Estatística e Probabilidade

Bacharelado em Sistemas de Informação

Aula 11: Regressão e Correlação

Prof. Dr. Samuel Sanches

REGRESSÃO

- ★ Muitas vezes queremos estabelecer relações que possibilitem prever uma ou mais variáveis em termos de outras.
- ★ Prever o resultado é muito difícil, o que se consegue é prever médias ou de valores esperados.
- ★ Tentar prever o valor médio de uma variável em termos do valor conhecido de outra variável é o que chamamos de problema de regressão.
- ★ https://phet.colorado.edu/sims/html/curve-fitting/latest/curve-fitting_en.html

AJUSTE DE CURVAS

- ★ Expressar ou aproximar a relação entre as grandezas em termos de equações matemáticas.
- ★ Decidir o tipo de curva, então o tipo de equação “de previsão” usar.
- ★ Encontrar a equação particular que é a melhor em algum sentido.
- ★ Investigar os méritos dessa equação escolhida e de previsões feitas a partir dela.

AJUSTE DE CURVAS

- ★ Inspecionar diretamente os dados, fazendo um gráfico, podemos utilizar: https://www.geogebra.org/calculator
- ★ Então verificar qual melhor curva, uma reta, uma parábola, ...
- ★ Equações lineares com duas incógnitas:

$$y = a + bx$$

a → corte no eixo y (o valor de y para x = 0)
b → inclinação da reta (variação de y que acompanha aumento em x)

- ★ Mais simples e muitas vezes é uma boa aproximação
- ★ https://phet.colorado.edu/sims/html/graphing-slope-intercept/latest/graphing-slope-intercept_en.html

MÉTODO DOS MÍNIMOS QUADRADOS

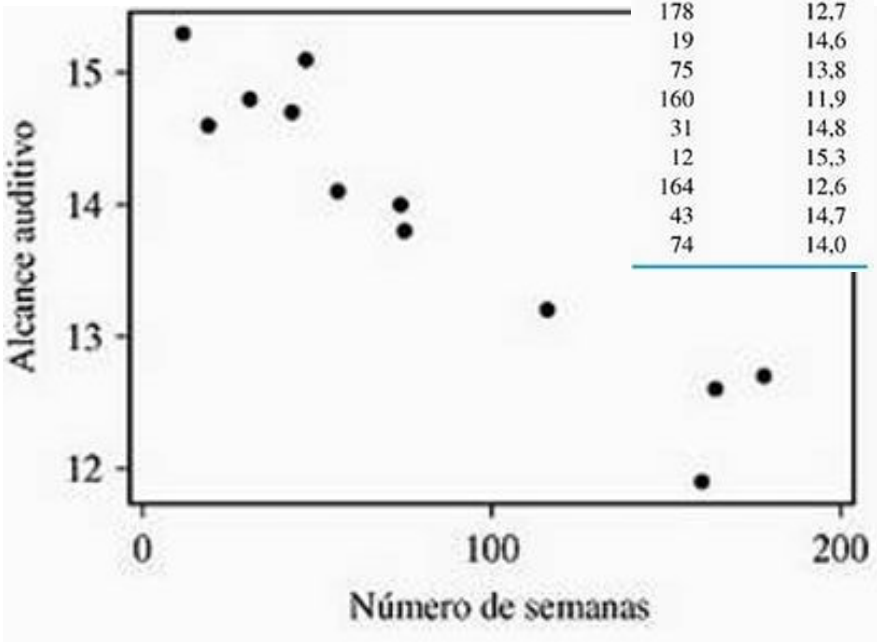
- ★ Como encontrar a melhor reta, ou seja, fazer o melhor ajuste?
- ★ Vamos usar os dados de exposição a sons altos e audição:
- ★ Total 12 **pontos de dados** (x, y), podemos montar o **diagrama de dispersão**

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

MÉTODO DOS MÍNIMOS QUADRADOS

★ Nem todos estão perfeitamente alinhados em uma reta, porém é possível notar que eles são lineares.

