

# CIÊNCIA DE DADOS COM LINGUAGEM R

Richard Guilherme dos Santos



# Contents

<b>1</b>	<b>Introdução</b>	<b>5</b>
<b>2</b>	<b>Introdução a Probabilidade</b>	<b>7</b>
<b>3</b>	<b>Introdução ao R</b>	<b>9</b>
<b>4</b>	<b>Medidas Descritivas</b>	<b>11</b>
4.1	Tipos de Variáveis . . . . .	11
4.2	Medidas de Posição . . . . .	12
4.3	Medidas de Dispersão . . . . .	13
4.4	Quantis Empíricos . . . . .	15
4.5	Box Plot . . . . .	15
4.6	Transformações . . . . .	16
4.7	Lab 01 - Conjunto de dados Iris . . . . .	16
<b>5</b>	<b>Tipos de Distribuições Discretas</b>	<b>21</b>
5.1	Valor Médio de uma Variável Aleatória . . . . .	21
<b>6</b>	<b>Tipos de Distribuições Contínuas</b>	<b>23</b>
<b>7</b>	<b>Introdução as bibliotecas do R</b>	<b>25</b>
7.1	Dplyr . . . . .	25
7.2	Tidyr . . . . .	25
7.3	GGPlot2 . . . . .	25
<b>8</b>	<b>Regressão Linear</b>	<b>27</b>

Bem vindos ao meu livro!



# Chapter 1

## Introdução

Este livro tem como objetivo servir como guia para as aulas do curso Ciência de Dados com R. Nele apresentaremos os conceitos de:

1. **Estatística Básica:** Nesta parte do curso abordaremos conceitos de estatística como variáveis, tipos de distribuições discretas e contínuas, medidas descritivas e distribuição normal.
2. **Manipulação de dados no R:** Neste tópico serão abordados as principais formas de manipulação de dados utilizando a linguagem R, com ênfase nas bibliotecas dplyr e tidyr. Além disso, abordaremos a criação de gráficos pelo pacote ggplot2.
3. **Modelos de Regressão Linear:** Parte final do curso, onde o aluno aprenderá sobre diagrama de dispersão, coeficiente de correlação linear, regressão linear simples, múltipla e regressão logística, ganhando a capacidade de começar a criar modelos utilizando a linguagem R.



## Chapter 2

# Introdução a Probabilidade





## Chapter 3

# Introdução ao R

Aqui introduziremos alguns comandos da linguagem R, onde utilizamos funções para realizar operações que vão desde leitura e manipulação de dados a operações matemáticas.

Começemos criando um vetor de números:

```
x <- c(1,3,2,5)
# x = c(1,3,2,5) # Também podemos utilizar "=" para atribuir variáveis
x

## [1] 1 3 2 5
```

O comando acima combina os números 1,3,2 e 5 em um vetor de números e os salva em um objeto denominado x. Escrevemos x para recebermos os atributos do vetor.

A partir disto podemos utilizar outras funções para calcularmos informações destes atributos, como o tamanho de um vetor:

```
length(x)
```

```
## [1] 4
```

ou sua média:

```
mean(x)
```

```
## [1] 2.75
```

Há outros tipos de objetos que podem ser criados quando trabalhamos com R. Os mais importantes para manipulação de dados são as matrizes:

```
mat = matrix(data = c(1,2,3,4), nrow = 2, ncol = 2,
              byrow = TRUE)
mat
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

Funções aceitam os mais diversos tipos de argumentos. Para termos uma ideia de quais utilizarmos e seus respectivos atributos devemos fazer consultas em suas bibliotecas:

```
help(matrix)
```

Além disso, para armazenamento de dados temos os `data.frames`, tabelas que aceitam dados de tipos distintos:

```
nomes = c('Carol', 'Alfredo', 'Godoberto')
idade = c(18, 23, 19)
peso = c(69, 75, 80)
altura = c(1.70, 1.80, 1.85)
ICM = peso/altura^2
df = data.frame(nomes, idade, peso, altura, ICM)
df
```

