CAPÍTULO 7
Exercícios Resolvidos

R7.1) O problema mais grave do estado do RJ

Considere uma pesquisa por amostragem feita em 1986 junto à população do Estado do Rio de Janeiro. Foram ouvidas 1230 pessoas que, entre outras coisas, apontaram qual era, em sua opinião, o problema mais grave do Estado naquele momento. Com base nos dados brutos, foi obtida a tabela a seguir.

Tabela 5.4 - Freqüências e Percentuais dos 1230 respondentes da Pesquisa junto à população do Estado do RJ em 1986, segundo o problema mais grave do Estado

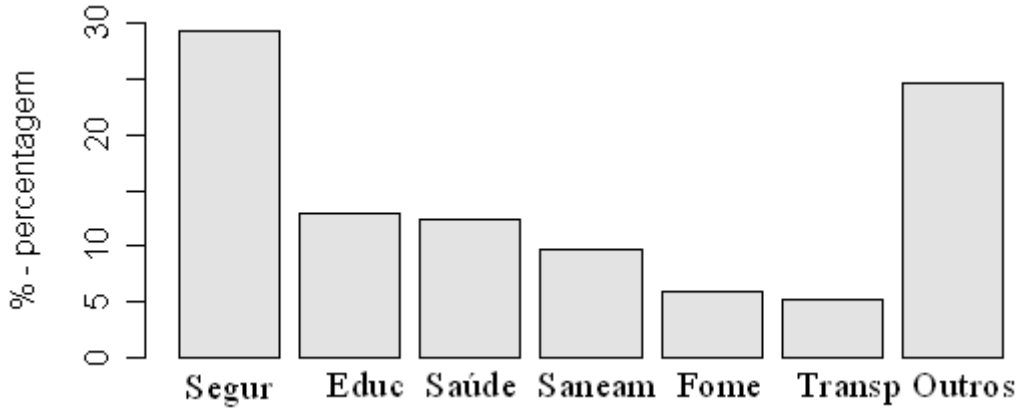
Problema mais grave do Estado	Freqüências	Percentuais
Segurança / Violência (S/V)	360	29,27
Educação	160	13,01
Saúde	152	12,36
Saneamento	118	9,59
Alimentação/Fome/Pobreza (A/F/P)	73	5,93
Transporte	63	5,12
Outros	304	24,72
Total	1230	100,00

Fonte: Pesquisa de Opinião sobre as Eleições do Rio de Janeiro 1986
IBASE / SERPRO / IM-UFRJ

Construa o gráfico de barras e o gráfico de setores (ou gráfico “pizza”) com base nessa tabela de freqüências.

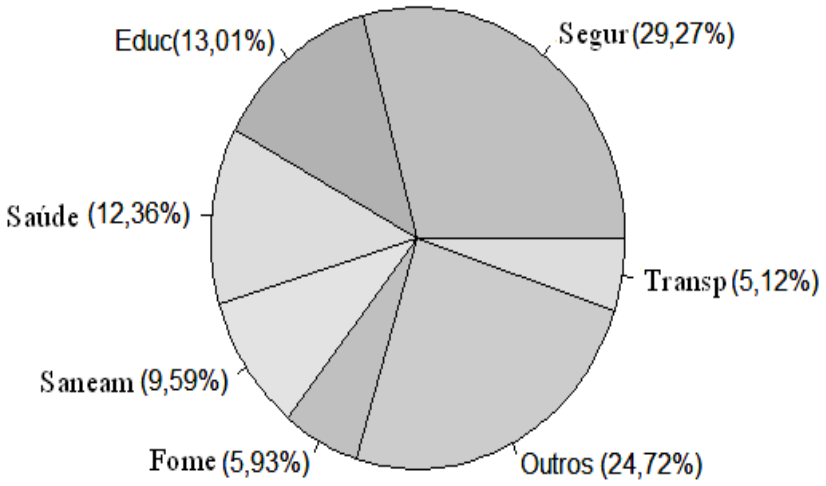
Solução:

Gráfico de barras correspondente aos percentuais dos 1230 respondentes da Pesquisa junto à população do Estado do RJ em 1986, segundo o “problema mais grave do Estado”



Fonte: Pesquisa de Opinião sobre as Eleições do Rio de Janeiro 1986
IBASE / SERPRO/IM-UFRJ

Gráfico de setores correspondente aos percentuais dos 1230 respondentes da Pesquisa junto à população do Estado do RJ em 1986, segundo o “problema mais grave do Estado”



Fonte: Pesquisa de Opinião sobre as Eleições do Rio de Janeiro 1986
IBASE / SERPRO / IM-UFRJ

R7.2) Variável mais constante e Variável vezes constante

Considere o conjunto de dados abaixo:

Obs.	X	Y	Z
1	1	2	4
2	2	4	5
3	3	6	6
4	4	8	7
5	5	10	8

(a) Para cada uma das variáveis X, Y, Z, calcule: média, variância, desvio padrão, coeficiente de variação (cv), mediana (Q2), Q1, Q3, DIQ=Q3-Q1.

(b) Faça um gráfico localizando no eixo horizontal (graduado de 1 a 10) as coordenadas dos pontos e no eixo vertical três níveis: X, Y e Z. Analise visualmente a relação entre as 3 variáveis em termos de centralidade e dispersão.

(c) Verifique que relação matemática existe entre as variáveis Y e X e faça o mesmo com relação às variáveis Z e X. Em seguida verifique que relação matemática existe entre os valores das medidas de centralidade e de dispersão relativas às variáveis Y e Z e as mesmas medidas para X.

(d) Como se poderia generalizar as constatações do item (c) para o caso de variáveis do tipo Y = cX e Z = c+X?

Solução:

(a) Exemplifiquemos os cálculos para a variável Y:

Média : $\bar{y} = \frac{2+4+6+8+10}{5} = 6$

Variância: $S_Y^2 = \frac{(2^2+4^2+6^2+8^2+10^2) - 5\times 6^2}{5-1} = 10$

Desvio Padrão: $S_Y = \sqrt{10} = 3{,}16$

Coeficiente de Variação: $cv_Y = \frac{3{,}16}{6} = 0{,}53$

Mediana: Dados Ordenados:

Posição:	1	2	3	4	5
Valor:	2	4	6	8	10

$$\text{Posição}(Q2) = \frac{1+5}{2} = 3 \qquad Q2(Y) = y_{(3)} = 6$$

1º Quartil: $$\text{Posição}(Q1) = \frac{1+3}{2} = 2 \qquad Q1(Y) = y_{(2)} = 4$$

3º Quartil: $$\text{Posição}(Q3) = \frac{3+5}{2} = 4 \qquad Q3(Y) = y_{(4)} = 8$$

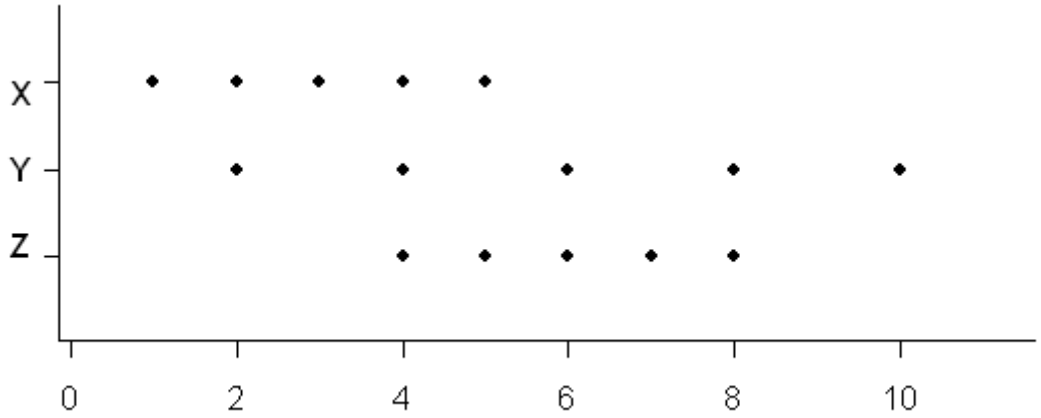
Distância Interquartil: $DIQ(Y) = Q3(Y) - Q1(Y) = 8 - 4 = 4$

O procedimento para as outras duas variáveis é análogo.

A tabela abaixo é um resumo de todos os resultados obtidos:

Var.	Média	Variância	DPadrão	Coef. var.	Q2	Q1	Q3	DIQ
X	3	2,5	1,58	0,53	3	2	4	2
Y	6	10	3,16	0,53	6	4	8	4
Z	6	2,5	1,58	0,26	6	5	7	2

(b) Gráfico simultâneo das variáveis X, Y e Z:



A figura sugere que:

- De X para Y tudo dobra, tanto em termos de locação como de dispersão;
- De X para Z tudo se desloca para a direita de 3 unidades.

