

CursoR: Problemas

1 Introdução

1. Se possível, instale o R no seu computador pessoal.

2 Dados: Leitura e Manipulação

1. Suponha que você anote a quilometragem do seu veículo toda vez que você enche o tanque. Suas últimas anotações foram : 65311 65624 65908 66219 66499 66821 67145 67447. Entre esses números no R e depois use a função `diff` nos dados. O que representa o resultado? Qual a média, mínimo e máximo de quilômetros rodados entre os abastecimentos?
2. As contas do seu celular variam de mês a mês. Suponha que no último ano os valores sejam: 46 33 39 37 46 30 48 32 49 35 30 48. Entre esses dados num objeto chamado `contas`. Quanto você gastou no ano? Qual o gasto mínimo e máximo? Em quantos meses a conta foi maior do que R\$40? Qual a proporção de meses em que a conta superou esse valor?
(Dica: Use a função `length` para saber o número de elementos de um vetor.)

3. Com os seguintes dados

```
x <- c(1, 8, 2, 6, 3, 8, 5, 5, 5, 5)
```

Calcule:

- (a) $(x_1 + x_2 + \dots + x_{10})/10$ (use o comando `sum`)
- (b) $\log_{10}(x_i)$ para todo i (o default da função `log` é base e . Use as opções da função para calcular na base 10.
- (c) Ache a diferença entre o menor e o maior número. Você pode usar as funções `min` e `max` ou achar uma função que calcule esse valor.

3 Análise de Dados

1. Escreva o código para produzir o gráfico da Figura 3. Use o seguinte código como base. Utilize função `rlm` para obter os parâmetros da regressão robusta

```
## ler os dados
data.al <- read.table(file="alagoas2.csv",header=TRUE,sep=",")
## inserir variavel com a proporção dos votos válidos do candidato Lula
```

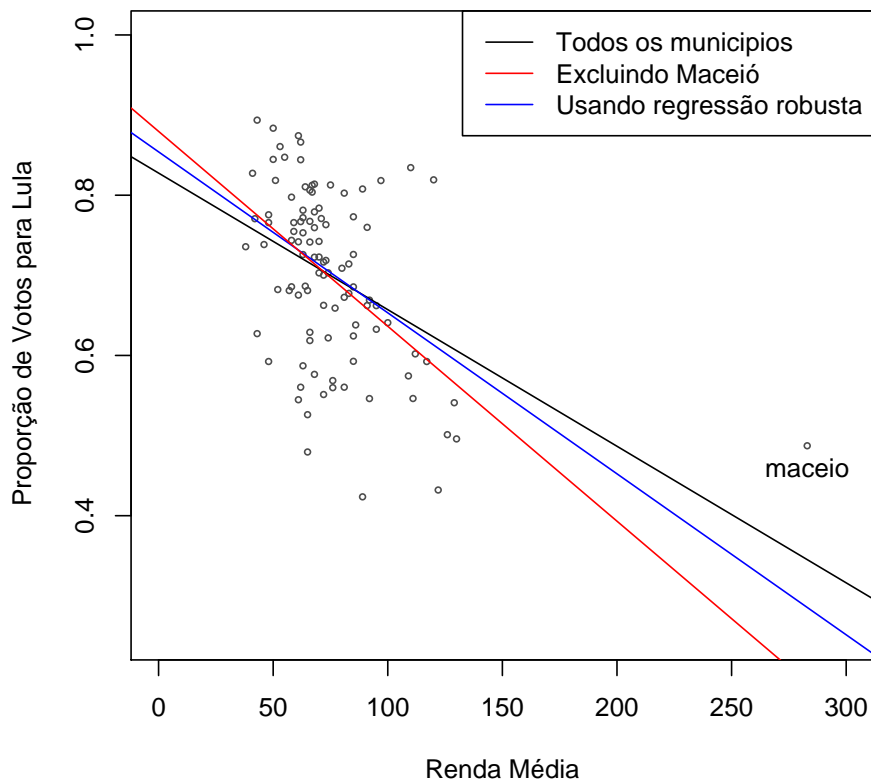


Figure 1: Regressão Robusta

```

data.al$p.lula <- with(data.al,
                      lula/(gerald.alckmin+lula))
## plotar os pontos
plot(p.lula ~ renda,ylim=c(0.25,1),xlim=c(0,300),cex=.5,
     data=data.al, col="gray30", xlab="Renda Média",
     ylab="Proporção de Votos para Lula")
## plotar maceio
with(subset(data.al,renda>200),text(renda,p.lula,
                                   labels=municipio,pos=1))
## plotar regressão linear
abline(lm(p.lula~renda,data=data.al))
abline(lm(p.lula~renda,data=subset(data.al,municipio!="maceio")),
       col="red")
legend("topright", legend=c("Todos os municípios",
                           "Excluindo Maceió"),
      lty=c(1,1), col=c("black","red"))

