7

ESTATÍSTICA

# MATERIAL-BASE DAS AULAS 25-28

UNIVESP
Universidade Virtual do Estado de São Paulo

FERNANDO TOBAL BERSSANETI

# 7

# ESTATÍSTICA
## MATERIAL-BASE DAS AULAS 25-28

### APRESENTAÇÃO

Na sexta etapa da disciplina evoluímos os estudos sobre Estatística Inferencial desenvolvendo os Testes de Hipóteses, ferramentas utilizadas para tomada de decisões sobre uma população a partir dos valores obtidos em experimentos realizados com amostras. Foram apresentadas três modalidades de Testes de Hipóteses: para a média quando a variância da população é conhecida, para a média quando a variância da população é desconhecida e para a proporção. Vale lembrar que ao tomarmos decisões com base nestes testes sempre estamos sujeitos a dois tipos de erros: erro do Tipo I (rejeitar a hipótese inicial sendo que a hipótese inicial é verdadeira) ou erro do Tipo II (manter a hipótese inicial sendo que a hipótese inicial é falsa). Note que os problemas que trabalhamos até agora envolviam apenas uma variável de interesse.

Nesta sétima etapa vamos evoluir ainda mais nossos conhecimentos sobre Estatística Inferencial, estudando situações que envolvem duas variáveis, *x* e *y*. Em especial estamos interessados em verificar se estas variáveis apresentam alguma tendência conjunta, ou seja, se o comportamento de uma variável está vinculado ao comportamento da outra. Para isto vamos buscar e analisar as respostas para as seguintes questões:

- Para um determinado *x* qual o valor esperado de *y*?
- Dado um modelo linear que explica o relacionamento entre as variáveis *x* e *y*, como podemos afirmar se este modelo linear é significativo?
- Ainda considerando o modelo linear obtido, como podemos testar estatisticamente as hipóteses sobre os diferentes parâmetros que descrevem esta reta?

Para responder estas perguntas vamos apresentar o diagrama de correlação e calcular seus coeficientes. Estes valores serão essenciais para compreendermos se realmente ocorre uma regressão linear entre as variáveis. Finalmente, utilizando os mesmos princípios apresentados nas semanas anteriores, vamos utilizar os principais parâmetros das análises de correlação e regressão para estabelecer intervalos de confiança para os parâmetros obtidos e, considerando um determinado nível de confiança, verificar se estes parâmetros são efetivamente significativos.

Esta semana está organizada com as seguintes atividades:

- 2 exercícios resolvidos (solução apresentada);
- 2 exercícios propostos (solução desenvolvida por você. O gabarito estará disponível ao final da semana de estudos. Não será critério de avaliação).

Lembre-se de conferir as dicas que selecionamos especialmente para você no tópico “Materiais de Apoio”.

Bons estudos!
Prof. Fernando

## OBJETIVOS DAS AULAS DA SEMANA

- Conceituar o diagrama de correlação, calcular o coeficiente de correlação de Pearson e testar se realmente existe regressão linear entre as variáveis (teste para o coeficiente de correlação);
- Apresentar aplicações dos cálculos do coeficiente de correlação de Pearson e testar se existe regressão linear entre as variáveis;
- Calcular a reta estimativa de regressão, calcular os intervalos de confiança para os parâmetros $\alpha$ e $\beta$ da reta e realizar os testes de hipóteses aplicado à Regressão;
- Apresentar aplicações dos cálculos da reta estimativa de regressão, apresentar aplicações dos cálculos de intervalos de confiança para os parâmetros reta e os testes de hipóteses aplicados à regressão.