(c) É fácil ver que $Y = 2X$ e $Z = 3 + X$.

Além disso, Média(Y) = 6 = 2 x 3 = 2 Média(X)

Variância(Y) =10 = 4 x 2,5 = 2^2 . Variância(X)

DPadrão (Y) = 3,16 = 2 x 1,58 = 2 DPadrão (X)

Coef. var.(Y) = 0,53 = Coef. var.(X)

Mediana(Y) = 6 = 2 x 3 = 2 Mediana(X)

Q1(Y) = 4 = 2 x 2 = 2 Q1(X)

Q3 (Y) = 8 = 2 x 4 = 2 Q3(X)

DIQ(Y) = 4 = 2 x 2 = 2 DIQ(X)

Isso quer dizer que, quando multiplicamos uma variável por 2, todas as medidas aqui consideradas também ficam multiplicadas por 2, exceto: a Variância, que fica multiplicada por 4; e o Coeficiente de Variação, que permanece inalterado.

Média(Z) = 6 = 3 + 3 = 3 + Média(X)
Variância(Y) =10 = 4 x 2,5 = 2^2 . Variância(X)

Variância(Z) = 2,5 = Variância(X)

DPadrão (Z) = 1,58 = DPadrão (X)

Coef. var.(Z) = 0,26 = $\frac{1{,}58}{3+3} = \frac{S_X}{3+\bar{x}}$

Mediana(Z) = 6 = 3 + 3 = 3 + Mediana(X)

Q1(Z) = 5 = 3 + 2 = 3 + Q1(X)

Q3 (Z) = 7 = 3 + 4 = 3 + Q3(X)

DIQ(Z) = 2 = DIQ(X)

Isso quer dizer que, quando somamos 3 unidades a uma variável, a média e os 3 quartis (Q1, Q2 e Q3) também aumentam de 3 unidades. Já a Variância, o Desvio Padrão e a DIQ não se alteram.

(d) Se Y = c.X, temos: Média(Y) = c. Média(X)
Variância(Y) = c^2 . Variância(X)

DPadrão (Y) = |c|. DPadrão (X)

Coef. var.(Y) = Coef. var.(X)

Mediana(Y) = c.Mediana(X)

Q1(Y) = c.Q1(X)

Q3 (Y) = c.Q3(X)

DIQ(Y) = |c|. DIQ(X)

Se Z = c + X: Média(Z) = c + Média(X)
Variância(Z) = Variância(X)

DPadrão (Z) = DPadrão (X)

Coef. var.(Z) = $\frac{S_X}{c+\bar{x}} = \frac{\text{Coef.var.(X)}}{\frac{c}{\bar{x}}+1}$

Mediana(Z) = c + Mediana(X)

Q1(Z) = c + Q1(X)

Q3 (Z) = c + Q3(X)

DIQ(Z) = DIQ(X)

R7.3) Tempo de permanência em hospital – Análise diretamente a partir da distribuição de freqüências

Há determinadas situações em que não se tem acesso aos dados individuais, mas está disponível uma distribuição de freqüências da variável de interesse, como na tabela abaixo.

Tabela – Distribuição de freqüências do tempo de permanência na última internação referente a uma amostra de pacientes do Hospital Espírita de Porto Alegre nos quatro primeiros meses de 1996

Permanência (dias)	Ponto médio (x_j)	Freqüência simples (f_j)
0 a 10	5	70
10 a 20	15	69
20 a 30	25	90
30 a 40	35	43
40 a 50	45	43
50 a 60	55	31
60 a 70	65	16
70 a 80	75	7
80 a 90	85	0
90 a 100	95	1
Total		370

Fonte: Jornal Brasileiro de Psiquiatria - Setembro de 1999

Como determinar as medidas de centralidade e de dispersão em uma tal situação?

Solução:

Para simplificar, vamos considerar que, para todas as observações que pertencem a uma determinada classe (intervalo), o valor da variável é exatamente igual ao ponto médio daquele intervalo.

Portanto, para obter **valores aproximados** para a média $\bar{x}$ e o desvio padrão S usam-se as expressões a seguir:

$$\bar{x} = \frac{\sum_{j=1}^{J} f_j x_j}{n} \qquad S = \sqrt{\frac{\sum_{j=1}^{J} f_j x_j^2 - \frac{\left(\sum_{j=1}^{J} f_j x_j\right)^2}{n}}{n-1}}$$

onde J é o número total de classes da tabela e

para cada classe j, j = 1,2,...,J,

f_j é a freqüência absoluta de observações naquela classe

x_j é o ponto médio do intervalo considerado

No caso do exemplo acima temos então

$$\bar{x} = \frac{70\times 5 + 69\times 15 + ... + 1\times 95}{370} = 28{,}22 \text{ dias}$$

$$S = \sqrt{\frac{\left(70\times 5^2 + 69\times 15^2 + ... + 1\times 95^2\right) - \frac{\left(70\times 5 + 69\times 15 + ... + 1\times 95\right)^2}{370}}{369}} = 18{,}66 \text{ dias}$$

Quanto à determinação da mediana e da distância interquartil, recomenda-se complementar a tabela original com mais algumas colunas, conforme a seqüência de passos abaixo:

i. Construir a partir dos dados uma coluna com as freqüências absolutas acumuladas. Por exemplo: $139 = 70 + 69$; $229 = 139 + 90$; etc.

ii. Construir a partir dos dados uma coluna com as freqüências relativas acumuladas y_j. Por exemplo: $0{,}189 = {}^{70}\!/_{370}$; $0{,}376 = {}^{139}\!/_{370}$; etc.

iii. Montar a tabela a seguir:

Tabela – Cálculos necessários para a determinação da mediana e dos quartis do Tempo de Permanência

Nº da classe	Classe de Permanência (dias)	Freq. abs. simples	Freq. abs. Acumulada	Freq. Rel. acumulada
1	0 a 10	70	70	0,189
2	10 a 20	69	139	0,376
3	20 a 30	90	229	0,619
4	30 a 40	43	272	0,735
5	40 a 50	43	315	0,851
6	50 a 60	31	346	0,935
7	60 a 70	16	362	0,978
8	70 a 80	7	369	0,997
9	80 a 90	0	369	0,997
10	90 a 100	1	370	1,000

Isso quer dizer que:

- 18,9% dos tempos de permanência são menores que 10 dias;
- 37,6% dos tempos de permanência são menores que 20 dias.
- 61,9% dos tempos de permanência são menores que 30 dias.
- 73,5% dos tempos de permanência são menores que 40 dias.
- 85,1% dos tempos de permanência são menores que 50 dias.
- 93,5% dos tempos de permanência são menores que 60 dias.
- 97,8% dos tempos de permanência são menores que 70 dias.
- 99,7% dos tempos de permanência são menores que 80 dias.
- 100,0% dos tempos de permanência são menores que 100 dias.

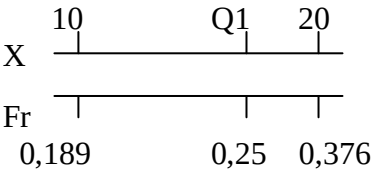
iv. Determinação do 1º quartil Q1

Sabemos que ¼ (ou 25%) das observações devem estar abaixo de Q1. Então os 25% menores tempos de permanência devem ser inferiores a Q1, ou seja, a freqüência relativa acumulada correspondente a Q1 tem que ser igual a 0,25.

Como 0,189 < 0,25 < 0,376, isso implica que necessariamente Q1 está entre 10 e 20 dias. A figura abaixo então nos mostra como podemos calcular o valor de Q1 através de uma Regra de três:

$$\frac{Q1-10}{0,25-0,189}=\frac{20-10}{0,376-0,189}\text{. Então,}$$

$$Q1=10+\frac{(20-10)\times(0,25-0,189)}{0,376-0,189}=13,26$$



v. Determinação do 2º quartil Q2 (mediana)

Sabemos que 1/2 (ou 50%) das observações devem estar abaixo de Q2. Então os 50% menores tempos de permanência devem ser inferiores a Q2, ou seja, a freqüência relativa acumulada correspondente a Q2 tem que ser igual a 0,50.

Como 0,376 < 0,50 < 0,619, isso implica que necessariamente Q2 está entre 20 e 30 dias. Analogamente, podemos escrever também:

$$Q2=20+\frac{(30-20)\times(0,50-0,376)}{0,619-0,376}=25,14$$

vi. Determinação do 3º quartil Q3

Sabemos que 3/4 (ou 75%) das observações devem estar abaixo de Q3. Então os 75% menores tempos de permanência devem ser inferiores a Q3, ou seja, a freqüência relativa acumulada correspondente a Q3 tem que ser igual a 0,75.

