Estatística Descritiva no Software R

Nome: Rafael Lima de Souza
RU: 1237272
Polo: Porto Alegre - Centro Histórico
e-mail: ls.rafael@icloud.com

26 January 2019

Introdução Sobre o R

O R é uma linguagem de programação estatística, trata-se de uma linguagem de programação especializada em computação de dados. Algumas de suas principais características são o seu caráter gratuito e sua disponibilidade para uma gama bastante variada de sistemas operacionais. É também altamente expansível com o uso de pacotes que são bibliotecas para áreas de estudos ou funções específicas, onde é possível executar cálculos complexos e ainda gerar uma infinidade de gráficos.

Ainda é possível contar com um ambiente de desenvolvimento integrado, o RStudio, onde é possível gerar relatórios e apresentações com alto nível de qualidade. Inclusive, este trabalho foi totalmente desenvolvido com o editor R Markdown.

O código fonte desde trabalho podem ser encontrado no repositório Estatística Descritiva com R no GitHub e o trabalho destaca os links que podem ser acessados para o leitor ver os detalhes dos termos usados neste trabalho.

Medição dos Dados

Para os cálculos deste relatório, usamos as funções nativas do R, que foram desenvolvidas préviamente pelos desenvolvedores. Abaixo analisaremos as medidas de tendência central conhecidas por Mediana, Média e as medidas de dispersão chamadas de Variância e o Desvio Padrão.

Mediana

A mediana de um conjunto de dados é o valor que ocupa a posição central, desde que estejam colocados em ordem crescente ou decrescente, ou seja, em um rol.

A fórmula da Mediana é: $Md = \frac{n+1}{2}$, desta forma é possível encontrar a posição da mediana.

No R, a função que calcula a Mediana é chamada de median, abaixo mostra o nome da coluna calculada após o símbolo #, depois o comando usado para o cálculo e na sequencia ao lado do símbolo ## mostra o resultado calculado.

Média

A média aritimética simples, ou simplesmente média, nada mais é que a soma dos resultados obtidos dividida pela quantidade dos resultados.

A fórmula da Média é: $\overline{X} = \frac{\sum X_i}{n}$, em que i varia de 1 até n.

A função utilizada para calcular a Média no R, é chamada de mean.

Variância

Como a soma dos desvios em relação à média é sempre igual a zero, é possível evitar este fato elevando cada desvio ao quadrado, pois sabemos que o quadrado de qualquer número real é sempre positivo.

Tabela 1: Tabela 1 - Número de Acidentes de Trânsito na Metrópole A

	Sem Vítimas	Com Ferimentos Graves	Com Ferimentos Leves	Com Mortos
Domingo	70	25	75	20
Segunda	42	15	54	5
Terça	45	22	32	6
Quarta	42	23	30	5
Quinta	50	24	42	8
Sexta	61	36	50	15
Sábado	72	40	52	18

Fonte:
Autor do portifolio.

Para calcular a variância de uma população, usamos a seguinte fórmula: $S^2 = \frac{\sum[(X_i - \overline{X})^2 . f_i]}{N}$

Quando se trata de calcular uma amostra, a seguinte fórmula deve ser aplicada: $S^2 = \frac{\sum[(X_i - \overline{X})^2 . f_i]}{N-1}$

A função utilizada para calcular a Variância no R, é chamada de var.

Desvio Padrão

O desvio padrão o quanto os dados estão dispersos em torno da média. Um desvio padrão alto, indica que os dados estão espalhados por uma ampla gama de valores. O desvio padrão, nada mais é do que a raíz quadrada da variância.

Sua fórmula para medir a variabilidade dos dados em uma população, é: $S^2 = \sqrt{\frac{\sum[(X_i - \overline{X})^2 . f_i]}{N}}$

E para fazermos a medição em uma amostra, usamos: $S^2 = \sqrt{\frac{\sum[(X_i - \overline{X})^2 . f_i]}{N-1}}$

A função utilizada para calcular o Desvio Padrão no R, é chamada de sd.