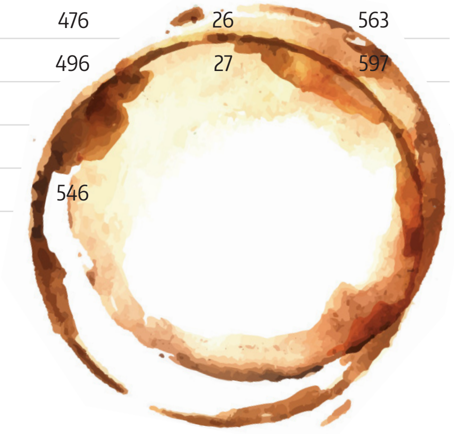
## 1. EXERCÍCIOS RESOLVIDOS

### *EXERCÍCIO 1*

O histórico de venda de uma empresa é o mostrado abaixo, onde *x* é o mês e *y* é a quantidade vendida. Infelizmente um funcionário descuidado deixou cair café sobre a tabela, perdendo alguns de seus valores. Imediatamente tentaram recuperar os dados perdidos, mas não existia nenhuma cópia, apenas algumas somatórias já feitas:

**HISTÓRICO DE VENDAS**

| x | Y | X | y | x | y |
|---|---|---|---|---|---|
| 1 | 331 | 11 | 435 | 21 | 501 |
| 2 | 334 | 12 | 427 | 22 | 523 |
| 3 | 386 | 13 | 450 | 23 | 549 |
| 4 | 347 | 14 | 451 | 24 | 542 |
| 5 | 387 | 15 | 462 | 25 | 573 |
| 6 | 380 | 16 | 476 | 26 | 563 |
| 7 | 389 | 17 | 496 | 27 | 597 |
| 8 | 381 | 18 | | | |
| 9 | 403 | 19 | | | |
| 10 | 415 | 20 | 546 | | |



$R^2 = 0{,}9597$

*Observações* 30

$\sum x = 465$

$\sum x^2 = 9455$

$\sum y = 14231$

$\sum xy = 242628$

Considerando as informações apresentadas determine a reta de regressão.

**Solução**

$$S_{XY} = \sum xy - \frac{\sum x \times \sum y}{n} = 22047{,}5$$

$$S_{XX} = \sum x^2 - \frac{(\sum x)^2}{n} = 2247{,}5$$

$$S_{YY} = \frac{(S_{xy})^2}{r^2 S_{xx}} = 225363{,}46$$

$$a = \frac{Sxy}{Sxx} = 9{,}8098$$

$$b = \bar{y} - a\bar{x} = 322{,}3149$$

$$y = ax + b = 9{,}80983x + 322{,}3149$$

## EXERCÍCIO 2

Um artigo no Jornal de Sons e Vibrações (*Journal of Sound and Vibration*, Vol. 151, 1991, pp. 383-394) descreveu um estudo investigando a relação entre a exposição ao barulho e hipertensão. Os seguintes dados foram reportados no estudo:

| AUMENTO DA PRESSÃO SANGUÍNEA (mmHg) | NÍVEL DE PRESSÃO SONORA (dB) |
|---|---|
| 1 | 60 |
| 0 | 63 |
| 1 | 65 |
| 2 | 70 |
| 5 | 70 |
| 1 | 70 |
| 4 | 80 |
| 6 | 80 |
| 2 | 80 |
| 3 | 80 |
| 5 | 85 |
| 4 | 89 |
| 6 | 90 |
| 8 | 90 |
| 4 | 90 |
| 5 | 90 |
| 7 | 94 |
| 9 | 100 |
| 7 | 100 |
| 6 | 100 |

**a.** Ajuste o modelo de regressão linear simples usando o método de mínimos quadrados e determine a equação da reta.
**b.** Encontre o aumento médio previsto para a pressão sanguínea, associado com o nível de pressão sonora de 85 dB.
**c.** Determine o coeficiente de correlação de Pearson.

**Solução**

**a.** Variável independente = pressão sonora

$$a = \frac{Sxy}{Sxx} = 0{,}1715$$

$$b = \bar{y} - a\bar{x} = -9{,}8131$$

$$y = ax + b = 0{,}1715x - 9{,}8131$$

**b.**

$$y = ax + b = 0{,}1715x - 9{,}8131$$

$$y = 0{,}1715 \times 85 - 9{,}8131$$

$$y = 4{,}7644$$

**c.**

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 124{,}2$$

$$S_{xy} = \sum x_i \times y_i - \frac{\sum x_i \times \sum y_i}{n} = 516{,}2$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 3010{,}2$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{516{,}2}{\sqrt{3010{,}2 \times 124{,}2}} = 0{,}8442$$