Como 0,735 < 0,75 < 0,851, isso implica que necessariamente Q3 está entre 40 e 50 dias. Analogamente, podemos escrever também:

$$Q3=40+\frac{(50-40)\times(0,75-0,735)}{0,851-0,735}=41,29$$

Logo, Mediana = 25,14 dias e DIQ = 41,29 – 13,26 = 28,03 dias.

R7.4) Critério para apontar outliers e o peso da cauda da distribuição mãe

Suponha que dispomos de uma amostra com n observações (dados reais) $x_1, x_2, \ldots, x_n$ relativas a uma determinada variável e desejamos usar o seguinte critério, proveniente da Análise Exploratória, para detectar observações discrepantes nesse conjunto de dados:

A observação x_j é discrepante, se $x_j \notin (\bar{x}-\alpha\cdot s, \bar{x}+\alpha\cdot s)$, onde

$\bar{x}$ e s são, respectivamente, a média e o desvio padrão amostrais

e α é uma constante positiva (a ser escolhida).

Nossa intenção é escolher o valor de α para que somente em 1% dos casos uma observação pertinente seja (erradamente) apontada como *outlier*.

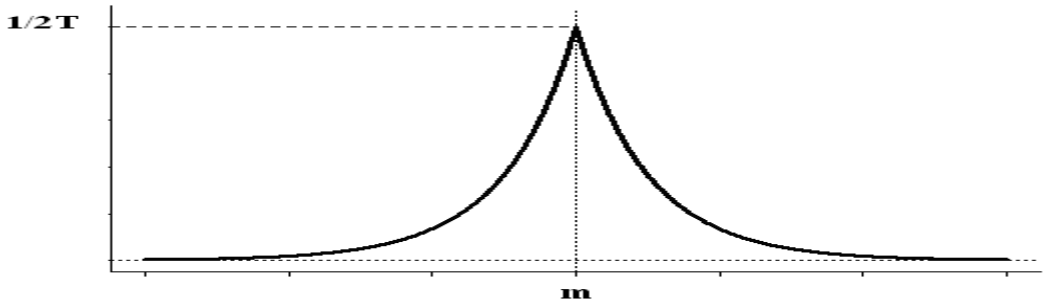
Formulando agora o problema em termos populacionais (e não amostrais), em cada um dos casos abaixo, calcule o valor da constante α (positiva) para que

$$P\left[\left|\frac{X-E(X)}{dp(X)}\right|<\alpha\right]=P\left[E(X)-\alpha\cdot DP(X)<X<E(X)+\alpha\cdot DP(X)\right]=0{,}99.$$

(a) Se X obedece a uma distribuição Uniforme em um intervalo [a,b].
(b) Se X obedece a uma distribuição Normal(μ,σ^2).
(c) Se X obedece a uma distribuição Exponencial dupla com densidade dada por

$$f(x)=\frac{1}{2T}\exp\left(-\frac{|x-m|}{T}\right),\forall x\in R,$$

onde m e T são parâmetros reais com $-\infty<m<\infty$ e $T>0$.
Neste caso, o gráfico da densidade fica com o seguinte aspecto:



(d) Que conclusões podem ser extraídas dos itens (a), (b) e (c) quanto ao valor da constante α a ser utilizado nesse critério?

Obs.: Use as seguintes propriedades matemáticas:

Distribuição de Probabilidade	Média	Desvio Padrão
Uniforme (a;b)	$\frac{a+b}{2}$	$\frac{b-a}{\sqrt{12}}$
Normal ($\mu;\sigma$)	μ	σ
Exp dupla (m;T)	m	$T\sqrt{2}$

Solução:

(a) Suponhamos que X~U[a,b]. Então $E(X)=\frac{a+b}{2}$ e $dp(X)=\frac{b-a}{\sqrt{12}}$.

Portanto, a condição a ser obedecida nesse caso se transforma em

$$0{,}99=P\left[\frac{a+b}{2}-\alpha\frac{b-a}{\sqrt{12}}<X<\frac{a+b}{2}+\alpha\frac{b-a}{\sqrt{12}}\right]. \qquad (I)$$

Lembremos que, no caso da distribuição Uniforme, a probabilidade em (I) coincide com a área do retângulo cuja base é igual a

$\left(\frac{a+b}{2}+\alpha\frac{b-a}{\sqrt{12}}\right)-\left(\frac{a+b}{2}-\alpha\frac{b-a}{\sqrt{12}}\right)=2\alpha\frac{b-a}{\sqrt{12}}$ e cuja altura é igual a $\frac{1}{b-a}$.

Conseqüentemente: $0{,}99=2\alpha\frac{b-a}{\sqrt{12}}\cdot\frac{1}{b-a}=\frac{2\alpha}{\sqrt{12}}$.

Daí, $\alpha = \dfrac{0{,}99\sqrt{12}}{2} = 1{,}715$.

(b) Suponhamos que $X \sim N(\mu;\sigma^2)$. Então $E(X) = \mu$, $DP(X) = \sigma$ e $Z = \dfrac{X-\mu}{\sigma} \sim N(0;1)$.

$$0{,}99 = P[\mu - \alpha\sigma < X < \mu + \alpha\sigma] = P\left[\frac{\mu - \alpha\sigma - \mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{\mu + \alpha\sigma - \mu}{\sigma}\right] = P[-\alpha < Z < \alpha]$$

.

Então, é fácil ver que $\alpha = 2{,}576$. (Basta usar a tabela da Normal padrão.)

(c) Suponhamos que X é exponencial dupla (m,T). Então sua densidade é

$f(x) = \dfrac{1}{2T}\exp\left(-\dfrac{|x-m|}{T}\right), \forall x \in R$. Além disso, $E(X) = m$ e $DP(X) = T\sqrt{2}$.

A condição a ser obedecida nesse caso se transforma em

$$0{,}99 = P[m - \alpha T\sqrt{2} < X < m + \alpha T\sqrt{2}] = \int_{m-\alpha T\sqrt{2}}^{m+\alpha T\sqrt{2}} \frac{1}{2T}\exp\left(-\frac{|x-m|}{T}\right)dx.$$

Fazendo a mudança de variável $u = \dfrac{x-m}{T}$, e usando o fato de que a função

[h: $u \to \exp(-|u|)$] é uma função par, (ou seja, $h(-u) = h(u)$, para todo u), temos

$$0{,}99 = \int_{-\alpha\sqrt{2}}^{\alpha\sqrt{2}} \frac{1}{2T}\exp(-|u|)T\,du = 2\int_0^{\alpha\sqrt{2}} \frac{1}{2}\exp(-u)du = [-\exp(-u)]_0^{\alpha\sqrt{2}} = 1 - \exp(-\alpha\sqrt{2}).$$

Daí, $\exp(-\alpha\sqrt{2}) = 0{,}01$, e conseqüentemente, $\alpha = -\dfrac{\ln 0{,}01}{\sqrt{2}} = 3{,}256$.

(d) Todas as três distribuições aqui consideradas são simétricas em torno da sua média. Porém, à medida que passamos da Uniforme para a Normal e desta para a Exponencial dupla, as caudas da distribuição vão se tornando cada vez mais "pesadas", ou seja, a densidade tende a zero cada vez mais lentamente, à medida que o módulo do seu argumento tende a infinito. Assim sendo, os itens (a), (b) e (c) acima nos mostram que quanto mais "pesadas" forem as caudas da distribuição de probabilidade que deu origem aos dados, maior deverá ser o valor da constante α que figura no critério para apontar *outliers*.

R7.5) Eleição - Intenção de voto em função da faixa etária do eleitor

Com base em uma pesquisa eleitoral relativa ao 2º turno da eleição para a Prefeitura de uma determinada cidade foi obtida a tabela abaixo, que informa a preferência do eleitorado por faixa etária.

Idade (em anos)	Candidato C1	Candidato C2	Brancos, Nulos e Indecisos
$15 \mapsto 25$	70%	10%	20%
$25 \mapsto 35$	50%	20%	30%
$35 \mapsto 45$	40%	40%	20%
$45 \mapsto 55$	30%	60%	10%
$55 \mapsto 65$	20%	70%	10%

Através dessa mesma pesquisa apurou-se também que a distribuição por faixa etária do eleitorado é a seguinte:

Faixa Etária (em anos)	$15 \mapsto 25$	$25 \mapsto 35$	$35 \mapsto 45$	$45 \mapsto 55$	$55 \mapsto 65$
Percentual	25%	30%	25%	15%	5%

Pergunta-se:

(a) Quantos por cento do eleitorado total (entre 15 e 65 anos de idade) pretende votar em C1? E em C2? Qual o percentual correspondente a Brancos, Nulos e Indecisos (BNI)?
(b) Quais são a média e o desvio padrão da idade do eleitorado do candidato C1?
(c) Quais são a mediana e o intervalo interquartil da idade do eleitorado do candidato C2?

