

Métodos Estatísticos Básicos

Aula 4 - Medidas de Dispersão

Regis A. Ely

Departamento de Economia
Universidade Federal de Pelotas

20 de julho de 2020

Conteúdo

Medidas de dispersão

- Amplitude

- Desvio quartil

- Boxplot

- Desvio médio absoluto

- Desvio-padrão

- Variância

Medidas de dispersão relativa

- Coeficiente de variação de Pearson

- Coeficiente de variação de Thorndike

Medidas de assimetria

Medidas de curtose

Momentos de uma distribuição

Medidas de dispersão

Medidas de dispersão são estatísticas que nos dão informação sobre a variabilidade dos dados. Podemos separá-las em 4 grupos:

- **Medidas de dispersão absolutas:** caracterizam a variabilidade de um conjunto de dados, porém não são comparáveis entre conjuntos com dados de magnitudes diferentes
- **Medidas de dispersão relativa:** possibilitam a comparação da variabilidade dos dados para conjuntos diferentes
- **Medidas de assimetria:** calculam a posição em que os maiores valores de um conjunto de dados se situam
- **Medidas de curtose:** calculam o grau de achatamento da distribuição dos dados

Amplitude

Amplitude total: valor máximo menos o valor mínimo da amostra

$$AT = X_{max} - X_{min}$$

- Medida de dispersão que não é centrada na média
- Para dados com intervalos de classe: $AT = L_{max} - l_{min}$
 - Sendo l_{min} o menor limite inferior das classes e L_{max} o maior limite superior
- A amplitude total não é afetada por valores intermediários

Amplitude

Vamos calcular a amplitude de uma amostra da altura de mulheres obtida na variável `women` do R. Antes de fazermos os cálculos, transformamos os dados de polegadas para centímetros e de libras para quilos utilizando o pacote `measurements`:

```
library(tidyverse)
library(measurements)
women_adj <- women %>%
  mutate(
    height = conv_unit(height, from = "inch", to = "cm"),
    weight = conv_unit(weight, from = "lbs", to = "kg")
  )
```

Amplitude

```
range(women_adj$height)
```

```
[1] 147.32 182.88
```

```
max(women_adj$height) - min(women_adj$height)
```

```
[1] 35.56
```

Note que a função `range` reporta o valor mínimo e máximo, enquanto que o segundo comando calcula a amplitude da amostra

Desvio quartil

Desvio quartil: é a média da diferença entre os quartis da distribuição, também chamado de *amplitude semi-interquartílica*

$$D_q = \frac{(Q_3 - Q_1)}{2}$$

- O desvio quartil é mais comum quando a medida de tendência central utilizada é a mediana
- O desvio quartil não é tao afetado por valores extremos como a amplitude

Ex: Calcule o desvio quartil de $\{40, 45, 48, 62, 70\}$

$Q_1 = 45$ e $Q_3 = 62$

$$D_q = \frac{62 - 45}{2} = 8,5$$

Desvio quartil

Vamos calcular os quartis e o amplitude semi-interquartílica dos dados do exemplo anterior no R:

```
quantile(women_adj$height)
```

0%	25%	50%	75%	100%
147.32	156.21	165.10	173.99	182.88

```
IQR(women_adj$height) / 2
```

```
[1] 8.89
```

