

Soluções para a Prova Escrita

1. Disserte sobre testes de hipóteses para uma única amostra sobre a média de uma população normal, com variância conhecida.

Resposta. Suponhamos que uma amostra aleatória

$$X_1, X_2, \dots, X_n \quad (1)$$

é obtida da população. Vamos supor conhecido o fato de que a média amostral $\bar{X}$ é um estimador pontual não tendencioso da média populacional μ com variância σ^2/n . Consideraremos três casos:

(i) Suponhamos que queremos testar a hipótese bilateral:

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0, \quad (2)$$

sendo μ_0 dada. Como a população é normal, $\bar{X}$ é uma variável aleatória normal, com desvio padrão $\sigma/\sqrt{n}$ e média μ_0 se a hipótese nula H_0 for verdadeira. Para testar a hipótese H_0 podemos calcular o chamado teste- z que descreveremos a seguir.

É conveniente padronizar a média amostral e usar um teste estatístico baseado na distribuição normal padrão. Ou seja, para testar $H_0: \mu = \mu_0$ usamos a estatística de teste

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (3)$$

Se a hipótese nula é verdadeira, $H_0: \mu = \mu_0$ é verdadeira, $E(\bar{X}) = \mu_0$ e Z_0 tem distribuição normal padrão. Para testar a hipótese computamos o valor $\bar{x}$ de uma amostra aleatória de n elementos. Padronizando $\bar{x}$ obtemos, usando a estatística de teste (3),

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad (4)$$

Para um teste bilateral, o valor-P associado a z_0 é a probabilidade associada às áreas sombreadas mostradas na Fig. 1, que mostra a distribuição normal padrão $N(0,1)$ e os valores $\pm z_0$.

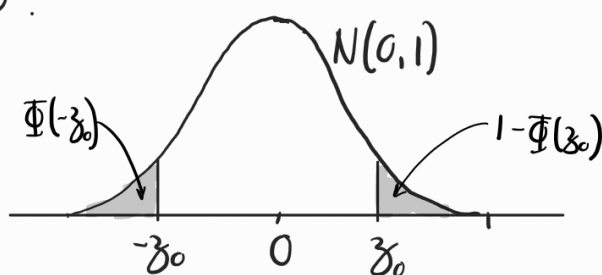


Fig. 1

O valor-P é a probabilidade de que a variável aleatória normal padrão seja menor que $-z_0$ e maior que z_0 (vale H_1). Se $\Phi(z)$ é a função de distribuição acumulativa padrão, então

$$P = \Phi(-z_0) + [1 - \Phi(z_0)] = 2[1 - \Phi(|z_0|)] \quad (5)$$

Como o teste é bilateral, somente valores de $\bar{x}$ que são menores maiores ou menores que μ_0 são consistentes com a hipótese alternativa H_1 , o que justifica a fórmula (5).

(ii) Suponhamos que temos um teste unilateral superior:

$$H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0. \quad (6)$$

Novamente computamos z_0 usando (4). Somente valores de $\bar{x}$ maiores que μ_0 são consistentes com a hipótese alternativa H_1 . Portanto, o valor-P é a probabilidade de que a variável aleatória normal padrão seja maior do que o valor da estatística de teste z_0 :

$$P = 1 - \Phi(z_0) \quad (7)$$

Tal valor de P corresponde à área sombreada mostrada na Fig. 2.

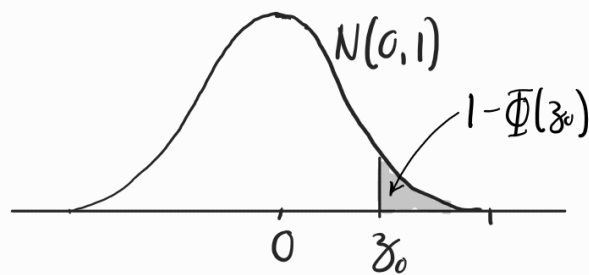


Fig. 2

(iii) Para o teste unilateral inferior:

$$H_0: \mu = \mu_0, \quad H_1: \mu < \mu_0 \quad (8)$$

temos, analogamente, o valor-P dado por

$$P = \Phi(z_0), \quad (9)$$

ou seja, somente valores de $\bar{x}$ que são menores que μ_0 são consistentes com a hipótese alternativa H_1 . Tal valor-P corresponde à área sombreada da Fig. 3.