Solução:

Os valores da 1ª tabela já nos fornecem uma primeira impressão de que o candidato C1 conta principalmente com o apoio do eleitorado mais jovem, enquanto o candidato C2 conta principalmente com a preferência dos mais idosos.

(a) Para calcular a proporção de intenção de voto em C1 trabalhamos com a coluna referente a este candidato na 1° tabela, bem como a distribuição por faixa etária que está na 2° tabela.
Int. voto C1 = 0,70x0,25+0,50x0,30+0,40x0,25+0,30x0,15+0,20x0,05 = 0,48
Analogamente podemos calcular também
Int. voto C2 = 0,10x0,25+0,20x0,30+0,40x0,25+0,60x0,15+0,70x0,05 = 0,31 e
Int. voto BNI = 0,20x0,25+0,30x0,30+0,20x0,25+0,10x0,15+0,10x0,05 = 0,21
Assim, as intenções de voto globais são: 48% para C1, 31% para C2, e 21% para BNI.

(b) As freqüências correspondentes a cada faixa etária no eleitorado de C1 são:

Faixa etária	Freqüência relativa	Ponto médio
$15 \mapsto 25$	0,365	20
$25 \mapsto 35$	0,313	30
$35 \mapsto 45$	0,208	40
$45 \mapsto 55$	0,094	50
$55 \mapsto 65$	0,021	60

onde $0{,}365 = 0{,}70 \times 0{,}25 / 0{,}48$ $0{,}313 = 0{,}50 \times 0{,}30 / 0{,}48$ ...

Então, trabalhando com o ponto médio de cada intervalo, para o eleitorado de C1, a média de idade (em anos) é

$$\bar{x} = 20 \times 0{,}365 + 30 \times 0{,}313 + 40 \times 0{,}208 + 50 \times 0{,}094 + 60 \times 0{,}021 = 30{,}94$$

e o desvio padrão da idade (em anos) é

$$S \cong (20^2 \times 0{,}365 + 30^2 \times 0{,}313 + 40^2 \times 0{,}208 + 50^2 \times 0{,}094 + 60^2 \times 0{,}021 - 30{,}94^2)^{1/2} = 10{,}61$$

(c) Analogamente, as freqüências correspondentes a cada faixa etária no eleitorado de C2 são

Faixa etária	Frequencia relativa simples	Frequencia relativa acumulada
$15 \mapsto 25$	0,081	0,081
$25 \mapsto 35$	0,194	0,275
$35 \mapsto 45$	0,323	0,598
$45 \mapsto 55$	0,290	0,888
$55 \mapsto 65$	0,113	1,001

Os quartis da variável idade relativa ao eleitorado de C2 são:

$$Q1 = 25 + \frac{0{,}25 - 0{,}081}{0{,}194} \times 10 = 33{,}71 \qquad Q2 = 35 + \frac{0{,}50 - 0{,}275}{0{,}323} \times 10 = 41{,}96$$

$$Q3 = 45 + \frac{0{,}75 - 0{,}598}{0{,}290} \times 10 = 50{,}24 \qquad Q3 - Q1 = 16{,}53$$

Obs.: Para uma explicação mais detalhada desse tipo de raciocínio, veja o Exercício 3 acima.

Portanto, mediana = 41,96 anos e distância interquartil = 16,53 anos.

Esses resultados corroboram a nossa impressão inicial de que C1 é o preferido do eleitor mais jovem, enquanto que o eleitorado de C2 já é formado principalmente pelos eleitores de idade mais avançada.

R7.6) Por que o módulo da correlação é menor ou igual a 1?

A desigualdade de Schwarz é uma propriedade da Álgebra Linear, segundo a qual,

$(\mathbf{u}^t\mathbf{v})^2 \le (\mathbf{u}^t\mathbf{u})(\mathbf{v}^t\mathbf{v})$, se $\mathbf{u}$ e $\mathbf{v}$ são dois vetores do R^n.

Usando a Desigualdade de Schwarz, justifique por que o coeficiente de correlação amostral é menor ou igual a 1 em módulo, ou seja, $|r_{xy}| \le 1$, para quaisquer duas variáveis quantitativas x e y.

Sugestão: Dados os vetores $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ e $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ do R^n, faça $\mathbf{u} = \mathbf{x} - \bar{x}\mathbf{1}$ e $\mathbf{v} = \mathbf{y} - \bar{y}\mathbf{1}$, onde $\bar{x}$ e $\bar{y}$ são as médias amostrais de $\mathbf{x}$ e $\mathbf{y}$ respectivamente, e $\mathbf{1}$ é o vetor do R^n cujas coordenadas são todas iguais a 1.

Solução:
Usando a sugestão, temos:
$[(x - \bar{x}1)^t(y - \bar{y}1)]^2 \leq \{(x - \bar{x}1)^t(x - \bar{x}1)\} \cdot \{(y - \bar{y}1)^t(y - \bar{y}1)\}$
Ou seja:

$$\left([x_1 - \bar{x} \quad \cdots \quad x_n - \bar{x}] \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \right)^2 \leq \cdots$$

$$\cdots \leq \left\{ [x_1 - \bar{x} \quad \cdots \quad x_n - \bar{x}] \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} \right\} \cdot \left\{ [y_1 - \bar{y} \quad \cdots \quad y_n - \bar{y}] \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \right\}$$

Efetuando os produtos matriciais:
$$\left[\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \right]^2 \leq \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \cdot \left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right)$$
Logo:
$$\frac{[\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})]^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2) \cdot (\sum_{i=1}^{n}(y_i - \bar{y})^2)} \leq 1$$
Extraindo a raiz quadrada, temos $|r_{xy}| \leq 1$

R7.7) Regressão e Correlação

A partir de uma massa de dados com n = 20 pares (x_i, y_i) calcularam-se:

- reta de regressão: y = 2627,82 – 37,15 x;
- α = quociente entre as médias amostrais = $\bar{y} / \bar{x}$ = 159,50;
- β = quociente entre os desvios padrão amostrais = s_Y / s_X = 39,123.

(a) Determine o coeficiente de correlação r_{xy}.
(b) Determine as médias amostrais $\bar{x}$ e $\bar{y}$.

Solução:
Vamos utilizar os seguintes símbolos:
$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \qquad SXX = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(a) Então podemos escrever:
$$-37{,}15 = b = \frac{SXY}{SXX} \quad \text{e} \quad 2627{,}82 = a = \bar{y} - b\bar{x} = \bar{y} + 37{,}15\bar{x}, \quad (*)$$
onde a e b são os coeficientes da reta de regressão.

Por outro lado:

$$r_{xy} = \frac{SXY}{\sqrt{SXX \cdot SYY}} \qquad s_X = \sqrt{\frac{SXX}{n-1}} \qquad s_Y = \sqrt{\frac{SYY}{n-1}} \qquad \beta = \frac{s_Y}{s_X} = \sqrt{\frac{SYY}{SXX}} = 39{,}123$$

Daí se deduz que

$$r_{xy} = \frac{SXY}{SXX} \cdot \sqrt{\frac{SXX}{SYY}} = \frac{b}{\beta} = \frac{-37{,}15}{39{,}123} = -0{,}94957$$

(b) Sabemos também que $159{,}50 = \alpha = \frac{\bar{y}}{\bar{x}}$. Substituindo em (*), obtemos

$a = \alpha\bar{x} - b\bar{x}$, o que implica que $\bar{x} = \frac{a}{\alpha - b} = \frac{2627{,}82}{159{,}50+37{,}15} = 13{,}363$

Finalmente, $\bar{y} = \alpha\bar{x} = 159{,}50 \times 13{,}363 = 2131{,}387$.

Exercícios Propostos

P7.1) Um erro grosseiro, mas (infelizmente) bastante comum

A partir dos dados $x_1, x_2, ..., x_n$, calcula-se a variância da variável x pela expressão

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Então, se $n = 5$, $x_1 = 3$, $x_2 = 5$, $x_3 = 0$, $x_4 = 2$, $x_5 = 6$, temos:

$$\sum_{i=1}^{5} x_i^2 = 3^2 + 5^2 + 0^2 + 2^2 + 6^2 = 74$$

Logo, $S^2 = \frac{74 - \frac{74}{5}}{5-1} = \frac{74 - \frac{74}{5}}{4} = 14{,}8$

(a) O que está errado nesse cálculo?
(b) Qual a solução correta?

