MAT02018 - Estatística Descritiva

Medidas de forma

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2021

Introdução

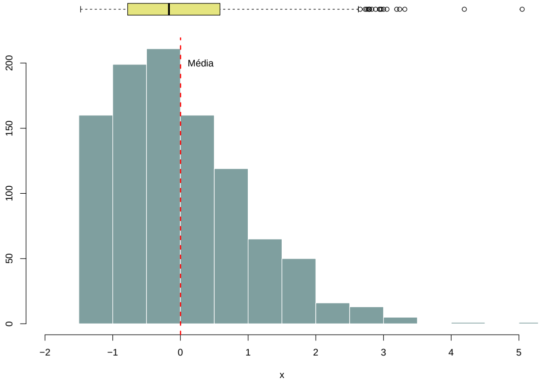
Introdução

- Nas últimas semanas dedicamos a nossa atenção para resumir numericamente certas características da distribuição de frequências.
- No entanto, as medidas de localização e variabilidade nem sempre dão conta de descrever todas as características de uma distribuição.
- Nestas breves notas de aula, vamos discutir duas **medidas de forma:** os coeficientes de **assimetria** e **curtose**.
- Além disso, vamos apresentar propriedades da média e variância.

Assimetria

Assimetria

- A medida denominada **coeficiente de assimetria** mede o grau de assimetria exibido pela distribuição dos dados.
- A figura abaixo apresenta uma **distribuição assimétrica**.



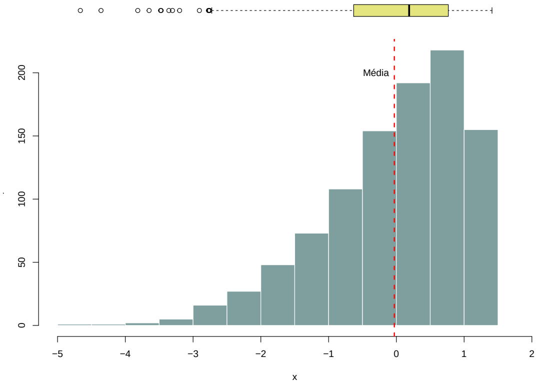
Assimetria

- O histograma revela que há mais observações abaixo da média.
- Veja que a mediana (apresentada no *boxplot*) está posicionada à esquerda da média (logo, mais de 50% dos valores da distribuição estão abaixo da média).
- Também é possível notar que a classe modal (faixa de maior frequência; maior retângulo do histograma) está à esquerda da mediana (e portanto, à esquerda da média).
- Este comportamento é conhecido como **assimetria positiva**[1].

[1] Quando $\bar{x} > Md > Mo$ dizemos que uma distribuição é assimétrica positiva.

Assimetria

- Já quando a média é menor que a mediana os dados apresentam **assimetria negativa**[2], como na figura a seguir.



[2] Quando $\bar{x} < Md < Mo$ dizemos que uma distribuição é assimétrica negativa.

Assimetria

- O **coeficiente de assimetria**[3] é calculado primeiro somando o **cubo dos desvios da média** e, então, dividindo pelo produto do **cubo do desvio padrão** pelo número de observações:

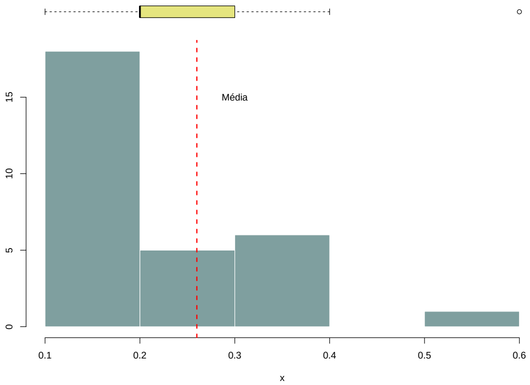
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}.$$

- Quando $b_1 = 0$ os dados são simétricos em torno da média;
- Quando $b_1 > 0$ os dados apresentam assimetria positiva;
- Quando $b_1 < 0$ os dados apresentam assimetria negativa.

[3] Existem outras propostas de mensuração da assimetria na literatura. Nestas notas vamos nos ater a apenas uma proposta.

Assimetria

- **Exemplo:** considere mais uma vez que x representa a largura de pétala de 30 flores do tipo Íris. Observando os gráficos de boxplot e histograma, é notável que a distribuição é assimétrica positiva.