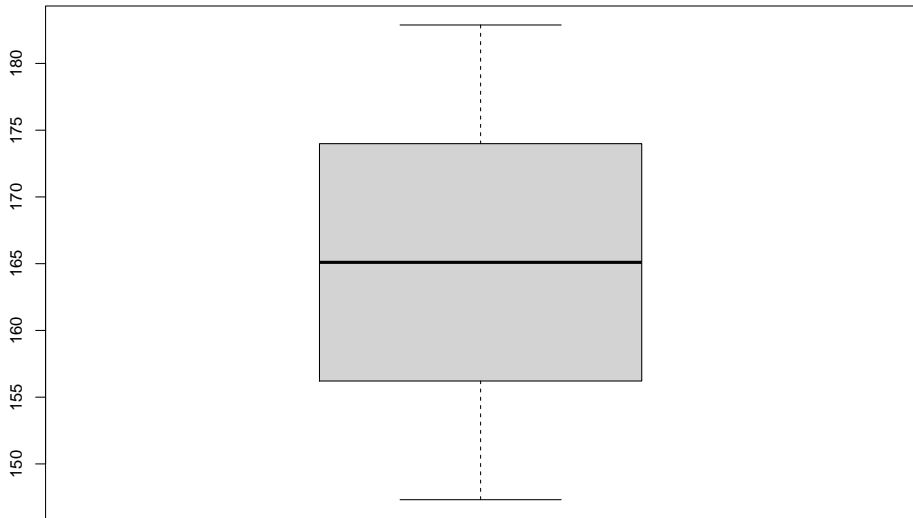
A função `quantile` reporta os quartis, enquanto que a função `IQR` calcula a amplitude interquartílica

Boxplot

- Uma maneira muito útil de visualizar os quartis e valores máximos e mínimos da distribuição dos dados é através de um **boxplot**
- O *boxplot* é um gráfico de caixa que marca com linhas horizontais a mediana, o primeiro e terceiro quartis, bem como o valor mínimo e máximo dos dados
- Podemos plotar este gráfico para a altura das mulheres utilizando a função `boxplot` no R:

```
boxplot(women_adj$height)
```

Boxplot



Desvio médio absoluto

Desvio médio absoluto: é a média aritmética dos valores absolutos dos desvios tomados em relação à média (ou à mediana)

$$D_m = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

- Podemos substituir a média, \bar{X} , pela mediana, \bar{M}_e para termos o desvio médio absoluto em relação à mediana

Ex: calcule o desvio médio absoluto de $\{-4, -3, -2, 3, 5\}$

$$\bar{X} = -0,2 \text{ e } M_e = -2$$

$$D_m = \frac{|-4+0,2|+|-3+0,2|+|-2+0,2|+|3+0,2|+|5+0,2|}{5} = 3,36$$

$$D_{me} = \frac{|-4+2|+|-3+2|+|-2+2|+|3+2|+|5+2|}{5} = 3$$

Desvio médio absoluto

Para dados agrupados devemos utilizar as frequências:

$$D_m = \frac{\sum_{i=1}^n f_i \cdot |X_i - \bar{X}|}{\sum_{i=1}^n f_i}$$

- Se os dados forem agrupados em intervalos de classe, então X_i será o ponto médio de cada classe

Desvio médio absoluto

No R podemos calcular o desvio médio absoluto através da função MeanAD do pacote DescTools:

```
library(DescTools)
MeanAD(women_adj$height, center = Mean)
```

```
[1] 9.482667
```

O argumento center pode ser utilizado para centrar os desvios ao redor da mediana, utilizando center = Median

Desvio-padrão

Desvio padrão: mede o grau de variação de um conjunto de elementos

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

$$\text{Ex: } \{-4, -3, -2, 3, 5\}$$

$$\bar{X} = -0,2$$

$$\sigma = \sqrt{\frac{(-4+0,2)^2 + (-3+0,2)^2 + (-2+0,2)^2 + (3+0,2)^2 + (5+0,2)^2}{5}} = \sqrt{12,56} = 3,54$$

Desvio-padrão

Para calcular o desvio-padrão a partir de uma amostra, fazendo um pequeno ajuste no denominador - Usaremos σ para denotar o desvio-padrão populacional e S para denotar o desvio-padrão amostral

Desvio padrão amostral: utilizamos uma pequena correção no caso de termos apenas uma amostra da população inteira

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

Desvio-padrão com dados agrupados

Se tivermos dados agrupados, ponderamos o desvio-padrão pelas frequências:

$$S = \sqrt{\frac{\sum_{i=1}^n f_i \cdot (X_i - \bar{X})^2}{\sum_{i=1}^n f_i}}$$

- Com intervalos de classe, a fórmula será a mesma, mas X_i será o ponto médio da classe