P7.2) Combinando duas amostras

Foram coletados os dados relativos a uma determinada variável para duas amostras distintas. Apresentam-se a seguir, para cada uma delas, os resultados obtidos quanto a: tamanho da amostra, média e desvio padrão.

Amostra	No de Obs.	Média	D Padrão
1	5	7,4	6,309
2	4	11,5	9,983

(a) Considerando agora a amostra combinada, ou seja, a amostra composta por todas as 9 observações, qual o valor da sua média?
(b) Qual o seu desvio padrão?

P7.3) Atualização da mediana

Temos um conjunto de dados com 11 observações já devidamente ordenadas:

$x_{(1)} < x_{(2)} < x_{(3)} < \ldots < x_{(11)}$.

Quais das seguintes afirmações estão corretas e quais estão erradas? Por que?

(a) A mediana desses dados é $Q2 = x_{(6)}$.

(b) Suponha que foi eliminada desse conjunto de dados original a maior observação $x_{(11)}$. Então a nova mediana é agora $= \frac{1}{2}\left(x_{(1)} + x_{(10)}\right)$.

(c) Suponha que foi adicionada ao conjunto de dados original mais uma observação, de modo que ele passou a ter 12 observações. Então a diferença entre o maior valor possível da nova mediana e o menor valor possível da nova mediana é $= \frac{1}{2}\left(x_{(7)} - x_{(5)}\right)$.

P7.4) Implantes mamários – Raciocínio equivocado

Um determinado fabricante produz implantes mamários utilizando gel de Silicone. Os dados a seguir se referem à tensão de ruptura desses implantes e foram obtidos através de testes físicos realizados com uma amostra de tamanho n = 20:

72,2	80,1	70,4	67,8	70,9	72,1	75,1	73,0	59,4	77,2
65,1	66,5	64,1	79,0	70,6	70,3	63,1	64,4	74,9	75,3

Com base nesses dados obtenha os quartis Q1, Q2 e Q3.

Foi apresentada a seguinte solução:

Posição	1	2	3	4	5	6	7	8	9	10
Valor	72,2	80,1	70,4	67,8	70,9	72,1	75,1	73	59,4	77,2
Posição	11	12	13	14	15	16	17	18	19	20
Valor	65,1	66,5	64,1	79	70,6	70,3	63,1	64,4	74,9	75,3

Posição (Q2) = $\frac{1+20}{2} = 10{,}5$ → $Q2 = \frac{1}{2} \times 77{,}2 + \frac{1}{2} \times 65{,}1 = 71{,}15$

Posição (Q1) = $\frac{1+10{,}5}{2} = 5{,}75$ → $Q1 = \frac{1}{4} \times 70{,}9 + \frac{3}{4} \times 72{,}1 = 71{,}8$

Posição (Q2) = $\frac{10{,}5+20}{2} = 15{,}25$ → $Q3 = \frac{3}{4} \times 70{,}6 + \frac{1}{4} \times 70{,}3 = 70{,}525$

a) Algo está errado nessa solução. O que é?

b) Qual a solução correta?

P7.5) Preços de automóveis

A tabela de freqüências a seguir se refere aos preços (em reais) pelos quais foram anunciados 2695 automóveis para venda em um determinado site.

Faixa de preço	Freqüência
Até R$ 7.000	344
De R$ 7.001 a R$ 10.000	419
De R$ 10.001 a R$ 15.000	530
De R$ 15.001 a R$ 20.000	443
De R$ 20.001 a R$ 25.000	320
De R$ 25.001 a R$ 30.000	229
De R$ 30.001 a R$ 40.000	220
De R$ 40.001 a R$ 50.000	99
De R$ 50.001 a R$ 100.000	80
Acima de R$ 100.000	11
Total	2695

(a) Determine a média e o desvio padrão dessa variável.

(b) Determine a mediana e a distância interquartil dessa variável.

P7.6) Telefonia fixa per capita

A Tabela de dados brutos a seguir reporta, o número linhas telefônicas por 1000 habitantes em cada estado do Brasil, em 2001.

Tabela – Telefonia fixa per capita em cada estado do Brasil em 2001
(em linhas telefônicas por 1000 habitantes)

Acre	183,8	**Maranhão**	86,1	**Rio Janeiro**	347,5
Alagoas	125,4	**M Grosso**	199,6	**R G Norte**	150,1
Amapá	193,3	**M G Sul**	235,3	**R G Sul**	236,9
Amazonas	162,0	**M Gerais**	218,6	**Rondônia**	214,6
Bahia	142,3	**Pará**	128,0	**Roraima**	214,1
Ceará	140,6	**Paraíba**	125,4	**Sta Catarina**	257,3
D Federal	456,8	**Paraná**	244,2	**S Paulo**	362,8
E Santo	228,7	**Pernambuco**	147,8	**Sergipe**	140,7
Goiás	231,4	**Piauí**	118,2	**Tocantins**	113,8

Fonte: Almanaque Abril 2002

(a) Construa um ramo-folha para essa variável.

(b) Determine os seus quartis.

(c) Construa o seu *Box-plot*.

(d) Há *outliers* entre essas observações? Quais?

P7.7) Habitantes por leito hospitalar nos estados do Brasil

A tabela a seguir contem o número de habitantes por leito hospitalar em cada estado do Brasil no ano de 2005.

Tabela – N^{o} de habitantes/leito no Brasil em 2005

Rio Janeiro	341,30	**Minas Gerais**	414,94	**D Federal**	469,48
Goiás	344,83	**Mato Grosso**	418,41	**Tocantins**	471,70
R G Sul	354,61	**R G Norte**	418,41	**Rondônia**	497,51
Paraná	362,32	**Acre**	421,94	**Alagoas**	507,61
M G Sul	364,96	**São Paulo**	436,68	**Pará**	520,83
Sta Catarina	375,94	**Maranhão**	440,53	**Sergipe**	552,49
Pernambuco	395,26	**E Santo**	446,43	**Amazonas**	641,03
Paraíba	398,41	**Bahia**	456,62	**Roraima**	653,59
Piauí	404,86	**Ceará**	467,29	**Amapá**	800,00

Fonte: IBGE - Pesquisa Assistência Médico-Sanitária

(a) Faça um *box plot* desses dados.

(b) Calcule a média, a mediana, o desvio padrão e a distância interquartil.

(c) Repita o item (a), porém excluindo o estado do Amapá.

(d) Compare as variações em cada uma dessas quatro medidas, com e sem o Amapá, e extraia conclusões a esse respeito.

P7.8) Propriedades da Distribuição Normal

Seja X uma variável aleatória com distribuição Normal de média μ e desvio padrão σ. Sejam $q_1(X)$, $q_2(X)$ e $q_3(X)$ os três quartis de X, ou seja, eles são tais que

$$P(X < q_1(X)) = 1/4 \quad P(X < q_2(X)) = 1/2 \quad P(X < q_3(X)) = 3/4 .$$

Finalmente, sejam $a = q_1(X) - 1{,}5(q_3(X) - q_1(X))$ e $b = q_1(X) + 1{,}5(q_3(X) - q_1(X))$.

(a) Obtenha expressões matemáticas para $q_1(X)$, $q_2(X)$ e $q_3(X)$ em função de μ e σ.

(b) Prove que $P[a < X < b] > 0{,}99$.

Obs.: Esta é uma avaliação probabilística do procedimento para apontar observações aberrantes (usando medidas resistentes) que foi exposto na teoria, para o caso de dados – sem *outliers* – provenientes de um modelo gaussiano.

P7.9) Proximidade entre medidas de centralidade e entre medidas de dispersão

Quando o ramo-folha, obtido com base em um conjunto de dados relativos a uma determinada variável, sugere que:

- existe simetria em torno de um valor central;
- não há observações discrepantes;

qual é a sua expectativa no que se refere:

- à proximidade entre a média e a mediana dessa variável?
- à proximidade entre o desvio padrão e a distância interquartil dessa variável?

Justifique as suas respostas.

Sugestão:

Imagine, por exemplo, que os dados foram gerados a partir de uma distribuição Normal.

P7.10) Número de hotéis nos municípios da Região Serrana do RJ

Os dados abaixo se referem ao número de estabelecimentos hoteleiros em cada um dos 37 municípios da Região Serrana do Estado de Rio de Janeiro no ano de 2001.