```
##      nomes idade peso altura      ICM
## 1    Carol    18   69   1.70 23.87543
## 2  Alfredo    23   75   1.80 23.14815
## 3 Godoberto    19   80   1.85 23.37473
```

## Chapter 4

# Medidas Descritivas

**Importante:** A partir deste capítulo utilizaremos a função `kable` do pacote `knitr` para visualização de conjuntos de dados. Na prática isto não é necessário, apenas o realizamos para efeitos de visualização.

### 4.1 Tipos de Variáveis

Antes de analisarmos conjuntos de dados, é necessário termos um conhecimento sobre tipos de variáveis. Para isto, consideremos a seguinte tabela:

```
nome = c('Djoko', 'Wilson', 'Jiraiya', 'Leon', 'Nilce')
est_civil = c('Solteiro', 'Casado', 'Solteiro', 'Casado', 'Casado')
escolaridade = c('Pós-graduação',
                 'Ensino médio completo',
                 'Ensino médio completo',
                 'Pós-graduação',
                 'Superior completo')
n_filhos = c(0, 0, 1, 0, 0)
salario = c(4500, 3000, 1500, 5000, 5500)
idade = c(29, 33, 33, 39, 32)
df = data.frame(nome, est_civil, escolaridade, n_filhos, salario, idade)
kable(df, align = 'c',
      caption = 'Dados sobre Youtubers.') # Melhor visualização dos dados para este PDF
```

Variáveis como sexo, escolaridade e estado civil apresentam realizações de uma qualidade ou atributo do indivíduo pesquisado, enquanto outras como número de filhos, salário e idade apresentam números como resultados de uma contagem ou mensuração. Chamamos as do primeiro tipo de **qualitativas** e as do segundo de **quantitativas**

Table 4.1: Dados sobre Youtubers.

nome	est_civil	escolaridade	n_filhos	salario	idade
Djoko	Solteiro	Pós-graduação	0	4500	29
Wilson	Casado	Ensino médio completo	0	3000	33
Jiraiya	Solteiro	Ensino médio completo	1	1500	33
Leon	Casado	Pós-graduação	0	5000	39
Nilce	Casado	Superior completo	0	5500	32

Cada uma das duas ainda pode ser dividida em dois tipos:

- **Variável qualitativa nominal:** atributos não apresentam uma ordem lógica;
- **Variável qualitativa ordinal:** atributos apresentam uma ordem lógica bem estabelecida;
- **Variável quantitativa discreta:** dados de contagem, assumem apenas valores inteiros;
- **Variável quantitativa contínua:** dados que podem assumir qualquer tipo de valor.



Muitas vezes queremos resumir estes dados, apresentando um ou mais valores que sejam representativos da série toda. Neste contexto entram às **medidas de posição e dispersão**.

## 4.2 Medidas de Posição

Usualmente utilizamos uma das seguintes medidas de posição (ou localização): **média, mediana ou moda**. Vamos as suas definições:

- **Moda:** valor mais frequente do conjunto de valores observados.
- **Mediana:** valor que ocupa a posição central das observações quando estas estão ordenadas em ordem crescente.

- Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais.

Na tabela 4.1 temos a seguinte mediana para uma coluna específica:

```
median(df$salario)
```

```
## [1] 4500
```

Para todas as colunas:

```
# Aplicaremos a função median para todas as colunas:  
apply(df, MARGIN = 2, FUN = median)
```

```
##          nome      est_civil  escolaridade      n_filhos      salario  
##      "Leon"      "Casado" "Pós-graduação"      "0"      "4500"  
##          idade  
##      "33"
```

- **Média:** soma de todos os elementos do conjunto dividida pela quantidade de elementos do conjunto

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Na tabela 4.1 temos a seguinte mediana para uma coluna específica:

```
mean(df$salario)
```

```
## [1] 3900
```

Para todas as colunas:

