Inferência Estatística com R

P. Maçaira, L. Bastos, S. Aguilar & I. Peres

Contents

Prefácio

Adicionar um vídeo nosso nos apresentando e o objetivo do livro.(pensar numa capa para o livro)

Sejam muito bem-vindos ao nosso livro on-line sobre Inferência Estatística.

Este livro destina-se a fornecer uma introdução abrangente aos métodos de inferência estatística e apresentar informações suficientes sobre cada método para que os leitores possam usá-los com sensatez. Desejamos que o leitor adquira o raciocínio necessário para, a partir dos dados, obter conclusões gerais acerca de uma população com base numa amostra.

O livro foi escrito para três públicos: (1) pessoas que se encontram utilizando estatística inferencial nos negócios quando podem não ter nenhum treinamento formal na área; (2) estudantes de graduação em Engenharia; (3) alunos de pós-graduação fazendo disciplina de estatística. Nós mesmos o usamos para alunos de pós-graduação e graduação da Pontifícia Universidade Católica do Rio de Janeiro, Brasil.

Para a maioria das seções, assumimos apenas que os leitores estão familiarizados com probabilidade estatística introdutória e com álgebra do ensino médio. Existem algumas seções que também exigem conhecimento de matrizes.

Usaremos o software R em todo o livro, um software gratuito e disponível em quase todos os sistemas operacionais. O R é uma ferramenta maravilhosa para todas as análises estatísticas e muito mais. Ao longo de todo o livro você se familiarizará com a linguagem e aprenderá a fazer inferência com o R.

A saída abaixo mostra as versões dos pacotes que usamos na compilação desta edição do livro. Alguns exemplos no livro não funcionarão com versões anteriores dos pacotes.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
```

```
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Finalmente, a saída lista alguns conflitos mostrando qual função será preferida quando uma função de mesmo nome estiver em vários pacotes.

O livro é diferente de outros livros de inferência estatística de várias maneiras.

1. É gratuito e online, tornando-o acessível a um grande público.
2. Ele usa R, que é um software gratuito, de código aberto e extremamente poderoso.
3. A versão online é continuamente atualizada. Você não precisa esperar até a próxima edição para que os erros sejam removidos ou novos métodos sejam discutidos. Atualizaremos o livro com frequência.
4. Existem dezenas de exemplos de dados reais retirados de nossa própria prática de consultoria.
5. Enfatizamos os métodos gráficos mais do que a maioria dos analistas. Usamos gráficos para explorar os dados, analisar a validade dos modelos ajustados e apresentar os resultados.

Boa leitura!

Paula Maçaira, Leonardo Bastos, Soraida Aguilar e Igor Peres.

Fevereiro 2022

Para citar a versão online deste livro, use o seguinte:

> Maçaira, P.; Bastos, L.; Aguilar, S. & Peres, I. (2022) Inferência Estatística com R, 1ª edição, OTexts: Melbourne, Australia. OTexts.com/fpp3. Acessado em 2022-02-04.

Chapter 1

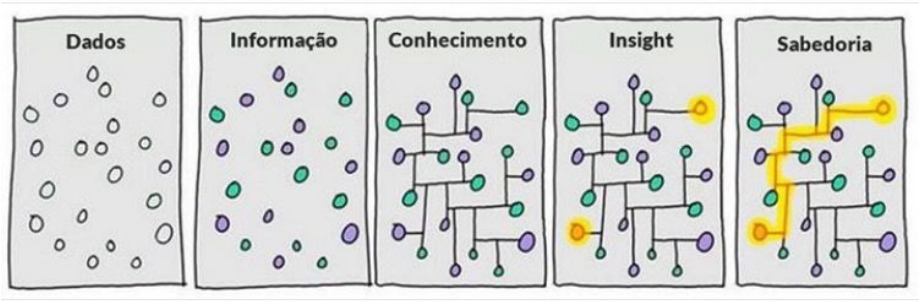
Análise Descritiva

1.1 Introdução

A indútria 4.0, ou quarta revolução industrial, é a continuação do aperfeiçoamento das máquinas, um processo que começou com a primeira Revolução Industrial e nunca mais parou. Podemos dizer que a indústria 4.0 é a realidade na qual a tecnologia industrial está cada vez mais eficiente: mais inteligente, mais rápida e mais precisa. O termo é utilizado para caracterizar a utilização do que há de mais moderno para produzir bens de consumo: big data, internet das coisas, inteligência artificial e muito mais.

O big data é um conceito muito importante e que vem ganhando bastante notoriedade nos últimos anos. É o termo em Tecnologia da Informação (TI) que trata sobre grandes conjuntos de dados que precisam ser processados e armazenados. É um conceito-chave para a quarta revolução industrial, pois são esses dados que permitem às máquinas trabalharem com maior eficiência.

Dados sozinhos não podem dizer muita coisa. Por exemplo, eu posso te dizer que minha cor favorita é verde. Isso não te leva a tirar muitas conclusões sobre minha personalidade. Porém, se eu te der uma pequena tabela em excel com alguns dos meus livros favoritos, provavelmente você terá conclusões mais significativas sobre mim. O que aconteceu foi que você, munido da sua capacidade mental e analítica, aliada a uma certa quantidade de dados foi capaz de tirar alguns insights sobre um assunto. E é exatamente aí que está a beleza e a grandiosidade da chamada “Era do Big Data”.



Big Data é basicamente análise de dados. De fato isso não é nenhuma novidade. Há muitos e muitos anos a humanidade coleta dados para serem analisados. A grande inovação está em aliar métodos antigos e limitados de análise de dados aos modernos recursos de hardware de alto processamento. Ou seja, agora é possível transitar todos esses cálculos e análises por meio de softwares desenvolvidos especificamente para trabalharem com enormes quantidades de dados. Uma solução de Big Data funciona com algoritmos complexos que trabalham a informação de modo a obter como saída os mais diversos tipos de insights.

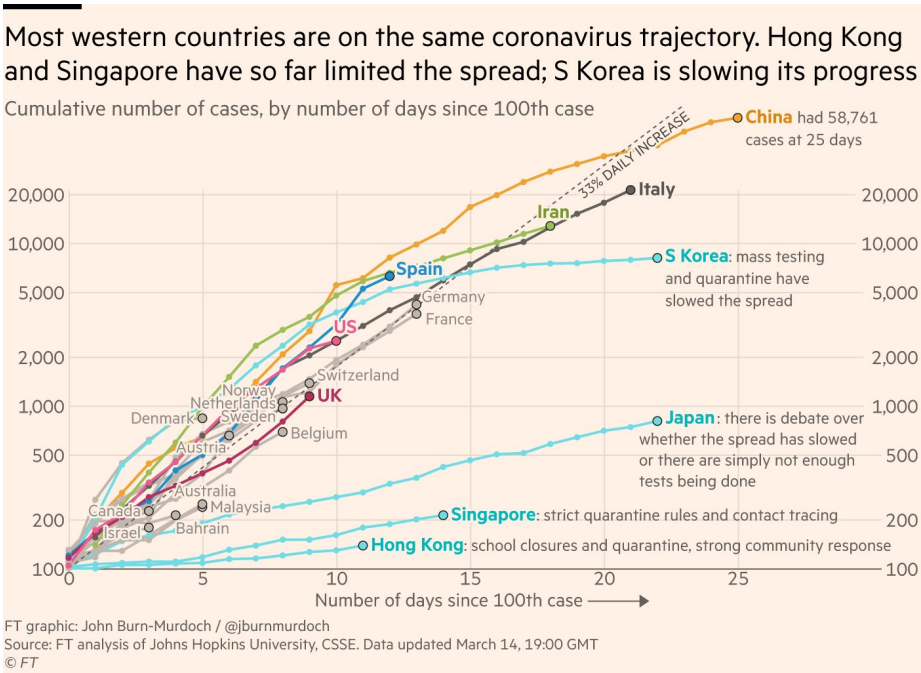
A Análise Descritiva é a fase inicial deste processo de estudo dos dados coletados. Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos. As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

Ao se condensar os dados, perde-se informação, pois não se têm as observações originais. Entretanto, esta perda de informação é pequena se comparada ao ganho que se tem com a clareza da interpretação proporcionada.

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais frequente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

Durante a pandemia da COVID-19 foram publicadas ótimas visualizações de dados de casos e mortes, mas a mais conhecida delas é o gráfico de John Burn-Murdoch no Financial Times. Esta é uma ótima visualização e ajudou a apresentar gráficos em escala logarítimica para um público amplo.



Ao mesmo tempo em que o uso das ferramentas estatísticas vem crescendo, aumenta também o abuso de tais ferramentas. É muito comum vermos em jornais e revistas, até mesmo em periódicos científicos, gráficos - voluntariamente ou intencionalmente - enganosos e estatísticas obscuras para justificar argumentos polêmicos.

Neste capítulo iremos mergulhar no mundo da Análise de Descritiva de Dados, explorando estatísticas descritivas e análises gráficas com os pacotes mais atuais do software `R` [R Core Team, 2020].

1.2 Coleta e Armazenamento de Dados

1.2.1 Exemplo inicial

O primeiro caso da COVID-19 no Brasil foi confirmado em 25 de fevereiro de 2020, em São Paulo. Desde então o Governo Federal do Brasil reporta diariamente o número total de casos e mortes no país para cada Unidade da Federação, que corresponde aos 26 estados mais o Distrito Federal. No github do Wesley Cota [Cota, 2020] é possível extrair tais informações atualizadas para o dia anterior de forma sistematizada numa planilha de dados de maneira a realizar a entrada dos dados num programa de computador.

Para exemplificação, iremos extrair do github do Wesley Cota a planilha que possui informações agregadas para o Brasil e suas UFs. Este é o formato mais comum de uma base de dados, composta por linhas e colunas. Cada linha

contém os dados de uma Unidade da Federação (elemento) e as informações (variáveis) esntão dispostas nas colunas. Dessa forma, a base de dados contém um número de linhas igual ao número de Unidades da Federação mais o total Brasil e um número de colunas igual ao número de variáveis sendo estudadas.

Estado	Casos	Mortes
TOTAL	26107894	630301
AC	105014	1881
AL	262735	6456
AM	541668	13991
AP	155288	2055
BA	1389564	28077
CE	1160180	25450
DF	625031	11208
ES	897437	13607
GO	1057811	25117
MA	390765	10510
MG	2805771	57575
MS	433787	9930
MT	643879	14363
PA	658777	17400
PB	504216	9755
PE	716733	20701
PI	349140	7419
PR	2042294	41334
RJ	1819607	70026
RN	434608	7768
RO	326272	6846
RR	144582	2101
RS	1885128	37041
SC	1477548	20690
SE	299127	6115
SP	4704552	158872
TO	276380	4013

Note que se usássemos uma base de dados que contém a evolução da doença por UF então teríamos uma combinação que seria uma linha para cada dia e cada estado, aumentando consideravelmente a dimensão da base.

1.3 Tipos de Variáveis

Variável é a característica de interesse que é medida em cada indivíduo da amostra ou população. Como o nome diz, seus valores variam de indivíduo para indivíduo. As variáveis podem ter valores numéricos ou não numéricos.

1.4 Variáveis Quantitativas

São as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.

- **Variáveis contínuas**: características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores não-inteiros (com casas decimais) fazem sentido. Usualmente devem ser medidas através de algum instrumento.
  - Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade.
- **Variáveis discretas**: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente, são o resultado de contagens.
  - Exemplos: número de filhos, número de bactérias por litro de leite, número de casos de uma doença.

1.5 Variáveis Qualitativas (ou categóricas)

São as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

- **Variáveis nominais**: não existe ordenação entre as categorias. Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
- **Variáveis ordinais**: existe uma ordenação entre as categorias. Exemplos: escolaridade (1o, 2o, 3o graus), estágio da doença (inicial,intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).

Uma variável originalmente quantitativa pode ser coletada de forma qualitativa. Por exemplo, a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos etc.), é qualitativa (ordinal). Outro exemplo é o peso dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado etc.).

Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa.

1.6 Medidas de Tendência Central

A tendência central de uma variável em um conjunto de dados é caracterizada pelo valor típico dessa variável. Essa é uma maneira de resumir a informação contida nos dados, pois escolheremos um valor para representar todos os outros. Assim, poderíamos perguntar, por exemplo, qual é a altura típica dos brasileiros adultos no final da década de 90 e compará-la com o valor típico da altura dos brasileiros no final da década de 80, a fim de verificar se os brasileiros estão se tornando, em geral, mais altos, mais baixos ou não sofreram nenhuma alteração em sua altura típica. Fazer essa comparação utilizando medidas-resumo (as alturas típicas em cada período) é bem mais sensato do que comparar os dois conjuntos de dados valor a valor, o que seria inviável. Mas, como identificar o valor típico de um conjunto de dados?

Existem três medidas que podem ser utilizadas para descrever a tendência central de um conjunto de dados: a média, a mediana e a moda. Apresentaremos essas três medidas e discutiremos suas vantagens e desvantagens.

1.6.1 Média Aritmética Simples

A média aritmética simples (que chamaremos apenas de média) é a medida de tendência central mais conhecida e usada para o resumo de dados. Essa popularidade pode ser devida à facilidade de cálculo e à idéia simples que ela nos sugere. De fato, se queremos um valor que represente a altura dos brasileiros adultos, por que não medir as alturas de uma amostra de brasileiros adultos, somar os valores e dividir esse "bolo" igualmente entre os participantes? Essa é a idéia da média aritmética.

Para apresentar a média, primeiramente vamos definir alguma notação. A princípio, essa notação pode parecer desnecessária, mas facilitará bastante nosso trabalho futuro.

$$\text{Notação} \begin{cases} n & \text{tamanho da amostra} \\ x_i & \text{valor da } i\text{-ésima observação} \\ \sum_{i=1}^{n} x & \text{soma de todas as observações} \\ \bar{x} & \text{símbolo que representa a média aritmética simples} \end{cases}$$

Assim,

$$\bar{x} = \frac{\text{soma de todas as observações}}{n} = \frac{\sum_{i=1}^{n} x}{n}$$

Exemplo: No conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2), temos $n = 5$, $x_1 = 1.3$, $x_2 = 0.7$, $x_3 = 5.8$, $x_4 = 2.4$ e $x_5 = 1.2$, portanto $\sum_{i=1}^{5} x_i = 1.3 + 0.7 + 5.8 + 2.4 + 1.2 = 11.4$ e assim $\bar{x} = \frac{11.4}{5} = 2.28$.

Se esses seis valores representassem, por exemplo, as quantidades de peixe pescado (em toneladas) durante cinco dias da semana, a quantidade típica pescada por dia, naquela semana, seria 2,28 toneladas. Como estamos representando o valor típico pela média aritmética, podemos falar em quantidade média diária naquela semana.

Fazendo no R:

```
mean(c(1.3, 0.7, 5.8, 2.4, 1.2))
```

```
## [1] 2.28
```

1.6.2 Mediana

A mediana de um conjunto de dados é definida como sendo o "valor do meio" desse conjunto de dados, dispostos em ordem crescente, deixando metade dos valores acima dela e metade dos valores abaixo dela.

Como calcular a mediana? Basta seguir sua definição. Vejamos:

- n **é ímpar**: Existe apenas um "valor do meio", que é a mediana.
  - Seja o conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2).
  - Ordenando os valores (0.7, 1.2, 1.3, 2.4, 5.8).
  - O valor do meio é o 1.3.
  - A mediana é o valor 1.3.
- n **é par**: Existem dois "valores do meio". A mediana é a média aritmética simples deles.
  - Seja o conjunto de dados (1.3, 0.7, 5.8, 2.4, 1.2, 2.1).
  - Ordenando os valores (0.7, 1.2, 1.3, 2.1, 2.4, 5.8).
  - Os valores do meio são 1.3 e 2.1.
  - A mediana é (1.3+2.1)/2=1.7.

Fazendo no R:

```
median(c(1.3, 0.7, 5.8, 2.4, 1.2))
```

```
## [1] 1.3
```

```
median(c(1.3, 0.7, 5.8, 2.4, 1.2, 2.1))
```

```
## [1] 1.7
```

Como medida de tendência central, a mediana é até mais intuitiva do que a média, pois representa, de fato, o centro (meio) do conjunto de valores ordenados. Assim como a média, o valor da mediana não precisa coincidir com algum dos valores do conjunto de dados. Em particular, quando os dados forem de natureza contínua, essa coincidência dificilmente ocorrerá.

1.6.3 Moda

Uma maneira alternativa de representar o que é "típico" é através do valor mais frequente da variável, chamado de moda.

Como calcular a moda? Basta verificar o valor que "aparece" mais vezes. Vejamos:

- No conjunto de dados (1, 2, 3, 3, 4, 5, 5, 5, 5, 5), há apenas uma moda, o valor 5, portanto o conjunto de dados é **unimodal**.
- No conjunto de dados (1, 2, 2, 2, 2, 3, 4, 5, 6, 6, 6, 6, 7, 9), existem duas modas, os valores 2 e 6, portanto o conjunto de dados é **bimodal**.
- Nem sempre a moda existe ou faz sentido, no conjunto de dados (1, 2, 3, 4, 5, 6, 7, 8, 9), não existe um valor mais frequente que os demais, portanto o conjunto de dados é **amodal**.

Para usar a função que calcula a moda (`Mode`) no `R` temos que instalar e carregar o pacote `pracma`:

```
library(pracma)
Mode(c(1,2,3,3,4,5,5,5,5,5))
```

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

```
Mode(c(1,2,2,2,2,3,4,5,6,6,6,6,7,9)) # escolherá o menor valor, caso haja empate
```

```
## [1] 2
```

```
Mode(c(1,2,3,4,5,6,7,8,9))
```

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

A moda é também a única das medidas de tendência central que faz sentido no caso de variáveis qualitativas. Assim, a categoria dessas variáveis que aparecer com maior freqüência é chamada de categoria modal.

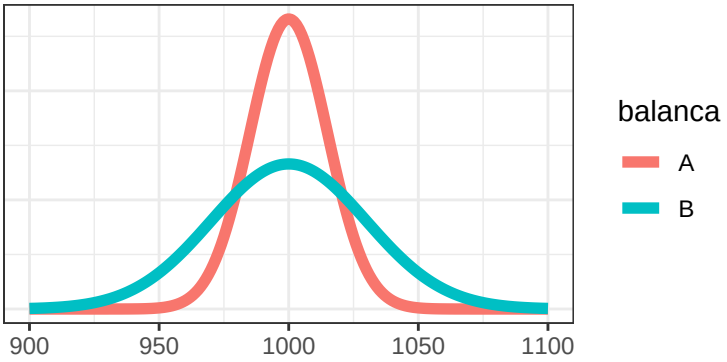
1.7 Medidas de Variabilidade

As medidas de tendência central (média, mediana, moda) conseguem resumir em um único número, o valor que é "típico" no conjunto de dados. Mas, será que, somente com essas medidas, conseguimos descrever adequadamente o que ocorre em um conjunto de dados?