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

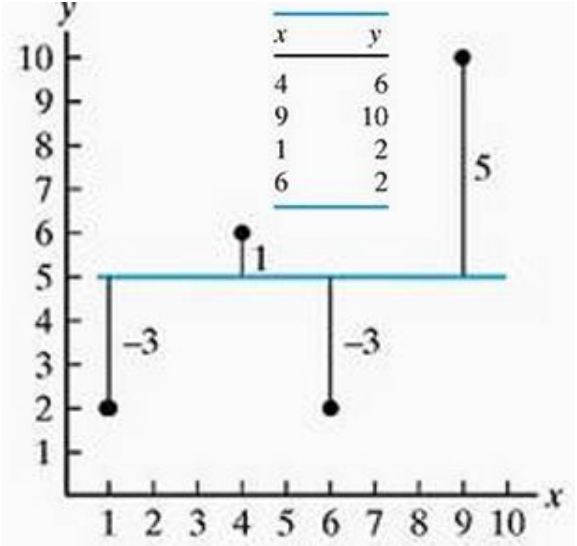


MÉTODO DOS MÍNIMOS QUADRADOS

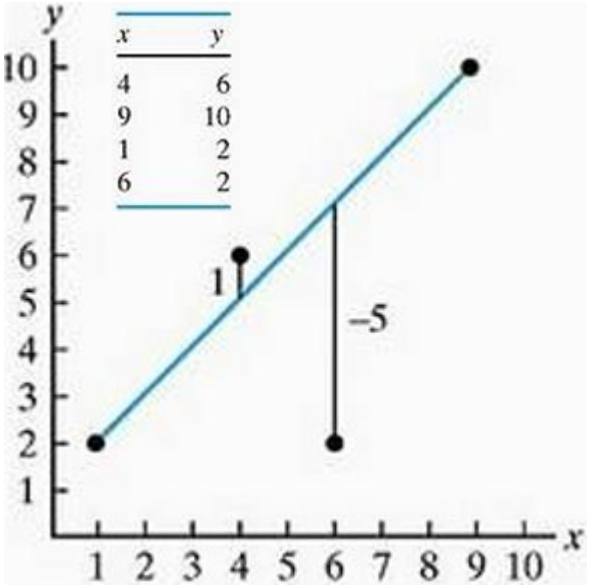
★ Uma das técnicas de encontrar a melhor reta, é o método dos mínimos quadrados, nele ele requer que a reta que ajustamos aos dados tenha a propriedade de que **seja mínima a soma dos quadrados das distâncias verticais dos pontos à reta**.

★ Tabela ao lado: vamos tentar uma reta horizontal, **y = 5**, os erros dessa "previsão" são 6 - 5 = 1, 10 - 5 = 5, 2 - 5 = -3 e 2 - 5 = -3.

★ Quadrado dos erros: $1^2 + 5^2 + (-3)^2 + (-3)^2 = 44$



MÉTODO DOS MÍNIMOS QUADRADOS



★ Outra tentativa:

★ Tabela ao lado: vamos tentar uma reta, **y = 1 + x**, para os erros, precisamos saber os valores de y para cada valor de x: 1 + 4 = 5, 1 + 9 = 10, 1 + 1 = 2 e 1+ 6 = 7

★ Desvio (erro): 6 - 5 = 1, 10 - 10 = 0, 2 - 2 = 0 e 2 - 7 = -5

★ Quadrado do erro: $1^2 + 0^2 + 0^2 + (-5)^2 = 26$, antes tínhamos 44, bem menor, ou seja, essa reta proporciona melhor ajuste, do que a anterior.

