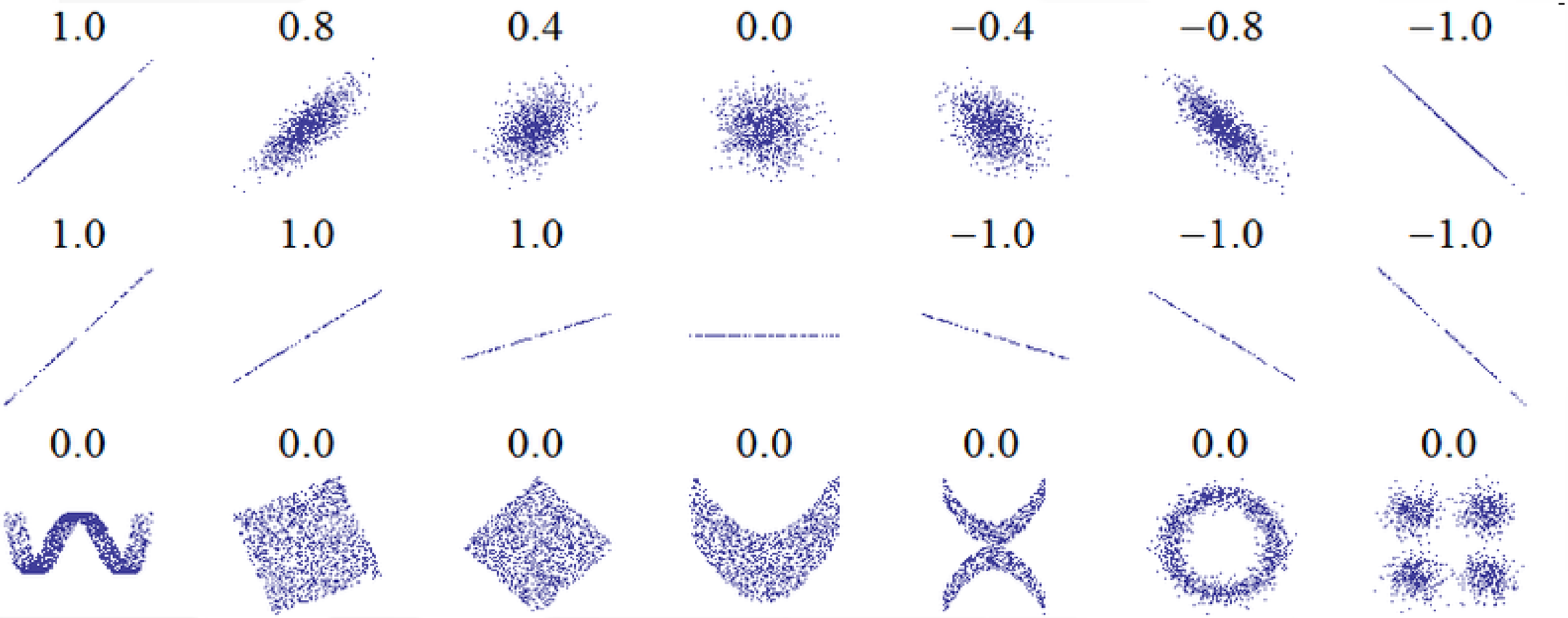
Coeficiente Linear de Correlação em amostra – Definição

$$r(X,Y) = \frac{Cov(X,Y)}{S_X * S_Y}$$

Onde S_X é o desvio padrão da variável X e S_Y é o desvio padrão da variável Y.

$$-1 < r(X, Y) < +1$$

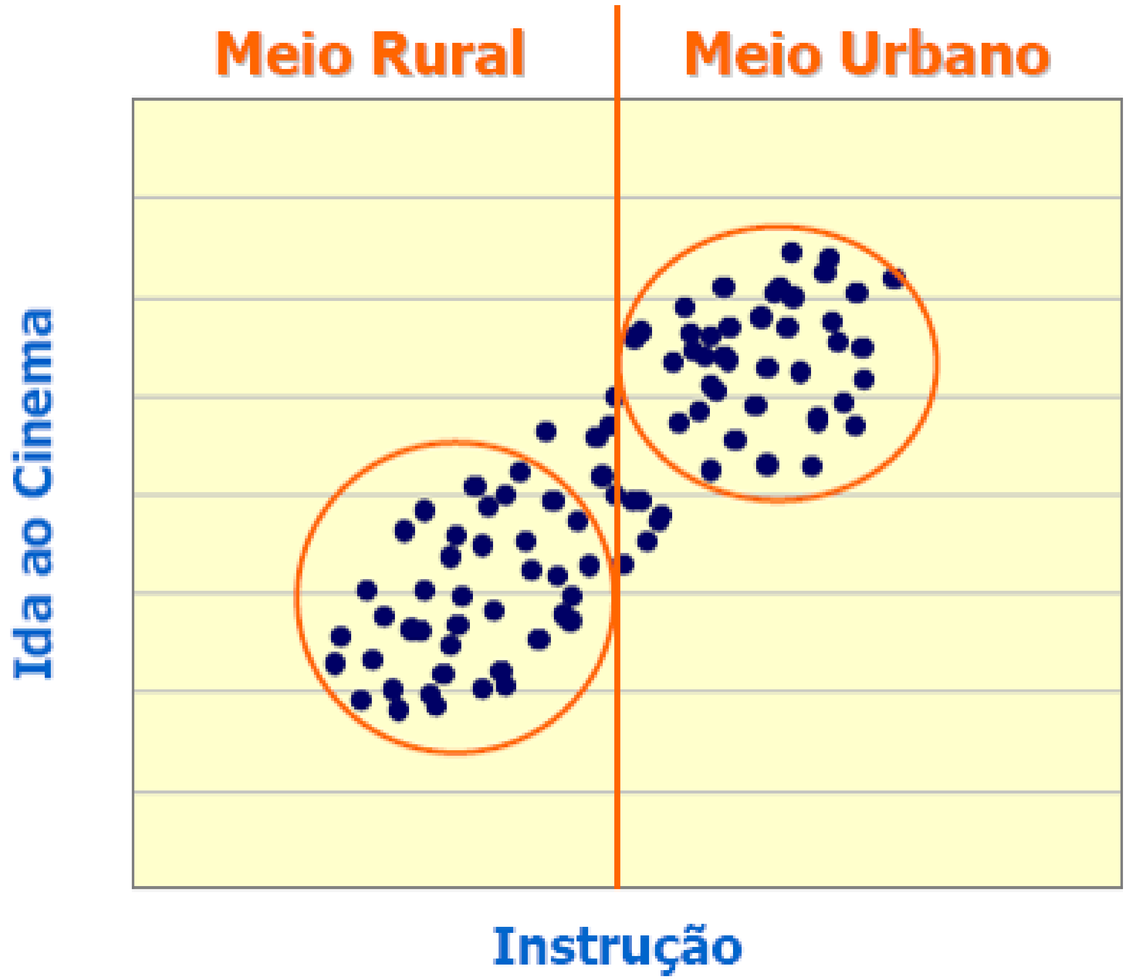
Importante: Correlação zero não significa independência entre variáveis!