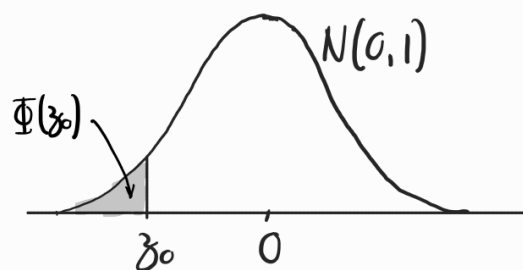


Fig. 3

Nesse caso, P é o valor da probabilidade de que o valor da variável aleatória normal padrão seja menor que o valor da estatística de teste z_0 .

Podemos também associar um nível de significância ao teste- z . Para isso devemos determinar as regiões críticas para H_1 nos casos de testes bilateral e unilaterais. Consideremos cada caso novamente:

(i) Teste bilateral: $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$. (10)

Neste caso, definimos o nível de significância α de modo que rejeitamos H_0 se

$$z_0 < -z_{\alpha/2} \quad \text{ou} \quad z_0 > z_{\alpha/2}$$

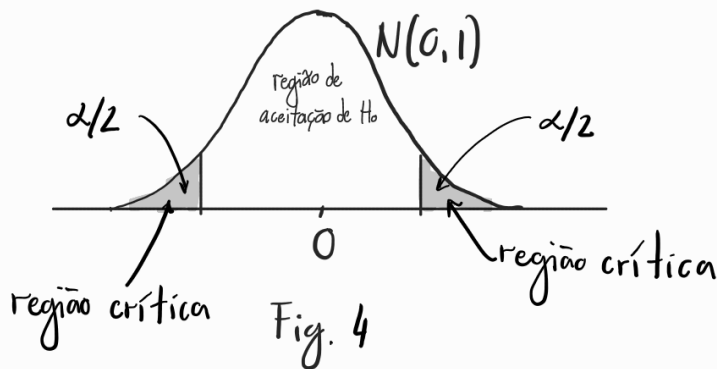
e não rejeitamos H_0 se

$$-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2},$$

sendo $\pm z_{\alpha/2}$ os valores tais que

$$\frac{\alpha}{2} = \Phi(z_{-\alpha/2}) = 1 - \Phi(z_{\alpha/2}) . \quad (12)$$

As regiões de rejeição de H_0 (críticas) e de aceitação de H_0 estão mostradas na Fig. 4.

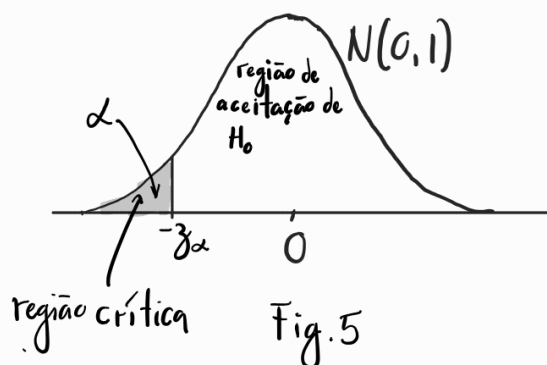


(ii) Teste unilateral inferior : $H_0 : \mu = \mu_0$, $H_1 : \mu < \mu_0$.