MÉTODO DOS MÍNIMOS QUADRADOS

★ Não vamos ficar tentando até encontrar a melhor:

$$\sum (y - \hat{y})^2 = \sum [y - (a + bx)]^2$$

★ **Equações normais**:

$$\sum y = na + b\left(\sum x\right)$$

$$\sum xy = a\left(\sum x\right) + b\left(\sum x^2\right)$$

★ n → número de pares de observações
Σx e Σy → soma dos valores observados x e y
Σx^2 → soma dos quadrados dos valores de x
Σxy → soma dos produtos determinados multiplicando cada x pelo y correspondente

MÉTODO DOS MÍNIMOS QUADRADOS

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

★ **Exemplo**: Usando a tabela obtenha as equações normais que determinam uma reta de mínimos quadrados.

$$\sum y = na + b\left(\sum x\right)$$

$$\sum xy = a\left(\sum x\right) + b\left(\sum x^2\right)$$

n = 12
Σx = 975
Σx^2 = 117397
Σy = 166,8
Σy^2 = 2331,54
Σxy = 12884,4

$$166{,}8 = 12a + 975b$$

$$12.884{,}4 = 975a + 117.397b$$

Agora precisamos resolver o sistema de equações, pode-se utilizar qualquer método, mas vamos simplificar um pouco.

MÉTODO DOS MÍNIMOS QUADRADOS

★ **Soluções de equações normais:**

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2 \qquad S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \frac{\sum y - b\left(\sum x\right)}{n}$$

MÉTODO DOS MÍNIMOS QUADRADOS

★ **Exemplo**: Usando a tabela obtenha as equações normais que determinam uma reta de mínimos quadrados.

n = 12
Σx = 975
Σx^2 = 117397
Σy = 166,8
Σy^2 = 2331,54
Σxy = 12884,4

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$S_{xx} = 117.397 - \frac{1}{12}(975)^2 = 38.178,25$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$S_{xy} = 12.884,4 - \frac{1}{12}(975)(166,8) = -668,1$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$b = \frac{-668,1}{38.178,25} \approx -0,0175$$

$$a = \frac{\sum y - b\left(\sum x\right)}{n}$$

$$a = \frac{166,8 - (-0,0175)(975)}{12} \approx 15,3$$

$$y = a + bx$$

$$\hat{y} = 15,3 - 0,0175x$$

Planilha Aula 11:
https://docs.google.com/spreadsheets/d/1mTkUCunLrP501B8wYcGMDX4Gr2zJMYCbJR15SgVoNxM/edit?usp=share_link

MÉTODO DOS MÍNIMOS QUADRADOS

Número de semanas x	*Alcance auditivo* y
47	15,1
56	14,1
116	13,2
178	12,7
19	14,6
75	13,8
160	11,9
31	14,8
12	15,3
164	12,6
43	14,7
74	14,0

★ **Exemplo**: Usando a equação de mínimos quadrados obtida no exercício anterior, estime o alcance auditivo de uma pessoa que foi exposta a ruído por: **a)** um ano; **b)** dois anos;

veja que temos unidades, no caso ano, então precisamos tomar cuidado para saber as unidades da nossa equação, veja que os dados nos informavam que x é em semanas e y o alcance auditivo, ou seja, ele quer que calcule y e temos que passar o valor um ano e dois anos para semanas, para poder utilizar a nossa equação.

$$\hat{y} = 15{,}3 - 0{,}0175x$$

1 ano ≃ 52 semanas → $\hat{y} = 15{,}3 - 0{,}0175(52) = 14{,}4$

2 anos ≃ 104 semanas → $\hat{y} = 13{,}5$

ANÁLISE DE REGRESSÃO

- ★ Obtemos um resultado de uma reta de regressão, ou seja, interpretamos como médias ou valores esperados, porém:

- ★ Qual é a precisão dos valores obtidos para a e b na equação dos mínimos quadrados?

$$\hat{y} = 15{,}3 - 0{,}0175x$$

- ★ Qual a precisão da estimativa y = 13,5 que encontramos quando utilizamos o x com o valor de 2 anos (104 semanas)?