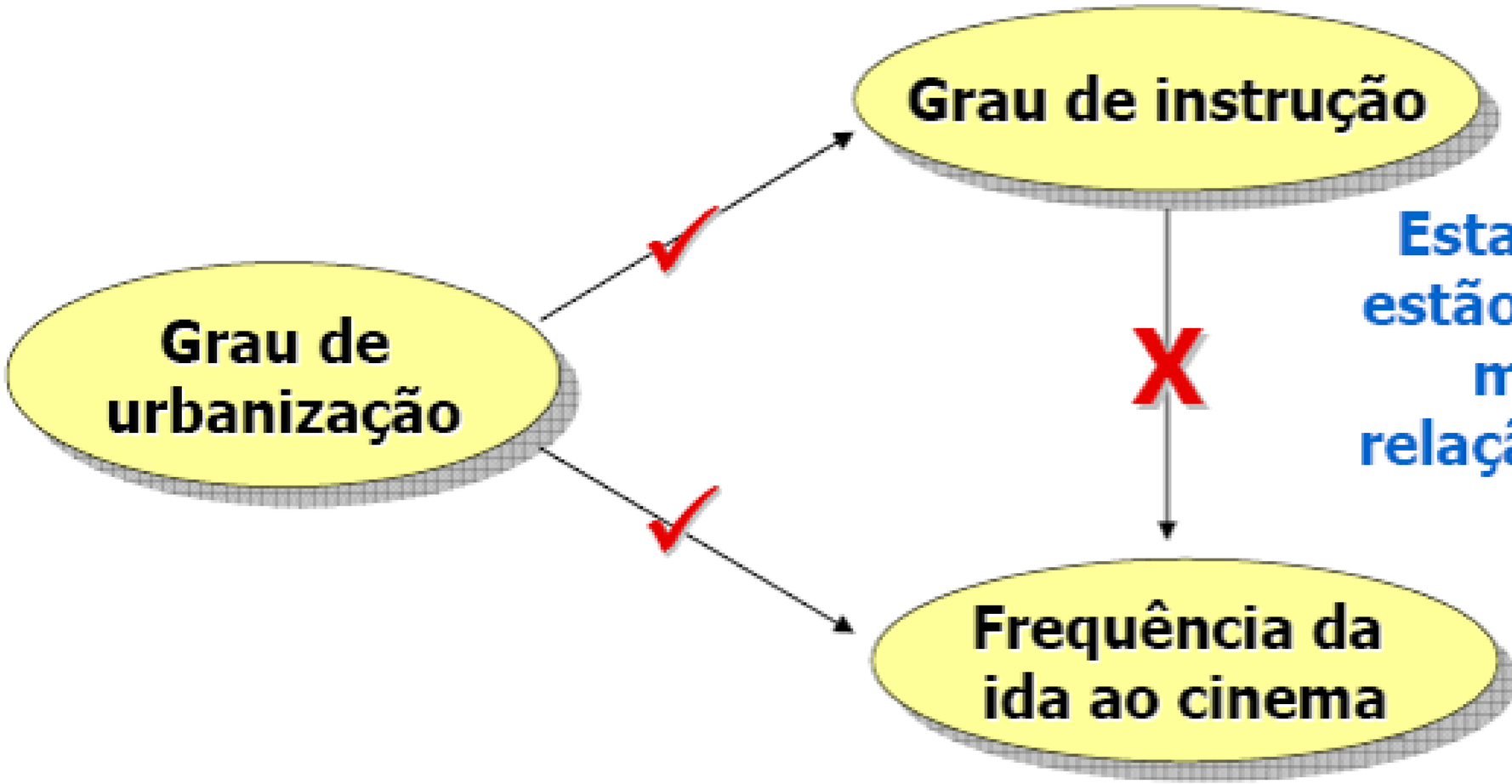
ESEG
FACULDADE DO GRUPO ETAPA

Importante: Correlação alta não significa causalidade!

Isso quer dizer que se uma pessoa é mais instruída ela vai mais ao cinema? Veja o que pode ter acontecido nessa pesquisa:



Separando as observações:



Estas duas variáveis estão correlacionadas, mas não existe relação de causalidade entre elas!