Nesse caso α é definido de modo que a região de rejeição de H_0 é aquela tal que $z_0 < -z_\alpha$, e não rejeição de H_0 , aquela tal que $z_0 \geq -z_\alpha$, sendo

$$\alpha = \Phi(-z_\alpha) . \quad (13)$$

As correspondentes regiões crítica e de aceitação são mostradas na Fig. 5



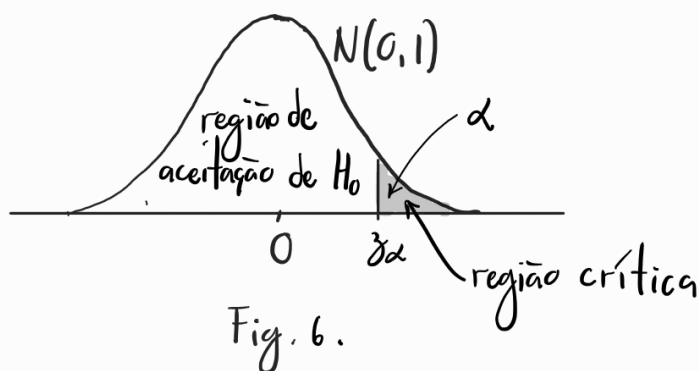
(iii) Teste unilateral superior : $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$.

Aqui rejeitamos H_0 se $z_0 > z_\alpha$, com

⑥

$$\alpha = 1 - \Phi(z_\alpha) \quad (14)$$

e não rejeitamos H_0 se $z_0 \leq z_\alpha$. As correspondentes regiões são mostradas na Fig. 6.



O valor de significância mais utilizado na prática é $\alpha = 0.05$.

2. Explique o conceito de probabilidade condicional e use-o para dissertar sobre o Teorema de Bayes.

Resposta, Em alguns casos devemos recalcular probabilidades à medida que informações se tornam disponíveis. Um modo de incorporar tal informação adicional no modelo de probabilidade, consiste em supor que o resultado deve ser necessariamente um membro de um certo evento. A probabilidade de um evento B , sob o conhecimento de que o resultado estará contido no evento A é denotada por

$$P(B|A) \quad (1)$$

Tal expressão é também denominada "probabilidade

condicional de B dado A". É importante notar que o fato do evento estar em A já é conhecido. Isso sugere a seguinte definição de probabilidade condicional: ⑦

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0. \quad (2)$$

Notemos que como $P(A \cap B) = P(B \cap A)$, podemos obter uma relação extremamente útil de (2):

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B). \quad (3)$$

Outra relação útil é a chamada lei de probabilidade total. Para quaisquer eventos A e B,

$$P(B) = P(B \cap A) + P(B \cap A') \quad (4)$$

$$= P(B|A)P(A) + P(B|A')P(A'). \quad (5)$$

A expressão (4) pode ser justificada pelo fato de que $B \cap A$ e $B \cap A'$ são mutuamente exclusivos, pois $A \cap A' = \emptyset$

$$(B \cap A) \cap (B \cap A') = \emptyset$$

Um caso particular é aquele no qual $A \cup A' = S$,

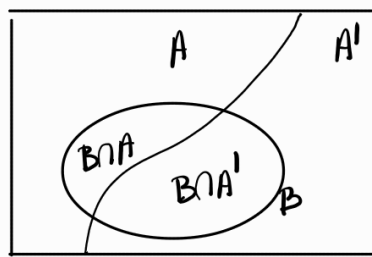


Fig. 1

sendo S o espaço amostral. Nesse caso, a coleção de conjuntos A e A' é dita exaustiva (Fig. 1) e o evento B foi particionado em dois mutuamente exclusivos conjuntos. De modo geral, uma coleção de conjuntos $E_1, E_2, \dots, E_n$ tal que

$$E_1 \cup E_2 \cup \dots \cup E_n = S \quad (6)$$

é dito exaustivo. Se, além disso $E_1, E_2, \dots, E_n$ são mutuamente exclusivos, obtemos a partição de um evento B em subconjuntos mutuamente exclusivos, como mostrada na Fig. 2, na qual