Vejamos um exemplo: quando pesamos algo em uma balança, esperamos que ela nos dê o verdadeiro peso daquilo que estamos pesando. No entanto, se fizermos várias medições do peso de um mesmo objeto em uma mesma balança, teremos diferentes valores para o peso deste objeto. Ou seja, existe variabilidade nas medições de peso fornecidas pela balança. Neste caso, quanto menor a variabilidade desses valores, mais precisa é a balança (considerando que a média das medidas de peso coincida como seu valor real). Observe na Figura abaixo,

onde estão representadas as distribuições das medições do peso de uma esfera de 1000g, feitas por duas balanças (A e B). As duas balanças registram o mesmo peso médio de 1000g (média dos pesos de todas as medições feitas). Isto é, as duas balanças tipicamente acertam o verdadeiro peso da esfera. Porém, pela Figura, podemos notar que

- as medições da balança A *variam pouco* em torno de 1000g: oscilam bastante entre cerca de 950g e 1050g (uma "imprecisão" de 50g)
- as medições da balança B *variam muito* em torno de 1000g: oscilam basicamente entre 900g e 1100g (uma "imprecisão" de 100g)



Dois conjuntos de dados podem ter a mesma medida de centro (valor típico), porém uma dispersão diferente em torno desse valor. Desse modo, além de uma medida que nos diga qual é o valor "típico" do conjunto de dados, precisamos de uma medida do grau de dispersão (variabilidade) dos dados em torno do valor típico.

O objetivo das medidas de variabilidade é quantificar esse grau de dispersão. Nesta seção, apresentaremos três dessas medidas (amplitude total, desvio-padrão e coeficiente de variação), discutindo suas vantagens e desvantagens. Em discussões posteriores, apresentaremos medidas de variabilidade alternativas.

1.7.1 Amplitude Total

A medida de variabilidade mais simples é a chamada amplitude total (AT), que é a diferença entre o valor máximo e o valor mínimo de um conjunto de dados.

$$AT = \text{Máximo} - \text{Mínimo}$$

Exemplo: Medição do peso de uma esfera de 1000g em duas balanças (A e B).

Estatísticas	Balança A	Balança B
Mínimo	945g	895g

Estatísticas	Balança A	Balança B
Máximo	1040g	1095g
AT	1040-945=95g	1095-895=200g

A variabilidade das medições de peso da balança B é maior que a variabilidade das medições de peso da balança A (apesar do valor médio ser igual).

Embora seja uma medida simples de variabilidade, a amplitude total é um tanto grosseira, pois depende somente de dois valores do conjunto de dados (máximo e mínimo), não captando o que ocorre com os outros valores.

Fazendo no R:

```
1040-945
```

```
## [1] 95
```

```
1095-895
```

```
## [1] 200
```

1.7.2 Desvio Padrão

Uma boa medida de dispersão deve considerar todos os valores do conjunto de dados e resumir o grau de dispersão desses valores em torno do valor típico.

Considerando a média como a medida de tendência central, podemos pensar em medir a dispersão (desvio) de cada valor do conjunto de dados em relação à ela. A medida mais simples de desvio entre duas quantidades é a diferença entre elas. Assim, para cada valor x_i, teremos o seu desvio em relação à $\bar{x}$ representado por $(x_i - \bar{x})$.

Exemplo: No conjunto de dados 1, 1, 2, 3, 4, 4, 5, 6, 7, 7, relativo ao número de filhos de 10 mulheres, temos $\bar{x} = 4$ filhos. Na tabela abaixo, a coluna 1 mostra esses 10 valores e a coluna 2 mostra o desvio de cada um deles até a média.

Coluna 1 (Xi)	Coluna 2 (Xi-Media)	Coluna 3 (Xi-Media)^2
1	-3	9
1	-3	9
2	-2	4
3	-1	1
4	0	0
4	0	0
5	1	1
6	2	4
7	3	9
7	3	9
Soma	0	46

A ideia do desvio padrão

Como temos um desvio para cada elemento, poderíamos pensar em resumi-los em um desvio típico, a exemplo do que fizemos com a média. Porém, quando somarmos esses desvios para o cálculo do desvio médio, a soma dará sempre zero, como pode ser visto na coluna 2 do exemplo anterior. Isto ocorre com qualquer conjunto de dados, pois os desvios negativos sempre compensaram os positivos.

No entanto, os sinais dos desvios não são importantes para nossa medida de dispersão, já que estamos interessados na quantidade de dispersão e não na direção dela. Portanto, eliminaremos os sinais elevando os desvios ao quadrado, como mostrado na coluna 3. A soma desses desvios ao quadrado pode ser, então, dividida entre os participantes do "bolo". Na verdade, por razões absolutamente teóricas, dividiremos essa soma pelo total de participantes menos 1 $(n-1)$. Assim, usando a notação definida anteriormente, teremos

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

Para os dados do exemplo, teremos $46/(10-1) = 5,11$. Esse valor pode ser visto como uma quase-média dos desvios ao quadrado e é chamado de **variância**.

A variância seria nossa medida de variabilidade se não fosse o fato de que ela está expressa em uma unidade diferente da unidade dos dados, pois, ao elevarmos os desvios ao quadrado, elevamos também as unidades de medida em que eles estão expressos. No caso dos dados do exemplo, medidos em número de filhos, a variância vale $5,11$ "filhos ao quadrado", algo que não faz nenhum sentido.

Para eliminar esse problema, extraímos a raiz quadrada da variância e, finalmente, temos a nossa medida de variabilidade, que chamaremos desvio-padrão (DP).

$$DP = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

O desvio-padrão, como o nome já diz, representa o desvio típico dos dados em relação à média, escolhida como medida de tendência central. No exemplo, temos que o desvio padrão vale $2,26$. Isto significa que a distância típica (padrão) de cada mãe até o número médio de filhos (4 filhos) é de 2,26 filhos. Quanto maior o desvio-padrão, mais diferentes entre si serão as quantidades de filhos de cada mãe.

O desvio-padrão, em alguns livros chamado de s, é uma medida sempre positiva. Se observarmos a maneira como ele é calculado, veremos que não há como obter um valor negativo.

Fazendo no `R`:

```
round(var(c(1,1,2,3,4,4,5,6,7,7)),2) # função round() serve para arrendondar o número
```

```
## [1] 5.11
```

```
round(sd(c(1,1,2,3,4,4,5,6,7,7)),2)
```

```
## [1] 2.26
```

Exemplo: Os agentes de fiscalização de certo município realizam, periodicamente, uma vistoria nos bares e restaurantes para apurar possíveis irregularidades na venda de seus produtos. A seguir, são apresentados dados de uma vistoria sobre os pesos (em gramas) de uma amostra de 10 bifes, constantes de um cardápio de um restaurante como "bife de 200 gramas".

```
##  [1] 170 175 180 185 190 195 200 200 200 205
```

Como podemos notar, nem todos os "bifes de 200 gramas" pesam realmente 200 gramas. Esta variação é natural e é devida ao processo de produção dos bifes. No entanto, esses bifes deveriam pesar cerca de 200 gramas e com pouca variação em torno desse valor. Com o auxílio do R, calcularemos a média e o desvio-padrão.

```
mean(c(170,175,180,185,190,195,200,200,200,205))
```

```
## [1] 190
```

```
sd(c(170,175,180,185,190,195,200,200,200,205))
```

```
## [1] 12.0185
```

Os bifes desse restaurante pesam, em média, 190 gramas, com um desvio-padrão de 12 gramas. Ou seja, os pesos dos "bifes de 200 gramas" variam tipicamente entre 178 e 202 gramas. Analisando esses valores, concluímos que esse restaurante pode estar lesando a maior parte de seus clientes.

Para casos como esse, os agentes fiscalizadores podem estabelecer parâmetros (valores) para saber até quanto a média pode se desviar do valor correto e o quanto de variação eles podem permitir numa amostra para concluir que o processo de produção de bifes não possui problemas. Por exemplo, a média da amostra não poderia ser inferior a 198 gramas, com um desvio-padrão que não seja superior a 5% dessa média.

Essas idéias são utilizadas no controle do processo de produção das indústrias, onde já se espera alguma variação entre as unidades produzidas. Porém, essa variação deve estar sob controle. Numa indústria farmacêutica, por exemplo, espera-se que os comprimidos de uma certa droga sejam produzidos com uma certa variação em sua composição (maior ou menor quantidade do princípio ativo), devido à própria maneira como os comprimidos são produzidos (máquinas, pessoas etc.). No entanto, esta variação deve ser pequena, para que não sejam produzidos comprimidos inócuos (com pouco do princípio ativo) ou

com extra-dosagem do princípio ativo, o que, em ambos os casos, pode causar sérias complicações à saúde do paciente.

O desvio-padrão nos permite distinguir numericamente conjuntos de dados de mesmo tamanho, mesma média, mas que são visivelmente diferentes. Usando o desvio-padrão, também conseguimos representar numericamente a variabilidade das medições das balanças A e B, que, apesar de possuírem a mesma média, possuem variabilidades bastante diferentes.

Quando os conjuntos de dados a serem comparados possuem médias diferentes, a comparação da variabilidade desses conjuntos deve levar em conta essa diferença. Por esta e outras razões, definiremos uma terceira medida de variabilidade, o **coeficiente de variação**.

1.7.3 Coeficiente de Variação

Ao analisarmos o grau de dispersão de um conjunto de dados, poderemos nos deparar com uma questão do tipo: um desvio-padrão de 10 unidades é pequeno ou grande?

Vejamos:

- Se estivermos trabalhando com um conjunto de dados cuja média é 10.000, um desvio típico de 10 unidades em torno dessa média significa pouca dispersão;
- Mas, se a média for igual a 100, um desvio típico de 10 unidades em torno dessa média significa muita dispersão.

Assim, antes de responder se um desvio-padrão de 10 unidades é grande ou pequeno, devemos avaliar sua magnitude em relação à média:

- No primeiro caso, o desvio-padrão corresponde a 0,1% da média
- No segundo caso, o desvio-padrão corresponde a 10% da média

À essa razão entre o desvio-padrão e a média damos o nome de **Coeficiente de Variação**:

$$CV = \frac{\text{Desvio Padrão}}{\text{Média}}$$

Quanto menor o Coeficiente de Variação de um conjunto de dados, menor é a sua variabilidade. O Coeficiente de Variação expressa o quanto da escala de medida, representada pela média, é ocupada pelo desvio-padrão.

O Coeficiente de Variação é uma medida adimensional, isto é, não depende da unidade de medida. Essa característica nos permite usá-lo para comparar a variabilidade de conjuntos de dados medidos em unidades diferentes, o que seria impossível usando o desvio-padrão.

Exemplo: Numa pesquisa na área de Saúde Ocupacional, deseja-se comparar a idade de motoristas e cobradores de ônibus da região metropolitana de Belo Horizonte. Algumas estatísticas descritivas são apresentadas na Tabela abaixo.

Grupo	n	Media	DP	CV
Motoristas	150	35.6	5.08	0.143
Cobradores	50	22.6	3.11	0.137

Os motoristas são, em média, 13 anos mais velhos do que os cobradores. Ao compararmos o grau de dispersão dos dois grupos usando o desvio-padrão, concluiríamos que os motoristas são menos homogêneos quanto à idade do que os cobradores. Ao fazermos isso, estamos esquecendo que, apesar de estarem em unidades iguais, as medidas de idade nos dois grupos variam em escalas diferentes. As idades dos motoristas variam em torno dos 35 anos e podem chegar até 18 anos (idade mínima para se conseguir a habilitação), numa amplitude de 17 unidades. Enquanto isso, as idades dos cobradores variam em torno de 22 anos e também só podem chegar até a 18 anos, uma amplitude de apenas 4 anos. Assim, os motoristas tem a possibilidade de ter um desvio-padrão maior do que o dos cobradores. Se levarmos em conta a escala de medida, usando o coeficiente de variação, veremos que os motoristas são somente um pouco mais heterogêneos (dispersos) quanto à idade do que os cobradores.

Fazendo no R:

```
media=c(35.6,22.6)
desvio=c(5.08,3.11)
round(desvio/media,3)
```

```
## [1] 0.143 0.138
```

1.8 Medidas de Posição

Quando falamos de posição ou colocação de um indivíduo em uma corrida ou em um teste como o Vestibular, frequentemente nos referimos ao seu posto, como 1º, 2º, 3º, 29º ou último lugar. Mas, para sabermos se uma dada colocação é ou não um bom resultado, precisamos informar quantos indivíduos participaram da corrida ou do Vestibular.

A medida de posição que veremos aqui, os percentis, solucionam este e outros problemas de posicionamento (ranking). A posição de um indivíduo no conjunto de dados é mostrada, pelo percentil, contando-se (em porcentagem) quantos indivíduos do conjunto têm valores menores que o deste indivíduo.

Como veremos, esta medida de posição pode ser usada para comparar a posição do indivíduo em diferentes conjuntos de dados, nos quais foram medidas as mesmas variáveis ou variáveis diferentes.

1.8.1 Percentis

Considere o trecho a seguir, sobre a posição do Brasil, entre os países do mundo, quanto à renda per capita:

O Brasil obviamente não é país rico, mas também não está entre os mais pobres. ... Mais de três quartos da população mundial vivem em países de renda per capita menor

Neste caso, o posição do Brasil é dada pela quantidade de países que têm renda per capita menores que o Brasil, a saber três quartos ou 75%. O mesmo tipo de raciocínio fazemos quando dizemos que certo aluno está entre os 5% melhores do colégio. Não precisamos nem saber quantos alunos tem o colégio ou em quantos países estão sendo consideradas as rendas. Aqui já houve uma padronização da posição usando-se a porcentagem de alunos ou países com desempenho ou renda abaixo do valor considerado. É este raciocínio que define os percentis.

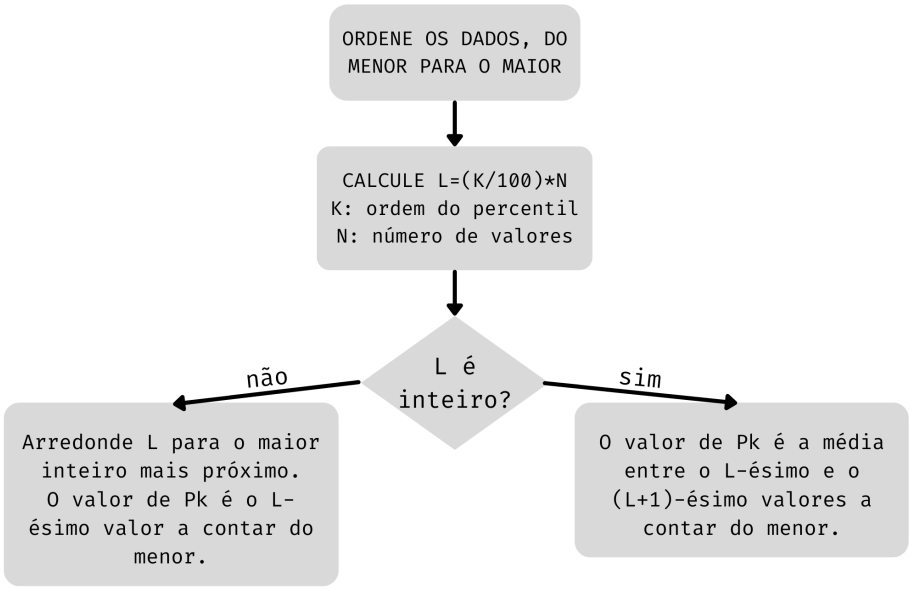
O percentil de ordem K (onde k é qualquer valor entre 0 e 100), denotado por P_k, é o valor tal que $K\%$ dos valores do conjunto de dados são menores ou iguais a ele.

Assim, o percentil de ordem 10, o P_{10}, é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele etc.

Os percentil de ordem $10, 20, 30, \ldots, 90$ dividem o conjunto de dados em dez partes com mesmo número de observações e são chamados de **decis**.

Os percentis de ordem $25, 50$ e 75 dividem o conjunto de dados em quatro partes com o mesmo número de observações. Assim, estes três percentis recebem o nome de quartis – **primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3)**, respectivamente. O segundo quartil é a já conhecida mediana.

Existem vários processos para calcular os percentis, usando interpolação. Vamos ficar com um método mais simples, mostrado na Figura a seguir. As diferenças serão muito pequenas e desaparecerão à medida que aumenta o número de dados.



Considere as notas finais dos 40 candidatos ao curso de Direito no Vestibular de certa faculdade, já colocadas em ordem crescente:

```
##  [1] 40 41 42 42 44 47 48 48 49 49 51 52 53 58 59 62 63 64 65 66 67 68 69 70 75
## [26] 76 83 83 85 86 86 87 87 88 92 93 94 95 97 98
```

Vamos calcular alguns percentis:

- Percentil de ordem 10: 10% de 40 = 4. Então o P_10 = média(4º e 5º valores)=$(42 + 44)/2 = 43$.
- Percentil de ordem 95: 95% de 40 = 38. Então o P_95 = média(38º e 39º valores)=$(95 + 97)/2 = 96$.
- Primeiro Quartil: 25% de 40 = 10. Então o Q1 = média(10º e 11º valores)=$(49 + 51)/2 = 50$.
- Terceiro Quartil: 75% de 40 = 30. Então o Q3 = média(30º e 31º valores)=$(86 + 86)/2 = 86$.
- Mediana: 50% de 40 = 20. Então mediana = média(20º e 21º valores)=$(66 + 67)/2 = 66,5$.

Fazendo no R:

```
dados_ex7=c(40,41,42,42,44,47,48,48,49,
            49,51,52,53,58,59,62,63,64,
            65,66,67,68,69,70,75,76,83,
```

```
            83,85,86,86,87,87,88,92,93,94,95,97,98)
quantile(dados,probs=c(0.1,0.95,0.25,0.75,0.5))
```

```
##     10%     95%     25%     75%     50%
## 174.50 202.75 181.25 200.00 192.50
```

Veja que os valores encontrados pelo R não coincidem com os calculados "na mão", isto porque, o R utiliza um método de interpolação para calcular valores que não estão presentes na amostra.

1.9 Referências

Chapter 2

Análise Gráfica

2.1 Introdução

Já sabemos que as variáveis de um estudo dividem-se em quatro tipos:

- Qualitativas:
  - nominais
  - ordinais
- Quantitativas:
  - discretas
  - contínuas

Os dados gerados por esses tipos de variáveis são de naturezas diferentes e devem receber tratamentos diferentes. Portanto, vamos estudar as ferramentas mais adequadas para cada tipo de dados, separadamente.

2.2 Variáveis Qualitativas - Nominais e Ordinais

Na base de dados `mtcars`, uma das duas variáveis qualitativas presentes é a categoria da transmissão (automática ou manual). Para organizar os dados provenientes de uma variável qualitativa, é usual fazer uma tabela de frequências, como a tabela abaixo, onde estão apresentadas as frequências com que ocorre cada um dos tipos de transmissão no total dos 32 carros observados. Cada categoria da variável transmissão (automática, manual) é representada numa linha da tabela. Há uma coluna com as contagens de carros em cada categoria (frequência absoluta) e outra com os percentuais que essas contagens representam no total de carros (frequência relativa). Esse tipo de tabela representa a distribuição de frequências dos carros segundo a variável transmissão.

Como a variável transmissão é qualitativa nominal, ou seja, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma.

É comum a disposição das linhas pela ordem decrescente das frequências das classes.

Transmissão	Frequência Absoluta	Frequência Relativa (%)
Automático	19	59.38
Manual	13	40.62
Total	32	100.00

Quando a variável tabelada for do tipo qualitativa ordinal, as linhas da tabela de frequências devem ser dispostas na ordem existente para as categorias. A tabela abaixo mostra a distribuição de frequências das coletas segundo o mês de observação na base de dados `airquality`, que é uma variável qualitativa ordinal. Nesse caso, podemos acrescentar mais duas colunas com as frequências acumuladas (absoluta e relativa), que mostram, para cada mês, a frequência das coletas até aquele mês. Por exemplo, até o mês de julho, foram coletadas 92 amostras, o que representa 60,13% do total de amostras coletadas.