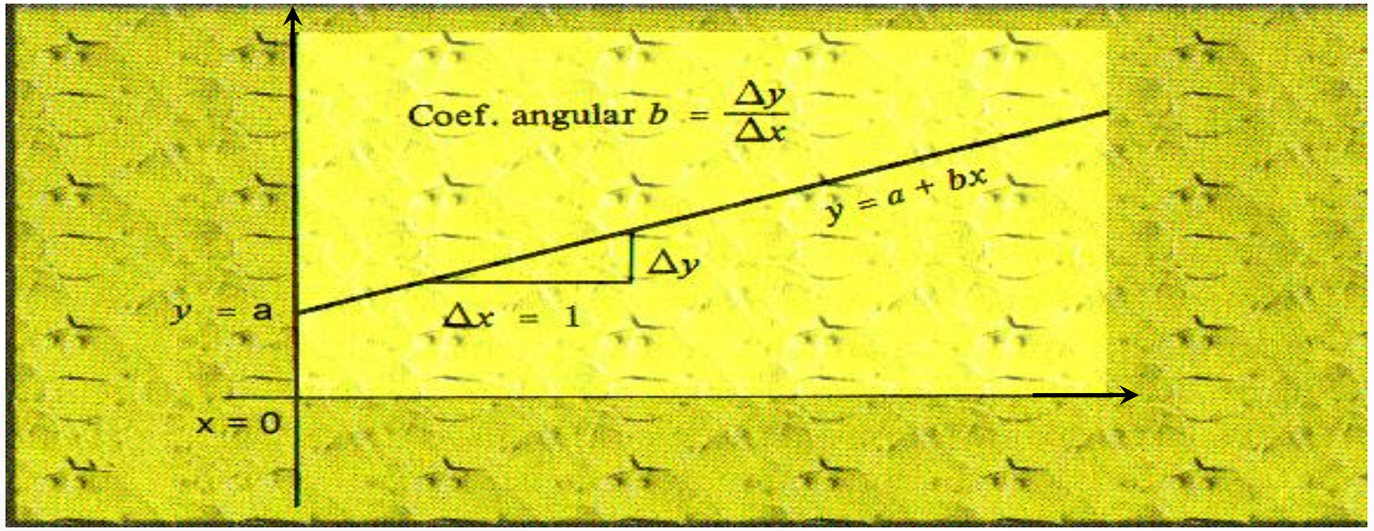
ESEG
FACULDADE DO GRUPO ETAPA

Correlação Linear e Regressão Linear

- A regressão e a correlação são duas técnicas relacionadas que envolvem uma forma de estimação.
- A análise da correlação e regressão compreende a análise de dados amostrais para saber se, e como, duas ou mais variáveis estão relacionadas uma com a outra em uma população.
- A correlação mede a força, ou o grau, de relacionamento linear entre duas variáveis; a regressão fornece uma equação que descreve o relacionamento entre essas variáveis, em termos matemáticos. A equação pode ser usada para estimar, ou predizer, valores futuros de uma variável quando se conhecem ou se supõem conhecidos os valores da outra variável.

Regressão Linear

- Regressão linear simples
  A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear (linha reta) que descreva o relacionamento entre 2 variáveis aleatórias.
- A função linear
  Uma equação linear tem a forma y = a + b.x onde a e b são valores que se determinam com base nos dados amostrais; a é a cota da reta em X = 0 e b é o coeficiente angular da reta.



ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

- Regressão linear simples
  Na regressão, os valores Y são preditos com base em valores dados ou conhecidos de X. A variável Y é chamada variável dependente, explicada ou endógena, e a variável x é a variável independente, explicativa ou exógena.

- O método dos mínimos quadrados
  O método dos mínimos quadrados é o método mais usado para ajustar uma linha reta a um conjunto de pontos. A reta resultante tem 2 características importantes: 1. A soma dos desvios verticais dos pontos em relação à reta é zero; 2. A soma dos quadrados desses desvios é mínima; isto é, nenhuma outra reta daria menor soma de quadrados de tais desvios.

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

Método dos Mínimos Quadrados

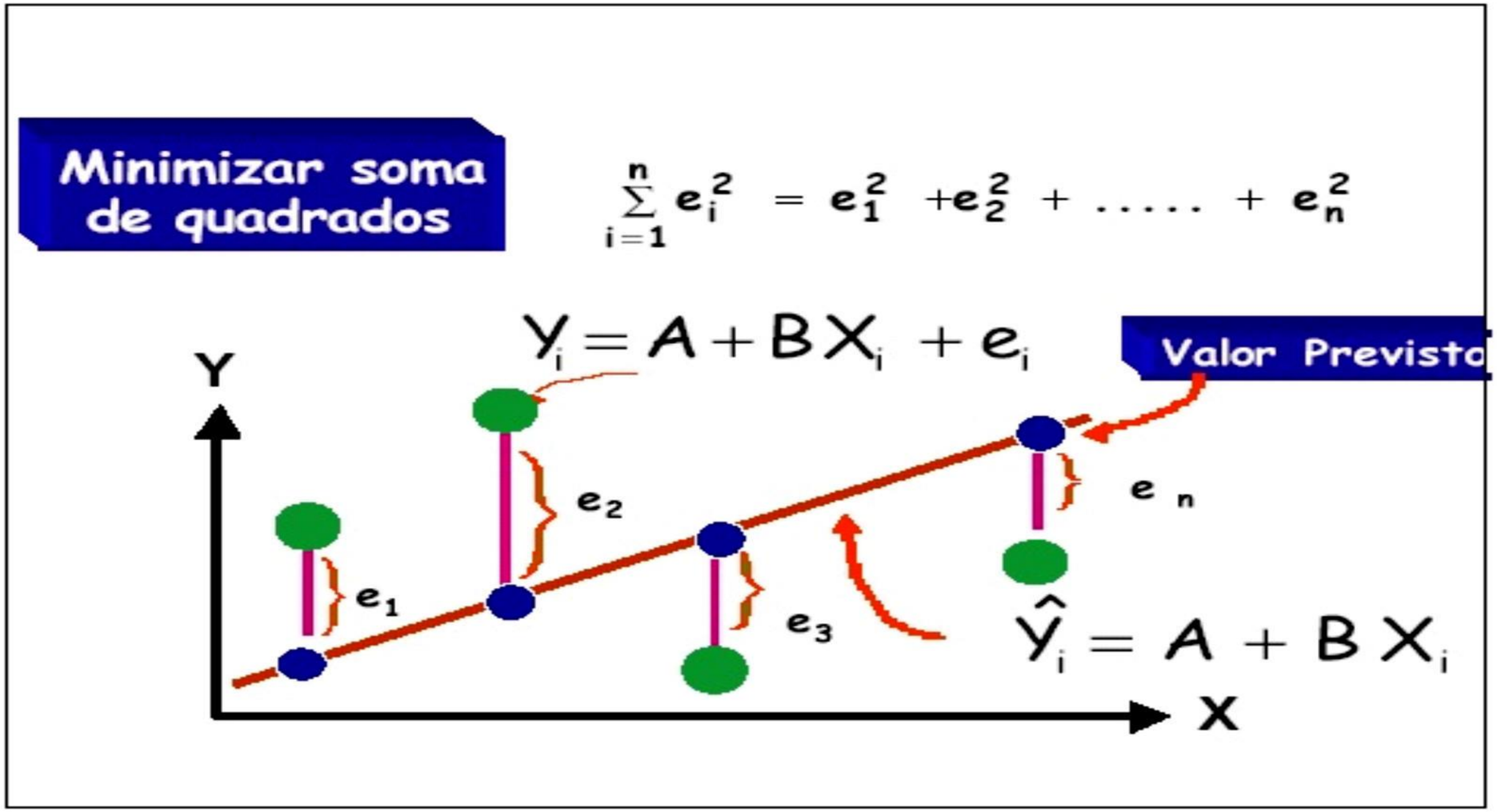
1. 'Melhor Ajuste' A diferença entre o valor Real (Y_i) e o valor previsto ($\hat{y}_i$) é mínima
   As diferenças positivas anulam diferenças negativas