$$B = B \cap E_1 \cup B \cap E_2 \cup \dots \cup B \cap E_n \quad (5)$$

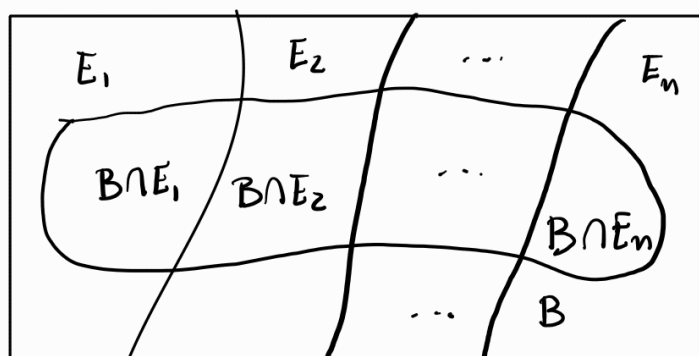


Fig. 2

Nesse caso temos

$$P(B) = P(B \cap E_1) + P(B \cap E_2) + \dots + P(B \cap E_n) \quad (6)$$

Usando (3) em cada termo do lado direito temos

$$P(B) = P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \dots + P(B|E_n)P(E_n) \quad (7)$$

Podemos agora obter o Teorema de Bayes. Da eq. (3)

temos

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad P(B) > 0. \quad (8)$$

9

Em particular, aplicando tal expressão aos eventos $E_1, E_2, \dots, E_n$, mutuamente exclusivos e exaustivos, com B particionado de acordo com (5) e $A = E_1$ (por exemplo), temos

$$P(E_1|B) = \frac{P(B|E_1)P(E_1)}{P(B|E_1)P(E_1) + P(B|E_2)P(E_2) + \dots + P(B|E_n)P(E_n)}, \quad (9)$$

sendo $P(B) > 0$.

Uma clássica aplicação do Teorema de Bayes envolve o cálculo da probabilidade de um indivíduo de uma população ter uma doença de acordo com um teste, dada a eficiência do teste e a prevalência da doença na população. Sejam os eventos:

D : um indivíduo da população tem a doença.

S : o teste aplicado a um indivíduo deu resultado positivo para a doença.

Queremos calcular a probabilidade $P(D|S)$, ou seja, a probabilidade de um indivíduo ter a doença, dado que o resultado do teste é positivo. A eficiência do teste é caracterizada pelas seguintes informações:

$P(S|D)$: probabilidade do teste indicar corretamente um indivíduo doente, se ele está realmente doente.

$P(S|D')$: probabilidade do teste indicar incorretamente um indivíduo doente, quando ele está saudável.

Nesse caso,

$$P(D|S) = \frac{P(S|D)P(D)}{P(D)} \quad (10)$$

Como os eventos D e D' são mutuamente exclusivos,

$$P(D) = P(S|D)P(D) + P(S|D')P(D') \quad (11)$$

e (10) pode ser reescrita como

$$P(D|S) = \frac{P(S|D)P(D)}{P(S|D)P(D) + P(S|D')P(D')} \quad (12)$$

Portanto, como em geral $P(D')$ é alto, $P(D|S)$ é muito baixo, mesmo que $P(S|D)$ seja alto (teste muito efetivo), que é um resultado que surpreende o senso comum.

3. Diserte sobre o uso de gráficos de barras, histogramas, diagramas de caixa (box plots) e de gráficos de dispersão (scatter plots) na análise exploratória de dados.

(i) Gráficos de barras. Tais gráficos são normalmente usados para descrever dados categóricos dos tipos nominais (tipos de produtos, nomes de empresas, etc.) ou ordinais (como escalas de categorias: "péssimo", "regular", "bom", "ótimo", etc.). O tamanho das barras (verticais ou horizontais) pode representar frequências, médias ou outras números de interesse de cada categoria. Como exemplo, ilustramos na Fig. 1 um gráfico de barras que representa o índice de satisfação de clientes em um portal de atendimento.