Experimento 1 - Acidentes de Trânsito

A tabela abaixo mostra o número de acidentes de trânsito durante uma semana em uma grande metrópole.

```
# Criação de um objeto do tipo matriz(7x4), para armazenar os dados dos acidentes.
mt_acidentes <- matrix(c(70,42,45,42,50,61,72,25,15,22,23,24,36,40,
                         75,54,32,30,42,50,52,20,5 ,6 ,5 ,8 ,15,18),
                       nrow = 7,
                       ncol = 4)

# Nomeando os nomes das variáveis e observações.
row.names(mt_acidentes) <- c('Domingo','Segunda','Terça','Quarta','Quinta','Sexta','Sábado')
colnames(mt_acidentes) <- c('Sem Vítimas','Com Ferimentos Graves','Com Ferimentos Leves','Com Mortos')
```

Resultado do Experimento 1

Como resultado dos comandos acima, obtemos a seguinte tabela com os dados sobre os acidentes ocorridos durante uma semana.

Análise dos Dados de Acidentes

Abaixo segue o código usado para calcular as medidas dos dados dos acidentes.

Tabela 2: Resultados das Medidas Centrais e Dispersão

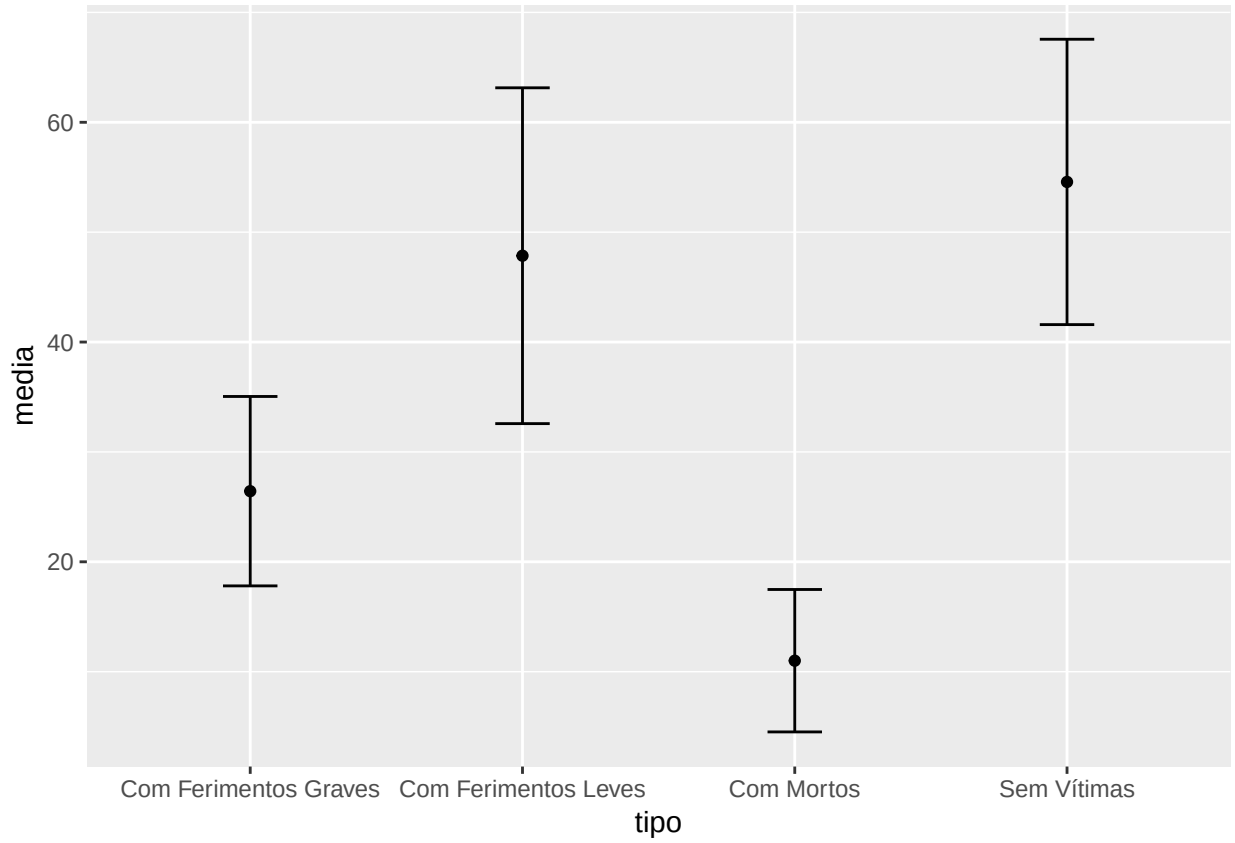
tipo	media	des.padr	var	mediana
Sem Vítimas	54.57143	12.985340	168.61905	50
Com Ferimentos Graves	26.42857	8.618916	74.28571	24
Com Ferimentos Leves	47.85714	15.279928	233.47619	50
Com Mortos	11.00000	6.480741	42.00000	8

```
# Comando usado para calcular as medidas e armazenar os dados em um dataframe
df_ferimentos_graves <-  data.frame(tipo    = c('Sem Vítimas','Com Ferimentos Graves','Com Ferimentos L
                                    media   = apply(mt_acidentes,2, mean),
                                    des.padr  = apply(mt_acidentes,2, sd),
                                    var     = apply(mt_acidentes,2, var),
                                    mediana = apply(mt_acidentes,2, median))
# Remove os nomes das linhas
row.names(df_ferimentos_graves) <- NULL
```