Tabela – Nº de hotéis na Região Serrana do RJ – 2001

Município	Hotéis	Município	Hotéis	Município	Hotéis
Areal	3	Resende	36	Nova Friburgo	84
Barra Mansa	22	Engenheiro Passos	4	Lumiar	16
Barra do Piraí	18	Visconde de Mauá	12	São Pedro da Serra	14
Engº. Paulo de Frontin	8	Rio Claro	10	Paty do Alferes	8
Itatiaia	121	Rio das Flores	2	Petrópolis	83
Maringá	17	Sapucaia	5	Petrópolis / arredores	58
Maromba	20	Três Rios	15	Rio Bonito	6
Penedo	55	Valença	34	SJ do Vale do Rio Preto	5
Mendes	5	Conservatória	20	Silva Jardim	6
Paraíba do Sul	12	Volta Redonda	14	Teresópolis	44
Piraí	7	Cach de Macacu	14	Vassouras	9
Porto Real	4	Guapimirim	7		
Quatis	5	Miguel Pereira	12		

Com base nesse conjunto de dados foram calculados:

Média	22, 027
Mediana	12
Q1	6
Q3	20
DIQ	14

Se for utilizado o critério para identificação de observações discrepantes que se baseia em medidas resistentes, teremos Cerca Superior = Q3 + 1,5 DIQ = 20 + 1,5 x 14 = 41. Sendo assim, 6 das 37 observações (16%) seriam apontadas como possíveis *outliers*, isto é, municípios onde haveria um número anormalmente alto de hotéis: Teresópolis (44), Penedo (55), Petrópolis/arredores (58), Petrópolis (83), Nova Friburgo (84) e Itatiaia (121).

Responda:

(a) Por que a média resultou ser tão maior que a mediana neste caso?

(b) Por que tantos municípios teriam sido apontados pelo critério que identifica *outliers*?

P7.11) Deficit habitacional no Estado do Rio de Janeiro

A tabela a seguir contem o número de domicílios rústicos em alguns municípios do Estado do Rio de Janeiro no ano 2000.

Tabela – N^{o} de domicílios rústicos no RJ – 2000

Angra dos Reis	572		Miracema	216
Araruama	117		Niterói	914
Barra do Piraí	741		Nova Friburgo	295
Barra Mansa	250		Nova Iguaçu	457
Belford Roxo	339		Petrópolis	1.839
Cabo Frio	566		Queimados	81
Campos dos Goytacazes	1.119		Resende	66
Duque de Caxias	556		Rio das Ostras	123
Guapimirim	51		Sto Antônio de Pádua	88
Itaboraí	132		São Gonçalo	394
Itaguaí	70		São João de Meriti	103
Itaperuna	74		São Pedro da Aldeia	77
Japeri	122		Saquarema	289
Macaé	143		Seropédica	159
Magé	567		Teresópolis	329
Maricá	64		Valença	229

Fonte: Fundação João Pinheiro (FJP), Centro de Estatística e Informações (CEI)

a. Determine os quartis
b. Obtenha um gráfico *Box-plot* para esses dados.
c. Seria o gráfico ramo-folha adequado para representar estes dados? Por que?

P7.12) Fundo de Participação dos Municípios

A tabela a seguir contem o valor total no 1º semestre da 2010 do Fundo de Participação dos Municípios por unidade da Federação, em milhões de reais:

Tabela – Fundo de participação dos municípios – 2010

AC	100,33	MA	846,76	RJ	615,30
AL	483,44	MG	2666,50	RN	512,34
AM	299,97	MS	309,41	RO	178,50
AP	71,81	MT	375,42	RR	76,00
BA	1843,72	PA	740,54	RS	1378,96
CE	1060,50	PB	656,62	SC	789,20
DF	33,82	PE	1021,47	SE	295,47
ES	350,06	PI	538,47	SP	2688,37
GO	742,54	PR	1379,80	TO	281,75

Fonte: Secretaria do Tesouro Nacional

(a) Determine os quartis Q1, Q2, Q3.

(b) Construa um Box plot para esses dados.

P7.13) Desemprego no Brasil

A tabela a seguir contem a taxa de desemprego em cada estado do Brasil no ano de 2007:

Tabela – Desemprego no Brasil - 2007

Rondônia	6,26		Ceará	6,89		Rio de Janeiro	10,16
Acre	4,07		R G do Norte	8,3		São Paulo	9,31
Amazonas	10,54		Paraíba	7,53		Paraná	5,62
Roraima	9,24		Pernambuco	11,41		Santa Catarina	4,69
Pará	7,03		Alagoas	7,6		R Gr do Sul	6,59
Amapá	14,46		Sergipe	8,81		M Gr do Sul	6,59
Tocantins	5,67		Bahia	9,23		Mato Grosso	5,83
Maranhão	6,41		Minas Gerais	7,43		Goiás	7,72
Piauí	3,76		Espírito Santo	10,27		Distrito Federal	7,63

Obtenha um gráfico ramo-folha e um *Box-plot* para esses dados.

P7.14) Densidade populacional em cada estado do Brasil

Os dados a seguir se referem à densidade populacional de cada estado do Brasil conforme apurado no Censo de 1980.

Tabela – Densidade Populacional no Brasil - 1980

Estado	Densidade (habit./km^2)	Estado	Densidade (habit./km^2)
Rondônia	2,02	Acre	1,97
Amazonas	0,92	Roraima	0,34
Pará	2,77	Amapá	1,26
Maranhão	12,31	Piauí	8,52
Ceará	36,02	R G do Norte	35,8
Paraíba	49,14	Pernambuco	62,49
Alagoas	71,7	Sergipe	51,84
Bahia	16,88	Minas Gerais	22,96
E Santo	94,37	R Janeiro	260,74
São Paulo	101,25	Paraná	38,33
S Catarina	38,00	R G Sul	29,06
M Grosso Sul	3,91	Mato Grosso	1,29
Goiás	6,01	Distrito Federal	203,94

(a) Calcule a média e o desvio padrão da variável densidade.
(b) Suponha que um determinado valor da variável pode ser considerado discrepante dos demais se a distância entre esse valor e a média for maior que 3 desvios padrão. Quais dos estados acima se enquadram nessa categoria de valores discrepantes?
(c) Construa um *box-plot* para esses dados.
(d) Construa um *box-plot* para as raízes quadradas desses dados.
(e) Construa um *box-plot* para os logaritmos desses dados.
(f) Compare esses *box-plots* quanto ao seu grau de simetria.

Obs.: Para os dados acima, $\sum x = 1153{,}84$ e $\sum x^2 = 150326{,}3774$.

P7.15) Dados Simulados a partir de uma distribuição conhecida

Os dados a seguir podem ser encarados como uma amostra aleatória de tamanho n = 15 da distribuição Normal com média populacional $\mu = 10$ e variância populacional $\sigma^2 = 4$. Eles foram obtidos por simulação usando um gerador de números aleatórios.

9,5	11,4	7,2	10,0	9,4	8,2	6,4	10,9	7,6	9,5	10,7	9,9	8,8	8,6	9,9

(a) Calcule a mediana q_2 e a distância interquartil populacionais diq (= $q_3 - q_1$) dessa distribuição de probabilidade.
(b) Obtenha a média, a variância, a mediana e a distância interquartil amostrais usando os dados aqui fornecidos.
(c) Repita o que foi feito no item anterior, porém acrescentando aos dados um *outlier* cujo valor é 100.
(d) Preencha a tabela a seguir e extraia as conclusões cabíveis.

	Média	Variância	Mediana	Dist. Interquartil
Medidas populacionais				
Medidas amostrais (sem o outlier)				
Medidas amostrais (com o outlier)				

P7.16) Escolha da carreira e suas motivações entre os vestibulandos

Foi realizada uma pesquisa junto aos alunos classificados no vestibular da UFRJ em 1993. A Tabela de contingência a seguir foi montada a partir dos dados que constavam em 810 questionários selecionados por amostragem.

Tabela de Contingência relativa às variáveis Opção de carreira e Fator predominante na escolha de carreira. Dados relativos a uma amostra do 810 classificados no vestibular da UFRJ em 1993

Áreas de Opção de Carreira	**Fatores Determinantes da Escolha da Carreira**					
	Mercado Trabalho	Prestígio	Aptidão Pessoal	Baixa conc. por vagas	Perspect. salariais	Total
Biomédica	13	2	113	4	5	137
Exatas/Tecnologia	24	1	176	2	5	208
Arquit./Artes Gráf.	0	1	49	1	1	52
Geo-econômica	11	0	61	0	1	73
Outras	33	5	286	6	10	340
Total	81	9	685	13	22	810

Fonte: "Perfil Sócio-Econômico dos Alunos Classificados na UFRJ no Vestibular de 1993", E.A.Simone, R.C.Gomes

(a) Com base nessa tabela de contingência obtenha uma nova tabela com percentuais:

- Promovendo eventualmente algumas fusões de linhas ou colunas de modo a garantir a confiabilidade do processo de extrapolação dos resultados da amostra para a população; e
- Comparando as 5 áreas de opções de carreira em termos do perfil de motivações para as escolhas características de cada área.

(b) Extraia as conclusões cabíveis.

P7.17) Será mera coincidência?