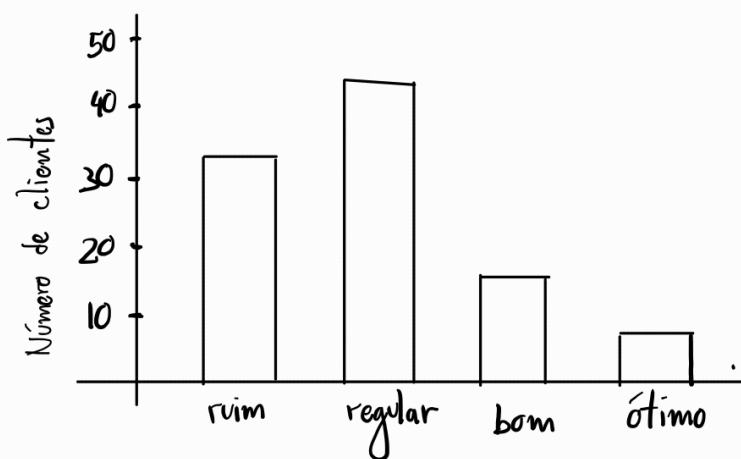


Fig. 1

(ii) Histogramas. Um histograma é uma representação gráfica de variáveis quantitativas (discretas ou contínuas). O correspondente gráfico mostra barras cuja largura inclui um intervalo numérico. A largura das barras é, em geral, constante, e a altura corresponde à frequência (absoluta ou relativa) de eventos em cada intervalo. Na Fig. 2 mostramos um histograma que dá a distribuição de frequências de alturas de 20 alunos em uma determinada sala de aula.

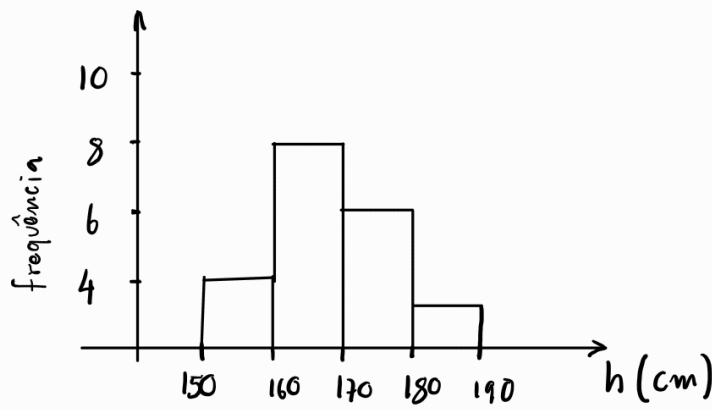


Fig. 2

O número de intervalos a ser usado depende do problema em particular. Um número de intervalos muito pequeno ou muito grande pode resultar em histogramas pouco informativos. Uma regra prática consiste em escolher o número de intervalos como aproximadamente a raiz quadrada do número de observações. Um histograma dá aproximadamente a forma da função de distribuição da variável estudada, de modo a fornecer uma visualização por inspeção de grandezas importantes como média e desvio padrão.

(iii) Diagramas de caixa (box plots). Um box plot é um modo gráfico de apresentar o sumário de cinco números de um conjunto de dados: mínimo, primeiro quartil, mediana, terceiro quartil e máximo. Essencialmente, ele permite uma rápida visualização do centro, da dispersão, afastamento da simetria e identificação de observações não usuais (outliers).

Para descrever a construção de um box plot a partir de um conjunto de dados usaremos a Fig. 3.