Desvio-padrão

No R podemos calcular o desvio-padrão amostral com a função `sd`:

```
sd(women_adj$height)
```

```
[1] 11.35923
```

Raramente teremos os dados da população inteira a nossa disposição, mas nesse caso, o desvio-padrão populacional pode ser calculado através do comando `sqrt(sum((x-mean(x))^2)/length(x))`

Propriedades do desvio-padrão

O desvio-padrão possui as seguintes propriedades:

1. O desvio-padrão nunca é um número negativo
2. Somando (ou subtraindo) uma constante a todos os valores de uma variável, o desvio-padrão não se altera
3. Multiplicando (ou dividindo) todos os valores de uma variável por uma constante (diferente de zero), o desvio-padrão será multiplicado (ou dividido) por essa constante

Exercício: checar estas propriedades no R utilizando os dados da variável `women$height`

Variância

Variância: é o desvio-padrão elevado ao quadrado

- A propriedade 1 continua válida para a variância, mas a propriedade 2 se altera, pois se multiplicarmos todos os valores por uma constante (diferente de zero), a variância será multiplicada por essa mesma constante elevada ao quadrado

Exemplo no R:

```
var(women_adj$height)
```

```
[1] 129.032
```

Coeficiente de variação de Pearson

Coeficiente de variação de Pearson (CVP): caracteriza a dispersão dos dados em relação ao seu valor médio

$$CVP = \frac{S}{\bar{X}} \times 100$$

- Onde S se refere ao desvio-padrão amostral
- Um desvio padrão de 2 pode ser grande para dados cuja média é 20, mas pequeno se a média é 200. O CVP padroniza as variações, possibilitando a comparação entre dados distintos

Coeficiente de variação de Pearson

Exemplo no R: embora o desvio-padrão da altura seja maior que do peso, quando normalizados através do coeficiente de variação, podemos observar que o peso varia mais entre as mulheres do que a altura

```
library(raster)
women_adj %>%
  summarise(
    sd_height = sd(height),
    sd_weight = sd(weight),
    cv_height = cv(height),
    cv_weight = cv(weight)
  )
```

	sd_height	sd_weight	cv_height	cv_weight
1	11.35923	7.03009	6.880209	11.33498

Coeficiente de variação de Thorndike

Coeficiente de variação de Thorndike (CVT): utilizamos a mediana para o cálculo do coeficiente de variação

$$CVT = \frac{s}{Me} \times 100$$

Exemplo no R: coeficiente de variação de Thorndike para o peso das mulheres

```
(sd(women$weight) / median(women$weight)) * 100
```

```
[1] 11.48051
```

Medidas de assimetria

- As medidas de assimetria calculam a posição em que os maiores valores de um conjunto de dados se situam
- O valor da assimetria depende da relação entre média e mediana, sendo que uma distribuição pode ser classificada como:

Simétrica: os dados têm uma distribuição simétrica quando Média = Mediana

Assimétrica à esquerda: os dados têm assimetria negativa quando Média < Mediana

Assimétrica à direita: os dados têm assimetria positiva quando Média > Mediana

Coeficiente de assimetria

Coeficiente de assimetria de Pearson: compara graus de assimetria entre distribuições diferentes¹

$$CAP = \frac{1}{n} \sum_{i=1}^n \left[\frac{x_i - \bar{X}}{S} \right]^3$$

Classificação:

- $CAP = 0 \Rightarrow$ Distribuição simétrica
- $CAP < 0 \Rightarrow$ Assimetria negativa (ou à esquerda)
- $CAP > 0 \Rightarrow$ Assimetria positiva (ou à direita)

¹Se os dados estiverem agrupados, devemos considerar os pesos na fórmula.