Considere o seguinte conjunto de dados:

No. obs	X	Y
1	2	6
2	4	2
3	7	2
4	3	9
5	6	0

(a) Calcule as variâncias amostrais de X e de Y, ou seja, S_X^2 e S_Y^2.

(b) Calcule a covariância amostral entre X e Y, ou seja, S_{XY}.

(c) Construa uma nova variável Z = X + Y e calcule a sua variância amostral S_Z^2.

(d) Compare S_Z^2 com $S_X^2 + S_Y^2 + 2S_{XY}$.

(e) Como você explica essa coincidência?

P7.18) Produção Industrial e Força de Trabalho no Brasil

A tabela a seguir fornece para cada estado do Brasil, o Valor total V da produção industrial (em milhões de cruzeiros), o Número total (P) de pessoas ocupadas na indústria, o logaritmo decimal de V e o logaritmo decimal de P, segundo o Censo Industrial de 1980.

Tabela – Prod. Industrial e Força de trabalho no Brasil - 1980

Estado	**V**	**P**	**$\log_{10}(V) = y$**	**$\log_{10}(P) = x$**
Amazonas	333	527	2,52	2,72
Pará	2655	2035	3,42	3,31
Maranhão	71	271	1,85	2,43
Piauí	882	1290	2,95	3,11
Ceará	8874	13776	3,95	4,14
Rio Grande do Norte	5989	9816	3,78	3,99
Paraíba	1469	2499	3,17	3,40
Pernambuco	9134	12720	3,96	4,10
Alagoas	924	1031	2,97	3,01
Sergipe	951	961	2,98	2,98
Bahia	2234	4154	3,35	3,62
Minas Gerais	17089	30002	4,23	4,48
Espírito Santo	2653	4402	3,42	3,64
Rio de Janeiro	39503	49256	4,60	4,69
São Paulo	172229	195756	5,24	5,29
Paraná	4364	7619	3,64	3,88
Santa Catarina	34335	28949	4,54	4,46
Rio Grande do Sul	64851	91813	4,81	4,96
Mato Grosso do Sul	59	222	1,77	2,35
Mato Grosso	32	83	1,51	1,92
Goiás	1196	2415	3,08	3,38
Distrito Federal	99	239	2,00	2,38

Com base nesses dados:

(a) Construa uma tabela no formato abaixo, onde em cada posição da tabela conste o percentual de ocorrências correspondentes àquela coluna - classe de V - dentro do total de ocorrências da linha - classe de P.

Classe de V - Valor da Produção
(em milhões de cruzeiros)

Classe de P – Pessoal Ocupado	Até 1000	Entre 1000 e 10000	Mais de 10000
Até 1000			
Entre 1000 e 10000			
Mais de 10000			

(b) Plote os pares (x,y) , onde $x = \log_{10} P$ e $y = \log_{10} V$, em um sistema de eixos coordenados e calcule o coeficiente de correlação entre essas variáveis.

(c) Obtenha as estimativas de mínimos quadrados dos coeficientes da reta de regressão $y = a + b.x$, onde $x = \log_{10} P$ e $y = \log_{10} V$.

(d) Qual seria a sua estimativa para o Valor Total da Produção Industrial (em milhões de Cruzeiros) em um estado onde houvesse 10000 pessoas ocupadas na indústria? Justifique a sua resposta.
(e) Identifique os valores discrepantes da variável log(P) usando o critério que se baseia nos quartis da variável.

Sabe-se que

$$\sum x = 78{,}25 \quad \sum y = 73{,}72 \quad \sum x^2 = 295{,}72 \quad \sum y^2 = 268{,}66 \quad \sum xy = 281{,}39$$

P7.19) Engenharia de Estruturas

Em seu livro "*Uncertainties analysis, loads and safety in Structural Engineering* (em português: Análise de incertezas, cargas e segurança em Engenharia de Estruturas), Prentice Hall, 1982", Gary C. Hart apresenta o conjunto de dados a seguir, que nos permite investigar a relação de dependência entre duas propriedades mecânicas do concreto: **X = módulo secante (em 10^6 psi) e Y = força de compressão. (em 10^3 psi)**

X	3,41	3,52	3,57	3,61	3,43	3,59	3,62	3,56	3,35	3,47
Y	8,20	7,10	7,30	8,60	6,80	7,60	8,50	6,90	5,40	6,20
X	3,53	3,33	3,54	3,22	3,49	3,25	3,79	3,64	3,67	3,72
Y	7,90	5,80	9,10	4,50	6,30	5,20	9,50	8,90	7,40	8,70

(a) Obtenha um diagrama de dispersão relativo a esses dados.

(b) Calcule o coeficiente de correlação entre x e y.

(c) Ajuste aos dados a reta de regressão y = a +bx

P7.20) Acidentes em Auto-estradas

Este conjunto de dados contem informações relativas a n = 39 trechos de grandes auto-estradas do estado de Minnesota, EUA. Somente algumas das variáveis originalmente disponíveis foram aqui consideradas.

Fonte: Weisberg, "Applied Linear Regression", Wiley, 1980

Descrição das variáveis:

Símbolo	Nome	Unidades ou Explicação
RATE	Taxa de Acidentes	Acidentes por milhão de veículos.milhas
ACPT	Pontos de Acesso	N^o de pontos de acesso por milha no trecho
FAI	Indicador de FAI	= 1, se auto-estrada interestadual secundária, senão = 0
PA	Indicador de PA	= 1, se artéria principal, senão = 0
MA	Indicador de MA	= 1, se artéria especial, senão = 0

Obs.: Para dois dos trechos na amostra FAI = PA = MA = 0. Eles correspondem a auto-estradas recebedoras de fluxo especiais.

Aqui estão os dados:

Rate	ACPT	FAI	PA	MA
4,58	4,6	1	0	0
2,86	4,4	1	0	0
3,02	4,7	1	0	0
2,29	3,8	1	0	0
1,61	2,2	1	0	0
6,87	24,8	0	1	0
3,85	11	0	1	0
6,12	18,5	0	1	0
3,29	7,5	0	1	0
5,88	8,2	0	1	0
4,2	5,4	0	1	0
4,61	11,2	0	1	0
4,8	15,2	0	1	0

Rate	ACPT	FAI	PA	MA
3,85	5,4	0	1	0
2,69	7,9	0	1	0
1,99	3,2	0	1	0
2,01	11	0	1	0
4,22	8,9	0	1	0
2,76	12,4	0	1	0
2,55	7,8	0	1	0
1,89	9,6	0	1	0
2,34	4,3	0	1	0
2,83	11,1	0	1	0
1,81	6,8	0	1	0
9,23	53	0	0	1
8,6	17,3	0	0	1

Rate	ACPT	FAI	PA	MA
8,21	27,3	0	0	1
2,93	18	0	0	1
7,48	30,2	0	0	1
2,57	10,3	0	0	1
5,77	18,2	0	0	1
2,9	12,3	0	0	1
2,97	7,1	0	0	1
1,84	14	0	0	1
3,78	11,3	0	0	1
2,76	16,3	0	0	1
4,27	9,6	0	0	1
3,05	9	0	0	0
4,12	10,4	0	0	0

(a) Obtenha um gráfico de setores para o tipo de estrada.

(b) Faça um gráfico de dispersão para RATE versus ACPT.

(c) Calcule o coeficiente de correlação entre essas duas variáveis.
(d) Obtenha a equação da reta de regressão RATE = a + b ACPT.

Pfessora, eu tenho uma pergunta...
Tem uma coisa nessa prova de Matemática que eu não entendo...
Todos esses números!

P7.21) Apartamentos de dois quartos em Botafogo

A tabela a seguir contem apartamentos de 2 quartos no bairro de Botafogo, Rio de Janeiro e oferecidos para venda no site www.zap.com.br/imoveis em 21/10/2010 com preço em milhares de reais:

Tabela – Área (m^2) e Preço (10^3 reais) de imóveis em 2010

Área (m^2)	**Preço**	**Área (m^2)**	**Preço**	**Área (m^2)**	**Preço**	**Área (m^2)**	**Preço**
69	400	84	480	68	435	85	550
92	400	59	520	70	439	97	557
68	410	55	520	75	440	85	560
69	416	85	525	70	440	85	570
68	420	76	530	75	450	110	580
75	420	80	530	60	450	72	589
75	420	75	530	70	450	80	590
79	430	60	540	77	455	80	600
64	430	74	550	68	455	80	600
70	430	85	550	68	460	75	600
66	430	76	550	75	460	100	600

(a) Construa um gráfico de dispersão para esses dados
(b) Calcule o coeficiente de correlação entre Área e Preço.
(c) Ajuste a esses dados uma reta de regressão expressando o Preço como função linear da Área.
(d) Extraia as conclusões cabíveis.