2. Minimizar a soma das diferenças ao quadrado (ou Erros)

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear



Regressão Linear

Equação da Regressão
Parâmetros da Amostra

$$\hat{Y}_i = A + B X_i$$

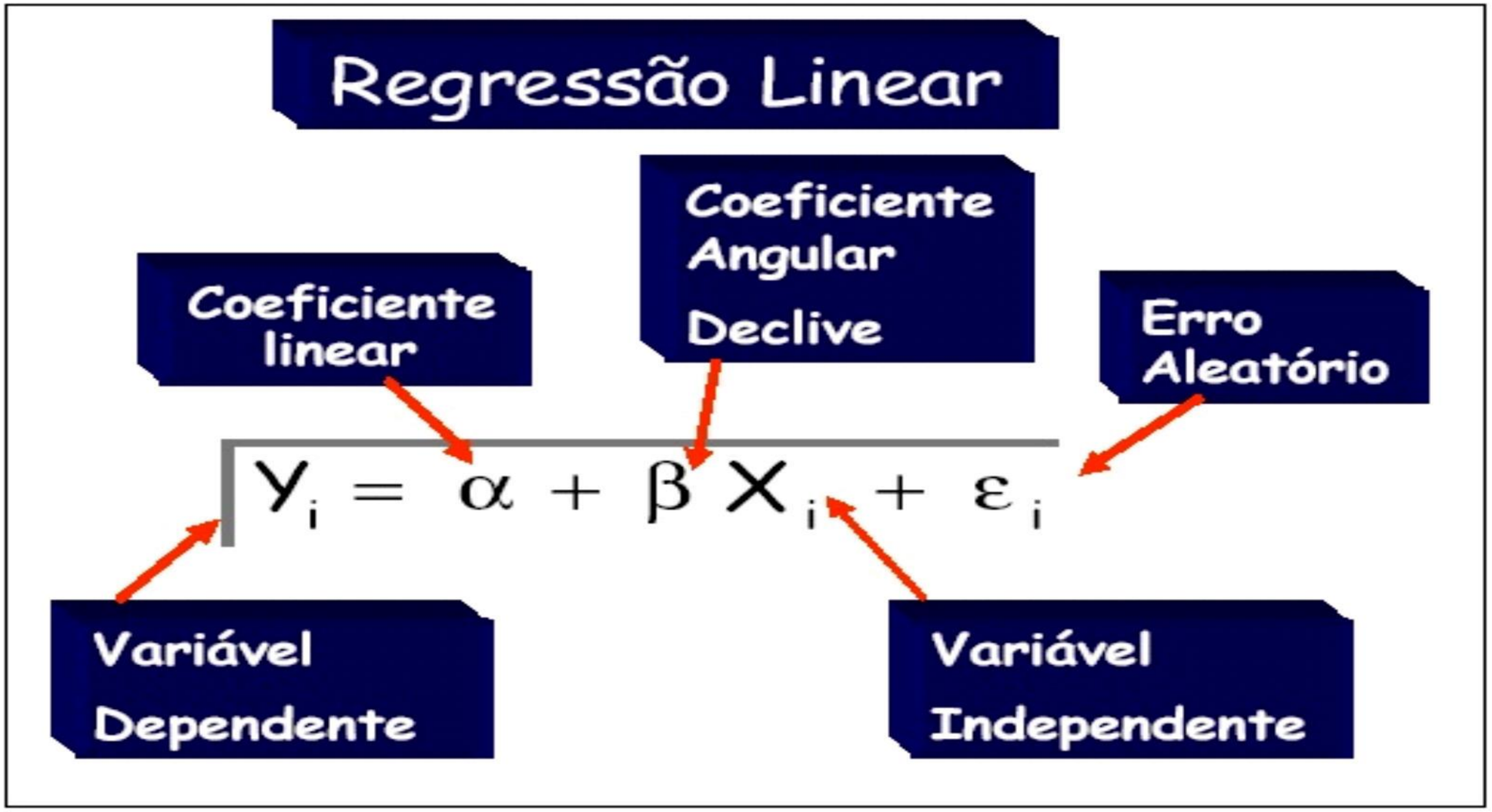
Coeficiente Angular Declive

$$B = \frac{S_{XY}}{S_{XX}}$$

Coeficiente linear

$$A = \overline{Y} - B\overline{X}$$

Regressão Linear

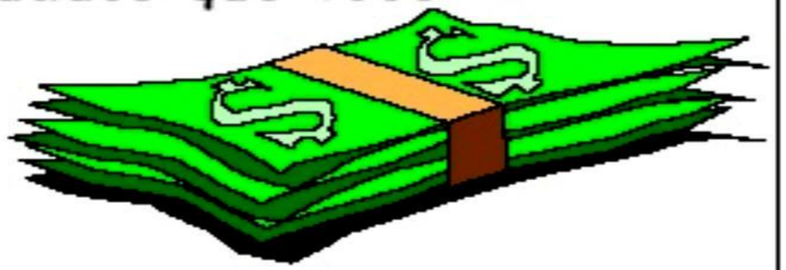


Regressão Linear

Exemplo: Receita $ x Quantidade

Admitindo que Y é a receita de uma empresa comercial em certo intervalo de tempo e que X é a quantidade vendida. Os dados que você tem são os seguintes

Quantidade	Receita
10	100
20	100
30	200
40	200
50	400



Qual a relação entre Receita e Quantidade?

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

Exemplo (Receita $ x Quantidade)
Tabela de Resultados

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
10	100	100	10000	1000
20	100	400	10000	2000
30	200	900	40000	6000
40	200	1600	40000	8000
50	400	2500	160000	20000
150	1000	5500	260000	37000

Regressão Linear

$$S_{xx} = \sum (X_i - \overline{X})^2 = \sum X^2 - \frac{1}{n}(\sum X)^2$$

$$S_{xx} = 5500 - \frac{1}{5}(150)^2 = 1000$$

$$S_{yy} = \sum (Y_i - \overline{Y})^2 = \sum Y^2 - \frac{1}{n}(\sum Y)^2$$

$$S_{yy} = 260000 - \frac{1}{5}(1000)^2 = 60000$$

$$S_{xy} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum XY - \frac{1}{n}\sum X \sum Y$$

$$S_{XY} = 37000 - \frac{1}{5}150.1000 = 7000$$

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

Exemplo (Receita $ x Quantidade)
Cálculos

$S_{XX}=1000$ $S_{YY}=60000$ $S_{XY}=7000$

$$B=\frac{S_{XY}}{S_{XX}}=\frac{7000}{1000}=7$$

Declive

$$A=\bar{Y}-B\bar{X}=$$

$$A=200-7.30=-10$$

Interseção

$$\hat{y}_i = -10 + 7X_i$$

ESEG
FACULDADE DO GRUPO ETAPA

Regressão Linear

- O coeficiente de inclinação da reta “b” = 7 indica que para um incremento unitário na variável X (1 unidade a mais vendida) implica em um incremento na variável Y de R$7,00 (7 unidades a mais na Receita).