Transmissão	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
5	31	20.26	31	20.26
6	30	19.61	61	39.87
7	31	20.26	92	60.13
8	31	20.26	123	80.39
9	30	19.61	153	100
Total	153	100.00	-	-

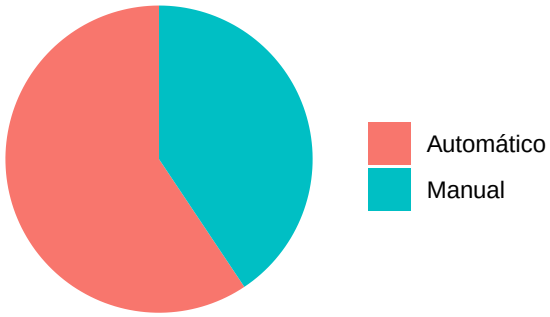
Note que as frequências acumuladas não fazem sentido em distribuição de frequências de variáveis para as quais não existe uma ordem natural nas categorias, as qualitativas nominais.

A visualização da distribuição de frequências de uma variável fica mais fácil se fizermos um gráfico a partir da tabela de frequências. Existem vários tipos de gráficos, dependendo do tipo de variável a ser representada. Para as variáveis do tipo qualitativas, abordaremos dois tipos de gráficos: os de setores e os de barras.

Os gráficos de setores, mais conhecidos como gráficos de pizza ou torta, são construídos dividindo-se um círculo (pizza) em setores (fatias), um para cada categoria, que serão proporcionais à frequência daquela categoria.

A Figura a seguir mostra um gráfico de setores para a variável transmissão, construído a partir da primeira tabela de frequências. Através desse gráfico, fica mais fácil perceber que os carros automáticos são a grande maioria dos carros. Como esse gráfico contém todas as informações da tabela, pode substituí-la com a vantagem de tornar análise dessa variável mais agradável.

```
library(ggplot2)
df = data.frame(table(mtcars$am))
levels(df$Var1) = c("Automático","Manual")
ggplot(df, aes(x="", y=Freq, fill=Var1)) +
  geom_bar(stat="identity") +
  coord_polar("y", start=0) +
  theme_void() +
  theme(legend.title = element_blank())
```

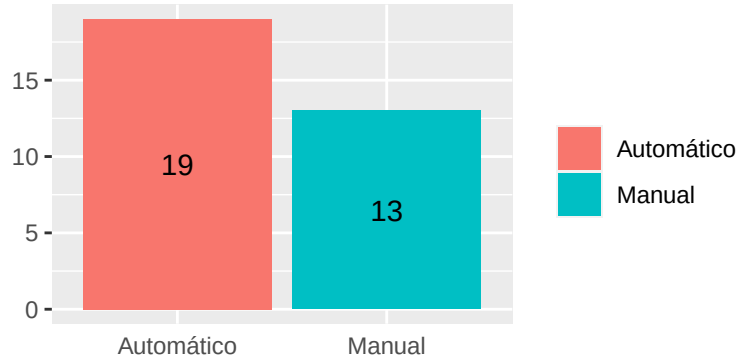


Quando houver mais de duas categorias de uma variável nominal, a disposição no gráfico de setores deve ser pela ordem decrescente das frequências, no sentido horário. A categoria "outros", quando existir, deve ser sempre a última, mesmo não seja a de menor frequência.

As vantagens da representação gráfica das distribuições de frequências ficam ainda mais evidentes quando há a necessidade de comparar vários grupos com relação à variáveis que possuem muitas categorias, como veremos mais adiante.

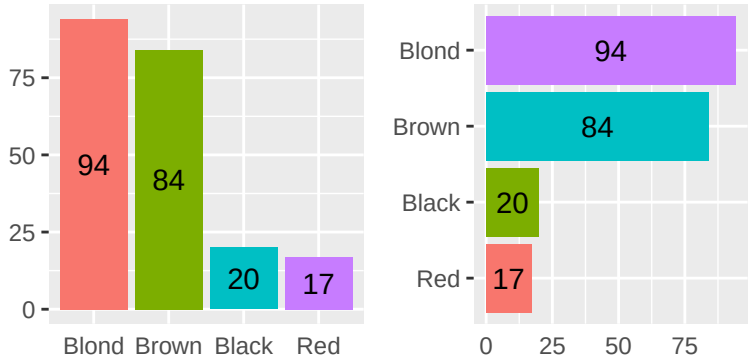
Uma alternativa ao gráfico de setores é o gráfico de barras (colunas) como o da figura a seguir. Ao invés de dividirmos um círculo, dividimos uma barra. Note que, em ambos os gráficos, as frequências relativas das categorias devem somar 100%. Aliás, esse é a idéia dos gráficos: mostrar como se dá a divisão (distribuição) do total de elementos (100%) em partes (fatias).

```
df = data.frame(table(mtcars$am))
levels(df$Var1) = c("Automático","Manual")
ggplot(df, aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Freq),position=position_stack(vjust = 0.5)) +
  theme(legend.title = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
```



Uma situação diferente ocorre quando desejamos comparar a distribuição de frequências de uma mesma variável em vários grupos, como por exemplo, a frequência de estudantes com olhos azuis entre todas as cores de cabelo na base de dados `HairEyeColor`. Se quisermos usar o gráfico de setores para fazer essa comparação, devemos fazer quatro gráficos, um para cada cor de cabelo, com duas fatias cada um (olhos azuis e olhos não azuis). Uma alternativa é a construção de um gráfico de colunas (barras) como os gráficos das figuras a seguir, onde há uma barra para cada cor de cabelo representando a frequência de estudantes com olhos azuis e aquela cor de cabelo. Além de economizar espaço na apresentação, permite que as comparações sejam feitas de maneira mais rápida.

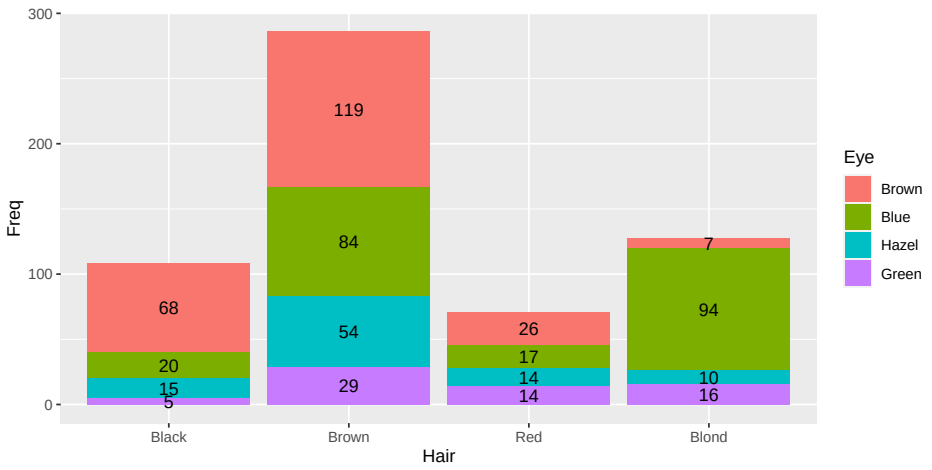
```
df = data.frame(HairEyeColor[,,1] + HairEyeColor[,,2])
df = df %>% filter(Eye == "Blue")
p1 = df %>% mutate(name = fct_reorder(Hair, desc(Freq))) %>%
  ggplot(aes(x=name, y=Freq, fill=name)) +
  geom_bar(position="dodge", stat="identity") +
  geom_text(aes(label = Freq),position=position_stack(vjust = 0.5)) +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
p2 = df %>% mutate(name = fct_reorder(Hair, Freq)) %>%
  ggplot(aes(x=name, y=Freq, fill=name)) +
  geom_bar(position="dodge", stat="identity") +
  geom_text(aes(label = Freq),position=position_stack(vjust = 0.5)) +
  coord_flip() +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank())
gridExtra::grid.arrange(p1,p2,ncol=2)
```



A ordem dos grupos pode ser qualquer, ou aquela mais adequada para a presente análise. Frequentemente, encontramos as barras em ordem decrescente, já antecipando nossa intuição de ordenar os grupos de acordo com sua frequência para facilitar as comparações. Caso a variável fosse do tipo ordinal, a ordem das barras seria a ordem natural das categorias, como na tabela de frequências.

A figura abaixo mostra um gráfico de barras que pode ser usado da comparação da distribuição de frequências de uma mesma variável em vários grupos. É também uma alternativa ao uso de vários gráficos de setores, sendo, na verdade, a junção de dois gráficos com os das figuras acima num só gráfico. Porém, esse tipo de gráfico só deve ser usado quando não houver muitos grupos a serem comparados e a variável em estudo não tiver muitas categorias.

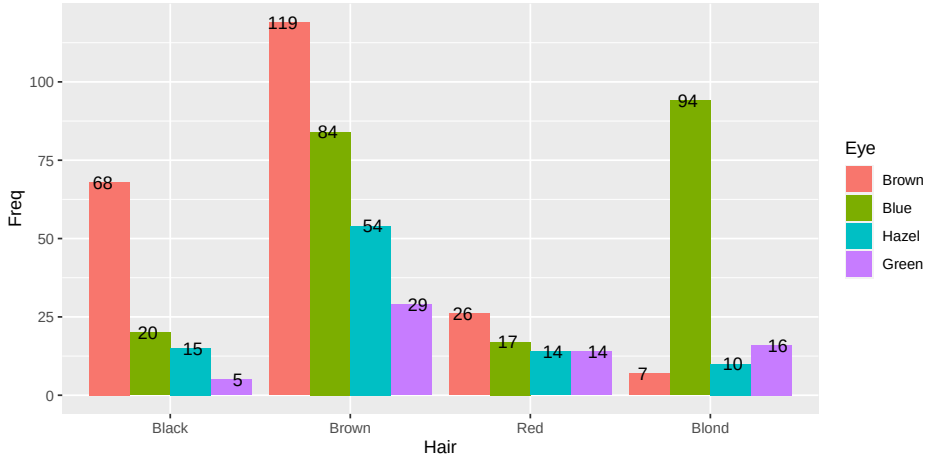
```
df = data.frame(HairEyeColor[,,1] + HairEyeColor[,,2])
ggplot(df,aes(x=Hair, y=Freq, fill=Eye)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = Freq),position=position_stack(vjust = 0.5))
```



Frequentemente, é necessário fazer comparações da distribuição de frequências de uma variável em vários grupos simultaneamente. Nesse caso, o uso de grá-

ficos bem escolhidos e construídos torna a tarefa muito mais fácil. Na figura abaixo, está representada a distribuição de frequências da cor dos olhos segundo a variável cor do cabelo.

```
df = data.frame(HairEyeColor[,,1] + HairEyeColor[,,2])
ggplot(df,aes(x=Hair, y=Freq, fill=Eye)) +
  geom_bar(position="dodge",stat="identity") +
  geom_text(aes(label = Freq),position=position_dodge(width = 1))
```



2.3 Variáveis Quantitativas Discretas

Quando estamos trabalhando com uma variável discreta que assume poucos valores, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.

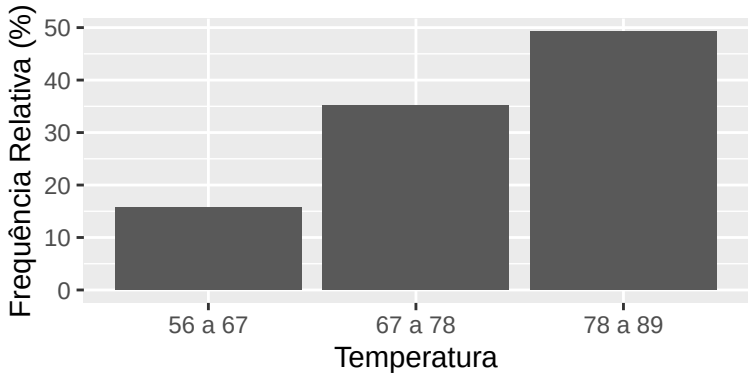
Quando trabalhamos com uma variável discreta que pode assumir um grande número de valores distintos como, por exemplo, parte inteira da temperatura máxima, a construção da tabela de frequências e de gráficos considerando cada valor como uma categoria fica inviável. A solução é agrupar os valores em classes ao montar a tabela, como mostra a tabela abaixo.

Temperatura	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
56 a 67	21	15.67	21	15.67
67 a 78	47	35.07	68	50.75
78 a 89	66	49.25	134	100
Total	134	99.99	-	-

A Figura abaixo mostra o gráfico da distribuição de frequências da temperatura

medida por 236 dias consecutivos.

```
dt %>% filter(dt$Variavel != "Total") %>%
  ggplot(aes(x=Variavel, y=Freq_rel)) +
  geom_bar(position="dodge",stat="identity") +
  labs(x="Temperatura",y="Frequência Relativa (%)") +
  theme(legend.position = "none")
```



2.4 Variáveis Quantitativas Contínuas

Quando a variável em estudo é do tipo contínua, que assume muitos valores distintos, o agrupamento dos dados em classes será sempre necessário na construção das tabelas de frequências. A tabela abaixo apresenta a distribuição de frequências para o peso dos carros.

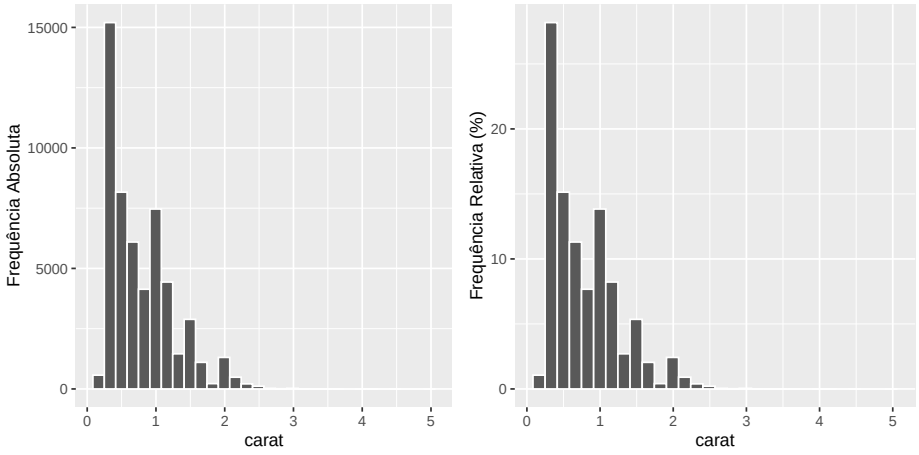
Temperatura	Frequência Absoluta	Frequência Relativa (%)	Frequência Absoluta Acumulada	Frequência Relativa Acumulada (%)
1.5 |- 2	0	0.00	0	0
2 |- 2.5	0	0.00	0	0
2.5 |- 3	3	9.38	3	9.38
3 |- 3.5	10	31.25	13	40.62
3.5 |- 4	12	37.50	25	78.12
4 |- 4.5	6	18.75	31	96.88
4.5 |- 5	1	3.12	32	100
5 |- 5.5	0	0.00	32	100
Total	32	100.00	-	-

2.5 Histograma

A representação gráfica da distribuição de frequências de uma variável contínua é feita através de um gráfico chamado histograma, mostrado nas figuras abaixo

com o peso de diamantes (`diamonds`). O histograma nada mais é do que o gráfico de barras verticais, porém construído com as barras unidas, devido ao caráter contínuo dos valores da variável.

```
library(ggplot2)
p1=ggplot(diamonds,aes(carat)) +
  geom_histogram(color = "white") +
  labs(y = "Frequência Absoluta")
p2=ggplot(diamonds,aes(carat)) +
  geom_histogram(aes(y = (..count..)/sum(..count..)*100),color = "white") +
  labs(y = "Frequência Relativa (%)")
gridExtra::grid.arrange(p1,p2,ncol=2)
```



Os histogramas das figuras acima têm exatamente a mesma forma, apesar de serem construídos usando as frequências absolutas e relativas, respectivamente. O objetivo dessas figuras é mostrar que a escolha do tipo de frequência a ser usada não muda a forma da distribuição. Entretanto, o uso da frequência relativa torna o histograma comparável a outros histogramas, mesmo que os conjuntos de dados tenham tamanhos diferentes (desde a mesma escala seja usada!)

Para o desenvolvimento do histograma:

- Calcule a amplitude dos dados
- Defina a quantidade de classes (as barras verticais). Não existe uma regra, porém uma boa aproximação seria calcular a raiz quadrada da quantidade de dados ($\sqrt{n}$)
- Calcule o intervalo das classes dividindo a amplitude pela quantidade de classes
- Determine os limites das classes. Selecione o valor mínimo dos dados (se for mais viável, ele pode ser arredondado para baixo) e soma o valor do intervalo de classe para obter o limite superior
- Faça o passo anterior para todas as classes

- Calcule a frequência dos dados que pertence a cada intervalo

Este é o passo a passo básico para a elaboração de um Histograma, que seja capaz de lhe trazer informações precisas sobre a frequência com que algo acontece em um determinado contexto.

Ao estudarmos a distribuição de frequências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

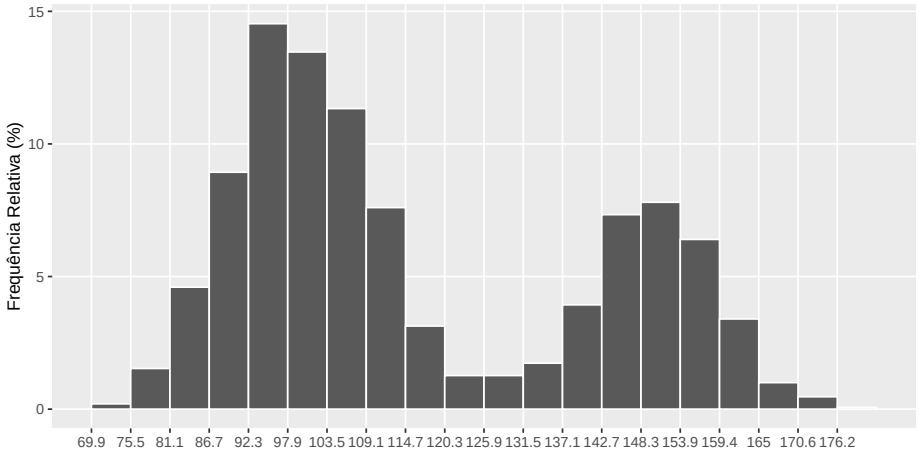
- Tendência Central;
- Variabilidade;
- Forma.

Tais características podem ser quantificadas através das medidas de síntese numérica ou visualizadas a partir do histograma.

2.5.1 Tendência Central

A tendência central da distribuição de frequências de uma variável é caracterizada pelo valor (ou faixa de valores) "típico" da variável.

Uma das maneiras de representar o que é "típico" é através do valor mais frequente da variável, chamado de moda. Ou, no caso da tabela de frequências, a classe de maior frequência, chamada de classe modal. No histograma, esta classe corresponde àquela com barra mais alta ("pico").



No exemplo acima, a classe modal é a que vai de 92.3 até 97.9. Assim, os dados repousam, tipicamente entre esses valores. Entretanto, temos dois picos.