```
colMeans(df[, c('idade', 'salario')])
```

```
##   idade salario  
##   33.2  3900.0
```

## 4.3 Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representativa de posição esconde toda a informação sobre a variabilidade de um conjunto de observações. Consideremos que cinco alunos realizaram cinco provas, obtendo as seguintes notas:

```
nomes = c('alunoA', 'alunoB', 'alunoC',  
          'alunoD', 'alunoE')  
notas = matrix(c(3,4,5,6,7,  
                1,3,5,7,9,  
                2,5,5,5,8,  
                3,5,5,5,7,
```

```
0,0,5,10,10), nrow = 5, ncol = 5, byrow = T)
df = data.frame(notas, row.names = nomes)
colnames(df) = c('P1', 'P2', 'P3', 'P4', 'P5')
kable(df, align = 'c')
```

	P1	P2	P3	P4	P5
alunoA	3	4	5	6	7
alunoB	1	3	5	7	9
alunoC	2	5	5	5	8
alunoD	3	5	5	5	7
alunoE	0	0	5	10	10

Temos as seguintes médias para os alunos:

```
rowMeans(df)
```

```
## alunoA alunoB alunoC alunoD alunoE
##      5      5      5      5      5
```

Cada aluno possui a mesma média de notas, porém, isto não informa nada sobre a diferença na **variabilidade das notas**. A partir disto, são criadas medidas que sumarizam a **variabilidade** de um conjunto de observações.

Em um primeiro momento podemos considerar a soma da diferença dos dados em relação a média:

$$x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_n - \bar{x}$$

Porém, em qualquer conjunto a soma destes desvios é igual a zero. Uma alternativa é então adicionar o valor absoluto em cada diferença:

$$|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|$$

Apesar de possuir uma boa interpretabilidade, tal métrica não possui propriedades matemáticas interessantes. Assim, trabalharemos com a diferença de quadrados de um conjunto de dados:

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

Como muitas vezes queremos comparar conjuntos de dados de diferentes tamanhos, realizamos a divisão destes valores pelo total de elementos em uma amostra:

$$\text{var}(X) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

A partir disto, definimos **desvio padrão** como sendo a raiz da variância:

$$dp = \sqrt{\text{var}(X)}$$

Realizamos isto pois caso os dados estejam em uma certa unidade de medida, como  $cm$ , ao calcularmos a variância passamos a trabalhar com  $cm^2$ , o que dificulta a interpretabilidade dos resultados.

## 4.4 Quantis Empíricos

Tanto a **média** como o **desvio padrão** podem não ser medidas adequadas para representar um conjunto de dados, uma vez que:

- São afetados por valores extremos;
- Apenas os dois valores não dão informação sobre a simetria ou assimetria da distribuição dos dados

Vimos que a **mediana** define uma divisão dos dados em duas metades. Além disto existem medidas chamadas de **quantil de ordem p** ou **p-quantil** indicado por  $q(p)$  onde  $p$  é uma proporção qualquer,  $0 < p < 1$  tal que 100% das observações sejam menores do que  $q(p)$ .

Abaixo temos alguns dos quantis mais utilizados:

- $q(0.25) = q_1$  : **1° Quartil** ou **25° Percentil**
- $q(0.50) = q_2$  : **2° Quartil**, **Mediana** ou **50° Percentil**
- $q(0.75) = q_3$  : **3° Quartil** ou **75° Percentil**
- $q(0.40)$  : **4° Decil**
- $q(0.95)$  : **95° Percentil**

## 4.5 Box Plot

A informação contida nos quantis pode ser confusa quando estamos observando vários conjuntos de dados. A partir disto traduzimos-a em um diagrama, qual é chamado de **box plot**:

Para construção dessa gráfico definimos por **intervalo interquartil** o valor:

$$IQR(X) = q_3 - q_1$$

Desenhamos um retângulo que parte do primeiro quartil até o terceiro, com a mediana sendo representada por uma linha em seu interior. A partir do retângulo desenhamos uma linha até o maior ponto que não exceda o valor  $q_3 + 1.5 \cdot IQR(X)$ , chamado de limite superior. De modo análogo fazemos o mesmo procedimento até a parte inferior do retângulo considerando o valor

$q_1 + 1.5 \cdot \text{IQR}(X)$  chamado de limite interior. As observações que estiverem acima do limite superior ou abaixo do limite inferior são chamados de pontos exteriores e representadas por asteriscos. Essas observações podem ser chamadas de outliers ou valores atípicos.

O **box plot** dá uma ideia de posição, dispersão, assimetria dos dados.

## 4.6 Transformações

Vários procedimentos estatísticos são baseados na posição que os dados possuem uma distribuição em forma de sino (oriundos de uma distribuição normal), ou que a distribuição seja mais ou menos simétrica.

Se quisermos utilizar tais procedimentos podemos efetuar transformações nas observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. As transformações mais frequentemente utilizadas são:

$$x = \begin{cases} \sqrt{x} \\ \ln(x) \\ \frac{1}{x} \end{cases}$$

para cada transformação obtemos gráficos apropriados para os dados originais e transformados, de modo a escolhermos o valor mais adequado de  $p$ .

## 4.7 Lab 01 - Conjunto de dados Iris

O conjunto de dados Iris é um dos mais utilizados quando introduzimos conceitos de ciência de dados. Este pode ser encontrado em UCI Machine Learning Repository. Tal conjunto consiste de 150 amostras de 4 tipos de espécies de flores distintas contendo os atributos:

- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm

Podemos acessá-lo no R sem nenhum carregamento prévio da seguinte forma:

```
# A função head() mostra os cinco primeiros itens de data.frame:
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1         3.5          1.4          0.2   setosa
## 2           4.9         3.0          1.4          0.2   setosa
## 3           4.7         3.2          1.3          0.2   setosa
## 4           4.6         3.1          1.5          0.2   setosa
```



```
## 5          5.0          3.6          1.4          0.2  setosa
## 6          5.4          3.9          1.7          0.4  setosa
```

Há certas boas práticas ao carregar um conjunto de dados, dentre elas temos:

- Visualização de sua dimensão:

```
# O primeiro valor é a quantidade de linhas do conjunto de dados
# e o segundo a sua quantidade de atributos
dim(iris)
```

```
## [1] 150   5
```

- Visualização do tipo de cada atributo:

```
str(iris) # Structure of an Arbitrary R Object
```

```
## 'data.frame':   150 obs. of  5 variables:
## $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
## $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
## $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
## $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
## $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

- Sumário de seus atributos:

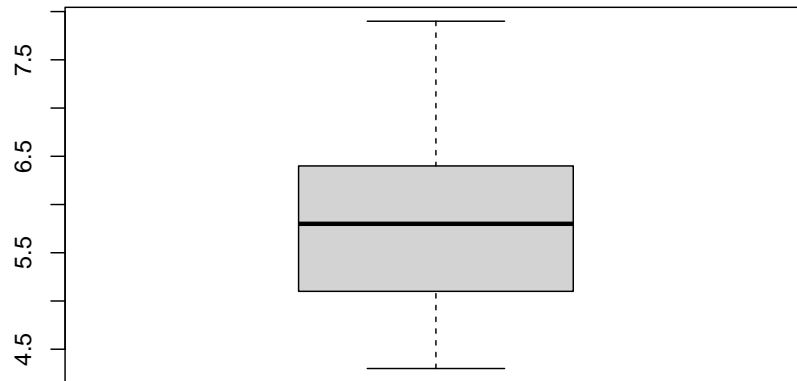
```
summary(iris)
```

```
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
## Min.    :4.300   Min.    :2.000   Min.    :1.000   Min.    :0.100
## 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
## Median :5.800   Median :3.000   Median :4.350   Median :1.300
## Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
## 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
## Max.    :7.900   Max.    :4.400   Max.    :6.900   Max.    :2.500
##      Species
## setosa    :50
## versicolor:50
## virginica :50
##
##
##
```