## 2. EXERCÍCIOS PROPOSTOS

### *EXERCÍCIO 1*

A indústria farmacêutica Farmex vende um remédio para combater resfriado. Após 20 meses de operação ela coletou as seguintes informações:

| MÊS | DESPESAS COM MARKETING (R$1.000,00) | VENDAS (10.000 unidades) |
|---|---|---|
| 1 | 11 | 25 |
| 2 | 5 | 13 |
| 3 | 3 | 8 |
| 4 | 9 | 20 |
| 5 | 12 | 25 |
| 6 | 6 | 12 |
| 7 | 5 | 10 |
| 8 | 9 | 15 |
| 9 | 4 | 10 |
| 10 | 7 | 16 |
| 11 | 8 | 18 |
| 12 | 10 | 20 |
| 13 | 13 | 26 |
| 14 | 11 | 22 |
| 15 | 7 | 14 |
| 16 | 10 | 15 |
| 17 | 13 | 27 |
| 18 | 4 | 13 |
| 19 | 6 | 16 |
| 20 | 2 | 6 |

**a.** Ajuste o modelo de regressão linear simples usando o método de mínimos quadrados e determine a equação da reta.

**b.** Encontre a venda esperada, associada com um investimento em da ordem de R$12.000,00 em Marketing.

**c.** Determine o coeficiente de correlação de Pearson.

### *EXERCÍCIO 2*

Uma montadora está tentando relacionar o consumo de combustível, em Km/l, associado à rotação média do motor (RPM). O resumo dos dados coletados na bancada durante 20 testes em laboratório está apresentado a seguir:

$$\sum x_i = 81.300$$

$$\sum y_i = 240{,}4$$

$$S_{yy} = 73{,}71$$

$$S_{xy} = -41.526$$

$$S_{xx} = 25.945.500$$

**a.** Determine a equação da reta.
**b.** Determine o coeficiente de correlação de Pearson.
**c.** Determine um intervalo com 95% de confiança para o coeficiente angular da reta (β).

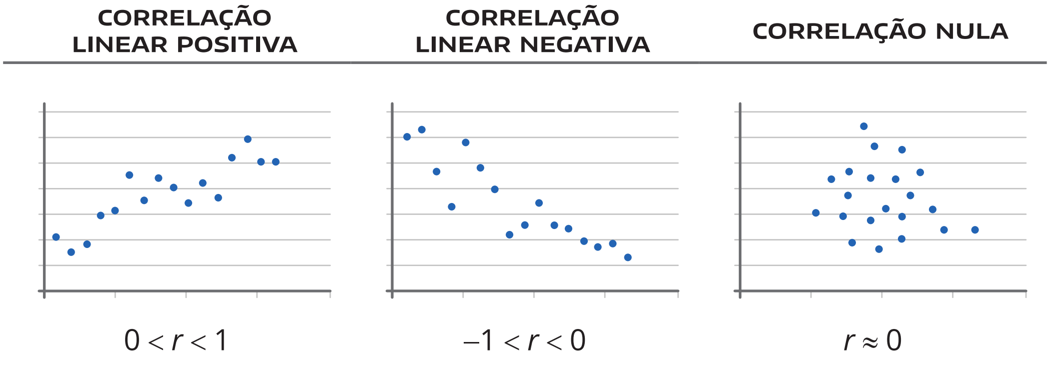
## 3. MATERIAIS DE APOIO

→ **Applet "Least-Squares Polynomial Approximation"**
http://www.chem.uoa.gr/applets/AppletPoly/Appl_Poly2.html
Indique alguns pontos no gráfico, e o applet fornecerá a aproximação polinomial de mínimos quadrados de acordo com o grau escolhido. (Note que, no caso da regressão linear simples, o grau é 1.) Também são mostrados o coeficiente de determinação e a soma dos quadrados dos resíduos.

→ **Applet "Regression by Eye"**
http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html
Teste sua capacidade de reconhecer o coeficiente de correlação apenas olhando para o gráfico de correlação.