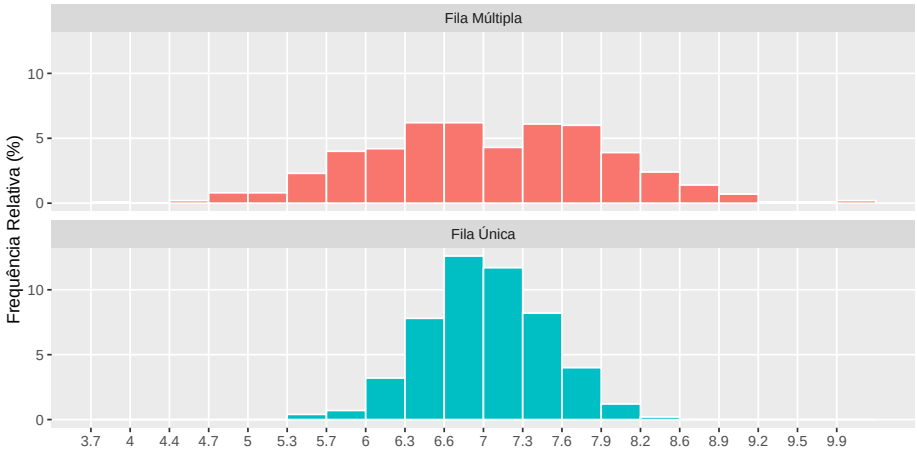
Geralmente, um histograma bimodal indica a existência de dois grupos, com valores centrados em dois pontos diferentes do eixo de valores. Uma distribuição de frequências pode também ser amodal, ou seja, todos os valores são igual-

mente frequentes. Ou também unimodal, quando os valores estão concentrados somente em um ponto/classe.

2.5.2 Variabilidade

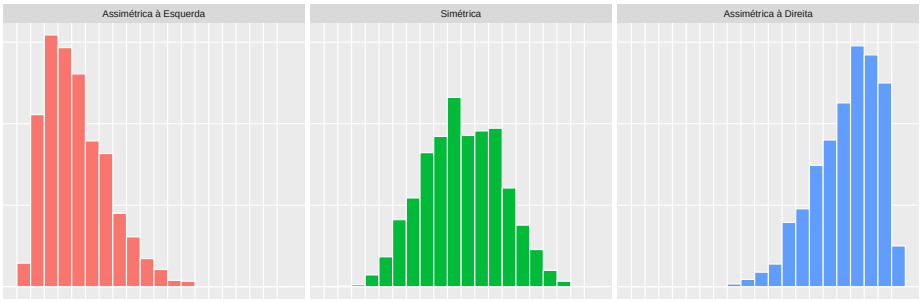
Para descrever adequadamente a distribuição de frequências de uma variável quantitativa, além da informação do valor representativo da variável (tendência central), é necessário dizer também o quanto estes valores variam, ou seja, o quão dispersos eles são.

De fato, somente a informação sobre a tendência central de um conjunto de dados não consegue representá-lo adequadamente. A Figura abaixo mostra um histograma para os tempos de espera de 1000 clientes de dois bancos, um com fila única e outro com fila múltipla, com o mesmo número de atendentes. Os tempos de espera nos dois bancos têm a mesma tendência central de 7 minutos. Entretanto, os dois conjuntos de dados são claramente diferentes, pois os valores são muito mais dispersos no banco com fila múltipla. Assim, quando entramos num fila única, esperamos ser atendidos em cerca de 7 minutos, com uma variação de, no máximo, meio minuto a mais ou a menos. Na fila múltipla, a variação é maior, indicando-se que tanto pode-se esperar muito mais ou muito menos que o valor típico de 7 minutos.



2.5.3 Forma

A distribuição de frequências de uma variável pode ter várias formas, mas existem três formas básicas, apresentadas na figura abaixo através de histogramas.



Quando uma distribuição é simétrica em torno de um valor (o mais frequente), significa que as observações estão igualmente distribuídas em torno desse valor (metade acima e metade abaixo).

A assimetria de uma distribuição pode ocorrer de duas formas:

- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas preferem se referir ao lado onde está "faltando" dados (cauda). As duas denominações são alternativas.

Em alguns casos, apenas o conhecimento da forma da distribuição de frequências de uma variável já nos fornece uma boa informação sobre o comportamento dessa variável. Por exemplo, o que você acharia se soubesse que a distribuição de frequências das notas da primeira prova da disciplina de Estatística que você está cursando é, geralmente, assimétrica com concentração à direita? Como você acha que é a forma da distribuição de frequências da renda no Brasil?

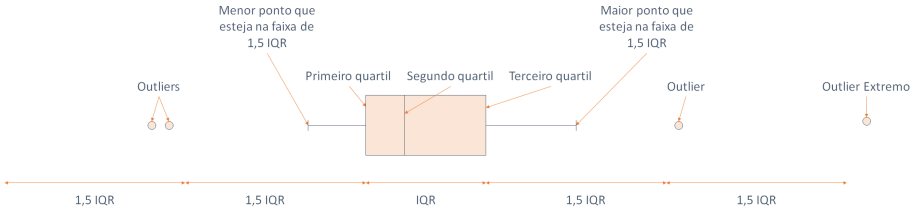
2.6 Boxplot

O Boxplot é um gráfico proposto para a detecção de valores discrepantes (outliers), que são aqueles valores muito diferentes do restante do conjunto de dados.

Esses valores discrepantes podem representar erros no processo de coleta ou de processamento dos dados, e, nesse caso, devem ser corrigidos ou excluídos do banco de dados. No entanto, os outliers podem ser valores corretos, que, por alguma razão, são muito diferentes dos demais valores. Nesse caso, a análise desses dados deve ser cuidadosa, pois, como sabemos, algumas estatísticas descritivas, como a média e o desvio-padrão, são influenciadas por valores extremos.

Na construção do Boxplot, utilizamos alguns percentis (mediana, primeiro e terceiro quartis), que são pouco influenciados por valores extremos. Além disso, precisamos saber quais são os valores mínimo e máximo do conjunto de dados.

O Boxplot é constituído por uma caixa atravessada por uma linha, construído usando um eixo com uma escala de valores, como mostra a figura abaixo. Como sabemos, entre o primeiro e o terceiro quartis, temos 50% dos dados. Podemos pensar, então, que essa caixa contém metade dos dados do conjunto.



Como um gráfico tem que representar todos os valores do conjunto de dados, precisamos representar os outros 50%, sendo 25% abaixo do Q1 e 25% acima do Q3. Esses valores serão representados pelas duas linhas que saem das extremidades da caixa. Cada uma das linhas é traçada, a partir das extremidades da caixa, até que encontre o valor máximo ou mínimo; ou atinja o comprimento máximo de 1,5 vezes a altura da caixa (IQR). O que acontecer primeiro.

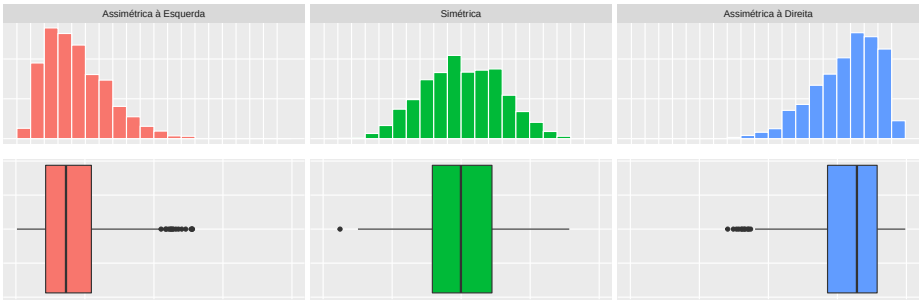
No exemplo da imagem acima o segundo caso aconteceu, assim os valores que ainda não foram representados devem ser devidamente marcados em suas respectivas posições na escala de valores. Esses valores são considerados outliers pelo critério do boxplot. Obviamente, o limite superior do boxplot não coincidiu com o valor máximo do conjunto de dados, que foi considerado um valor discrepante (outlier).

Além da detecção de valores discrepantes, o boxplot pode ser muito útil na análise da distribuição dos valores de um conjunto de dados. Através do boxplot, podemos:

- identificar a forma da distribuição (simétrica ou assimétrica);
- avaliar e comparar a tendência central (mediana) de dois ou mais conjuntos de dados;
- comparar a variabilidade de dois ou mais conjuntos de dados

Para avaliar a forma da distribuição, devemos observar o deslocamento da caixa em relação a linha do boxplot. Lembrando que a caixa do boxplot contém 50% dos dados, o seu deslocamento na linha nos informa onde estão concentrados os dados.

Se a caixa está mais deslocada para um dos lados da linha, significa que metade dos dados estão concentrados naquele lado da escala de valores e, assim, a distribuição é assimétrica. Se a caixa está praticamente no meio da linha, dividindo-a em duas partes iguais, é distribuição será considerada simétrica.

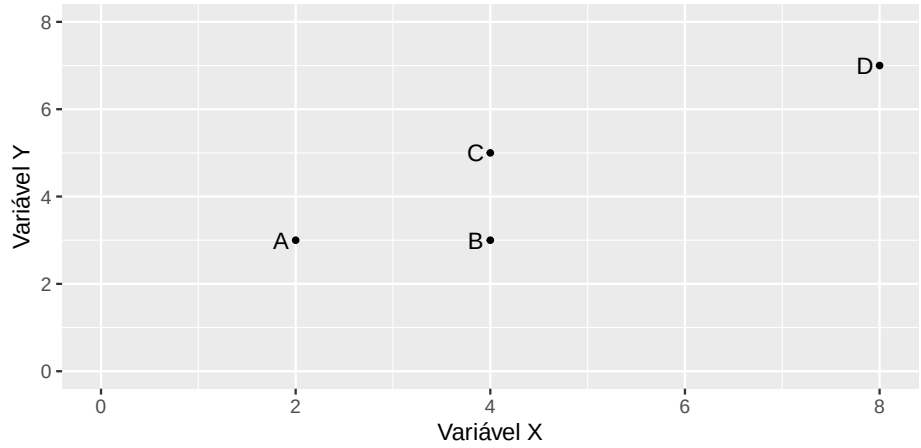


2.7 Diagrama de Dispersão

O diagrama de dispersão é um gráfico onde pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada indivíduo do conjunto de dados.

O Quadro e a Figura abaixo mostram um esquema do desenho do diagrama de dispersão. Neste exemplo, foram medidos os valores de duas variáveis quantitativas, X e Y, em quatro indivíduos. O eixo horizontal do gráfico representa a variável X e o eixo vertical representa a variável Y.

Indivíduos	Variável X	Variável Y
A	2	3
B	4	3
C	4	5
D	8	7



O diagrama de dispersão é usado principalmente para visualizar a relação/associação entre duas variáveis, mas também para é muito útil para:

- Comparar o efeito de dois tratamentos no mesmo indivíduo

Planta	Quantidade												
1001	5	15.15	15.45	15.63	15.65	16.38	NA	NA	NA	NA	NA	NA	NA
1002	6	14.00	14.50	15.35	15.86	15.94	16.13	NA	NA	NA	NA	NA	NA
1003	7	13.67	13.76	14.06	14.11	14.54	14.89	15.50	NA	NA	NA	NA	NA
1004	8	11.00	11.50	12.39	12.39	12.90	14.50	15.50	16.56	NA	NA	NA	NA
1005	9	10.24	11.12	12.05	12.37	13.48	13.80	14.04	15.39	16.00	NA	NA	NA
1006	10	9.00	9.32	10.67	11.56	11.67	12.56	12.83	12.84	13.43	15.09	NA	NA
1007	11	7.82	8.56	8.74	9.57	11.08	11.92	12.13	12.50	14.14	14.20	14.00	NA
1008	12	7.25	9.41	10.15	10.33	10.80	10.95	11.13	11.48	11.49	12.86	13.37	15.04
1009	13	6.95	7.61	8.53	10.00	10.94	11.04	11.43	11.63	11.97	12.02	12.74	13.53
1010	14	7.00	8.00	9.00	10.00	10.00	10.50	11.00	11.16	11.17	11.70	12.45	12.89

- Verificar o efeito tipo antes/depois de um tratamento

A seguir, veremos dois exemplos da utilização do diagrama de dispersão. O primeiro refere-se ao estudo da associação entre duas variáveis. O segundo utiliza o diagrama de dispersão para comparar o efeito da aplicação de um tratamento, comparando as medidas antes e depois da medicação.

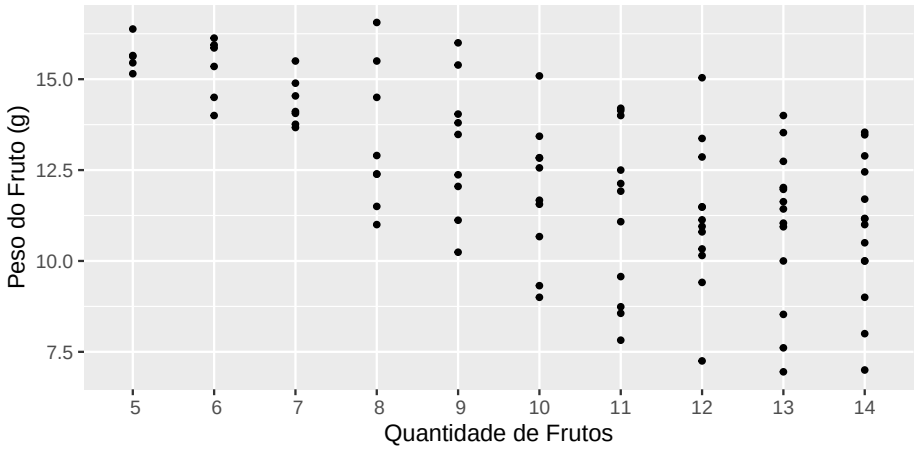
2.7.1 Exemplo 1

Um produtor de morangos para exportação deseja produzir frutos grandes, pois frutos pequenos têm pouco valor mesmo no mercado interno. Além disso, os frutos, mesmo grandes, não devem ter tamanhos muito diferentes entre si. O produtor suspeita que uma dos fatores que altera o tamanho dos frutos é o número de frutos por árvore.

Para investigar a relação entre o número de frutos que uma planta produz e o peso destes frutos, ele observou dados de 10 morangueiros na primeira safra (Quadro abaixo) e gerou o Diagrama de Dispersão apresentado abaixo.

```
library(ggplot2)
library(tidyr)
aux = matrix(c(15.15,15.45,15.63,15.65,16.38,NA,NA,NA,NA,NA,NA,NA,NA,NA,
               14,14.5,15.35,15.86,15.94,16.13,NA,NA,NA,NA,NA,NA,NA,NA,
               13.67,13.76,14.06,14.11,14.54,14.89,15.5,NA,NA,NA,NA,NA,NA,NA,
               11,11.5,12.39,12.39,12.9,14.5,15.5,16.56,NA,NA,NA,NA,NA,NA,
               10.24,11.12,12.05,12.37,13.48,13.8,14.04,15.39,16,NA,NA,NA,NA,NA,
               9,9.32,10.67,11.56,11.67,12.56,12.83,12.84,13.43,15.09,NA,NA,NA,NA,
               7.82,8.56,8.74,9.57,11.08,11.92,12.13,12.5,14.14,14.2,14,NA,NA,NA,
               7.25,9.41,10.15,10.33,10.8,10.95,11.13,11.48,11.49,12.86,13.37,15.04,NA
               6.95,7.61,8.53,10,10.94,11.04,11.43,11.63,11.97,12.02,12.74,13.53,14,NA
               7,8,9,10,10,10.5,11,11.16,11.17,11.7,12.45,12.89,13.47,13.54),
             nrow=10,ncol = 14,byrow = T)
dados = data.frame(ID_Morangueiro = 1001:1010,
                   Qtd_Frutos = 5:14,aux)
dados %>%
  pivot_longer(starts_with("X")) %>%
  ggplot(aes(x = Qtd_Frutos, y = value)) +
  geom_point(size = 1) +
```

```
  scale_x_discrete(limits=5:14) +
  labs(x = "Quantidade de Frutos",y = "Peso do Fruto (g)")
```



O diagrama de dispersão mostra-nos dois fatos. O primeiro, que há um decréscimo no valor médio do peso do fruto por árvore à medida que cresce o número de frutos na árvore. Ou seja, não é vantagem uma árvore produzir muitos frutos, pois eles tenderão a ser muito pequenos.

O segundo fato que percebemos é que, com o aumento no número de frutos na árvore, cresce também a variabilidade no peso, gerando tanto frutos muito grandes, como muito pequenos.

Assim, conclui-se que não é vantagem ter poucas plantas produzindo muito frutos, mas sim muitas plantas produzindo poucos frutos, mas grandes e uniformes. Uma análise mais detalhada poderá determinar o número ideal de frutos por árvore, aquele que maximiza o peso médio e, ao mesmo tempo, minimiza a variabilidade do peso.

2.7.2 Exemplo 2

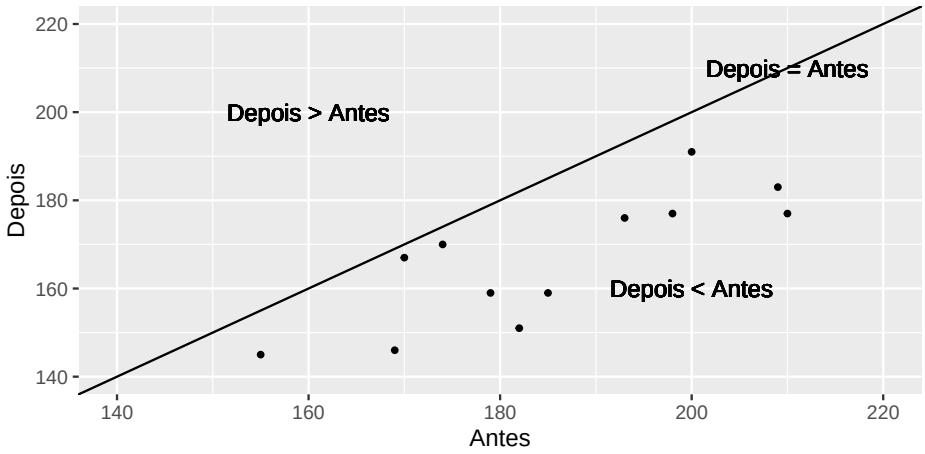
Captopril é um remédio destinado a baixar a pressão sistólica. Para testar seu efeito, ele foi ministrado a 12 pacientes, tendo sido medida a pressão sistólica antes e depois da medicação.

Paciente	Antes	Depois
A	200	191
B	174	170
C	198	177
D	170	167
E	179	159
F	182	151
G	193	176
H	209	183
I	185	159
J	155	145
K	169	146
L	210	177

Os mesmos indivíduos foram utilizados nas duas amostras (Antes/depois). Assim, é natural compararmos a pressão sistólica para cada indivíduo, comparando a pressão sistólica depois e antes. Para todos os pacientes, a pressão sistólica depois do Captopril é menor do que antes da medicação. Mas como podemos "ver" se estas diferenças são grandes? Com a ajuda do diagrama de dispersão mostrado na figura abaixo.

```
dados = data.frame(Paciente = LETTERS[1:12],
                   Antes = c(200, 174, 198, 170, 179, 182, 193, 209, 185, 155, 169, 210
                   Depois = c(191, 170, 177, 167, 159, 151, 176, 183, 159, 145, 146, 17

dados %>%
  ggplot(aes(x = Antes, y = Depois)) +
  geom_point(size = 1) +
  ylim(140,220) + xlim(140,220) +
  geom_abline(slope = 1, intercept = 0) +
  geom_text(aes(x = 200, y = 160), label = "Depois < Antes") +
  geom_text(aes(x = 160, y = 200), label = "Depois > Antes") +
  geom_text(aes(x = 210, y = 210), label = "Depois = Antes")
```



Cada ponto no diagrama de dispersão corresponde às medidas de pressão sistólica de um paciente, medida antes e depois da medicação. A linha marcada no diagrama corresponde à situação onde a pressão sistólica não se alterou depois do paciente tomar o Captopril. Veja que todos os pontos estão abaixo desta linha, ou seja para todos os pacientes o Captopril fez efeito. Grande parte destes pontos está bem distante da linha, mostrando que a redução na pressão sistólica depois do uso do medicamento não foi pequena.