Medidas de assimetria

Utilizamos a função `skewness` do pacote `e1071` para calcular o coeficiente de assimetria:

```
library(e1071)
skewness(women_adj$height)
```

```
[1] -4.731446e-16
```

```
skewness(women_adj$weight)
```

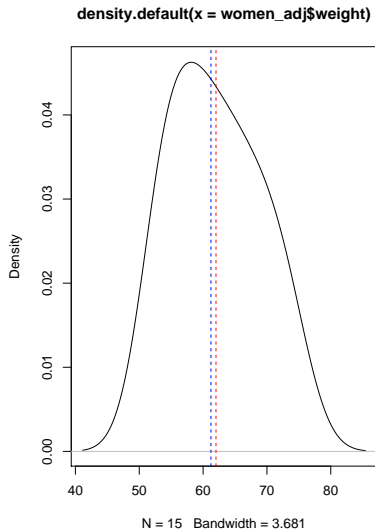
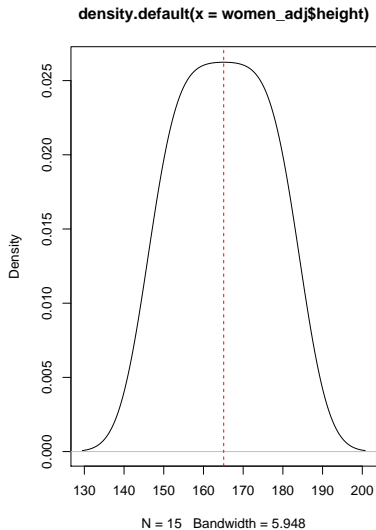
```
[1] 0.2276454
```

Medidas de assimetria

Podemos plotar a densidade da distribuição das variáveis de altura e peso das mulheres e checar a assimetria comparando a média com a mediana dos dados no R:

```
par(mfrow = c(1,2))  
plot(density(women_adj$height))  
abline(v=mean(women_adj$height), lty=2, col="red")  
plot(density(women_adj$weight))  
abline(v=mean(women_adj$weight), lty=2, col="red")  
abline(v=median(women_adj$weight), lty=2, col="blue")
```

Medidas de assimetria



Medidas de curtose

A curtose é o grau de achatamento de uma distribuição em relação à distribuição normal (em forma de sino)²

$$K = \frac{1}{n} \sum_{i=1}^n \left[\frac{x_i - \bar{X}}{S} \right]^4$$

- A fórmula acima se refere ao *momento de curtose*
- É comum também a utilização do *excesso de curtose*, definido por $K - 3$

²Se os dados estiverem agrupados, devemos considerar os pesos na fórmula.

Medidas de curtose

Classificação:

- $K = 3 \Rightarrow$ **Mesocúrtica**: distribuição não é nem achatada nem alongada (igual a distribuição normal)
- $K > 3 \Rightarrow$ **Leptocúrtica**: apresenta uma distribuição mais alongada do que a normal
- $K < 3 \Rightarrow$ **Platicúrtica**: apresenta uma distribuição mais achatada do que a normal

Medidas de curtose

Podemos calcular o excesso de curtose utilizando a função `kurtosis` no R. Para calcular o momento de curtose basta somar 3 ao resultado:

```
kurtosis(women_adj$height) + 3
```

```
[1] 1.558667
```

```
kurtosis(women_adj$weight) + 3
```

```
[1] 1.6553
```

Momentos de uma distribuição

- O n -ésimo momento de uma distribuição é definido como $E[X^n]$
- As estatísticas que vimos até aqui descrevem quatro momentos da distribuição de um conjunto de dados:
 1. A **média** é o primeiro momento de uma distribuição: $\mu = E[X]$
 2. A **variância** é o segundo momento centrado de uma distribuição:
 $\sigma^2 = E[(X - \mu)^2]$
 3. A **assimetria** é o terceiro momento padronizado de uma distribuição: $CAP = E[(X - \mu)^3]/\sigma^3$
 4. A **curtose** é o quarto momento padronizado de uma distribuição:
 $K = E[(X - \mu)^4]/\sigma^4$

Momentos de uma distribuição