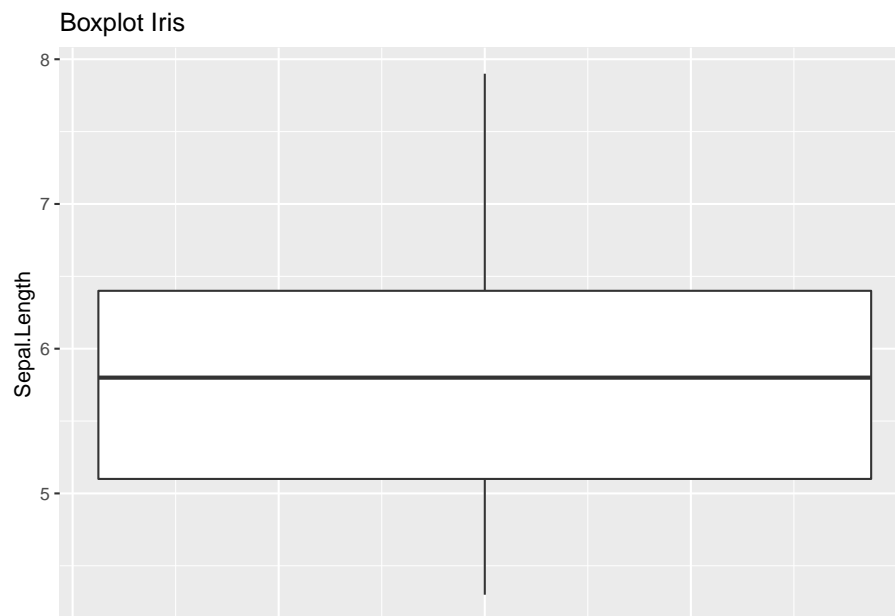
Dessa maneira poderemos contatar valores errôneos no conjunto de dados, distribuições de variáveis categóricas e ter um melhor contato com o conjunto de dados.

Há ainda diversas maneiras de realizarmos visualizações desse conjunto no R, observemos o boxplot da variável Sepa.Length:

```
boxplot(iris$Sepal.Length)
```

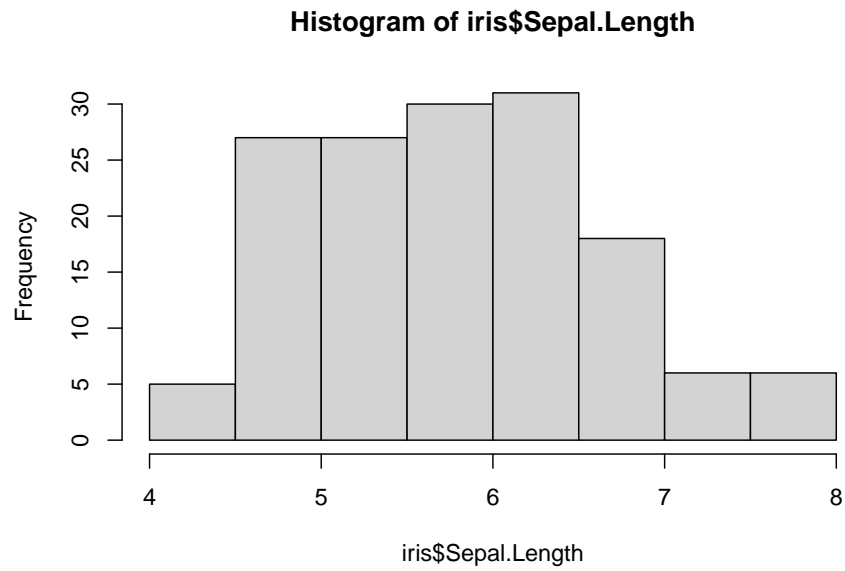


```
ggplot(data = iris, aes(y = Sepal.Length)) +  
  geom_boxplot() +  
  labs(title = 'Boxplot Iris') +  
  theme(axis.ticks.x = element_blank(),  
        axis.text.x = element_blank())
```