2.8 Séries Temporais

Séries temporais (ou séries históricas) são um conjunto de observações de uma mesma variável quantitativa (discreta ou contínua) feitas ao longo do tempo. O conjunto de novos casos da COVID-19 é um exemplo de série temporal.

Um dos objetivos do estudo de séries temporais é conhecer o comportamento da série ao longo do tempo (aumento, estabilidade ou declínio dos valores). Em alguns estudos, esse conhecimento pode ser usado para se fazer previsões de valores futuros com base no comportamento dos valores passados.

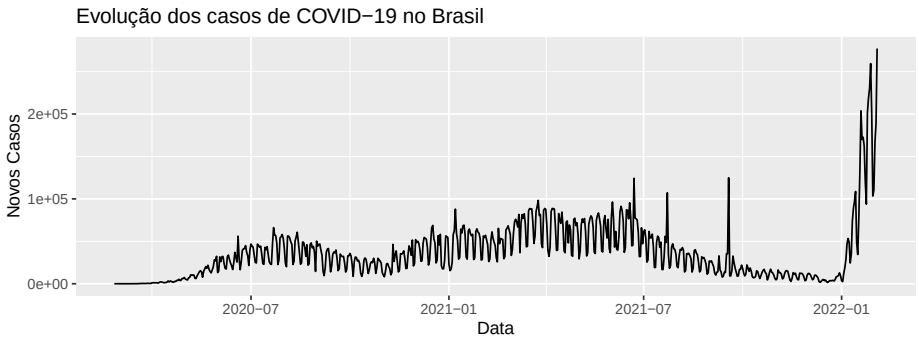
A representação gráfica de uma série temporal é feita através do gráfico de linha, como exemplificado nas figuras abaixo. No eixo horizontal do gráfico de linha, está o indicador de tempo e, no eixo vertical, a variável a ser representada.

```
library(tsibble)
library(lubridate)
library(dplyr)

dados = read.csv("https://raw.githubusercontent.com/wcota/covid19br/master/cases-brazil-states.cs

dados %>%
  mutate(date = as_date(date)) %>%
  as_tsibble(key = state, index = date) %>%
```

```
  filter(state == "TOTAL") %>%
  select(date, newCases) %>%
  ggplot(aes(x = date, y = newCases)) +
  geom_line() +
  labs(title = "Evolução dos casos de COVID-19 no Brasil",
       y = "Novos Casos", x = "Data")
```



De maneira geral, um gráfico de linhas deve ser construído de modo que:

- O início do eixo vertical seja o valor mínimo possível para a variável que está sendo representada, para evitar as distorções
- O final do eixo vertical seja tal que a série fica centrada em relação ao eixo vertical
- Os tamanhos dos eixos sejam o mais parecidos possível

2.9 Gráfico de Radar

O gráfico de radar é um diagrama e/ou gráfico que consiste de uma sequência de raios equi-angulares, com cada raio representando uma das variáveis. O comprimento de cada raio é proporcional à magnitude da variável para o ponto de dados em relação à máxima magnitude da variável em todos os pontos. Uma linha é desenhada ligando os valores de cada raio. Isso dá ao diagrama uma aparência de estrela, o que deu origem a um dos nomes mais populares para este gráfico. O gráfico de estrela pode ser usado para responder as seguintes questões:

- Que observações são mais semelhantes, por exemplo, existem clusters de observações?
- Existem exceções?

Gráficos de radar oferecem uma maneira útil de exibir observações multivariáveis com um número arbitrário de variáveis. Cada estrela representa uma única observação. Normalmente, os gráficos de radar são gerados em um formato multi-diagrama com muitas estrelas em cada página, cada estrela repre-

sentando uma observação. É um pouco mais fácil de ver padrões em dados se as observações forem organizadas em alguma ordem não-arbitrária (se as variáveis forem atribuídas aos raios da estrela em uma ordem significativa).

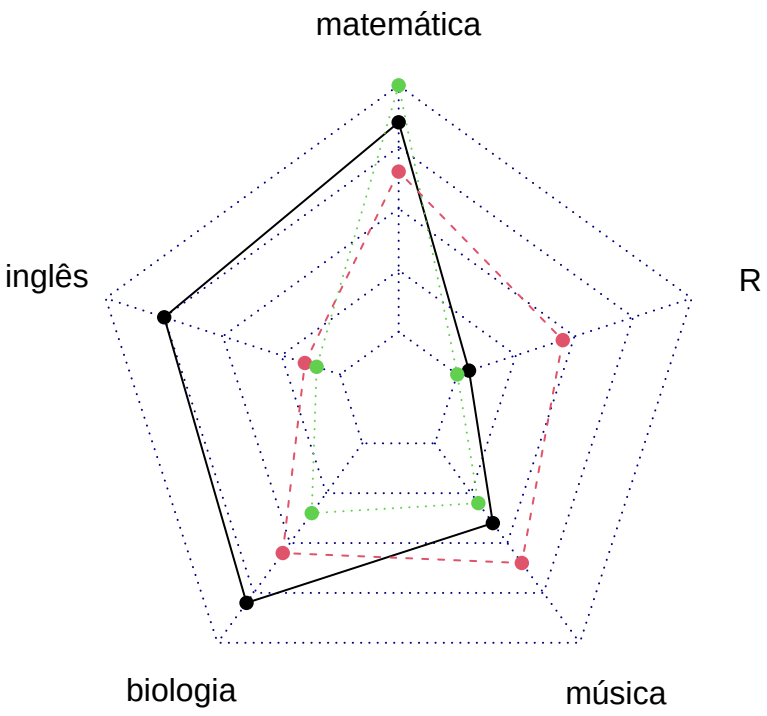
Uma aplicação de gráficos de radar é o controle de melhoria de qualidade para apresentar as métricas de desempenho de qualquer programa em curso. Eles também são usados em esportes para representar os pontos fortes e fracos de jogadores, onde eles são geralmente chamados de gráficos de aranha. No exemplo abaixo vemos a performance de três alunos (a,b,c) em relação as dimensões matemática, inglês, biologia, música e R.

```
library(fmsb)

dados = as.data.frame(matrix(sample(0:20, 15, replace=F), ncol=5))
colnames(dados) = c("matemática", "inglês", "biologia", "música" ,"R")
rownames(dados) = paste("aluno", letters[1:3])

dados = rbind(rep(20,5), rep(0,5), dados)

radarchart(dados)
```



Chapter 3

Conceitos Básicos

3.1 Introdução

No estudo da inferência estatística, o objetivo principal é obter informações sobre uma população a partir das informações de uma amostra e aqui vamos precisar de definições mais formais de população e amostra. Para facilitar a compreensão destes conceitos, apresentamos o exemplo abaixo a título de ilustração

Em um estudo antropométrico em nível nacional, uma amostra de 5000 adultos é selecionada dentre os adultos brasileiros e uma das variáveis de estudo é a altura.

Neste exemplo, a população é o conjunto de todos os brasileiros adultos. No entanto, o interesse (um deles, pelo menos) está na altura dos brasileiros. Assim, nesse estudo, a cada sujeito da população associamos um número correspondente à sua altura. Se determinado sujeito é sorteado para entrar na amostra, o que nos interessa é esse número, ou seja, sua altura.

Como sabemos, essa é a definição de variável aleatória: uma função que associa a cada ponto do espaço amostral um número real. Dessa forma, a nossa população pode ser representada pela variável aleatória $X =$ "altura do adulto brasileiro". Como essa é uma v.a. contínua, a ela está associada uma função de densidade de probabilidade f e da literatura, sabemos que é razoável supor que essa densidade seja a densidade normal. Assim, nossa população, nesse caso, é representada por uma v.a. $X \sim N(\mu; \sigma^2)$. Conhecendo os valores de μ e σ teremos informações completas sobre a nossa população.

Uma forma de obtermos os valores de μ e σ é medindo as alturas de todos os brasileiros adultos. Mas esse seria um procedimento caro e demorado. Uma solução, então, é retirar uma amostra (subonjunto) da população e estudar essa amostra. Suponhamos que essa amostra seja retirada com reposição e que os sorteios sejam feitos de forma independente, isto é, o resultado de cada ex-

tração não altera o resultado das demais extrações. Ao sortearmos o primeiro elemento, estamos realizando um experimento que dá origem à v.a. X_1 = altura do primeiro elemento; o segundo elemento dá origem à v.a. X_2 = "altura do segundo elemento" e assim por diante. Como as extrações são feitas com reposição, todas as v.a. $X_1, X_2, ...$ têm a mesma distribuição, que reflete a distribuição da altura de todos os brasileiros adultos. Para uma amostra específica, temos os valores observados $x_1, x_2, ...$ dessas variáveis aleatórias.

3.2 População

A inferência estatística trata do problema de se obter informação sobre uma população a partir de uma amostra. Embora a população real possa ser constituída de pessoas, empresas, animais etc., as pesquisas estatísticas buscam informações sobre determinadas características dos sujeitos, características essas que podem ser representadas por números. Sendo assim, a cada sujeito da população está associado um número, o que nos permite apresentar a seguinte definição.

População: *A população de uma pesquisa estatística pode ser representada por uma variável aleatória X que descreve a característica de interesse.*

Os métodos de inferência nos permitirão obter estimativas dos parâmetros (característica de interesse) de tal variável aleatória, que pode ser contínua ou discreta.

3.3 Amostra

Embora existam vários métodos de seleção de amostras, nosso foco é a amostragem aleatória simples. Segundo tal método, toda amostra de mesmo tamanho n tem igual chance (probabilidade) de ser sorteada. É possível extrair amostras aleatórias simples com e sem reposição. No entanto, para populações grandes - ou infinitas - extrações com e sem reposição não levam a resultados muito diferentes. Assim, no estudo da Inferência Estatística, estaremos lidando sempre com amostragem aleatória simples com reposição. Este método de seleção atribui a cada elemento da população a mesma probabilidade de ser selecionado e esta probabilidade se mantém constante ao longo do processo de seleção da amostra (se as extrações fossem sem reposição isso não aconteceria). No restante desse curso omitiremos a expressão "com reposição", ou seja, o termo amostragem (ou amostra) aleatória simples sempre se referirá à amostragem com reposição. Por simplicidade, muitas vezes abreviaremos o termo amostra aleatória simples por aas.

Uma forma de se obter uma amostra aleatória simples é escrever os números ou nomes dos elementos da população em cartões iguais, colocar estes cartões em uma urna misturando-os bem e fazer os sorteios necessários, tendo o cuidado de colocar cada cartão sorteado na urna antes do próximo sorteio. Na prática, em

geral são usados programas de computador, uma vez que as populações tendem a ser muito grandes.

Agora vamos formalizar o processo de seleção de uma amostra aleatória simples, de forma a relacioná-lo com os problemas de inferência estatística que iremos estudar. Seja uma população representada por uma variável aleatória X. De tal população será sorteada uma amostra aleatória simples com reposição de tamanho n. Como visto nos exemplos anteriores, cada sorteio dá origem a uma variável aleatória X_i e, como os sorteios são com reposição, todas essas variáveis têm a mesma distribuição de X. Isso nos leva à seguinte definição.

Amostra: *Uma amostra aleatória simples (aas) de tamanho n de uma v.a. X (população) é um conjunto de n v.a.* $X_1, X_2, ..., X_n$ *independentes e identicamente distribuídas (i.i.d.).*

É interessante notar a convenção usual: o valor observado de uma v.a. X é representado pela letra minúscula correspondente. Assim, depois do sorteio de uma aas de tamanho n, temos valores observados $x_1, x_2, ..., x_n$ das respectivas variáveis aleatórias.

3.4 Estatísticas e Parâmetros

Obtida uma aas, é possível calcular diversas características desta amostra, como, por exemplo, a média, a mediana, a variância, etc. Qualquer uma destas características é uma função de $X_1, X_2, ..., X_n$ e, portanto, o seu valor depende da amostra sorteada. Sendo assim, cada uma dessas características ou funções é também uma v.a.. Por exemplo, a média amostral é a v.a. definida por

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}.$$

Temos, então, a seguinte definição:

Estatística: *Uma estatística amostral ou estimador T é qualquer função da amostra* $X_1, X_2, ..., X_n$*, isto é,* $T = g(X_1, X_2, ..., X_n)$ *onde g é uma função qualquer.*

As estatísticas amostrais que estaremos considerando neste curso são:

- média amostral: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- variância amostral: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Para uma amostra específica, o valor obido para o estimador será denominado estimativa e, em geral, serão representadas por letras minúsculas. Por exemplo, temos as seguintes notações correspondentes à média amostral e à variância: $\bar{x}$ e s^2.

Outras estatísticas possíveis são o mínimo amostral, o máximo amostral, a amplitude amostral, etc.

De forma análoga, temos as características de interesse da população. No entanto, para diferenciar entre as duas situações (população e amostra), atribuimos nomes diferentes.

Parâmetro: *Um parâmetro é uma característica da população.*

Assim, se a população é representada pela v.a. X, alguns parâmetros são a esperança $E(X)$ e a variância $V(X)$. Com relação às características mais usuais, vamos usar a seguinte notação:

Característica	Parâmetro (população)	Estatística (amostra)
Média	μ	$\bar{X}$
Variância	σ^2	S^2
Número de elementos	N	n

3.5 Distribuições Amostrais

Nos problemas de inferência, estamos interessados em estimar um parâmetro θ da população (por exemplo, a média populacional) através de uma aas $X_1, X_2, \ldots, X_n$. Para isso, usamos uma estatística T (por exemplo, a média amostral) e, com base no valor obtido para T a partir de uma particular amostra, iremos tomar as decisões que o problema exige. Já foi dito que T é uma v.a., uma vez que depende da amostra sorteada; amostras diferentes fornecerão diferentes valores para T.

Consideremos o seguinte exemplo, onde nossa população é o conjunto $\{1, 3, 6, 8\}$, isto é, este é o conjunto dos valores da característica de interesse da população em estudo. Assim, para esta população, ou seja, para essa v.a. X temos $E(X) = 4.5$ e $V(X) = 7.25$.

Suponha que dessa população iremos extrair uma aas de tamanho 2 e a estatística que iremos calcular é a média amostral. Algumas possibilidades de amostra são $\{1, 1\}$, $\{1, 3\}$, $\{6, 8\}$, para as quais os valores da média amostral são 1, 2 e 7, respectivamente. Podemos ver, então, que há uma variabilidade nos valores da estatística e, assim, seria interessante que conhecêssemos tal variabilidade. Conhecendo tal variabilidade, temos condições de saber "quão infelizes" podemos ser no sorteio da amostra. No exemplo acima, as amostras $\{1, 1\}$ e $\{8, 8\}$ são as que têm média amostral mais afastada da verdadeira média populacional. Se esses valores tiverem chance muito mais alta do que os valores mais próximos de $E(X)$, podemos ter sérios problemas.

Para conhecer o comportamento (distribuição) da média amostral, teríamos que conhecer todos os possíveis valores de X, o que equivaleria a conhecer todas as possíveis amostras de tamanho 2 de tal população. Nesse exemplo, como só temos 4 elementos na população, a obtenção de todas as aas de tamanho 2 não é difícil. No entanto, para um número um pouco maior dele elementos, ou

quando não conhecemos todo o espaço amostral, essa tarefa se torna complexa.

Distribuição de Probabilidade: *A função de distribuição amostral de uma estatística T é a função de distribuição de probabilidades de T ao longo de todas as possíveis amostras de tamanho n.*

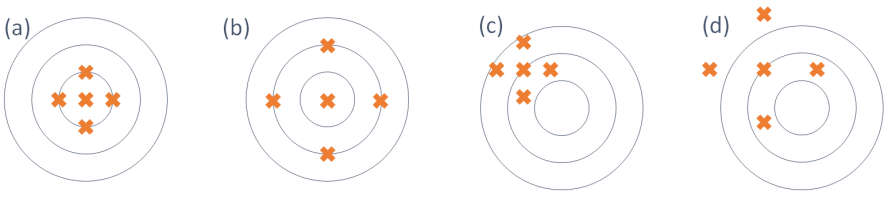
Em outras palavras, uma distribuição amostral é a distribuição de probabilidades de uma medida estatística baseada em uma amostra aleatória. Ao retirar uma amostra aleatória de uma população estaremos considerando cada valor da amostra como um valor de uma variável aleatória cuja distribuição de probabilidade é a mesma da população no instante da retirada desse elemento para a amostra. Em consequência do fato de os valores de amostra serem aleatórios, decorre que qualquer quantidade calculada em função dos elementos da amostra também será uma variável aleatória.

Podemos ver que a obtenção da distribuição amostral de qualquer estatística T é um processo tão ou mais complicado do que trabalhar com a população inteira. Na prática, o que temos é uma única amostra e com esse resultado é que temos que tomar as decisões pertinentes ao problema em estudo. Esta tomada de decisão, no entanto, será facilitada se conhecermos resultados teóricos sobre o comportamento da distribuição amostral.

3.6 Propriedades dos Estimadores

Vimos anteriormente que, dada uma população, existem muitas e muitas aas de tamanho n que podem ser sorteadas. Cada uma dessas amostras resulta em um valor diferente da estatística de interesse ($\bar{X}$ e S^2, por exemplo). O que esses resultados estão mostrando é como esses diferentes valores se comportam em relação ao verdadeiro (mas desconhecido) valor do parâmetro.

Considere a Figura abaixo, onde o alvo representa o valor do parâmetro e os "tiros", indicados pelos símbolo x, representam os diferentes valores amostrais da estatística de interesse.



Nas partes (a) e (b) da figura, os tiros estão em torno do alvo, enquanto nas partes (c) e (d) isso não acontece. Comparando as partes (a) e (b), podemos ver que na parte (a) os tiros estão mais concentrados em torno do alvo, isto é, têm menor dispersão. Isso reflete uma pontaria mais certeira do atirador em (a). Analogamente, nas partes (c) e (d), embora ambos os atiradores estejam com a mira deslocada, os tiros do atirador (c) estão mais concentrados em torno de um alvo; o deslocamento poderia até ser resultado de um desalinhamento da

arma. Já o atirador (d), além de estar com o alvo deslocado, ele tem os tiros mais espalhados, o que reflete menor precisão.

Traduzindo esta situação para o contexto de estimadores e suas propriedades, temos o seguinte: nas partes (a) e (b), temos dois estimadores que fornecem estimativas centradas em torno do verdadeiro valor do parâmetro, ou seja, as diferentes amostras fornecem valores distribuídos em torno do verdadeiro valor do parâmetro. A diferença é que em (b) esses valores estão mais dispersos e, assim, temos mais chance de obter uma amostra "infeliz", ou seja, uma amostra que forneça um resultado muito afastado do valor do parâmetro. Essas duas propriedades estão associadas à esperança e à variância do estimador, que são medidas de centro e dispersão, respectivamente. Nas partes (c) e (d), as estimativas estão centradas em torno de um valor diferente do parâmetro de interesse e na parte (d), a dispersão é maior.

Temos, assim, ilustrados os seguintes conceitos:

Viés: *Um estimador T é dito um estimador não-viesado do parâmetro θ se $E(T) = \theta$.*

Essa esperança é calculada ao longo de todas as possíveis amostras, ou seja, é a esperança da distribuição amostral de T. Nas partes (a) e (b) da Figura os estimadores são não-viesados e nas partes (c) e (d), os estimadores são viesados.

