Introdução à datascience com R

Cleuton Sampaio

Lição 2: Estatística básica: Medidas de tendência central

Tipos de dados

Antes de entrarmos em estudos estatísticos é necessário conceituarmos e classificarmos os tipos de dados (o domínio) das variáveis que trabalharemos:

- **Discreto**: Domínio dos inteiros (Z): Por exemplo, quantos filhos cada família possui: 1, 2, 3 etc;
- **Contínuo**: Domínio dos números reais (R): Por exemplo, a altura de uma pessoa: 1,75m, 1,68m etc;
- **Categoria**: Não possui significado matemático, mas classifica os dados: Por exemplo, o estado civil de uma pessoa: Casada, Solteira, Divorciada, Viúva etc. Um tipo de categoria especial é a binomial ou binária, que só admite dois valores, por exemplo, se a pessoa trabalha (sim e não), ou se é solteira (verdadeiro, falso);
- **Ordinal**: É um tipo de categoria, que implica ordenação: Por exemplo, a classificação dos participantes de uma maratona: Primeiro, Segundo, Terceiro;

Estatística descritiva

Estatística descritiva serve para analisar e sumarizar um dataset, e é, geralmente, a primeira coisa que fazemos quando recebemos um novo trabalho.

População e amostra

Em estatística, "**população**" é o conjunto dos dados que desejamos analisar. Por exemplo, se queremos analisar o desempenho escolar dos alunos brasileiros do ensino fundamental, então a população será o conjunto de TODOS os alunos brasileiros do ensino fundamental.

A população pode ser muito grande, logo, faz sentido extrair subconjuntos de dados, desde que sejam representativos, para podermos analisar. Isto se chama "**amostra**".

Tendência central

As medidas de tendência central são: média, mediana e moda. Eu sei que você sabe o que é média, mas existem alguns detalhes que talvez desconheça.

Média (ou média da população)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Conhecemos a média (em inglês: "*mean*" ou "*average*") da população pela letra grega “mi” e a quantidade de elementos da população pela letra “N” (maiúscula).

Média da amostra

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Conhecemos a média da população pelo x com uma barra “x barra” e a quantidade de elementos da população pela letra “n” (minúscula).

A média é uma medida pouco confiável pois é afetada por **valores espúrios** (*outliers*), assunto que veremos mais adiante.

Mediana

A mediana (em inglês: "*median*") é o valor que está no centro de uma amostra. Vamos imaginar o seguinte dataset:

[1,3,4,**6**,7,9,10]

O valor da mediana é "6", que é o elemento central desse dataset, isto porque há um número ímpar de elementos. Se houver um número par de elementos, a mediana é calculada como a média dos dois valores centrais, por exemplo:

[1,3,**4,6**,7,9]

Neste caso, a mediana é "5" ((4 + 6) / 2).

Moda

A moda (em inglês: "mode") é o valor que mais se repete em uma amostra, por exemplo:

[1,**2,2**,3,4,5,6,7]

Neste caso, a moda é o número 2, que é o elemento que mais se repete.

Caso os dados sejam categóricos ou estejam agrupados por faixas, a moda é a faixa que possui maior número de ocorrências.

Quantidade de alunos	Média obtida
13	até 5,0
8	entre 5,0 (exclusive) até 8,0
3	entre 8,0 (exclusive) até 10

Existem técnicas de amostragem (“sampling”) que formar subconjuntos representativos de uma população. Estas técnicas são utilizadas para evitar o “**viés**” (em inglês: “bias”), que é uma tendência indesejável nos dados coletados.

Neste exemplo, vemos que a maior quantidade de alunos obteve média até 5,0, logo, esta é a "classe modal".

Viés também é conhecido como erro sistemático em uma ou mais características de uma amostra, representando uma distorção entre o valor da característica e o valor real. Vieses podem ser introduzidos por erros no cálculo ou por contaminação da amostra.

Voltando às técnicas de amostragem, temos:

- Amostragem aleatória: Utilizamos uma função para selecionar elementos aleatórios da População, de modo a evitar a introdução de viés;
- Amostragem sistemática: Pegamos elementos em intervalos selecionados, por exemplo, a cada 10 elementos pegamos 1;

- Amostragem estratificada: Dividimos a população em estratos (ou camadas) e retiramos alguns dados de cada estrato para formar a amostra. É importante manter a representatividade da população.