- A interseção “a” indica que quando não houver venda de produtos (X = 0 unidades vendidas), a Receita será negativa de R$ 10,00.

Testes de hipótese

- O objetivo de um teste de hipótese é verificar se calculado um parâmetro estatístico em determinada característica de uma amostra, esse parâmetro poderá representar o mesmo parâmetro da mesma característica de toda a população de observações, com um nível de significância α (normalmente α = 5% = 0,05).

- Existem muitos testes de hipótese para diversas finalidades. Por exemplo, o teste *t de student* para verificar a significância de uma média amostral de determinada característica; o teste *t de student* para verificar se duas médias amostrais são iguais; o teste *t de student* para verificar a significância dos coeficientes de uma regressão.

Teste de hipótese $t_{student}$

- O teste *t de student* para verificar a significância de uma média amostral de determinada característica, ou de cada coeficiente de uma regressão linear multivariada.

- Teste de uma amostra:

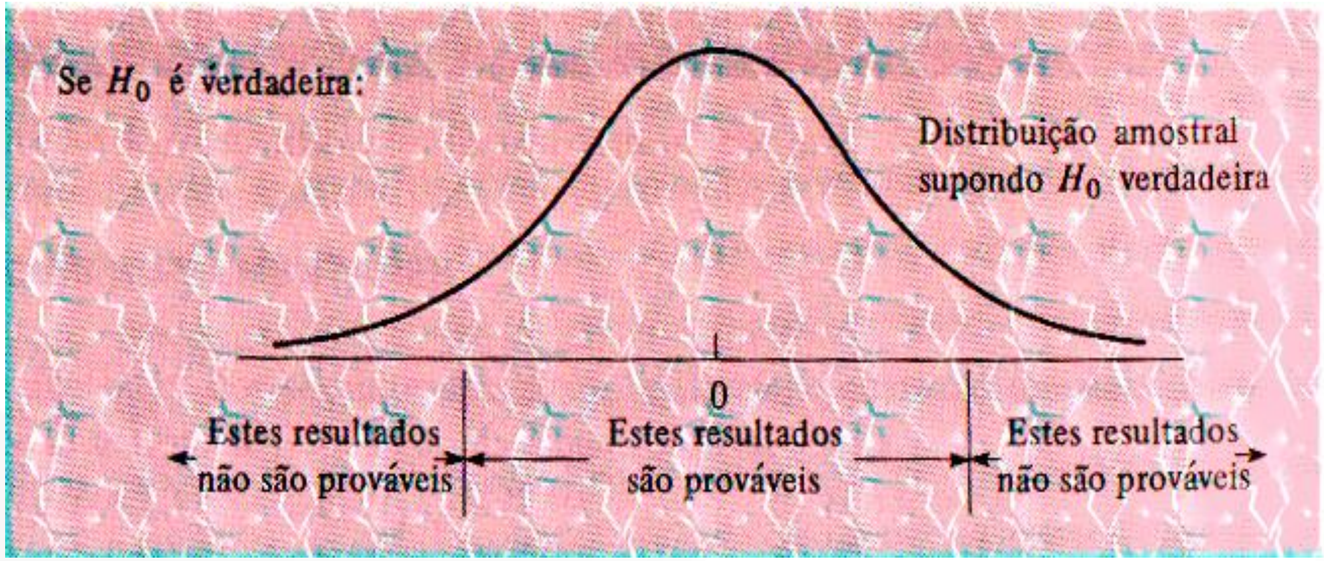
  - 1. Estabelecer a hipótese nula e a alternativa.
  - 2. Identificar uma distribuição amostral adequada – a maior parte dos testes envolve a distribuição normal ou a *t de student*.
  - 3. Particionar a distribuição amostral em regiões de aceitação – variações provavelmente causais – e de rejeição – variações provavelmente não causais.

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- 4. Calcular a estatística teste.
- 5. Comparar a estatística teste amostral com o valor crítico tabelado. Rejeitar H_0 se o valor da estatística teste calculado for maior que o módulo do valor crítico tabelado.

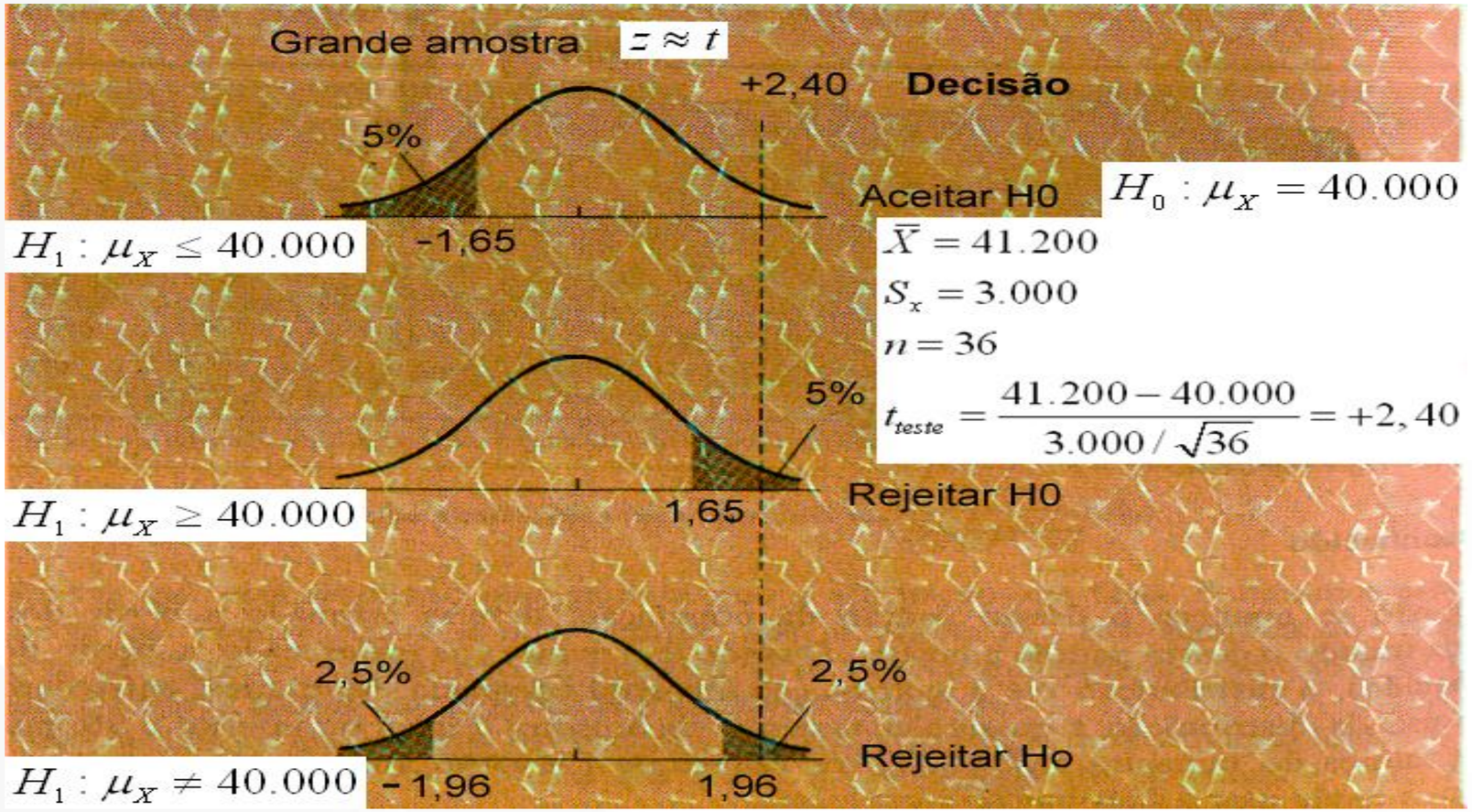
- Exemplo: Um fabricante de pneus afirma que seus pneus rodam em média 40.000 milhas. Testar essa hipótese.



ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- A estatística teste = (média amostral – média alegada) / desvio padrão da distribuição amostral



ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- Teste de duas amostras para médias:
  - Os testes de duas amostras são usados para decidir se as médias de duas populações são iguais.
  - Exige-se duas amostras independentes, uma de cada população.
  - Os testes de duas amostras são frequentemente usados para comparar dois métodos de ensino, duas marcas, duas cidades, dois distritos escolares etc.
  - A hipótese nula pode ser a de que as duas populações têm médias iguais:

    $H_0: \mu_1 = \mu_2 \leftrightarrow \mu_1 - \mu_2 = 0$

    $H_1: \mu_1 \neq \mu_2 \leftrightarrow \mu_1 - \mu_2 \neq 0$; $H_1: \mu_1 > \mu_2$; $H_1: \mu_1 < \mu_2$

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- O teste focaliza a diferença relativa entre as médias de duas amostras, uma de cada população.

- Esta diferença é dividida pelo desvio padrão de uma distribuição amostral.

- Calcula-se primeiro o desvio padrão, supondo H_0 verdadeira. Em tal caso, as duas amostras podem ser consideradas como provenientes da mesma população, e mediante combinação (*pooling*) das variâncias das duas populações (ou das duas amostras, se as variâncias da população são desconhecidas), pode-se determinar a variância global da população.

Testes de hipótese $t_{student}$

- Quando σ_1 e σ_2 são conhecidos: $$Z_{teste} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Quando H_0 é verdadeira, pode-se assumir que o verdadeiro valor de z seja distribuído normalmente com média zero e desvio padrão um – distribuição normal padronizada – para os casos em que a soma $n_1 + n_2$ é maior que 30.
- Para menores amostras, Z só terá distribuição normal se as duas populações em estudo forem normais.

ESEG
FACULDADE DO GRUPO ETAPA

Exemplo:

Uma fábrica de tintas está interessada em reduzir o tempo de secagem de uma tinta. Duas formulações de tinta são testadas:

a formulação 1 tem uma química padrão e a formulação 2 tem um novo ingrediente para secagem que deve reduzir o tempo

de secagem. Da experiência, sabe-se que o desvio padrão do tempo de secagem é 8 min e essa variabilidade não deve ser afetada pelo novo ingrediente.

Dez produtos são pintados com a formulação 1 e outros dez são pintados com a formulação 2. os 20 produtos são pintados em uma ordem aleatória. Os tempos médios de secagem das duas amostras são de 121 min e 112 min, respectivamente. Quais as conclusões que a fábrica de tintas pode tirar sobre a eficácia do novo ingrediente usando $\alpha = 0,05$?

ESEG
FACULDADE DO GRUPO ETAPA

Exemplo:

Deseja-se testar se existe diferença entre os tempos de secagem das tintas. Então,

H_0: $\mu_1 = \mu_2 \leftrightarrow \mu_1 - \mu_2 = 0$

H_1: $\mu_1 > \mu_2 \leftrightarrow \mu_1 - \mu_2 > 0$

$\alpha = 0{,}05$

A estatística de teste é $Z_{teste} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{121-112}{\sqrt{\dfrac{8^2}{10}+\dfrac{8^2}{10}}} = 2{,}52$

Rejeita-se H_0 se $Z_{teste} > 1{,}645 = Z_{0,05}$

Como $Z_{teste} = 2{,}52 > 1{,}645$, rejeita-se H_0 e conclui-se que a adição do novo ingrediente à tinta reduz o tempo de secagem.

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- Quando os desvios padrão populacionais são desconhecidos, a estatística toma a forma:

$$t_{teste} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{X1}^2}{n_1} + \frac{S_{X2}^2}{n_2}}}$$

- O valor de t, supondo H_0 verdadeira, pode ser bem aproximado por Z se $n_1 + n_2$ excede 30.

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese $t_{student}$

- Quando os tamanhos das duas amostras não são iguais e sua soma é menor que 30, a fórmula da estatística de teste é:

$$t_{teste} \sim \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1 - 1)S_{X1}^{\ 2} + (n_2 - 1)S_{X2}^{\ 2}}{n_1 + n_2 - 2}\right]\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- O valor de t quando H_0 é verdadeiro tem distribuição t com $n_1 + n_2 - 2$ graus de liberdade, desde que se possa admitir que ambas as populações sejam aproximadamente normais.

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese – F de Snedecor

- O teste F de Fisher-Snedecor – é um teste empregado na Análise de Variância (ANOVA), que tem como objetivo verificar a igualdade de médias amostrais de três ou mais amostras independentes. É usada também na verificação da significância do R^2 ajustado de uma regressão linear multivariada, com vistas a demonstrar a significância ou não da regressão.