Com relação aos estimadores $barX$, S^2 e $\hat{\sigma}^2$, veremos formalmente que os dois primeiros são não-viesados para estimar a média e a variância populacionais, respectivamente, enquanto $\hat{\sigma}^2$ é viesado para estimar a variância populacional. Essa é a razão para se usar S^2, e não $\hat{\sigma}^2$.

Eficiência: *Se T_1 e T_2 são dois estimadores não-viesados do parâmetro θ, diz-se que T_1 é mais eficiente que T_2 se $V(T_1) < V(T_2)$.*

Na Figura, o estimador da parte (a) é mais eficiente que o estimador da parte (b). Uma outra propriedade dos estimadores está relacionada à idéia bastante intuitiva de que à medida que se aumenta o tamanho da amostra, mais perto devemos ficar do verdadeiro valor do parâmetro.

Consistência: *Um estimador T de um parâmetro θ é consistente se:*

- $\lim_{n\to\infty} E(\hat{\theta}) = \theta$
- $\lim_{n\to\infty} V(\hat{\theta}) = 0$

Chapter 4

Distribuições Amostrais

4.1 Introdução

Uma distribuição amostral é a distribuição de probabilidades de uma medida estatística baseada em uma amostra aleatória. Ao retirar uma amostra aleatória de uma população estaremos considerando cada valor da amostra como um valor de uma variável aleatória cuja distribuição de probabilidade é a mesma da população no instante da retirada desse elemento para a amostra. Em consequência do fato de os valores de amostra serem aleatórios, decorre que qualquer quantidade calculada em função dos elementos da amostra também será uma variável aleatória

4.2 Distribuição Amostral da Média Amostral

Considere $X_1, ..., X_n$ uma amostra aleatória de uma distribuição normal com média μ e variância σ^2.

Tomamos, por exemplo, o problema de estimar quantas horas adicionais de sono são garantidas a um indivíduo após ingerir uma determinada droga. Além disso, suponha que a droga é testada em 20 indivíduos de modo que a média amostral seja $\bar{X} = 0,8$ horas. Porém, se o estudo for repetido com outros 20 participantes podemos ter outros resultados para a média amostral. Por exemplo, podemos ter $\bar{X} = 1,3$. E, repetindo o estudo novamente, poderíamos ter $\bar{X} = -0,2$. Em termos estatísticos, haverá variação entre as médias amostrais.

Este problema poderia ser resolvido se repetíssemos o estudo infinitas vezes, porém isto é inviável.

Quando as observações são amostradas aleatoriamente de uma distribuição normal, a média amostral também tem uma distribuição normal. Isto é, quando n observações são amostradas aleatoriamente de uma distribuição normal com

média μ e variância σ^2, a média amostral tem distribuição normal com média μ e variância $\frac{\sigma^2}{n}$. Ou seja, se

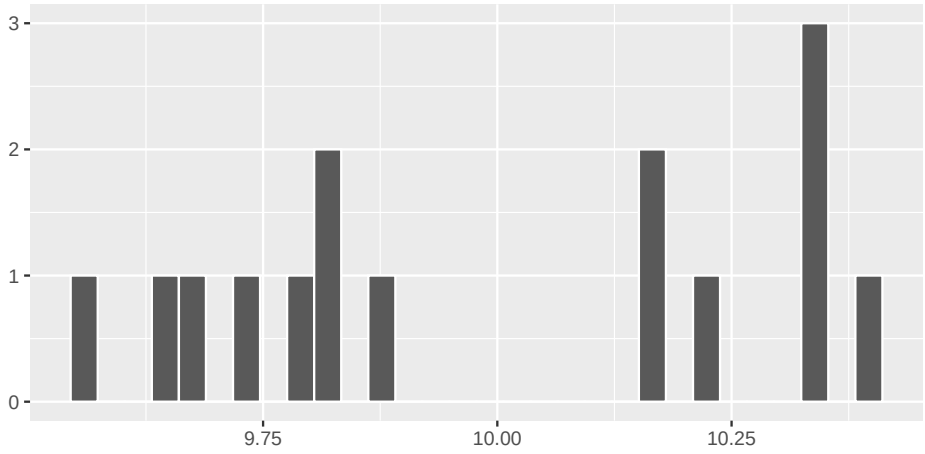
$$X \sim N(\mu, \sigma^2), \text{ então} \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

4.2.1 Visualizando

Considere uma população normal com média $\mu = 10$ e variância $\sigma^2 = 4$. Vamos realizar um estudo de simulação para a distribuição da média amostral considerando amostras de tamanho 20 dessa população.

Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.

```
library(ggplot2)
n = 15
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```

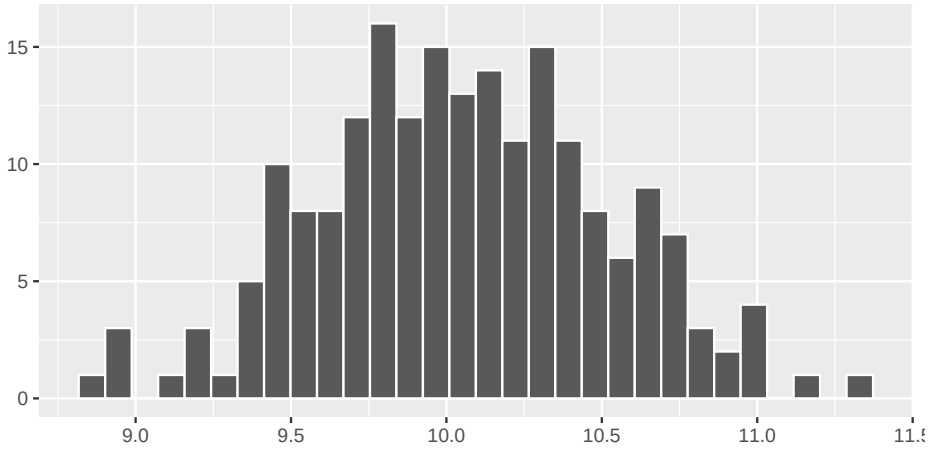


Nessa simulação obtivemos $\bar{x} = 9.99$ e $s = 0.3$.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 200 amostras.

```
n = 200
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
```

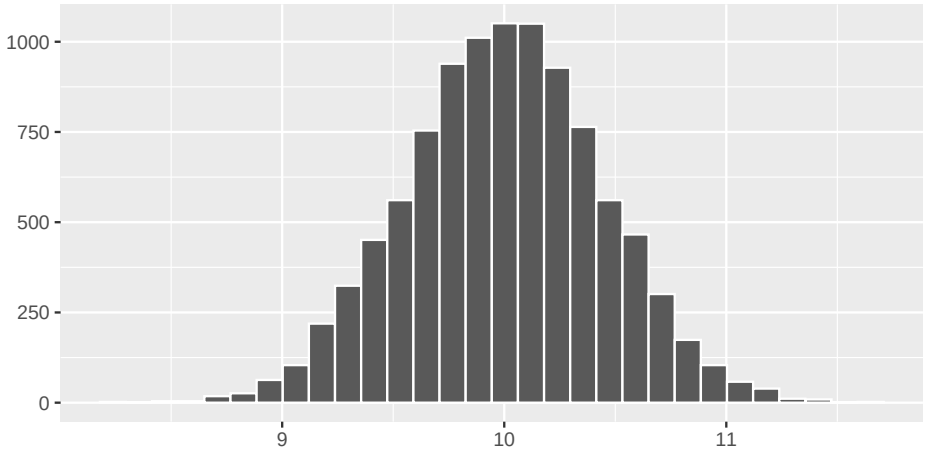
```
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```



Agora, obtivemos $\bar{x} = 10.05$ e $s = 0.46$.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da média amostral pode ser vista segundo o histograma abaixo.

```
n = 10000
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```



Para este caso, a média das médias amostrais foi $\bar{x} = 10$ e o desvio padrão foi $s = 0.44$. Então, empiricamente, podemos perceber que a distribuição da média amostral se aproxima de uma distribuição normal com média $\mu = 10$ e desvio padrão $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{20}} = 0,4472$.

4.2.2 Prova

Para provar que se $X \sim N(\mu, \sigma^2)$, então $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, vamos precisar de somente três passos.

1. Dado que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, onde $X_1, ..., X_n$ é uma amostra aleatória de tamanho n de uma população Normal sabemos que combinação linear de Normais resulta em também uma distribuição Normal, portanto daqui temos que $\bar{X} \sim Normal$;
2. Se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, então

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{n\mu}{n} = \mu$$

3. Por fim, se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, então como os X_i's são independentes

$$V\left[\bar{X}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V\left[X_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

4.3 Teorema Central do Limite

Os resultados vistos anteriormente são válidos para populações normais, isto é, se uma população é normal com média μ e variância σ^2, então a distribuição amostral de $\bar{X}$ é também normal com média μ e variância $\frac{\sigma^2}{n}$, onde n é o tamanho da amostra. O Teorema Central do Limite (TCL) nos fornece um

resultado análogo para qualquer distribuição populacional, desde que o tamanho da amostra seja suficientemente grande.

Seja $X_1, \ldots, X_n$ uma amostra aleatória simples de uma população X tal que $E(X) = \mu$ e $V(X) = \sigma^2$. Então, a distribuição de $\bar{X}$ converge para a distribuição Normal com média μ e variância $\frac{\sigma^2}{n}$ quando $n \to \infty$. Equivalentemente,

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0, 1)$$

A interpretação prática do TCL é a seguinte: para amostras "grandes" de qualquer população, podemos aproximar a distribuição amostral de $\bar{X}$ por uma distribuição normal com a mesma média populacional e variância igual à variância populacional dividida pelo tamanho da amostra.

Quão grande deve ser a amostra para se obter uma boa aproximação depende das características da distribuição populacional. Se a distribuição populacional não se afastar muito de uma distribuição normal, a aproximação será boa, mesmo para tamanhos pequenos de amostra. Em termos práticos, esse teorema é de extrema importância, daí ser chamado de Teorema Central e, em geral, amostras de tamanho $n > 30$ já fornecem uma aproximação razoável.

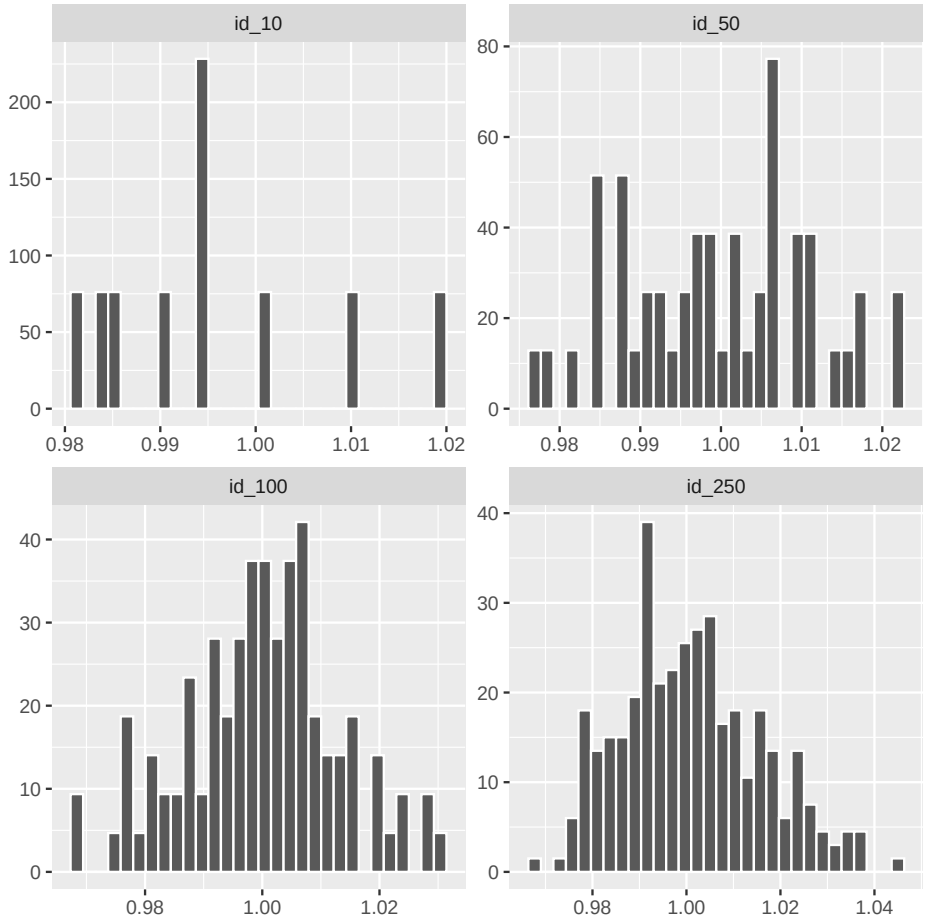
4.3.1 Visualizando

Considere uma população exponencial com média $\mu = 1$, ou seja, uma população distribuída segundo uma exponencial com parâmetro $\lambda = 1$.

O gráfico superior representa a distribuição populacional e os histogramas representam a distribuição amostral de $\bar{X}$ ao longo de 5000 amostras de tamanhos 10, 50, 100 e 250. Assim, podemos ver que, embora a população seja completamente diferente da normal, a distribuição amostral de $\bar{X}$ vai se tornando cada vez mais próxima da normal à medida que n aumenta.

```
matriz_aux_10 = matrix(NA,nrow=5000,ncol=10)
matriz_aux_50 = matrix(NA,nrow=5000,ncol=50)
matriz_aux_100 = matrix(NA,nrow=5000,ncol=100)
matriz_aux_250 = matrix(NA,nrow=5000,ncol=250)
for(j in 1:ncol(matriz_aux_10)){matriz_aux_10[,j]=rexp(5000,rate=1)}
for(j in 1:ncol(matriz_aux_50)){matriz_aux_50[,j]=rexp(5000,rate=1)}
for(j in 1:ncol(matriz_aux_100)){matriz_aux_100[,j]=rexp(5000,rate=1)}
for(j in 1:ncol(matriz_aux_250)){matriz_aux_250[,j]=rexp(5000,rate=1)}
medias_10 = data.frame(x=1:ncol(matriz_aux_10),y=colMeans(matriz_aux_10),label="id_10")
medias_50 = data.frame(x=1:ncol(matriz_aux_50),y=colMeans(matriz_aux_50),label="id_50")
medias_100 = data.frame(x=1:ncol(matriz_aux_100),y=colMeans(matriz_aux_100),label="id_100")
medias_250 = data.frame(x=1:ncol(matriz_aux_250),y=colMeans(matriz_aux_250),label="id_250")
tudo = rbind(medias_10,medias_50,medias_100,medias_250)
tudo$label = factor(tudo$label, levels = c("id_10", "id_50", "id_100", "id_250"))
```

```
ggplot(tudo,aes(y)) +
  geom_histogram(aes(y=..density..),color = "white") +
  facet_wrap(~label, scales = "free") +
  theme(axis.title = element_blank())
```



4.4 Distribuição Amostral da Variância Amostral

Já vimos que a estatística

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

é um estimador não viciado da variância σ^2, portanto

$$E(S^2) = \sigma^2$$

. Vamos estudar agora a distribuição amostral de S^2 e para isso precisamos estudar a distribuição qui-quadrado.

4.4.1 Distribuição Qui-quadrado

Se X é uma variável aleatória com densidade

$$f_X(x) = \frac{1}{\Gamma(\frac{k}{2})}(\frac{1}{2})^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, k > 0, x > 0,$$

onde $\Gamma(w) = \int_0^\infty x^{w-1}e^{-x}dx, w > 0$. Então, X tem uma distribuição qui-quadrado com k graus de liberdade, onde o parâmetro k é um número inteiro.

Para entender a ideia de graus de liberdade, consideremos um conjunto de dados qualquer. Graus de liberdade é o número de valores deste conjunto de dados que podem variar após terem sido impostas certas restrições a todos os valores. Por exemplo, consideremos que 10 estudantes obtiveram em um teste média 8. Assim, a soma das 10 notas deve ser 80 (restrição). Portanto, neste caso, temos um grau de liberdade de $10-1 = 9$, pois 9 notas podem variar livremente desde que a soma seja 80, no entanto 1 nota sempre será [80-(soma das 9 outras notas)].

Temos também que se as variáveis aleatórias X_i, $i = 1, 2, \dots, n$ são independentes e normalmente distribuídas com médias μ_i e variâncias σ_i^2, isto é $X_i \sim N(\mu_i, \sigma_i^2)$, então

$$U = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma^2} \right)^2$$

tem uma distribuição qui-quadrado com n graus de liberdade.

Além disso, se $X_1, \dots, X_n$ é uma a.a. de uma distribuição normal padrão, então, valem as seguintes propriedades:

(i) $\bar{X}$ e $\sum_{i=1}^n (X_i - \bar{X})^2$ são independentes;

(ii) $\sum_{i=1}^n (X_i - \bar{X})^2$ tem uma distribuição qui-quadrado com $n-1$ graus de liberdade.

Assim, chegamos que se S^2 é a variância amostral de uma amostra aleatória $X_1, \dots, X_n$ de uma distribuição normal com média μ e variância σ^2, então

$$U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

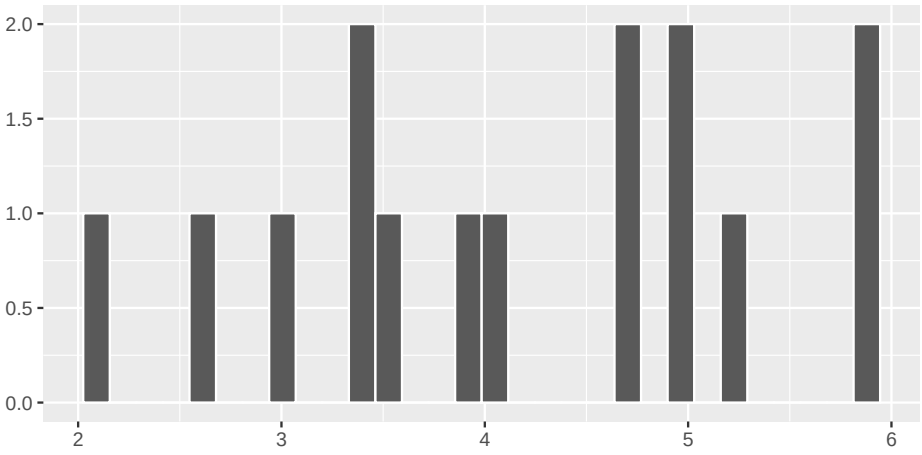
ou seja, U tem uma distribuição qui-quadrado com $n-1$ graus de liberdade.

4.4.2 Visualizando

Analogamente ao estudo de simulação realizado no caso da média amostral, considere uma população normal com média $\mu = 10$ e variância $\sigma^2 = 4$.

Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.

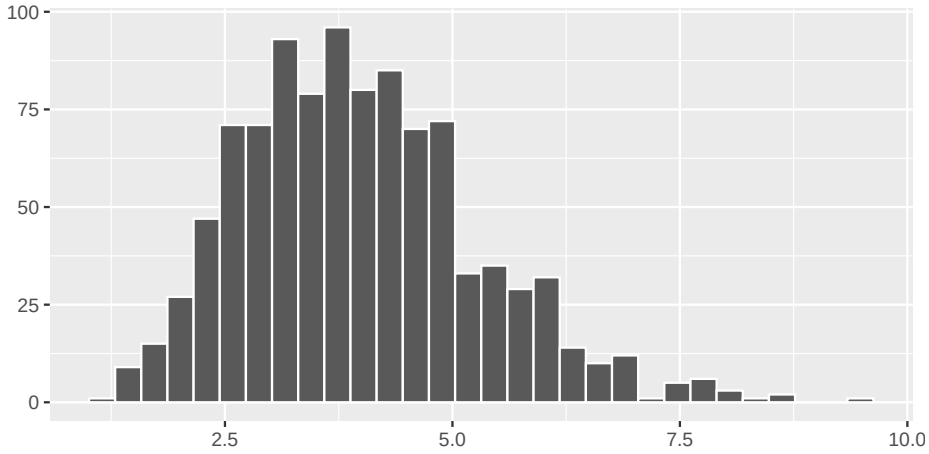
```
n = 15
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
variancias = data.frame(x=1:ncol(matriz_aux),y=apply(matriz_aux,2,var))
ggplot(variancias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```



Nessa simulação obtivemos a média das variâncias igual a 4.17 e a variância das variâncias igual a 1.33.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 1000 amostras.