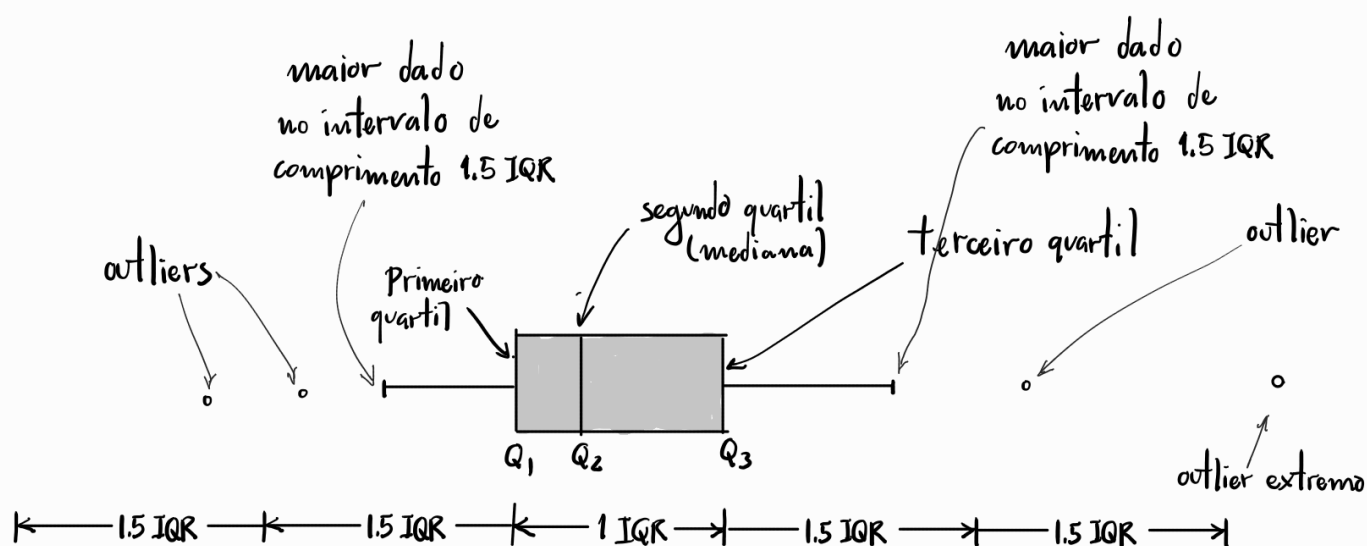


Fig. 3

- O primeiro quartil Q_1 (25º percentil), é o ponto para o qual 25% dos pontos correspondem a valores abaixo dele.
- O segundo quartil Q_2 (mediana do conjunto de dados), é o ponto para o qual 50% dos dados têm valores abaixo dele.
- O terceiro quartil Q_3 (75º percentil), é o ponto para o qual 75% dos pontos estão abaixo dele.
- O intervalo de tamanho 1 IQR é a distância entre o primeiro quartil Q_1 e o terceiro quartil Q_3 .
- As duas barras são chamadas "whiskers" (bigodes). O whisker à esquerda na Fig. 3 se estende de Q_1 até o dado de valor mais baixo, dentro intervalo de tamanho 1.5 IQR, medido a partir de Q_1 . Similarmente para o whisker à direita.
- Dados que estão além dos whiskers, mas ainda dentro de

dos intervalos de 3IQR, medidas a partir de Q_1 e Q_3 são chamados "outliers". Dados que estão além desses intervalos são chamados "outliers extremos".

Boxplots podem ser exibidos de modo horizontal e vertical e em grupos de vários boxplots para fins comparativos, tal como ilustrado na Fig. 4, que mostra dados de altura em cm de alunos de três salas diferentes.