Podemos constatar abaixo, que a média dos feridos gravemente foi de 26 e o desvio padrão foi 8. Este desvio padrão indica que existe boa aderência da variabilidade dos dados com a média.

No seguinte gráfico podemos analisar visualmente os dados de cada tipo de acidente e constatar que o desvio padrão possui relativa aderência à média.

```
# Biblioteca GGPLOT2 para visualização de dados
ggplot(df_ferimentos_graves, aes(x=tipo, y=media)) +
  geom_errorbar(aes(ymin=media-des.padr, ymax=media+des.padr), width=.2) +
  geom_line() +
  geom_point()
```



Experimento 2 - Gráficos Bioma Pampa

```r
# Criação de um objeto do tipo matriz(7x4), para armazenar os dados dos acidentes.
mt_bioma <- matrix(c(176496, 6210, 3340, 1607, 122682, 20974, 7658, 14025,
                         103835, 0   , 0   , 428 , 10980 , 2033 , 394 , 0    ,
                         58636 , 6210, 3340, 1179, 21702 , 18940, 7264, 0   ),
                       nrow = 8,
                       ncol = 3)

# Nomeando os nomes das variáveis e observações.
row.names(mt_bioma) <- c('Área Total',
                             'Floresta Estacional Semidecidual',
                             'Floresta Estacional Decidual',
                             'Savana Estépica',
                             'Estepe',
                             'Formações pioneiras',
                             'Contatos entre tipos de vegetação',
                             'Superfície com água')
colnames(mt_bioma) <- c('Total do bioma (Km²)','Área remanescente (Km²)','Área antroponizada (Km²)')
```

Apresentação dos Gráficos das Áreas Antropizadas

Abaixo seguem os gráficos relativos a tabela apresentada anteriormente.

Percetuais das Áreas Remanescentes e Antroponizadas

Tabela 3: Tabela 2 - Áreas Remanescentes e Áreas Antropizadas, no Bioma Pampa, Segundo os Tipos de Vegetação

	Total do bioma (Km²)	Área remanescente (Km²)	Área antroponizada (Km²)
Área Total	176496	103835	58636
Floresta Estacional Semidecidual	6210	0	6210
Floresta Estacional Decidual	3340	0	3340
Savana Estépica	1607	428	1179
Estepe	122682	10980	21702
Formações pioneiras	20974	2033	18940
Contatos entre tipos de vegetação	7658	394	7264
Superfície com água	14025	0	0