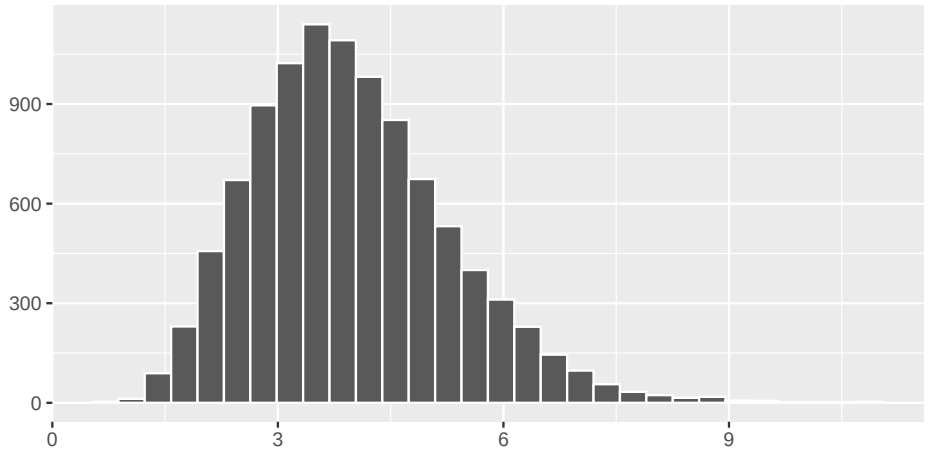
```
n = 1000
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
variancias = data.frame(x=1:ncol(matriz_aux),y=apply(matriz_aux,2,var))
ggplot(variancias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```



Neste caso a média das variâncias foi igual a 3.99 e a variância das variâncias igual a 1.69.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da variância amostral pode ser vista segundo o histograma abaixo.

```
n = 10000
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rnorm(20,mean=10,sd=2)}
variancias = data.frame(x=1:ncol(matriz_aux),y=apply(matriz_aux,2,var))
ggplot(variancias,aes(y)) +
  geom_histogram(color = "white") +
  theme(axis.title = element_blank())
```



Neste caso, a média das variâncias é 3.98 e a variância é 1.73. Então, realmente, podemos perceber que a distribuição da variância amostral se aproxima de uma

distribuição qui-quadrado com média $\mu = 4$ e variância $\frac{2\sigma^4}{n-1} = \frac{2\times 16}{19} = 1,684$.

Destas propriedades temos que

$$V(S^2) = \frac{2\sigma^4}{n-1}.$$

4.5 Distribuição Amostral da Proporção

Considere uma população em que cada elemento é classificado de acordo com a presença ou ausência de determinada característica. Por exemplo, podemos pensar em eleitores escolhendo entre 2 candidatos, pessoas classificadas de acordo com o sexo, trabalhadores classificados como trabalhador com carteira assinada ou não, e assim por diante. Em termos de variável aleatória, essa população é representada por uma v.a. de Bernoulli, isto é:

$$X = \begin{cases} 1, & \text{se elemento possui a característica de interesse} \\ 0, & \text{se elemento não possui a caracaterística de interesse} \end{cases}$$

Vamos denotar por p a proporção de elementos da população que possuem a característica de interesse. Então, $P(X = 1) = p$, $E(X) = p$ e $V(X) = p(1-p)$. Em geral, esse parâmetro é desconhecido e precisamos estimá-lo a partir de uma amostra.

Suponha, então, que dessa população seja extraída uma amostra aleatória simples $X_1, X_2, ..., X_n$ com reposição. Essas n extrações correspondem a n variáveis aleatórias de Bernoulli independentes e, como sabemos, $S_n = \sum_{i=1}^{n} X_i$ tem distribuição binomial com parâmetros n e p.

Note que S_n dá o número total de "sucessos" nas n repetições, onde "sucesso", neste caso, representa a presença da característica de interesse. Os valores possíveis de S_n são $0, 1, 2, ..., n$. Com relação à proporção $\hat{P}$ de elementos na amostra que possuem a característica de interesse, temos que

$$\hat{P} = \frac{S_n}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

e os valores possíveis de $\hat{P}$ são $0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$ com

$$P\left(\hat{P} = \frac{k}{n}\right) = P(S_n = k).$$

Como a proporção amostral é uma média de uma amostra aleatória simples de uma população com distribuição de Bernoulli com parâmetro p, o Teorema Limite Central nos diz, então, que a distribuição da proporção amostral se aproxima de uma normal com média p e variância $\frac{p(1-p)}{n}$. Isto é,

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

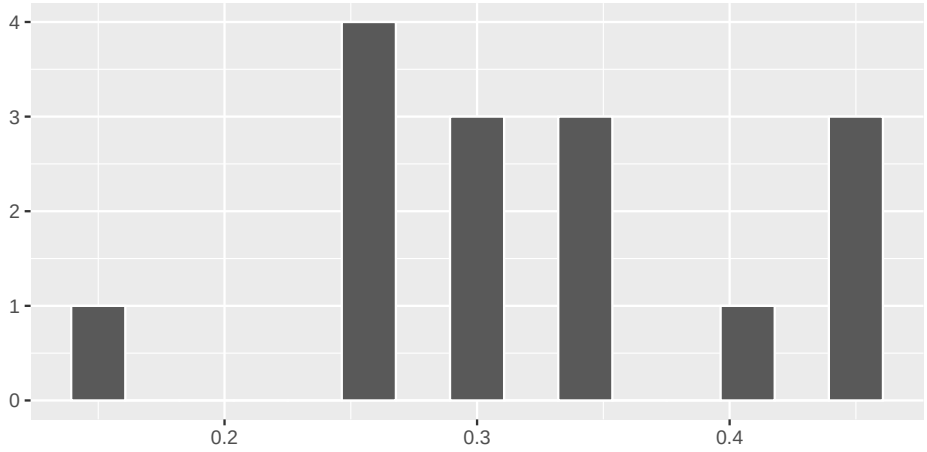
Como essa aproximação é uma conseqüência direta da aproximação normal da binomial, as mesmas regras continuam valendo: a aproximação deve ser feita se $np \geq 5$ e $n(1-p) \geq 5$.

4.5.1 Visualizando

Considere uma população bernoulli com $p = 0,3$. Vamos realizar um estudo de simulação para a distribuição da proporção amostral considerando amostras de tamanho 20 dessa população.

Primeiramente, considere que são retiradas 15 amostras de tamanho 20 dessa população.

```
library(ggplot2)
n = 15
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rbinom(20,1,prob=0.3)}
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white", bins = 15) +
  theme(axis.title = element_blank())
```

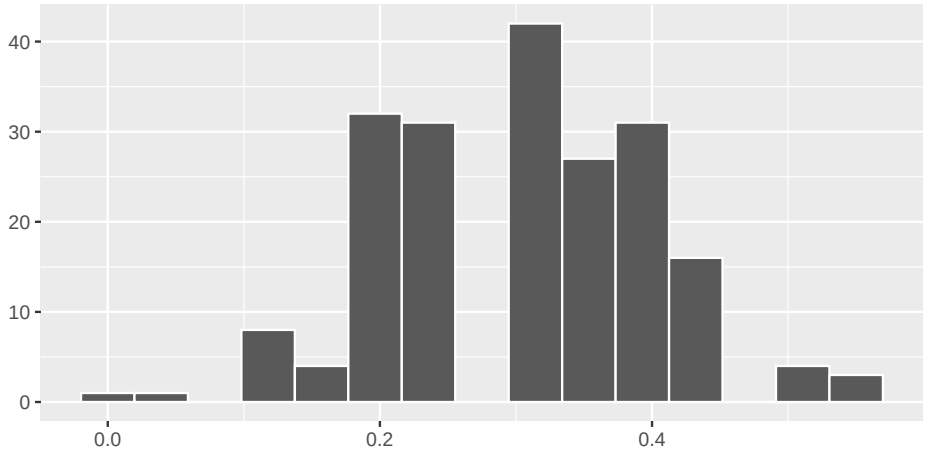


Nessa simulação obtivemos $\hat{p} = 0.32$.

Suponha agora que façamos o mesmo processo, porém ao invés de considerarmos 15 amostras de tamanho 20, consideramos 200 amostras.

```
n = 200
matriz_aux = matrix(NA,nrow=20,ncol=n)
```

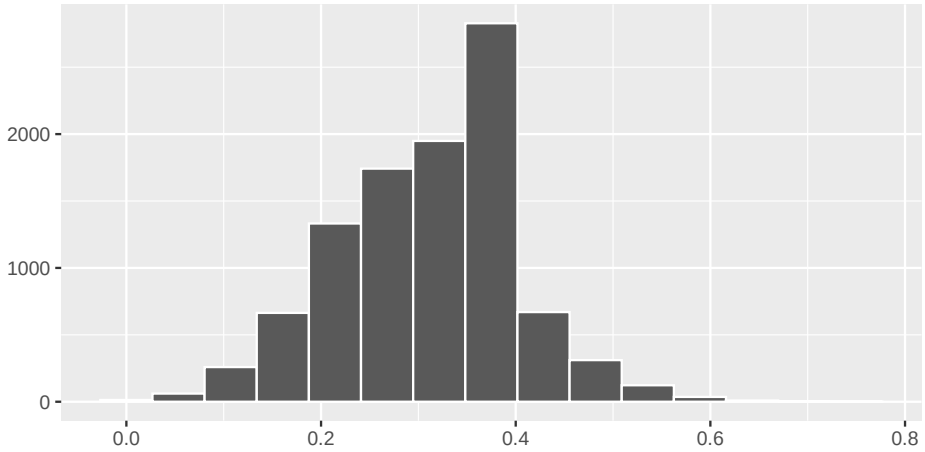
```
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rbinom(20,1,prob=0.3)}
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white", bins = 15) +
  theme(axis.title = element_blank())
```



Agora, obtivemos $\hat{p} = 0.3$.

Realizando o mesmo experimento, porém agora considerando 10000 amostras de tamanho 20, a distribuição da média amostral pode ser vista segundo o histograma abaixo.