Exemplo

Comparando Média, Mediana e Moda

Abra o arquivo "R-course/lesson2/medidasCentrais.R", no nosso repositório do Github:

https://github.com/cleuton/datascience/tree/master/R-course/lession1

Vamos imaginar um exemplo simples: Os pesos dos alunos adolescentes de uma turma:

[52,52,54,56,57,60,61,65,70,120]

- Peso médio: 64.7;
- Mediana: 58.5;
- Moda: 52.

Se tomarmos a média como descritor desse dataset, assumiremos que os alunos dessa turma pesam quase 65 kg, o que é um erro grosseiro. Note que, dos 10 alunos, 7 pesam menos que isso. O peso médio é 10% maior que a mediana dos pesos, e muito superior à moda.

Se você quiser saber quanto tipicamente pesa cada aluno, qual métrica usará? A moda simples, baseada em elementos que se repetem, pode não existir (caso não haja elementos repetidos), mas podemos dividir os pesos em faixas e calcular a classe modal, por exemplo:

Faixa de peso	Quantidade de alunos
Até 55 kg	3
Até 80 kg	6
Acima de 80 kg	1

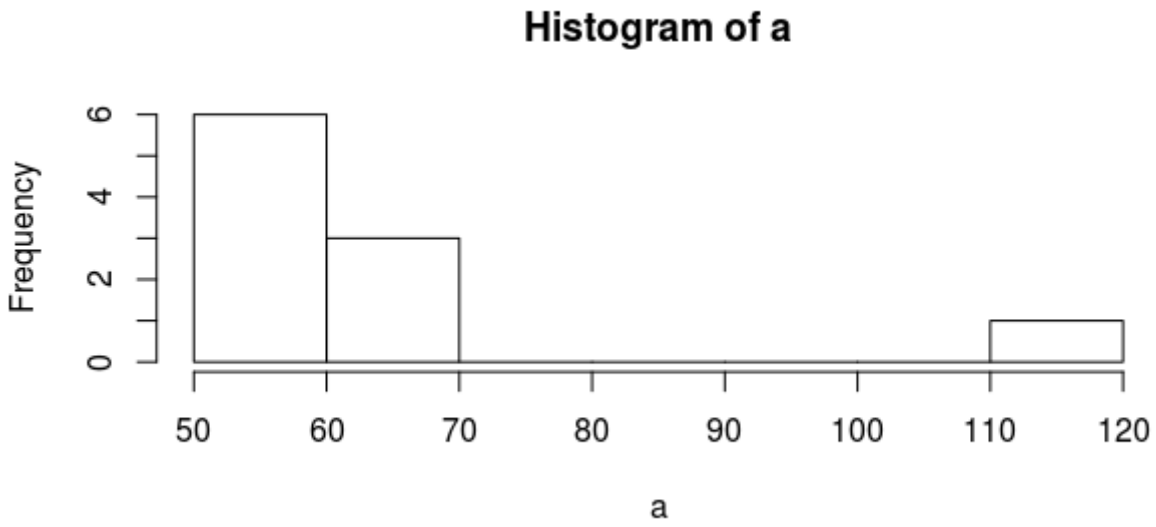
Podemos ver que a faixa de peso que tem mais alunos, ou a "classe modal" é a dos que pesam entre 55 e 80 kg.

Histograma

Wikipedia:

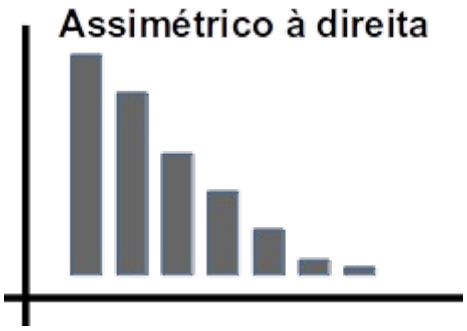
O histograma, também conhecido como distribuição de frequências, é a representação gráfica em colunas ou em barras (retângulos) de um conjunto de dados previamente tabulado e dividido em classes uniformes ou não uniformes. A base de cada retângulo representa uma classe. A altura de cada retângulo representa a quantidade ou a frequência absoluta com que o valor da classe ocorre no conjunto de dados para classes uniformes ou a densidade de frequência para classes não uniformes. Importante ferramenta da estatística, o histograma também é uma das chamadas sete ferramentas da qualidade.