P7.22) Densidade e/ou Viscosidade como preditoras do BMCI

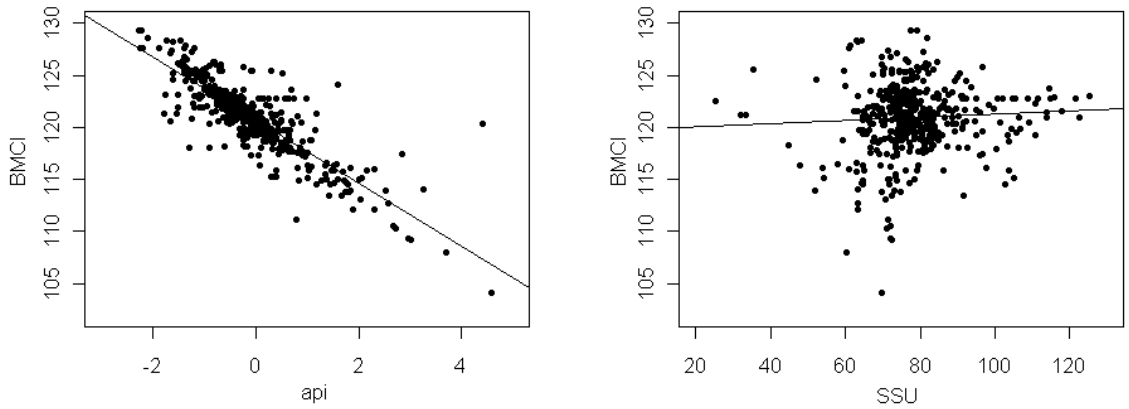
O Raro é um dos produtos do processo de Craqueamento Catalítico do petróleo. O BMCI é uma medida de aromaticidade que, em princípio, depende tanto da Densidade como da Viscosidade do Raro e é aqui a principal variável de interesse.

Os resultados dos ajustes por mínimos quadrados são:

BMCI = 120,6 – 3,03 api e BMCI =119,7 + 0,0158 SSU

As correlações amostrais são corr(SSU, BMCI) = 0,0561 e corr(api, BMCI)= – 0,848

Apresentamos nos gráficos a seguir o diagrama de dispersão da densidade (api) versus o BMCI e também o diagrama de dispersão da viscosidade (SSU) versus o BMCI, em cada um foi traçado a reta de regressão sendo a variável a ser explicada a BMCI.



O que os resultados obtidos evidenciam no que se refere à força da relação entre BMCI e densidade (api)? E entre BMCI e viscosidade (SSU)?

P7.23) Difusividade Térmica

Os dados a seguir mostram como a Difusividade Térmica de uma fibra varia em função da temperatura. Quatro situações diferentes são consideradas:

Carb sem = Fibra de carbono sem envelhecimento
Vidro sem = Fibra de vidro sem envelhecimento
Carb com = Fibra de carbono com envelhecimento
Vidro com = Fibra de vidro com envelhecimento

Tabela – Temperatura (^oC) e Difusividade Térmica (mm^2/s)

Carb sem		Vidro sem		Carb com		Vidro com	
Temp (ºC)	Dif Térm (mm²/s)	Temp (ºC)	Dif Térm (mm²/s)	Temp (ºC)	Dif Térm (mm²/s)	Temp (ºC)	Dif Térm (mm²/s)
30,2	0,459	30,3	0,304	29,9	0,422	30,1	0,339
50,3	0,445	50,4	0,297	50,2	0,408	50,3	0,341
60,1	0,442	60,2	0,296	60,1	0,404	60,3	0,337
120,2	0,414	119,9	0,283	120,0	0,365	120,1	0,290
130,1	0,414	130,2	0,281	130,0	0,362	130,1	0,288
170,0	0,371	170,1	0,281	169,9	0,334	170,0	0,276
180,0	0,366	180,0	0,284	179,8	0,326	180,0	0,273
27,8	0,449	30,1	0,336	30,1	0,436	29,9	0,299
90,0	0,497	90,0	0,317	90,2	0,392	90,0	0,268
110,0	0,436	110,0	0,310	110,1	0,383	110,0	0,260
114,8	0,436	115,0	0,313	115,4	0,380	115,1	0,267
139,9	0,425	139,9	0,306	140,1	0,370	140,0	0,254
179,9	0,379	179,8	0,299	180,0	0,336	179,9	0,247
190,0	0,372	189,9	0,296	190,0	0,328	189,9	0,244
209,9	0,367	209,9	0,285	210,0	0,330	209,9	0,236

(a) Para cada uma das 4 situações aqui consideradas: Carbono sem, Vidro sem, Carbono com, Vidro com, ajuste aos dados uma reta de regressão

$$\text{Difusividade Térmica} = \beta_0 + \beta_1 \text{Temperatura}$$

(b) No caso da fibra de carbono, o decrescimento da Difusividade Térmica em função da temperatura é mais rápido com ou sem envelhecimento? Por que?

(c) No caso da fibra de vidro, o decrescimento da Difusividade Térmica em função da temperatura é mais rápido com ou sem envelhecimento? Por que?

(d) Comparando as duas situações onde não há envelhecimento, o decrescimento da Difusividade Térmica em função da temperatura é mais rápido no caso da fibra de carbono ou no caso da fibra de vidro? Por que?

(e) Comparando as duas situações onde há envelhecimento, o decrescimento da Difusividade Térmica em função da temperatura é mais rápido no caso da fibra de carbono ou no caso da fibra de vidro? Por que?

Obs.: Para facilitar os cálculos, são fornecidos:

	Σx	Σy	Σx^2	Σy^2	Σxy
Carbono sem	1803,2	6,272	265974,1	2,644600	725,962
Vidro sem	1805,7	4,488	266095,4	1,346240	532,775
Carbono com	1805,8	5,576	266229,3	2,090994	642,041
Vidro com	1805,6	4,219	266148,3	1,203391	484,294

onde: **x** = Temperatura **y** = Difusividade Térmica

P7.24) Duas retas ou uma só?

O conjunto de dados a seguir se refere a um experimento para avaliação catalítica em uma refinaria. Ele contem n = 135 observações e p = 3 variáveis, a saber:

T1 = Temperatura (em graus Celsius)
T2 = Temperatura (em graus Farenheit)
MAD = Massa de água deslocada

Tabela – Resultados de experimento para avaliação catalítica

T1	T2	MAD	T1	T2	MAD	T1	T2	MAD
-7,7	18,14	1958	-13,1	8,42	1928	-15,2	4,64	2325
-6,5	20,3	1946	-7,1	19,22	1820	-15,1	4,82	2257
-9,4	15,08	1937	-11,5	11,3	1919	-10,8	12,56	2313
-11,3	11,66	1923	-15	5	2316	-15,1	4,82	2305
-9,6	14,72	1906	-15,1	4,82	2341	-14	6,8	2325
-8,2	17,24	1905	-15,1	4,82	2331	-15,5	4,1	2396
-9,9	14,18	1921	-15	5	2327	-15,7	3,74	2562
-8,8	16,16	1830	-15	5	2289	-12,9	8,78	2340
-13,1	8,42	1928	-15,2	4,64	2302	-13,2	8,24	2298

(a) Fazendo y = T2 e x = T1:
   i. Obtenha um gráfico de dispersão de x contra y.
   ii. Calcule o coeficiente de correlação r_{xy} entre x e y.
   iii. Ajuste por mínimos quadrados a reta de regressão y = a + bx.
   iv. Ajuste por mínimos quadrados a reta de regressão x = c + dy.
   v. Compare os valores obtidos de b e 1/d.
   vi. Compare os valores obtidos de a e $-\frac{c}{d}$.

(b) Fazendo y = MAD e x = T1, repita a mesma seqüência de passos.

(c) Tomando agora como ponto de partida um conjunto qualquer de n pares de observações (x_i,y_i), podem ser ajustadas aos dados duas retas de regressão:

- uma considerando y como a variável a explicar e x como a variável explicativa, como é usual, ou seja, y = a + bx;
- a outra invertendo os papéis, ou seja, considerando x como a variável a explicar e y como a variável explicativa, ou seja, x = c + dy.

Se explicitarmos y como função de x nessa segunda equação teremos

$$y = -\frac{c}{d} + \frac{1}{d}x.$$

Então, se essas duas retas forem traçadas no mesmo gráfico, em geral elas só coincidirão entre si se tivermos $b = \frac{1}{d}$ e $a = -\frac{c}{d}$. Evidentemente, em geral isso não acontecerá. Para que valores de r_{xy} podemos garantir que ao invés de duas retas distintas teremos uma só? Como ficaria o aspecto do gráfico de dispersão neste caso? Por que?