```
n = 10000
matriz_aux = matrix(NA,nrow=20,ncol=n)
for(j in 1:ncol(matriz_aux)){matriz_aux[,j]=rbinom(20,1,prob=0.3)}
medias = data.frame(x=1:ncol(matriz_aux),y=colMeans(matriz_aux))
ggplot(medias,aes(y)) +
  geom_histogram(color = "white", bins = 15) +
  theme(axis.title = element_blank())
```



Para este caso, a média das proporções amostrais foi $\hat{p} = 0.3$. Então, empiricamente, podemos perceber que a distribuição da média amostral se aproxima de uma distribuição normal com média $\mu = 0,3$ e desvio padrão $\frac{p(1-p)}{\sqrt{n}} = \frac{0.21}{\sqrt{20}} = 0,0105$.

Chapter 5

Intervalo de Confiança

5.1 Introdução

Neste capítulo você aprenderá um método muito importante de estimação de parâmetros. Vimos anteriormente que a média amostral $\bar{X}$ é um bom estimador da média populacional μ. Mas vimos, também, que existe uma variabilidade nos valores de $\bar{X}$, ou seja, cada amostra dá origem a um valor diferente do estimador. Uma maneira de informar sobre esta variabilidade é através da estimação por intervalos.

5.2 Ideias Básicas

O objetivo central da Inferência Estatística é obter informações para uma população a partir do conhecimento de uma única amostra. Em geral, a população é representada por uma variável aleatória X, com função de distribuição ou densidade de probabilidade f_X . Dessa população, então, extrai-se uma amostra aleatória simples com reposição, que dá origem a um conjunto $X_1, X_2, ..., X_n$ de n variáveis aleatórias independentes e identicamente distribuídas, todas com a mesma distribuição f_X.

Se f_X depende de um ou mais parâmetros, temos que usar a informação obtida a partir da amostra para estimar esses parâmetros, de forma a conhecermos a distribuição. Vimos, por exemplo, que a média amostral $\bar{X}$ é um bom estimador da média populacional μ, no sentido de que ela tende a "acertar o alvo" da verdadeira média populacional, isto é, a média amostral é um estimador não-viesado da média populacional. Mas vimos, também, que existe uma variabilidade nos valores de $\bar{X}$, ou seja, cada amostra dá origem a um valor diferente do estimador. Para algumas amostras, $\bar{X}$ será maior que μ, para outras será menor e para outras será igual.

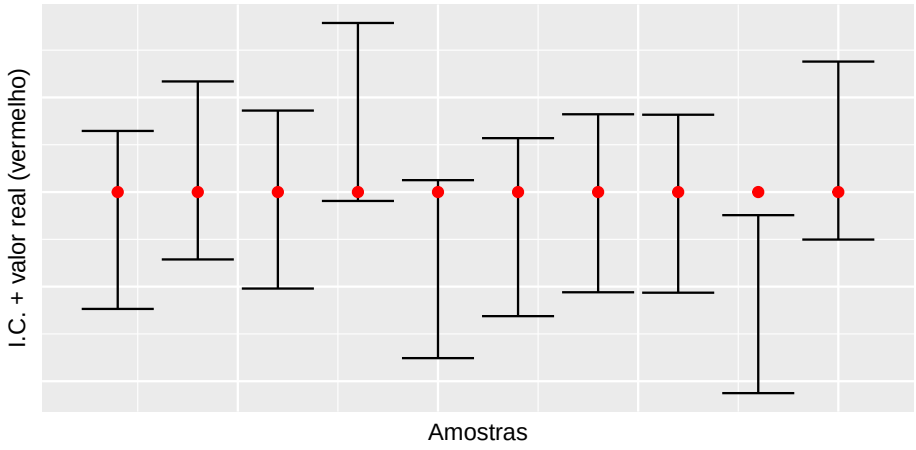
Na prática, temos apenas uma amostra e, assim, é importante que se dê alguma informação sobre essa possível variabilidade do estimador. Ou seja, é importante informar o valor do estimador $\hat{\theta}$ obtido com uma amostra específica, mas é importante informar também que o verdadeiro valor do parâmetro θ poderia estar num determinado intervalo, digamos, no intervalo $\left[\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon\right]$. Dessa forma, estamos informando a nossa margem de erro no processo de estimação; essa margem de erro é consequência do processo de seleção aleatória da amostra.

O que vamos estudar agora é como obter esse intervalo, de modo a "acertar na maioria das vezes", isto é, queremos um procedimento que garanta que, na maioria das vezes (ou das amostras possíveis), o intervalo obtido conterá o verdadeiro valor do parâmetro. A expressão "na maioria das vezes" será traduzida como "probabilidade alta". Dessa forma, estaremos lidando com afirmativas do seguinte tipo:

Com probabilidade alta (em geral, indicada por $1-\alpha$), o intervalo $\left[\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon\right]$ conterá o verdadeiro valor do parâmetro θ.

A interpretação correta de tal afirmativa é a seguinte: se $1-\alpha = 0,95$, por exemplo, então isso significa que o procedimento de construção do intervalo é tal que, em 95% das possíveis amostras, o intervalo $\left[\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon\right]$ obtido conterá o verdadeiro valor do parâmetro.

Note, no exemplo abaixo, que cada amostra resulta em um intervalo diferente, mas em média, em 95% das amostras, o intervalo contém o verdadeiro valor do parâmetro (em vermelho).



O valor $1-\alpha$ é chamado nível de confiança, enquanto o valor α é conhecido como nível de significância. O intervalo $\left[\hat{\theta}-\varepsilon; \hat{\theta}+\varepsilon\right]$ é chamado de intervalo de confiança de nível de confiança $1-\alpha$.

Tendo clara a interpretação do intervalo de confiança, podemos resumir a frase acima da seguinte forma:

$$P\left(\theta \in [\hat{\theta} - \varepsilon; \hat{\theta} + \varepsilon]\right) = 1 - \alpha$$

Mais uma vez, a probabilidade se refere à probabilidade dentre as diversas possíveis amostras, ou seja, a probabilidade está associada à distribuição amostral de $\hat{\theta}$. Note que os limites do intervalo dependem de $\hat{theta}$, que depende da amostra sorteada, ou seja, os limites do intervalo de confiança são variáveis aleatórias. Cada amostra dá origem a um intervalo diferente, mas o procedimento de obtenção dos intervalos garante probabilidade $1 - \alpha$ de acerto.

5.3 Média da $N(\mu, \sigma^2)$, σ^2 conhecido

Vamos agora introduzir os métodos para obtenção do intervalo de confiança para a média de uma população. Como visto, a média populacional é um parâmetro importante que pode ser muito bem estimado pela média amostral $\bar{X}$. Para apresentar as idéias básicas, vamos considerar um contexto que é pouco frequente na prática. O motivo para isso é que, em termos didáticos, a apresentação é bastante simples. Como o fundamento é o mesmo para contextos mais gerais, essa abordagem se justifica.

Consideremos uma população descrita por uma variável aleatória normal com média μ e variância σ^2: $X \sim N(\mu, \sigma^2)$. Vamos supor que o valor de σ^2 seja conhecido e que nosso interesse seja estimar a média μ a partir de uma amostra aleatória simples $X_1, \ldots, X_n$. Como visto anteriormente, a distribuição amostral de $\bar{X}$ é normal com média μ e variância $\frac{\sigma^2}{n}$, ou seja

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Da definição de distribuição amostral, isso significa que os diferentes valores de $\bar{X}$ obtidos a partir das diferentes possíveis amostras se distribuem normalmente em torno de μ com variância $\frac{\sigma^2}{n}$.

Das propriedades da distribuição normal, resulta que

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1),$$

ou equivalentemente,

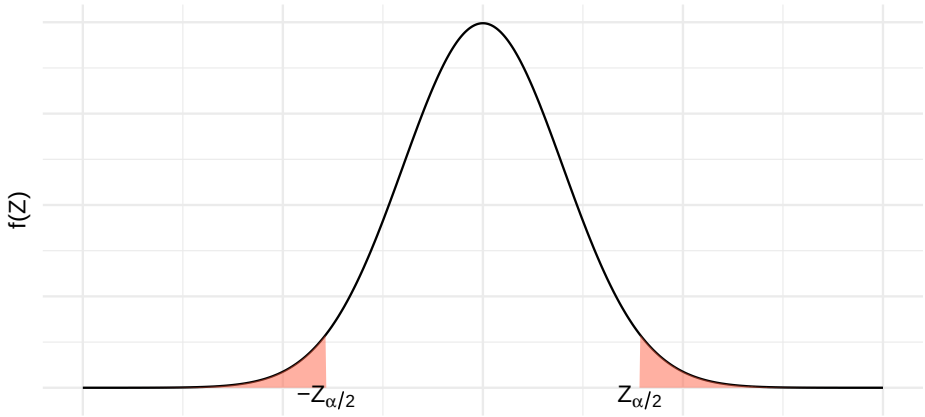
$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1).$$

5.3.1 Notação

Vamos estabelecer a seguinte notação: vamos indicar por z_α a abscissa da curva normal padrão que deixa probabilidade (área) igual a α acima dela. Veja a Figura abaixo. Temos, então, que $P(Z > z_\alpha) = \alpha$. Essa abscissa z_α é normalmente chamada de valor crítico.

Consideremos, agora, o valor crítico $z_{\frac{\alpha}{2}}$, veja a Figura abaixo. Daí podemos ver que, se $Z \sim N(0,1)$, então

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$



Note que isso vale para a distribuição normal padrão, em geral. Então, usando os resultados anteriories obtemos que

$$P\left(-z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Mas isso é equivalente a

$$P\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Note que a última expressão nos diz que

$$P\left(\mu \in \left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]\right) = 1-\alpha$$

Mas essa é exatamente a forma geral de um intervalo de confiança, conforme explicitado. Temos, então, a seguinte conclusão:

Seja $X \sim N(\mu, \sigma^2)$ uma população normal com variância σ^2 conhecida. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória dessa população, então o intervalo de confiança de nível de confiança $1-\alpha$ para a média populacional μ é dado por

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right],$$

onde $z_{\frac{\alpha}{2}}$ é a abcissa da curva normal padrão que deixa a área $\frac{\alpha}{2}$ acima dela.

5.4 Margem de Erro

O intervalo de confiança para μ pode ser escrito na forma $\left[\bar{X} - \varepsilon; \bar{X} + \varepsilon\right]$ onde $\varepsilon = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ é a margem de erro.

Analisando a margem de erro, podemos ver que ela depende diretamente do valor crítico e do desvio padrão populacional e é inversamente proporcional ao tamanho da amostra.

Na Figura abaixo ilustra-se a relação de dependência da margem de erro em relação ao desvio padrão populacional σ. Temos aí duas distribuições amostrais centradas na mesma média e baseadas em amostras de mesmo tamanho. Nas duas distribuições a área total das caudas sombreadas é α, de modo que o intervalo limitado pelas linhas verticais é o intervalo de confiança de nível de confiança $1-\alpha$. Para a distribuição mais dispersa, isto é, com σ maior, o comprimento do intervalo é maior. Esse resultado deve ser intuitivo: se há mais variabilidade na população, a nossa margem de erro tem que ser maior, mantidas fixas as outras condições (tamanho de amostra e nível de confiança).

Por outro lado, se mantivermos fixos o tamanho da amostra e o desvio padrão populacional, é razoável também esperar que a margem de erro seja maior para um nível de confiança maior. Ou seja, se queremos aumentar a probabilidade de acerto, é razoável que o intervalo seja maior. Aumentar a probabilidade de acerto significa aumentar o nível de confiança, o que acarreta em um valor crítico $z_{\frac{\alpha}{2}}$ maior.

Finalmente, mantidos o mesmo desvio padrão populacional e o mesmo nível de confiança, quanto maior o tamanho da amostra, mais perto vamos ficando da população e, assim, vai diminuindo a nossa margem de erro.

5.5 Introdução

Neste capítulo você completará seu estudo básico sobre intervalos de confiança para a média de uma população, analisando o problema de estimação da média de uma população normal quando não se conhece a variância desta população. Você verá que, neste caso, é necessário estimar essa variância e isso introduz mais uma fonte de variabilidade nas nossas estimativas: com uma única amostra, temos que estimar a média e a variância da população. O procedimento é simples e análogo aos casos anteriores vistos nos capítulos amteriores; o que muda é a distribuição amostral do estimador $\bar{X}$. Em vez de usarmos a distribuição normal para determinar os valores críticos, usaremos a distribuição t de Student.

5.6 Ideias Básicas

Considere uma população descrita por uma variável aleatória normal com média μ e variância σ^2: $X \sim N(\mu, \sigma^2)$. Nosso interesse é estimar a média μ a partir de uma amostra aleatória simples $X_1, X_2, \ldots, X_n$. Como visto anteriormente, a distribuição amostral de $\bar{X}$ é normal com média μ e variância $\frac{\sigma^2}{n}$, ou seja

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Asso, se o valor de σ é conhecido, resulta que

$$Z = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$$

e esse resultado foi utilizado na construção do intervalo de confiança para a média de uma população normal com variância conhecida, fornecendo o seguinte intervalo:

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

Suponhamos, agora, que a variância σ^2 não seja conhecida. Neste caso, temos que estimá-la (S^2) com os dados amostrais. Foi demostrado que

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

é um estimador não viesado para σ^2. Isso significa que, se calculássemos o valor de S^2 para cada uma das possíveis amostras aleatórias simples de tamanho n, a média desses valores seria igual a σ^2. Dessa forma, S^2 é um "bom" estimador de σ^2 e podemos usá-lo como uma estimativa pontual de σ^2. Como o desvio padrão é a raiz quadrada da variância, é natural perguntar: S é um "bom" estimador de σ, ou seja, S é um estimador não-viesado de σ? A resposta é NÃO, mas, para grandes amostras, o viés é pequeno, de modo que, em geral, usa-se S como estimador de σ. Sendo assim, é natural pensarmos em substituir o valor de σ por S na expressão da Normal Z e utilizarmos a estatística

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S}$$

na construção de intervalos de confiança para μ. Isso é exatamente o que faremos, mas, ao introduzirmos S no lugar de σ, a distribuição amostral de T deixa de ser normal e passa a ser uma distribuição t de Student.

5.6.1 Média da $N(\mu, \sigma^2)$, σ^2 desconhecido

O intervalo de confiança para a média de uma população normal com variância desconhecida é obtido com base no seguinte resultado:

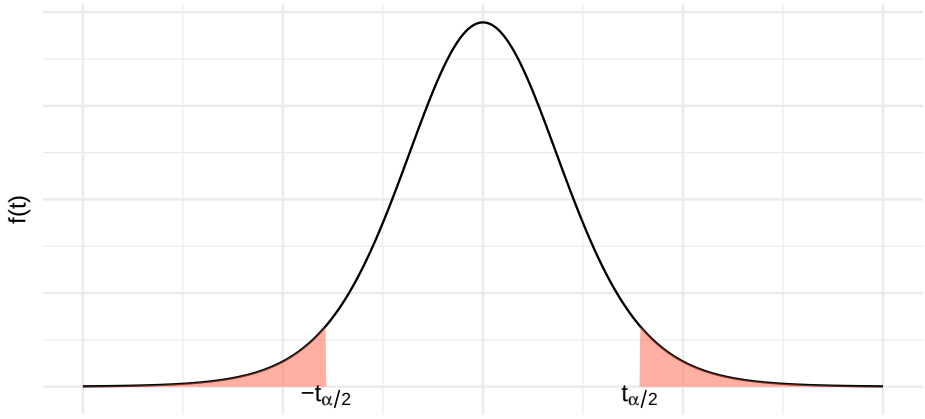
$$T = \sqrt{n}\frac{\bar{X} - \mu}{S} \sim t_{n-1},$$

onde $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$.

O número de graus de liberdade $gl = n-1$ resulta do fato de que, na soma que define S^2, há apenas $n-1$ parcelas independentes, ou seja, dados S^2 e $n-1$ das parcelas $(X_i - \bar{X})^2$, a n-ésima parcela fica automaticamente determinada.

Usando a simetria da densidade t, temos o seguinte resultado:

$$P\left(-t_{n,\alpha/2} \leq t_n \leq t_{n,\alpha/2}\right) = 1-\alpha$$



Note que isso vale para a distribuição t de student, em geral. Então, usando os resultados anteriories obtemos que

$$P\left(-t_{n-1,\alpha/2} \leq \sqrt{n}\frac{\bar{X}-\mu}{S} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

Mas isso é equivalente a

$$P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X}-\mu \leq t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Note que a última expressão nos diz que

$$P\left(\mu \in \left[\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}; \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]\right) = 1-\alpha$$

Mas essa é exatamente a forma geral de um intervalo de confiança, conforme explicitado. Temos, então, a seguinte conclusão:

Seja $X \sim N(\mu, \sigma^2)$ uma população normal com variância σ^2 desconhecida. Se $X_1, X_2, ..., X_n$ é uma amostra aleatória dessa população, então o intervalo de confiança de nível de confiança $1 - \alpha$ para a média populacional μ é dado por

$$\left[\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right],$$

onde $t_{n-1,\alpha/2}$ é o valor crítico da distribuição t de student com $n-1$ graus de liberdade que deixa a área $\frac{\alpha}{2}$ acima dela.

5.7 Margem de Erro

Note, mais uma vez, a forma do intervalo de confiança:

$$\bar{X} \pm \varepsilon,$$

onde a margem de erro ε, agora é definida em termos do valor crítico da distribuição t e do erro padrão estimado de $\bar{X}$:

$$\varepsilon = t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

.

5.8 Introdução

Anteriormente, foram apresentadas as idéias básicas da estimação por intervalos de confiança. Para ilustrar o princípio utilizado na construção de tais intervalos, consideramos a situação especial de estimação da média de uma população normal com variância conhecida. Neste caso, a distribuição amostral da média amostral é normal e foi com base nessa distribuição amostral normal que obtivemos o intervalo de confiança.

No primeiro momento desta aula usaremos o teorema limite central, que garante que a distribuição amostral da proporção amostral pode ser aproximada por uma distribuição normal, desde que utilizemos amostras grandes.

5.9 Intervalo de confiança para a proporção populacional

5.9.1 Ideias Básicas

O contexto de interesse é o seguinte: temos uma população em que cada elemento é classificado de acordo com a presença ou ausência de determinada

característica. Em termos de variável aleatória, essa população é representada por uma v.a. de Bernoulli, isto é:

$$X = \begin{cases} 1, \text{ se elemento possui a característica de interesse} \\ 0, \text{ se elemento não possui a caracaterística de interesse} \end{cases}$$

Então, $P(X = 1) = p$, $E(X) = p$ e $V(X) = p(1-p)$. O parâmetro p é também a proporção de elementos da população que possuem a caracterísitca de interesse. Em geral, esse parâmetro é desconhecido e precisamos estimá-lo a partir de uma amostra.

Suponha, então, que dessa população seja extraída uma amostra aleatória simples $X_1, X_2, ..., X_n comreposio$. Vimos que a proporção $\hat{P}$ de elementos na amostra que possuem a característica de interesse, definida por

$$\hat{P} = \frac{S_n}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

é um estimador não-viesado para p com variância $\frac{p(1-p)}{n}$. Mais precisamente,

$$E(\hat{P}) = p$$

$$V(\hat{P}) = \frac{p(1-p)}{n}$$

Como a proporção amostral é uma média de uma amostra aleatória simples de uma população com distribuição de Bernoulli com parâmetro p, o Teorema Limite Central nos diz que a distribuição de $\hat{P}$ se aproxima de uma nornal com média p e variância $\frac{p(1-p)}{n}$.

Resumindo, temos o seguinte resultado:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Usando as propriedades da distribuição normal, temos que

$$\frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

ou equivalentemente

$$\sqrt{n}\frac{\hat{P} - p}{\sqrt{p(1-p)}} \sim N(0, 1)$$

Vamos ver, agora, como usar esse resultado para obter um intervalo de confiança para a verdadeira proporção populacional p.

5.9.2 Construção do IC

O procedimento de construção do intervalo de confiança para a proporção populacional é totalmente análogo ao do intervalo de confiança para a média de uma população normal com variância conhecida, visto anteriormente. Assim, iremos usar a mesma notação, a saber: vamos representar por z_α a abscissa da curva normal padrão que deixa probabilidade (área) α acima dela. Como visto, temos o seguinte resultado, onde $Z \sim N(0,1)$:

$$P\left(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Como o resultado acima vale para qualquer variável aleatória $N(0,1)$, podemos usar para obter

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

e, portanto

$$P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq \hat{P}-p \leq z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

$$P\left(-\hat{P}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq -p \leq -\hat{P}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

$$P\left(\hat{P}-z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P}+z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

Como no caso da média, chegamos a uma expressão do seguinte tipo:

$$P(\hat{P}-\varepsilon \leq p \leq \hat{P}+\varepsilon) = 1-\alpha$$

onde $\varepsilon = z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$

Analisando a margem de erro, podemos ver uma diferença fundamental entre o IC para a proporção e para a média: a margem de erro da proporção amostral depende do parâmetro desconhecido p. Na prática, para construir o intervalo de confiança, temos que substituir esse valor por alguma estimativa.

Existem 3 abordagens possíveis:

1. Usar a própria proporção amostral observada; nesse caso, o intervalo de confiança seria:
$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. Usar o intervalo de confiança conservador, ou seja, usar o maior valor possível para $z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ para um dado n, o que equivale a obter o intervalo de confiança com o maior comprimento possível. Como o comprimento do intervalo é diretamente proporcional a $\sqrt{p(1-p)}$ ou equivalentemente a $p(1-p)$, $p = 0.5$ maxima esta função.

Neste caso, o o intervalo de confiança se torna:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{0.5 \times 0.5}{n}} = \hat{p} \pm z_{\frac{\alpha}{2}} \frac{0.5}{\sqrt{n}}$$

3. Usar algum valor auxiliar $\hat{p}_0$ ou estimativa prévia, obtida de outras fontes ou de uma amostra piloto:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}}$$

Seja $X \sim Bernoulli(p)$. Se $X_1, X_2, ..., X_n$ é uma amostra aleatória dessa população, então o intervalo de confiança de nível de confiança $1-\alpha$ para a proporção populacional p é dado por

$$\left[\hat{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}}; \hat{P} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}} \right],$$

onde $z_{\frac{\alpha}{2}}$ é a abcissa da curva normal padrão que deixa a área $\frac{\alpha}{2}$ acima dela e $\hat{p}_0$ é alguma estimativa para o verdadeiro valor de p.

5.9.3 Determinação do tamanho da amostra

Uma questão que se coloca freqüentemente é: qual o tamanho da amostra necessário para se estimar uma proporção p com uma margem de erro ε e nível de confiança $1-\alpha$? Vamos analisar a expressão da margem de erro:

$$\varepsilon = z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Resolvendo para n, obtemos que

$$\sqrt{n} = z_{\frac{\alpha}{2}} \frac{\sqrt{p(1-p)}}{\varepsilon}$$

ou

$$n = p(1-p)\left(\frac{z_{\frac{\alpha}{2}}}{\varepsilon}\right)^2$$

Vemos, então, que n é diretamente proporcional a $p(1-p)$, ou seja, quanto maior $p(1-p)$, maior será o tamanho da amostra n. Na prática, não conhecemos p (na verdade, estamos querendo estimar esse parâmetro). Então, para determinar o tamanho de amostra necessário para uma margem de erro e um nível de confiança dados, podemos considerar o pior caso, ou seja, podemos tomar o maior valor possível de $p(1-p)$ e calcular o tamanho da amostra com base nesse pior caso, que ocorre quando $p = 0,5$. É claro que essa é uma escolha conservadora, que em alguns casos pode levar a um tamanho de amostra desnecessariamente grande. Usando esta estimativa para p, obtemos que

$$n = \left(0,5 \times \frac{z_{\frac{\alpha}{2}}}{\varepsilon}\right)^2$$

5.10 Intervalo de confiança para a variância da $N(\mu, \sigma^2)$

Nesta parte você completará seu estudo básico sobre intervalos de confiança, analisando o problema de estimação da variância de uma população normal. Como antes, este intervalo se baseará na distribuição amostral de um estimador não-viesado para σ^2, a saber, S^2. Como a variância é um número não negativo, essa distribuição não é simétrica e está definida apenas para valores não-negativos.

5.10.1 Ideias básicas

O contexto subjacente é o seguinte: a partir de uma amostra aleatória simples $X_1, X_2, \ldots, X_n$ retirada de uma população normal com média μ e variância σ^2 queremos construir um intervalo de confiança para σ^2. A hipótese de normalidade da população é fundamental aqui. Assim como no caso da média, temos que usar a distribuição amostral de algum estimador. Neste caso, o estimador é S^2 e o resultado importate é o seguinte: $\frac{(n-1)S^2}{\sigma^2}$ tem distribuição qui-quadrado com $n-1$ graus de liberdade:

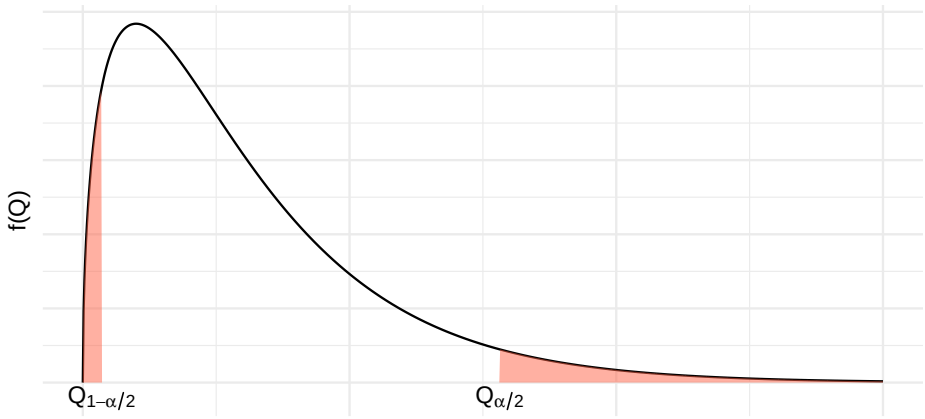
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

5.10.2 Construção do IC

Como no caso da distribuição t-Student e também da distribuição Normal, vamos definir o valor crítico $\chi^2_{n;\alpha}$ como a abscissa da distribuição qui-quadrado

com n graus de liberdade que deixa probabilidade α acima dela.

Com essa definição, podemos ver que a abscissa $\chi^2_{n;\alpha/2}$ deixa probabilidade $\alpha/2$ acima dela e a abscissa $\chi^2_{n;1-\alpha/2}$ deixa probabilidade $\alpha/2$ abaixo dela.



Logo,

$$P\left(\chi^2_{n;\frac{\alpha}{2}} \leq \chi^2_n \leq \chi^2_{n;1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Como o resultado acima vale para qualquer distribuição qui-quadrado, podemos usar o resultado anterior para escrever

$$P\left(\chi^2_{n-1;\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Daí resulta que

$$P\left(\frac{\chi^2_{n-1;\frac{\alpha}{2}}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{n-1;1-\frac{\alpha}{2}}}{(n-1)S^2}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\frac{\alpha}{2}}}\right) = 1-\alpha$$

e esse é o intervalo de confiança para a variância de uma população normal.

Seja $X \sim N(\mu, \sigma^2)$ uma população normal. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória dessa população, então o intervalo de confiança de nível de confiança $1-\alpha$ para a variância populacional σ^2 é dado por

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}; \frac{(n-1)S^2}{\chi^2_{n-1;\frac{\alpha}{2}}}\right],$$

onde $\chi^2_{n-1;\frac{\alpha}{2}}$ é o valor crítico da distribuição qui-quadrado com $n-1$ graus de liberdade que deixa a área $\frac{\alpha}{2}$ abaixo dela e $\chi^2_{n-1;1-\frac{\alpha}{2}}$ é o valor crítico da distribuição qui-quadrado com $n-1$ graus de liberdade que deixa a área $\frac{\alpha}{2}$ acima dela.

Bibliography

Wesley Cota. Monitoring the number of COVID-19 cases and deaths in brazil at municipal and federative units level. *SciELOPreprints:362*, May 2020. doi: 10.1590/scielopreprints.362. URL https://doi.org/10.1590/scielopreprints.362.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.