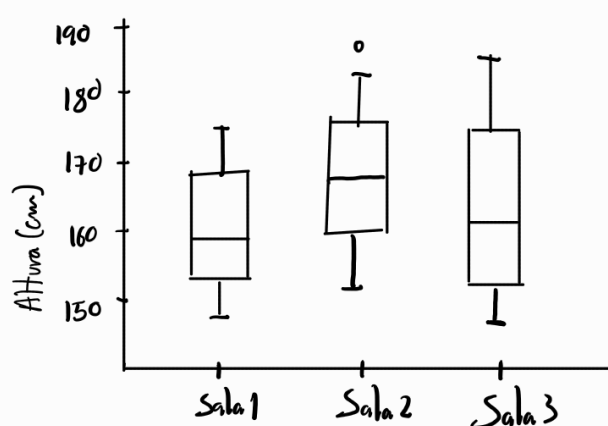


Fig. 4

(iv) Gráficos de dispersão (scatter plots). Gráficos de dispersão ou "scatter plots" são gráficos de pontos no plano que tentam estabelecer uma relação entre dois conjuntos de dados. Cada ponto corresponde a uma medida que coleta um par de dados. Um exemplo é mostrado na Fig. 5., que mostra no eixo horizontal o número médio de horas de dedicação ao estudo por dia de alunos de uma escola, e o eixo vertical é a nota final num exame final. Tal gráfico mostra uma entre variáveis que pode indicar uma relação de causa e efeito. Frequentemente correlações não implicam relações de causa-efeito. Correlação é necessária para causação, mas não suficiente.

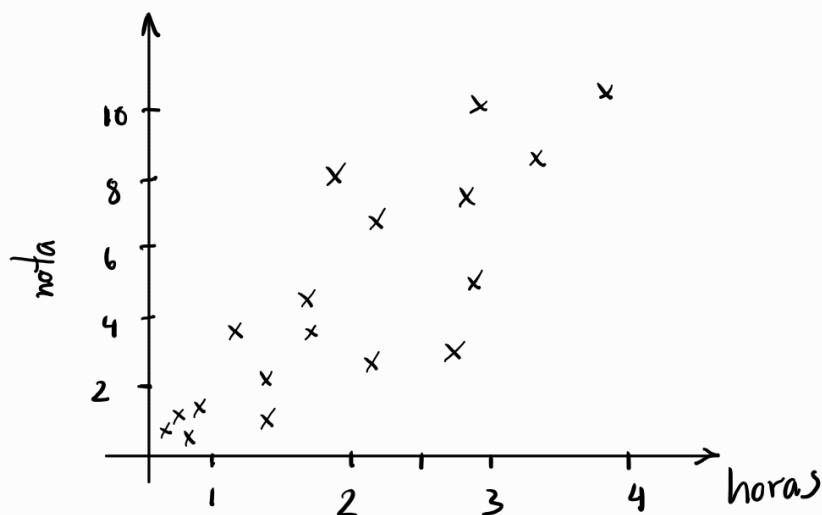


Fig. 5

O gráfico da Fig. 5 parece indicar uma relação linear entre as variáveis horas e nota. Quanto menos dispersos estiverem os pontos em torno da reta que otimiza os dados, mais forte é a correlação. Uma das medidas de correlação linear mais utilizadas é o coeficiente de correlação amostral de Pearson r_{xy} , definido para um par de variáveis aleatórias (x, y) , por

$$r_{xy} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}}$$

Uma correlação forte e positiva implica $r_{xy} \approx 1$ (Fig. 6),

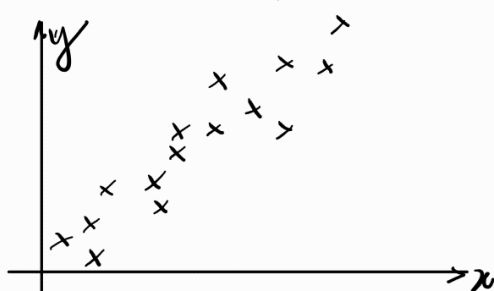


Fig. 6

Uma correlação fraca implica $r_{xy} \approx 0$ (Fig. 7)

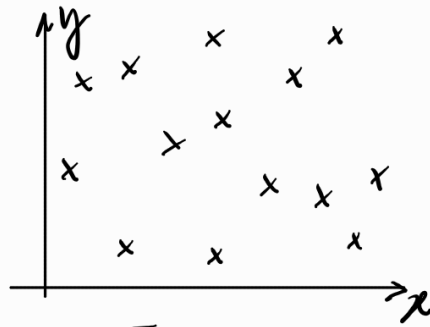


Fig. 7

Uma correlação negativa $r_{xy} < 0$ é ilustrada na Fig. 8.