Um histograma nos permite visualizar nossos dados e como eles se agrupam em torno da média, que é o nosso valor esperado. No caso dos pesos, este seria o histograma plotado como gráfico de barras:

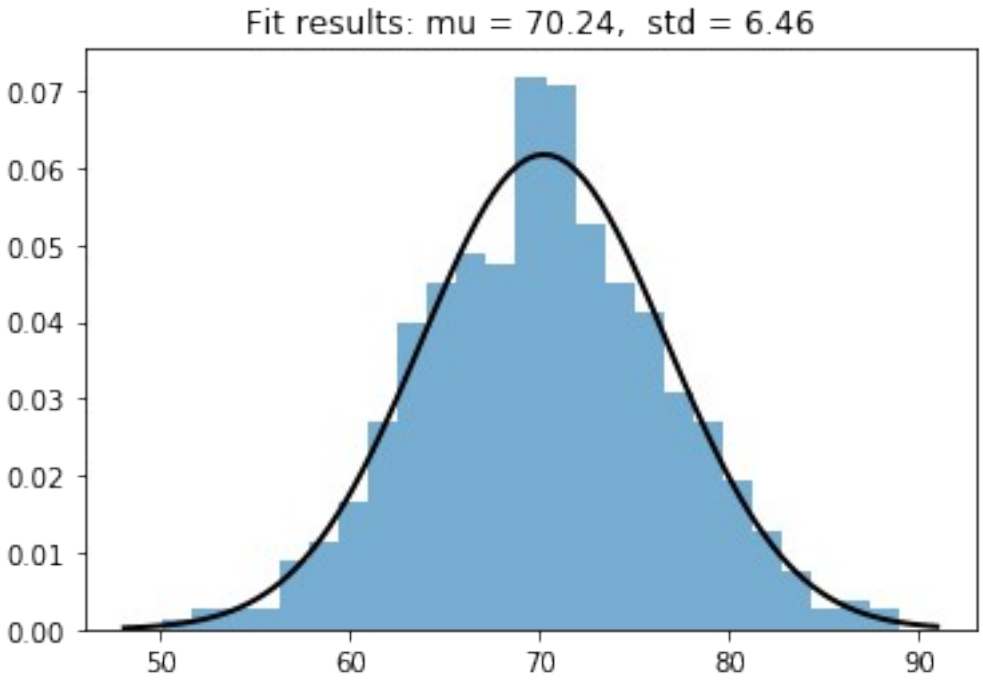


A média é 64,7 kg, e há mais valores abaixo dela. Temos alguns entre 60 e 70 kg e pelo menos 1 entre 110 e 120 kg.

Nosso histograma apresenta uma assimetria à direita (cauda mais comprida à direita), algo assim:

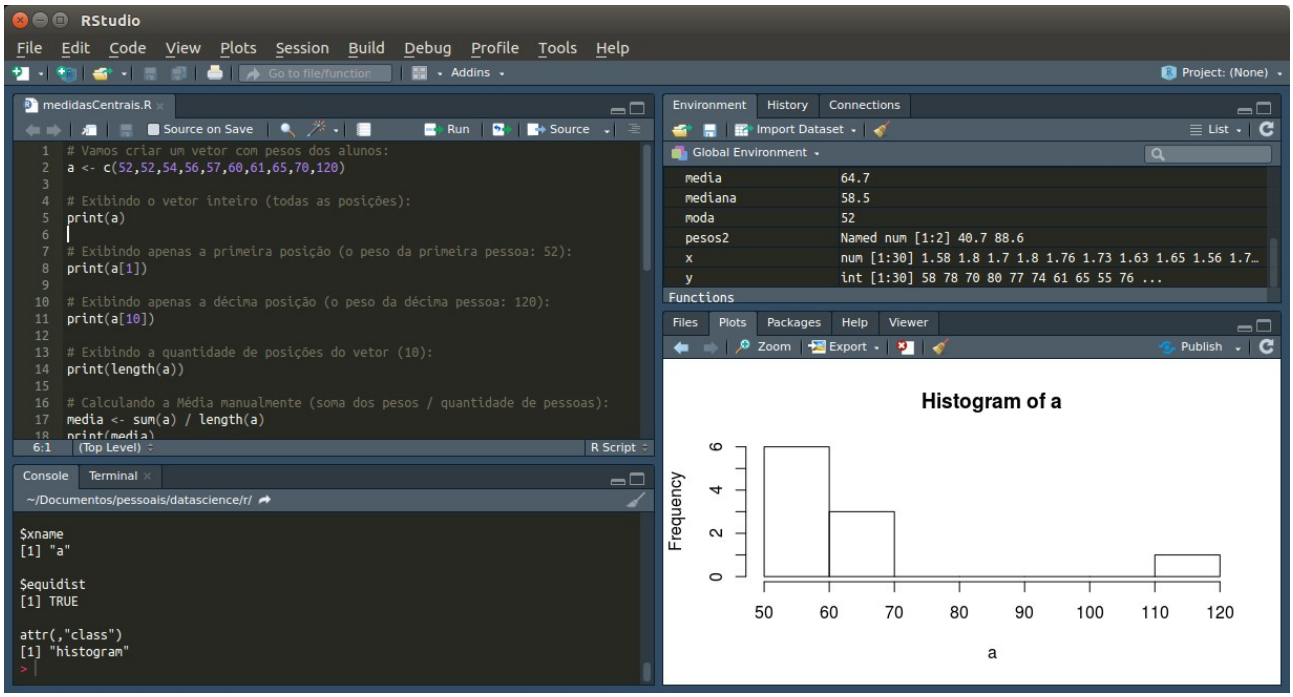


Isto denota outro fenômeno, que discutiremos posteriormente, que é a dispersão dos dados, com relação à média. Um histograma equilibrado seria como um sino:



Calculando as medidas centrais em R

Abra o RStudio. Podemos executar comandos de duas maneiras: Imediata ou scripts. Na forma imediata, digitamos os comandos na console (janela inferior esquerda) e os executamos imediatamente. Na forma de script, criamos (ou abrimos) um script R (um arquivo com extensão ".R") e digitamos os comandos, executando o script inteiro com o comando "source" (menu: "Code / Source", ou botão "source").



Abra o script "medidasCentrais.R" (menu: "File / Open File") e acompanhe:

```
# Vamos criar um vetor com pesos dos alunos:
a <- c(52,52,54,56,57,60,61,65,70,120)

# Exibindo o vetor inteiro (todas as posições):
print(a)

# Exibindo apenas a primeira posição (o peso da primeira pessoa: 52):
print(a[1])

# Exibindo apenas a décima posição (o peso da décima pessoa: 120):
print(a[10])

# Exibindo a quantidade de posições do vetor (10):
print(length(a))

# Calculando a Média manualmente (soma dos pesos / quantidade de pessoas):
media <- sum(a) / length(a)
print(media)

# Usando a função "mean()":
print(mean(a))

# Calculando a mediana dos pesos:
mediana <- median(a)
print(mediana)

# Moda (R não tem uma função nos pacotes padrões):
# Criando uma função para calcular a moda:
calcMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
```

```
}

# Invocando a função e calculando a moda do vetor (54):
moda <- calcMode(a)
print(moda)

# Mostrando um histograma com a distribuição dos pesos:
print(hist(a))
```

Criamos uma variável (um espaço na memória) chamado "a" e atribuímos a ela um vetor, contendo os pesos dos dez alunos, separados por vírgulas. Em R, usamos o operador seta ("<-") para inicializar variáveis. É possível substituir pelo sinal de igual ("="), mas é melhor se acostumar com a seta:

```
a <- c(52,52,54,56,57,60,61,65,70,120)
```

A partir deste momento, "a" é um vetor, como um vetor matemático, ela conterá posições. Cada posição contém o peso de um aluno associado:

- a[1] : Peso do primeiro aluno (52 kg);
- a[2] : Peso do segundo aluno (52 kg);
- a[5] : Peso do quinto aluno (57 kg);

Para calcular a média dos pesos, precisamos da soma dos pesos e da quantidade de alunos:

```
media <- sum(a) / length(a)
print(media)
```

Atribuímos à outra variável, chamada "media" (sem acentos mesmo), uma expressão matemática, formada pelo resultado da função "sum()" dividido pelo resultado da função "length()". A primeira, retorna o somatório dos pesos de todas as posições do vetor, cujo nome foi passado como parâmetro (a), e a segunda, retorna a quantidade de posições do vetor, cujo nome foi passado (a).

O comando "print()" mostra na console o resultado.

Funções

Além do "sum()" e do "length()", R possui várias funções prontas, como estas:

- mean(): Calcula a média;
- median(): Calcula a mediana;
- hist(): Desenha o histograma;

Porém, ele carece de uma função pronta para calcular a moda. Então, criamos uma função simples:

```
calcMode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Não importam os detalhes agora, apenas entenda que esta função calcula a moda de um vetor. E podemos invocá-la como fazemos com qualquer outra função:

```
moda <- calcMode(a)
```

A diferença é que esta função "calcMode()" só existe dentro do nosso código-fonte.

Execute o código e estude-o muito bem. Tente calcular as medidas de tendência central de outras amostras de dados. Experimente com valores contínuos também.

Cleuton Sampaio, M.Sc.