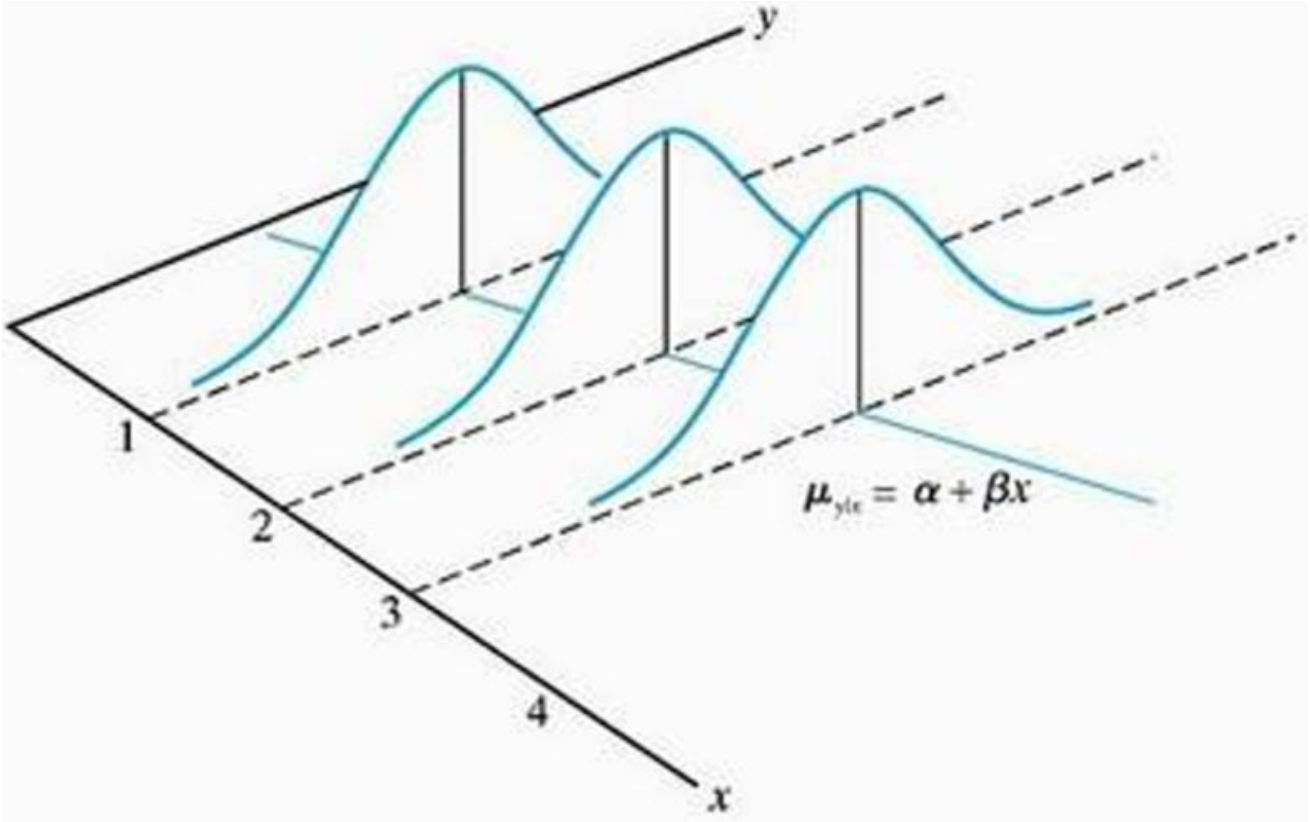
ANÁLISE DE REGRESSÃO

- ★ Para essa equação foram usados os dados daquela tabela, e se utilizássemos outra amostra (outra quantidade), teríamos a mesma equação? Provavelmente não, seria próxima, mas não igual, ou seja, os resultados também seriam um pouco diferentes. Então:

- ★ É possível estabelecer um intervalo para o qual possamos afirmar, com algum grau de confiança, que contém o real valor calculado?