## SÍNTESE

Nesta unidade você evoluiu seus conhecimentos sobre Estatística Inferencial estudando situações que envolvem duas variáveis, *x* e *y* e compreendendo se estas variáveis apresentam tendência conjuntas (correlação), como apresentado na figura a seguir:




### Coeficiente de correlação linear de Pearson

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

Em que:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

| | |
|---|---|
| r < 0 | correlação linear negativa |
| r > 0 | correlação linear positiva |
| \|r\| próximo de 1 | correlação linear forte |

**Regressão linear simples**

Modelo teórico de regressão linear simples:

$$y = \alpha + \beta x + \varepsilon$$

**Reta estimativa de regressão**

$$\hat{y} = a + bx$$

$$\begin{cases} b = \dfrac{S_{xy}}{S_{xx}} \\ a = \bar{y} - b\bar{x} \end{cases}$$

$\sigma_R^2$ pode ser estimado por:

$$S_R^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

**Distribuições das estimativas**

- $b$ ~ normal

$$\mu(b) = \beta$$

$$\sigma^2(b) = \frac{\sigma_R^2}{S_{xx}}$$

$$S^2(b) = \frac{S_R^2}{S_{xx}}$$

- $a$ ~ normal

$$\mu(a) = \alpha$$

$$\sigma^2(a) = \frac{\sigma_R^2 \sum x_i^2}{nS_{xx}}$$

$$S^2(a) = \frac{S_R^2 \sum x_i^2}{nS_{xx}}$$

**a.** Intervalo de confiança para $\alpha$ com coeficiente de confiança $\gamma$:

$$a \pm t_{n-2, \frac{1-\gamma}{2}} S_R \sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

Em que:

$$a = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

**b.** Intervalo de confiança para $\beta$ com coeficiente de confiança $\gamma$:

$$b \pm t_{n-2, \frac{1-\gamma}{2}} S_R \sqrt{\frac{1}{S_{xx}}}$$

Em que:

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

**Testes de hipóteses**

Modelo de regressão teórico:

$$y = \alpha + \beta x + \varepsilon$$

Com:

$$e \sim N(0, \sigma_R^2)$$

Reta experimental:

$$\hat{y} = a + bx$$

Em que:

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

$$a = \bar{y} - b\bar{x}$$

A tabela a seguir resume as fórmulas, os limites da região de rejeição e os critérios de rejeição para cada um dos parâmetros em estudo.

| PARÂMETRO | HIPÓTESE $H_0$ | ESTATÍSTICA DE TESTE | HIPÓTESE $H_1$ | CRITÉRIO DE REJEIÇÃO |
|---|---|---|---|---|
| $\beta$ | $H_0\colon\ \beta = \beta_0$ | $t_{calc} = \dfrac{b - \beta_0}{S(b)}$ | $H_1 : \beta > \beta_0$ | $t_{calc} > t_{n-2;\alpha}$ |
| | | | $H_1 : \beta < \beta_0$ | $t_{calc} < t_{n-2;\alpha}$ |
| | | | $H_1 : \beta \neq \beta_0$ | $\lvert t_{calc}\rvert > t_{n-2;\alpha/2}$ |
| $\alpha$ | $H_0\colon\ \alpha = \alpha_0$ | $t_{calc} = \dfrac{a - \alpha_0}{S(a)}$ | $H_1 : \alpha > \alpha_0$ | $t_{calc} > t_{n-2;\alpha}$ |
| | | | $H_1 : \alpha < \alpha_0$ | $t_{calc} < t_{n-2;\alpha}$ |
| | | | $H_1 : \alpha \neq \alpha_0$ | $\lvert t_{calc}\rvert > t_{n-2;\alpha/2}$ |

# GABARITO – EXERCÍCIOS PROPOSTOS

## EXERCÍCIO 1

**a.** $y = ax + b = 17{,}345x + 31.076$

**b.** $y = 239.216$ unidades

**c.** $r = \dfrac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \dfrac{3.707.500.000}{\sqrt{213.750.000 \times 72.495.000.000}} = 0{,}94183$

## EXERCÍCIO 2

**a.** $y = ax + b = -0{,}0016x + 18{,}526$

**b.** $r = \dfrac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \dfrac{-41.526}{\sqrt{25.945.500 \times 73{,}71}} = -0{,}9496$

**c.** $\beta = \mathrm{b} \pm t_{n-2;\,\alpha/2} \times S_b = 18{,}526 \pm 2{,}101 \times 0{,}0406 = 18{,}526 \pm 0{,}0853$