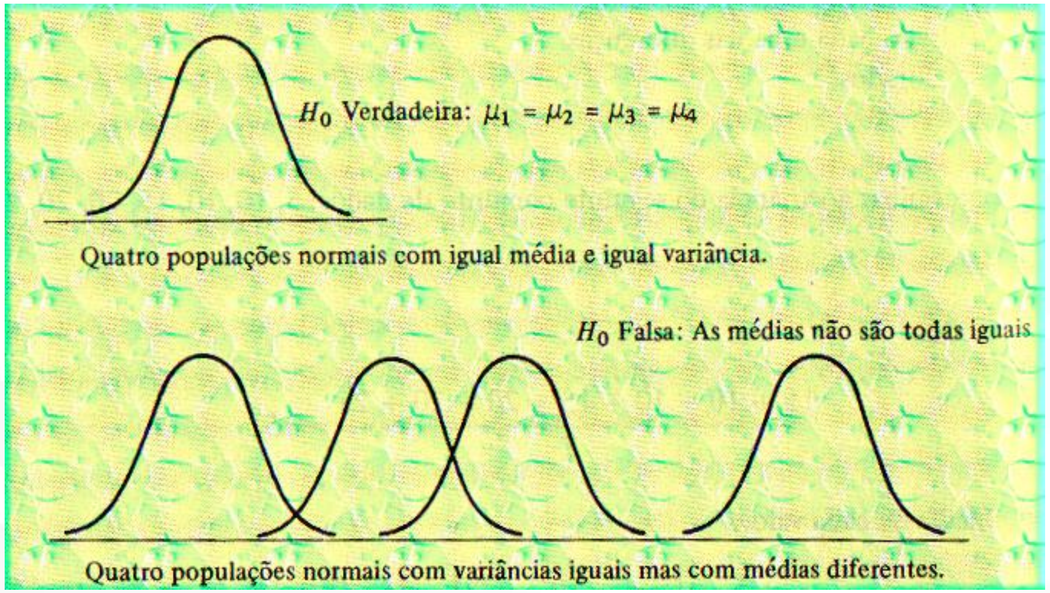
- Outros testes de hipótese: Teste de qui-quadrado, Teste de aderência, Teste de Kolmogorov-Smirnov, Teste de Anderson-Darling, Teste de independência, Comparação de variâncias, Teste de taxa de correlação etc.

Testes de hipótese – F de Snedecor

- A análise da variância é uma técnica que pode ser usada para determinar se as médias de duas ou mais populações são iguais. O teste se baseia numa amostra extraída de cada população.
- Suposições básicas que devem ser satisfeitas para que se possa aplicar a análise de variância:
  1. As amostras devem ser aleatórias e independentes;
  2. As amostras devem ser extraídas de populações normais;
  3. As populações devem ter variâncias iguais.
- Se a hipótese nula é verdadeira, então todas as amostras provêm de populações com médias iguais. E como se supõe que todas as populações sejam normais e tenham variâncias iguais, quando H_0 é verdadeira isto é conceitualmente idêntico a uma situação em que todas as amostras tenham sido extraídas de uma única população.

Testes de hipótese – F de Snedecor

- Se H_0 é falsa, então as amostras provêm de populações com médias diferentes.
- Um modo de estimar a variância populacional é por meio da média das variâncias amostrais. Como cada variância amostral reflete apenas a variação “dentro” daquela amostra em particular, a estimativa da variância baseada na média das variâncias amostrais é chamada de estimativa da variância “dentro”.



ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese – F de Snedecor

- A estimativa “dentro” serve como padrão de comparação pelo qual se pode julgar uma segunda estimativa chamada de estimativa “entre” da variância. Essa segunda estimativa é sensível às diferenças entre as médias populacionais.
- S_d^2 é o padrão de comparação porque não é afetado pela veracidade ou não de H_0. Enquanto S_e^2 é afetado, sendo aproximadamente igual a S_d^2 quando H_0 é verdadeira, porém maior do que S_d^2 quando H_0 é falsa.
- A essa altura fica claro que S_e^2 é na verdade maior do que S_d^2. Não sabemos se a variação causal devido ao processo de amostragem pode ser inteiramente responsável por isso, ou não.