- ★ a e b são “apenas estimativas baseados em dados amostrais”, os valores reais são chamados de alfa e beta, α e β denominados **coeficientes de regressão**, e a **reta de regressão verdadeira**: $\mu_{y|x} = \alpha + \beta x$

ANÁLISE DE REGRESSÃO



ANÁLISE DE REGRESSÃO

★ **Erro-padrão da Estimativa**:

$$s_e = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$$

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

ANÁLISE DE REGRESSÃO

★ **Exemplo**: Utilizando o exercício em que calculamos a reta, calcular o erro-padrão dela.

n = 12
Σx = 975
Σx^2 = 117397
Σy = 166,8
Σy^2 = 2331,54
Σxy = 12884,4

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

$$S_{yy} = 2.331{,}54 - \frac{1}{12}(166{,}8)^2 = 13{,}02$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$S_{xy} = 12.884{,}4 - \frac{1}{12}(975)(166{,}8) = -668{,}1$$

$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$S_{xx} = 117.397 - \frac{1}{12}(975)^2 = 38.178{,}25$$

$$s_e = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$b = \frac{-668{,}1}{38.178{,}25} \approx -0{,}0175$$

$$s_e = \sqrt{\frac{13{,}02 - (-0{,}0175)(-668{,}1)}{10}}$$

$$\approx 0{,}3645$$

CORRELAÇÃO

- ★ Temos a nossa reta, como saber quão bom é esse ajuste?
- ★ A correlação nos informará, como estão os valores dos dados e os valores calculados.
- ★ Poderíamos ver dado com dado, mas não é algo muito prático

O COEFICIENTE DE CORRELAÇÃO

★ A ideia é analisar o valor calculado (ŷ) juntamente com a média dos valores (y barra) e o valor original (y) com a média dos valores (y barra) → **Soma de quadrados residual**

$$\sum (y - \hat{y})^2$$

★ **Coeficiente de determinação**:

$$\frac{\sum (\hat{y} - \overline{y})^2}{\sum (y - \overline{y})^2}$$

★ Coeficiente de correlação é a raiz quadrada do coef. de determinação, ou:

$$S_{yy} = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

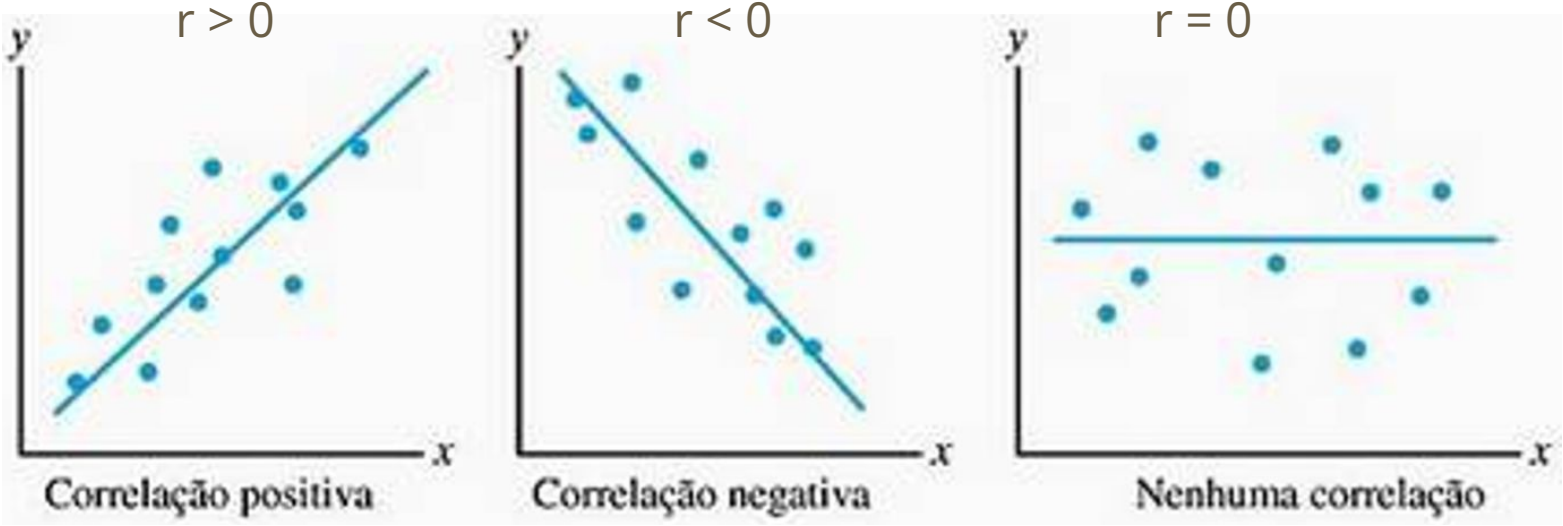
$$S_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