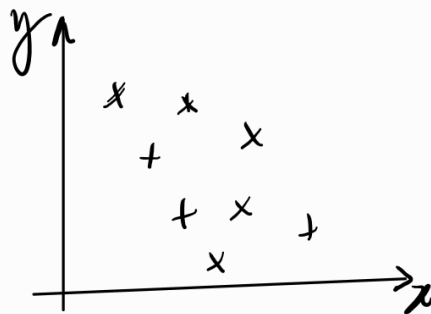


Fig. 8

É importante notar que o coeficiente de Pearson não deve ser aplicado a correlações não lineares, como a mostrada na Fig. 9, que indica uma correlação talvez quadrática entre x e y .

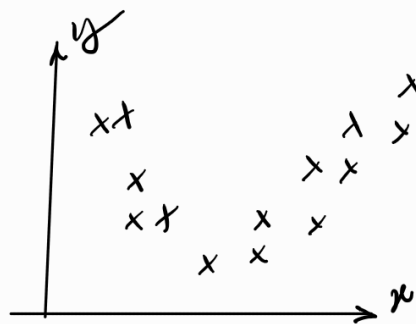


Fig. 9

Quando temos mais de duas variáveis associadas a uma observação ou medida é comum a análise das diferentes combinações de pares, na investigação de possíveis correlações.

Referências para as questões:

Questão 1 Sec. 9.2, p. 322 de [3], Seções 12.3 a 12.5 de [2]. Sec. 8.8 de [1].

Questão 2. Seções 4.3, 4.4 e 4.5 de [1]. Seções 5.3 e 5.4 de [2]. Seções 2.4 a 2.7 de [3].

Questão 3. Seções 2.3, 3.1 a 3.4 de [2]. Seções 6.3, 6.4 e 6.6 de [3]. Seções 3.1 a 3.4 de [1].

1. Barbetta et al., Probabilidade e Estatística para Cursos de Engenharia e Informática. Ed. Atlas, 2004.
2. Bussab et. al., Estatística Básica, 5^a ed., Saraiva, 2002.
3. Montgomery et. al., Estatística Aplicada e Probabilidade para Engenheiros, 6^a ed., LTC, 2016.



Assinaturas do documento



Código para verificação: **5TL39B8D**

Este documento foi assinado digitalmente pelos seguintes signatários nas datas indicadas:



JARBAS CLEBER FERRARI (CPF: 951.XXX.489-XX) em 05/02/2024 às 13:53:11

Emitido por: "SGP-e", emitido em 30/03/2018 - 12:36:56 e válido até 30/03/2118 - 12:36:56.

(Assinatura do sistema)



FERNANDO DEEKE SASSE (CPF: 545.XXX.419-XX) em 05/02/2024 às 13:59:02

Emitido por: "SGP-e", emitido em 30/03/2018 - 12:43:04 e válido até 30/03/2118 - 12:43:04.

(Assinatura do sistema)



JOSE RAFAEL SANTOS FURLANETTO em 05/02/2024 às 15:20:28

Emitido por: "SGP-e", emitido em 30/03/2018 - 12:36:36 e válido até 30/03/2118 - 12:36:36.

(Assinatura do sistema)

Para verificar a autenticidade desta cópia, acesse o link <https://portal.sgpe.sea.sc.gov.br/portal-externo/conferencia-documento/VURFU0NfMTlwMjJfMDAwMDMyNDIfMzI1MF8yMDI0XzVUTDM5QjhE> ou o site

<https://portal.sgpe.sea.sc.gov.br/portal-externo> e informe o processo **UDESC 00003249/2024** e o código **5TL39B8D** ou aponte a câmera para o QR Code presente nesta página para realizar a conferência.