```

2. A base de dados `mammals` (entre `data(mammals)` para coloca-la na memória) contem o peso do corpo versus o peso dos cérebros de 62 animais. Calcule o coeficiente de correlação (usando a função `cor`) de Pearson e Spearman. Há alguma diferença? Construa um gráfico com as duas medidas. Experimente também plotar na escala logarítmica.
3. Use a função `hist` para plotar o histograma das variáveis contínuas da base `data.a1`. Varie o número de barras (opção `breaks`) e veja se aspectos dos dados são realçados ou escondidos. Use a função `plot(density(x))` (com `x` sendo a variável de interesse) para plotar a densidade kernel dos dados. Qual gráfico você acha mais elucidativo?

4 Classes e Objetos

1. Os nomes do objeto `data.a1` são:

```
> names(data.a1)
 [1] "SP_ID"           "municipio"      "AREA_1997"
 [4] "ELEITORES"      "x"              "y"
 [7] "ESPER_2000"     "ALFAB_2000"    "FREQU_2000"
[10] "RENDA_2000"     "median.leg"    "voteRight"
[13] "geraldo.alckmin" "lula"
```

Dê novos nomes para as colunas 7 a 10.

5 Funções

1. A variância é uma medida de dispersão. Uma das fórmulas para calcular a variância de uma população a partir de uma amostra é:

$$s^2 = \frac{1}{N-1} \sum_i^N (y_i - \bar{y})^2 \quad (1)$$

Onde \bar{y} é a média de y .

- (a) Crie uma função para calcular a variância.
- (b) Crie outra função para calcular o desvio padrão s .
- (c) Crie um vetor com 30 números simulados de uma distribuição normal com média 1 e desvio padrão 2 do vetor:
`x <- rnorm(30,1,2).`
- (d) Calcule a variância e desvio padrão do vetor criado. Compare o resultado com as funções do R `var` e `sd`.

6 Simulando

1. Gere 10 números aleatórios de uma distribuição normal com média 5 e desvio padrão 5. Quantos são menores que 0?
2. Role um dado (usando R!) 10, 100 e 1000 vezes. Construa um gráfico com as três distribuições no mesmo gráfico (com cores diferentes). Esse código pode te ajudar a começar.

```
plot(0,0,ylim=c(0,1),xlim=c(1,6),xlab="",ylab="",type="n")
lines(prop.table(table(rolaDado(10))), col="black")
legend("topright",legend=c("n=10"),lty=1,col="black")
```

7 Estatística Básica

1. Crie 15 números aleatórios de uma distribuição normal com média 10 e desvio padrão 5. Faça um teste t com intervalo de confiança de 95%. O intervalo de confiança encobre o valor verdadeiro?
2. O vetor x tem o tempo de sobrevivência de 20 ratos expostos à radiação.

```
x <- c(152, 152, 115, 109, 137, 88, 94, 77, 160, 165, 125, 40, 128,
       123, 136, 101, 62, 153, 83, 69)
```

Faça um teste de t nos dados. Depois, faça um histograma de x (usando a função `hist`). A distribuição parece normal? Por que isso é importante?

3. Estes dados são as notas do dever de casa em escolas públicas e privadas. Se você tem acesso à internet, leia os dados com:

```
dc <- read.csv("http://cursor.cluelessresearch.com/dc.csv")
```

Privada	Pública
21.3	15.3
16.8	17.4
8.5	12.3
12.6	10.7
15.8	16.4
19.3	11.3
18.5	17.6
24.6	13.9
18.3	20.2
12.9	16.8
15.7	23.6
18.4	14.2
18.7	5.7
22.6	18.8
20.5	9.4

Faça um gráfico das duas distribuições. Você pode usar a função `hist(x)`, `plot(density(x))`, onde `x` é o vetor da variável representando escolas públicas ou privadas. Tente também um `boxplot` da data frame `boxplot(dc)`.

Agora compare as médias das duas variáveis. Elas são significativamente diferentes?

4. Numa pesquisa antes da eleição de 2006 apareceu seguinte relação entre classe e voto.

```
lula2
classe 0 1
ab 12 14
c 26 54
de 18 37
```

A classe é independente do voto para lula?