O COEFICIENTE DE CORRELAÇÃO

★ r > 0 | r < 0 | r = 0



★ Valor máximo r = 1 e mínimo r = -1 que indica ajuste perfeito!

Planilha Aula 11:
https://docs.google.com/spreadsheets/d/1mTkUCunLrP501B8wYcGMDX4Gr2zJMYCbJR15SgVoNxM/edit?usp=share_link

O COEFICIENTE DE CORRELAÇÃO

★ **Exemplo**: Na tabela temos 12 notas, calcule r.

n = 12
Σx = 850
Σx^2 = 65230
Σy = 927
Σy^2 = 74883
Σxy = 69453

$$S_{xx} = 65.230 - \frac{1}{12}(850)^2 \approx 5.021,67$$

$$S_{yy} = 74.883 - \frac{1}{12}(927)^2 = 3.272,25$$

$$S_{xy} = 69.453 - \frac{1}{12}(850)(927) = 3.790,5$$

$$r = \frac{3.790,5}{\sqrt{(5.021,67)(3.272,25)}} \approx 0,935$$

Economia	*Antropologia*
51	74
68	70
72	88
97	93
55	67
73	73
95	99
74	73
20	33
91	91
74	80
80	86

Planilha Aula 11:
https://docs.google.com/spreadsheets/d/1mTkUCunLrP501B8wYcGMDX4Gr2zJMYCbJR15SgVoNxM/edit?usp=share_link

A INTERPRETAÇÃO DE r

- ★ Caso r seja +1, -1 ou 0, não tem problemas de interpretação, +1 ou -1 é quando todos os pontos estão efetivamente sobre a reta e 0 quando o ajuste da reta é tão pobre que conhecendo x nada contribui para a previsão de y.

- ★ Fazendo $100*r^2$ temos a percentagem da variação total dos y que é explicada por sua relação com x, ou devida à relação.

- ★ **Cuidado**! r = 1 ou -1 não quer dizer que temos causa e efeito, só que os dados estão alinhados!!! Podemos também não ter uma relação linear, sempre ver o gráfico de dispersão!!!!

A INTERPRETAÇÃO DE r

★ **Exemplo**: Se r = 0,80 num estudo e r = 0,40 num outro, estaria correto dizer que a correlação de 0,80 é duas vezes mais forte do que a correlação de 0,40?

Não!
$r = 0,80 \rightarrow 100*(0,80)^2 = 64\%$
$r = 0,40 \rightarrow 100*(0,40)^2 = 16\%$

0,80 nos fornece 64% de que uma variação em y corresponde a variação em x, assim podemos dizer que a correlação de 0,80 é quatro vezes mais forte que a de 0,40.

ANÁLISE DE CORRELAÇÃO

★ Dois dados temos a tabela:

Dado vermelho x	*Dado verde* y
4	5
2	2
4	6
2	1
6	4

★ Calculando temos r = 0,66 que é extremamente alto!!! Faz sentido?