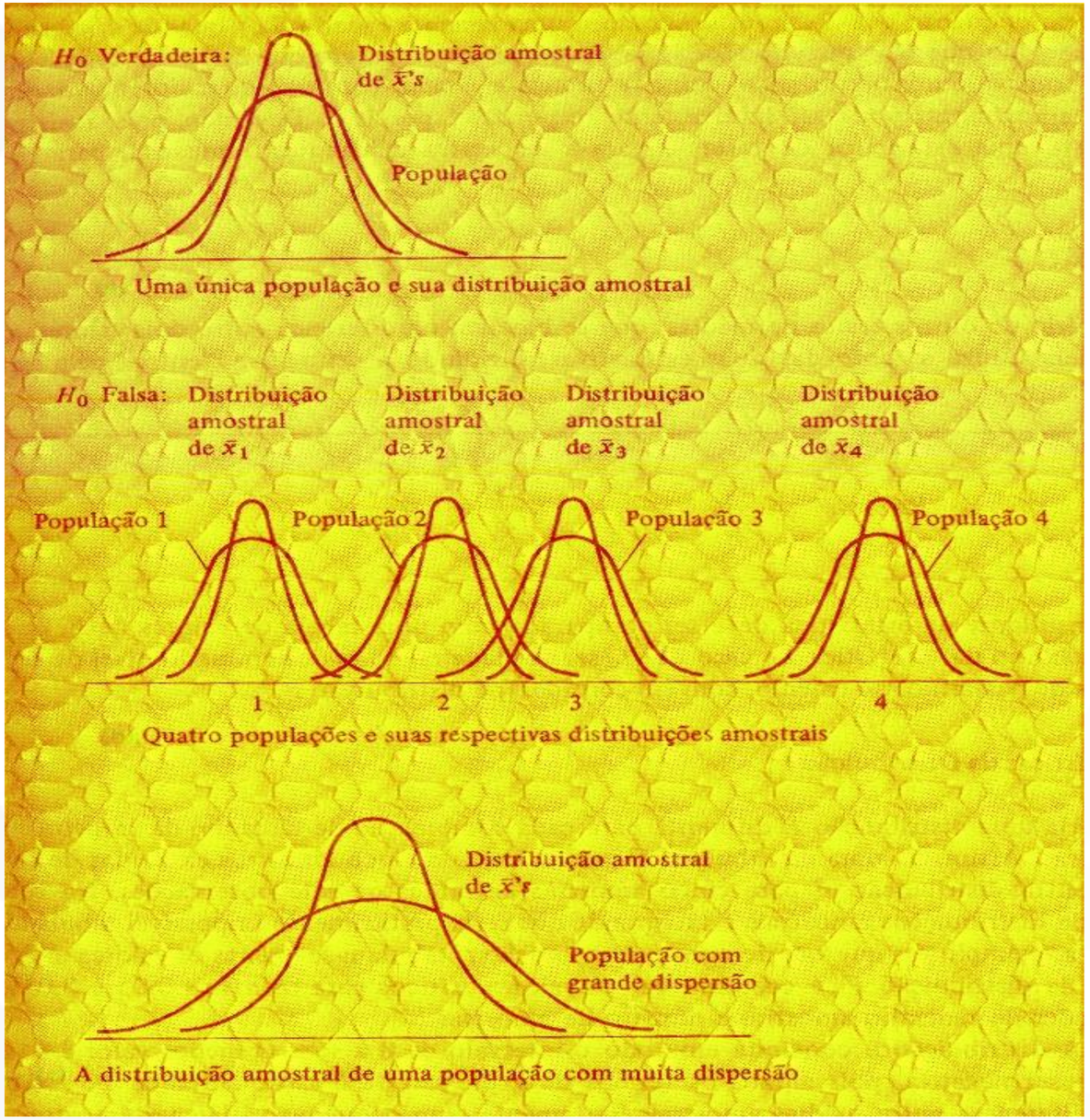
Testes de hipótese – F de Snedecor

- A distribuição F de Snedecor utiliza a razão das estimativas da variância. O valor resultante da estatística deve ser comparado com uma tabela de valores F que indica o valor máximo da estatística no caso de H_0 ser verdadeira, a determinado nível de significância.

$$F = \frac{S_e^2}{S_d^2} = \frac{nS_{\bar{X}}^2}{\left(S_1^2 + S_2^2 + \ldots + S_k^2\right)/k} = \frac{n\left[\dfrac{\sum\left(\bar{x}_j - \bar{\bar{x}}\right)^2}{k-1}\right]}{\dfrac{1}{k(n-1)}\left[\sum\left(x_i - \bar{x}_1\right)^2 + \left(x_i - \bar{x}_2\right)^2 + \ldots + \left(x_i - \bar{x}_k\right)^2\right]}$$

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese – F de Snedecor



ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese – F de Snedecor

- Existe uma distribuição F para cada combinação de tamanho da amostra, número de amostras e nível de significância. Por isso, é costume usar as tabelas para os níveis de 0,05 e 0,01 e certas combinações de tamanho amostral e número de amostras.

- Exemplo:
  Testa-se 4 tipos de combustíveis diferentes para verificar se eles têm o mesmo rendimento, ou seja, fazem a mesma média de quilometragem por litro. Cada combustível é testado por seis vezes num mesmo automóvel. Os resultados são apresentados a seguir:

Testes de hipótese – F de Snedecor

Tipo de gasolina				
Observação	I	II	III	IV
1	15,1	14,9	15,4	15,6
2	15,0	15,2	15,2	15,5
3	14,9	14,9	16,1	15,8
4	15,7	14,8	15,3	15,3
5	15,4	14,9	15,2	15,7
6	15,1	15,3	15,2	15,7
Médias amostrais	15,2	15,0	15,4	15,6
Variâncias amostrais	0,088	0,040	0,124	0,032

Pode-se, então, formular como hipóteses:

H_0: As médias das populações são todas iguais.

H_1: As médias das populações não são iguais.

ESEG
FACULDADE DO GRUPO ETAPA

Testes de hipótese - F de Snedecor

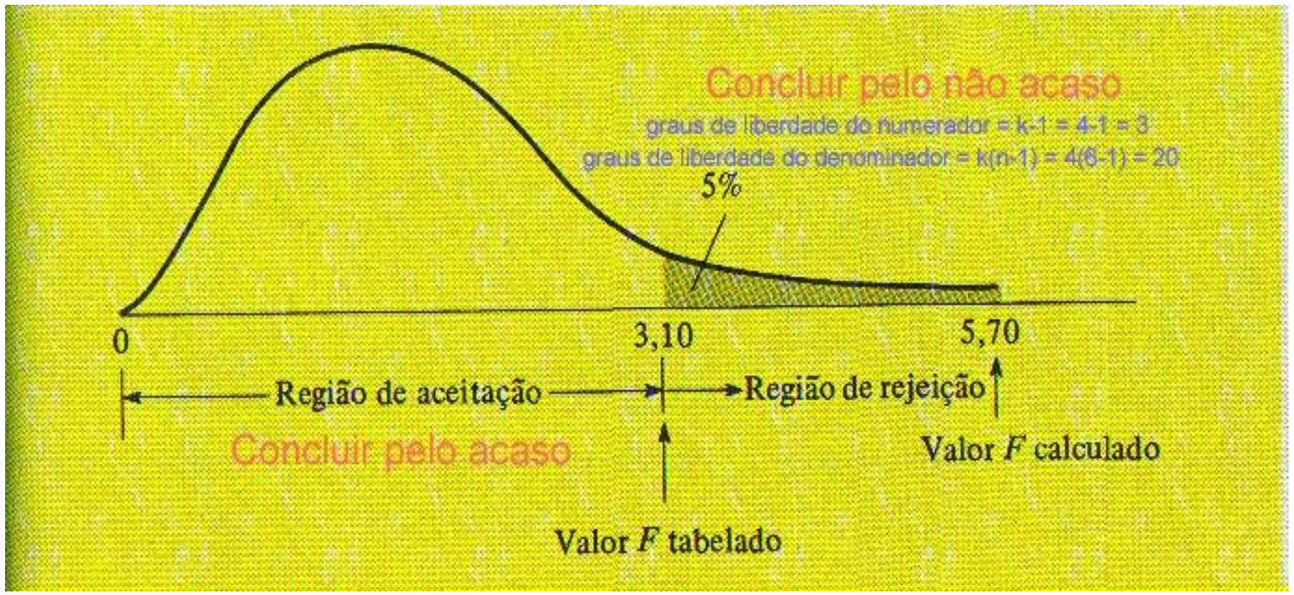
$$S_d^2 = \frac{0,088+0,040+0,124+0,032}{4} = \frac{0,284}{4} = 0,071$$

$$\bar{\bar{X}} = \frac{15,2+15,0+15,4+15,6}{4} = 15,3$$

$$S_{\bar{x}}^2 = \frac{(15,2-15,3)^2+(15,0-15,3)^2+(15,4-15,3)^2+(15,6-15,3)^2}{4-1} = 0,067$$

$$S_e^2 = nS_{\bar{x}}^2 = 6(0,067) = 0,402$$

$$F_{teste} = \frac{S_e^2}{S_d^2} = \frac{0,402}{0,071} = 5,70$$



ESEG
FACULDADE DO GRUPO ETAPA

Obrigado pela atenção!

ESEG
FACULDADE DO GRUPO ETAPA