Fonte:
FIGUEIREDO, 2016

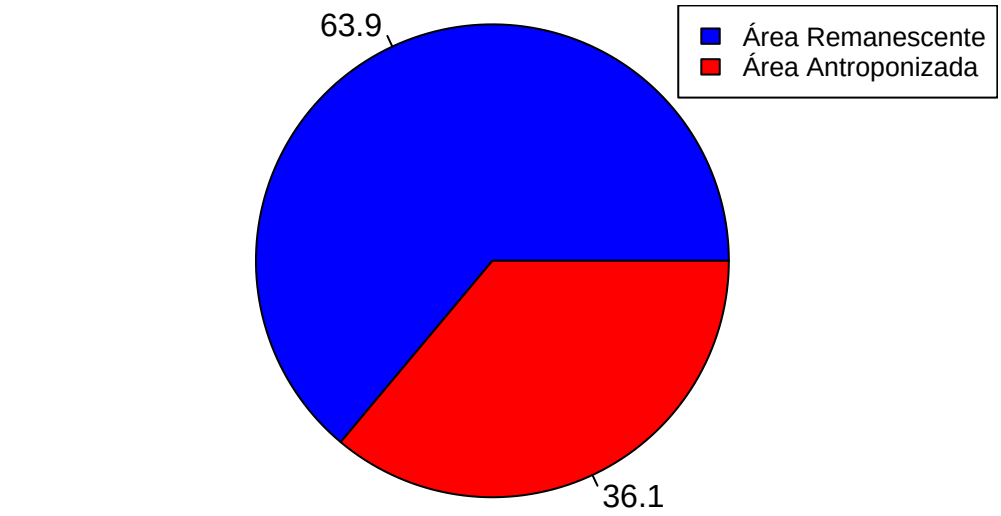
No gráfico que segue, podemos analisar o percentual dos totais do bioma pampa.

```
# Nomando os rótulos do gráfico
labels <- c('Área Remanescente','Área Antroponizada')

# Cálculo do percentual dos setores
perc <- round(100*mt_bioma[1,2:3]/sum(mt_bioma[1,2:3]), 1)

# Plot do gráfico
pie(mt_bioma[1,2:3],
    labels = perc,
    main = 'Percentual de Áreas Remanescentes e Antroponizadas',
    radius = 1,
    col = c('blue','red'))

# Ajuste das legendas
legend("topright", c('Área Remanescente','Área Antroponizada'), cex = 0.8,
       fill = c('blue','red'))
```

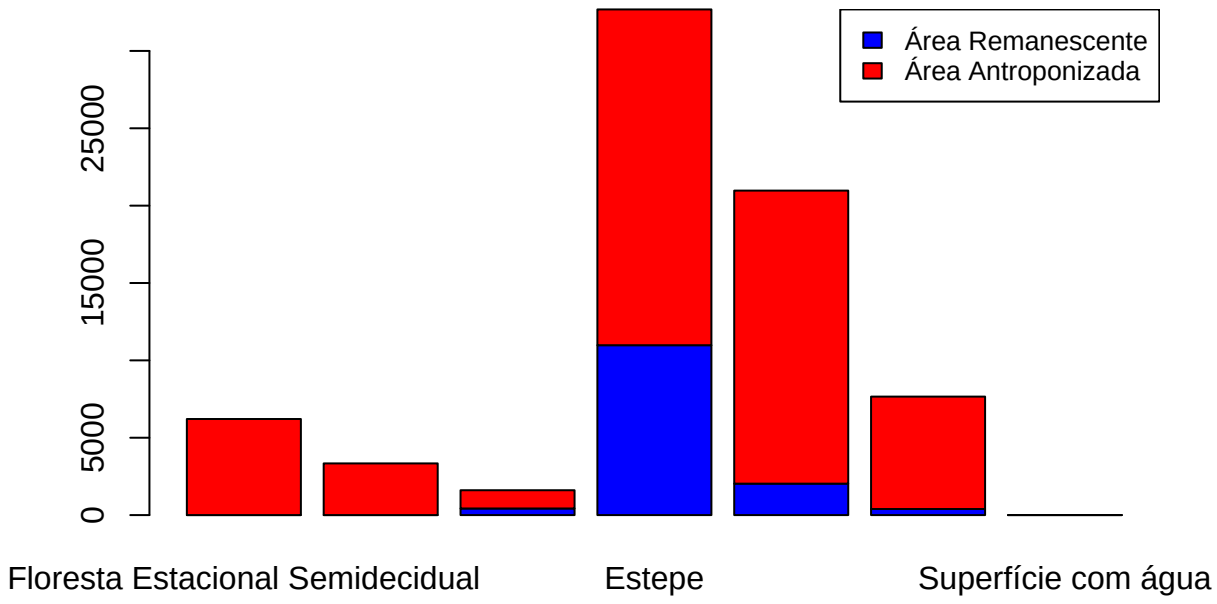


Quantitativos das Áreas Remanescentes e Antroponizadas

No gráfico seguinte podemos analisar os valores quantitativos das áreas fo bioma pampa.

```
# Criaçãodo Gráfico de colunas
barplot(t(mt_bioma[2:8,2:3]), col = c('blue','red'))

# Ajustes das legendas
legend("topright", c('Área Remanescente','Área Antroponizada'), cex = 0.8,
       fill = c('blue','red'))
```


Área Remanescente
Área Antroponizada
0
5000
15000
25000
Floresta Estacional Semidecidual
Estepe
Superfície com água