★ r calculado de uma amostra é uma estimativa do **coeficiente de correlação populacional** (ρ → rô), então podemos testar hipóteses sobre esse parâmetro, conhecendo o **coeficiente de correlação amostral** (r)

Planilha Aula 11:
https://docs.google.com/spreadsheets/d/1mTkUCunLrP501B8wYcGMDX4Gr2zJMYCbJR15SgVoNxM/edit?usp=share_link

ANÁLISE DE CORRELAÇÃO

★ **Estatística para inferências sobre ρ**:

$$z = (Z - \mu_Z)\sqrt{n-3}$$

★ transformação Z de Fisher:

$$Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r} \qquad \mu_Z = \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho} \qquad \sigma_Z = \frac{1}{\sqrt{n-3}}$$

ANÁLISE DE CORRELAÇÃO

TABELA X Valores de $Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$

r	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,000	0,010	0,020	0,030	0,040	0,050	0,060	0,070	0,080	0,090
0,1	0,100	0,110	0,121	0,131	0,141	0,151	0,161	0,172	0,182	0,192
0,2	0,203	0,213	0,224	0,234	0,245	0,255	0,266	0,277	0,288	0,299
0,3	0,310	0,321	0,332	0,343	0,354	0,365	0,377	0,388	0,400	0,412
0,4	0,424	0,436	0,448	0,460	0,472	0,485	0,497	0,510	0,523	0,536
0,5	0,549	0,563	0,576	0,590	0,604	0,618	0,633	0,648	0,662	0,678
0,6	0,693	0,709	0,725	0,741	0,758	0,775	0,793	0,811	0,829	0,848
0,7	0,867	0,887	0,908	0,929	0,950	0,973	0,996	1,020	1,045	1,071
0,8	1,099	1,127	1,157	1,188	1,221	1,256	1,293	1,333	1,376	1,422
0,9	1,472	1,528	1,589	1,658	1,738	1,832	1,946	2,092	2,298	2,647

Para valores negativos de *r*, coloque um sinal de menos na frente do *Z* correspondente, e vice-versa.

ANÁLISE DE CORRELAÇÃO

★ **Exemplo**: Ao nível 0,05 de significância, teste a hipótese nula de ausência de correlação (isto é, a hipótese nula $\rho = 0$) para o caso dos dados, onde obtemos r = 0,66.

1) H_0: $\rho = 0$ e H_A: $\rho \neq 0$ com $\alpha = 0,05$

2) $\mu_Z = 0$ quando $\rho = 0$, rejeitar H_0 se $z \leq -1,96$ ou $z \geq 1,96$ $Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$ $\mu_Z = \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho}$

3) Com n = 5 e Z = 0,793 (tabela slide anterior r = 0,66): $z = (Z - \mu_Z)\sqrt{n-3}$

4) Como z = 1,12 (entre -1,96 e 1,96), a hipótese nula não pode ser rejeitada, ou seja, o valor de r obtido não é significativo, como esperávamos.

ANÁLISE DE CORRELAÇÃO

★ **Exemplo**: Ao nível 0,01 de significância, teste a hipótese nula ρ = -0,80 contra a alternativa ρ < -0,80, para o caso r = -0,95 com n = 12.

1) H_0: ρ = -0,80 e H_A: ρ < -0,80 com α = 0,01

$$Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r} \qquad \mu_Z = \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho}$$

2) Rejeitar H_0 se z ≥ -2,33

$$z = (Z - \mu_Z)\sqrt{n-3}$$

3) Com n = 12 e Z = -1,832 (tabela slide anterior r = -0,95) e μ_Z = -1,099:

4) Como z = -2,20 (maior que -2,33), a hipótese nula não pode ser rejeitada.

EXERCÍCIOS

★ **Lista 4 de Exercícios → Parte 2**

https://drive.google.com/file/d/128w6Pc-oGISlvHcNmGggwQgsOXPHdTaB/view?usp=share_link

★ Muito obrigado pela atenção!