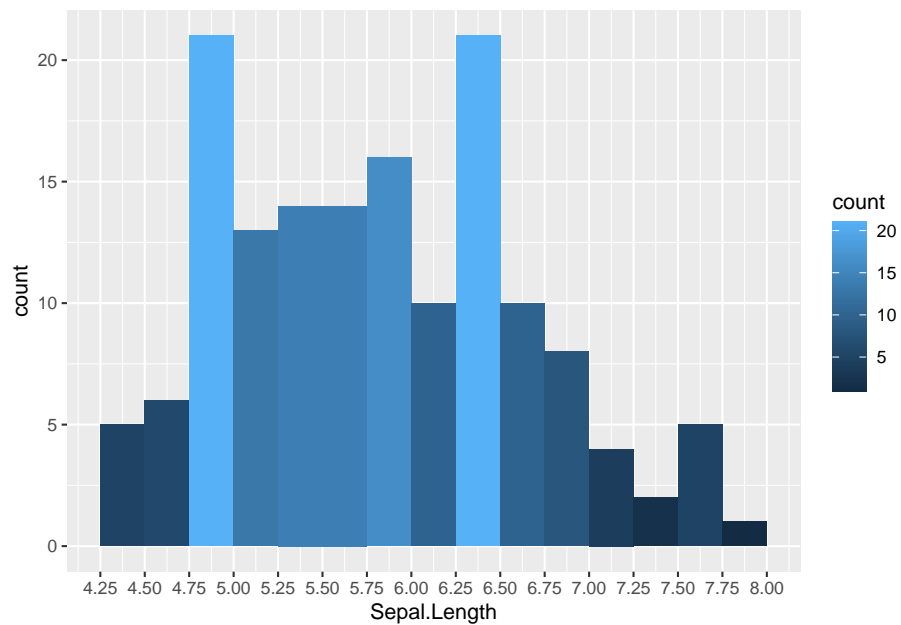


Observamos que não há presença de outliers, além disso, como a parte debaixo do retângulo separado pela linha que representa a mediana é menor, isto indica que a distribuição dos dados é ligeiramente assimétrica, o qual é confirmado pelo histograma:

```
hist(iris$Sepal.Length)
```



```
ggplot(data = iris, aes(x = Sepal.Length, fill = ..count..)) +  
  geom_histogram(binwidth = 0.25, boundary = 0) +  
  scale_x_continuous(breaks = seq(1, 10, by = 0.25))
```



## Chapter 5

# Tipos de Distribuições Discretas

Para atender a situações mais práticas, é necessário expandir os conceitos relacionados a probabilidade de forma que tenhamos modelos probabilísticos que representem todos os tipos de variáveis. Neste capítulo trabalharemos com variáveis quantitativas discretas.

Exemplo (Bussab):

Chamamos de **variável aleatória discreta** uma função  $X$  definida no espaço amostral  $\Omega$  que assume valores em um conjunto de números finito.

Neste contexto vimos como associar a cada valor  $x_i$  da variável aleatória  $X$  a sua probabilidade de ocorrência. Matematicamente, escrevemos

Além disso, chamamos de **função de probabilidade** da variável aleatória discreta  $X$  a função que a cada valor de  $x_i$  associa a sua probabilidade de ocorrência

$$p(x_i) = P(X = x_i) = p_i, i = 1, 2, \dots$$

### 5.1 Valor Médio de uma Variável Aleatória

Dada uma variável aleatória  $X$  discreta, assumindo os valores  $x_1, \dots, x_n$ , chamamos de valor médio ou esperança de  $X$  o valor

$$E[X] = \sum_{i=1}^n x_i P(X = x_i) = \sum_{i=1}^n x_i p_i.$$

Chamamos de variância da variável aleatória  $X$  o valor

$$\text{var}[X] = \sum_{i=1}^n [x_i - E[X]]^2 p_i$$

## Chapter 6

# Tipos de Distribuições Contínuas





## Chapter 7

# Introdução as bibliotecas do R

7.1 Dplyr

7.2 TidyR

7.3 GGPlot2



## Chapter 8

# Regressão Linear