Assimetria

- Vamos utilizar uma planilha para auxiliar na organização das operações necessárias para o cálculo do coeficiente de assimetria b_1.

-	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
	0,4	0,14	0,020	0,0027
	0,6	0,34	0,116	0,0393
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,3	0,04	0,002	0,0001
	0,1	-0,16	0,026	-0,0041
	0,4	0,14	0,020	0,0027
	0,2	-0,06	0,004	-0,0002
	0,3	0,04	0,002	0,0001
	0,4	0,14	0,020	0,0027
	0,1	-0,16	0,026	-0,0041
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,3	0,04	0,002	0,0001
	0,3	0,04	0,002	0,0001

Assimetria

	0,4	0,14	0,020	0,0027
	0,2	-0,06	0,004	-0,0002
	0,3	0,04	0,002	0,0001
	0,1	-0,16	0,026	-0,0041
	0,4	0,14	0,020	0,0027
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,2	-0,06	0,004	-0,0002
	0,4	0,14	0,020	0,0027
Soma	7,8	0,00	0,372	0,0406

Assimetria

Assim, temos

- $\bar{x} = 7,8/30 = 0,26$;
- $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{0,372}{29} = 0,0128$;
- $s = \sqrt{s^2} = 0,1133$;
- $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3} = \frac{0,0406}{30 \times (0,1133)^3} = 0,93 > 0$, confirmando a assimetria positiva que o gráfico já indicava.

Curtose

Curtose

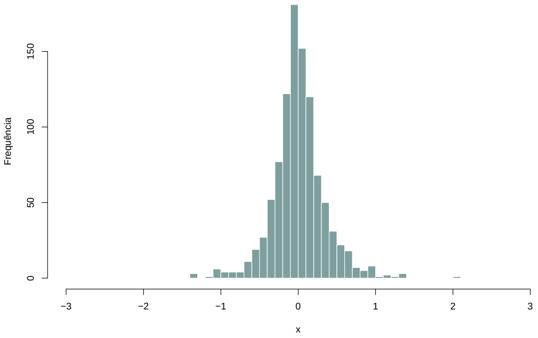
- A **curtose** mede o **alongamento da distribuição**.
- Sua definição é semelhante à da assimetria[4], com a ressalva de que a **quarta potência** é usada em vez da terceira:

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4}.$$

[4]Mais uma vez, chamamos a atenção para o fato que existem outras propostas de medir a curtose, porém vamos discutir apenas uma delas.

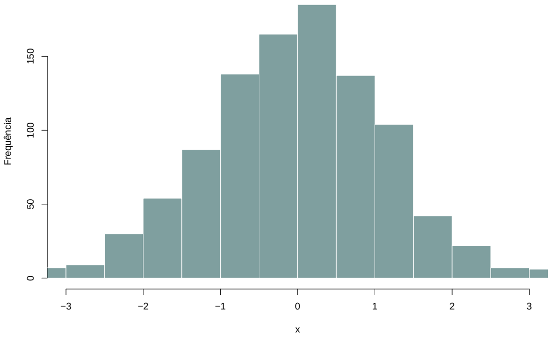
Curtose

- Distribuições com um elevado grau de alongamento são chamadas de **leptocúrticas** e têm valores de **curtose acima de 3**.



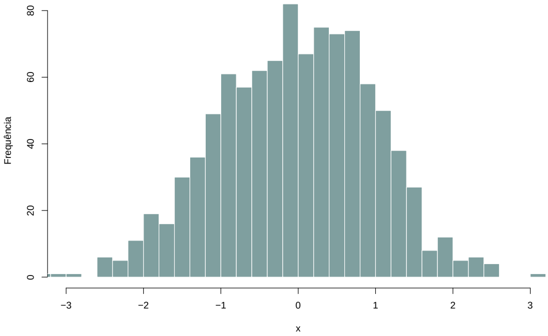
Curtose

- Distribuições achatadas são chamadas **platicúrticas** e têm valores para **curtose inferiores a 3**.



Curtose

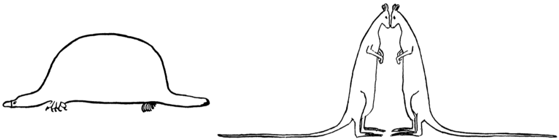
- No caso em que a **curtose é igual a 3**, a distribuição é chamada **mesocúrtica**.



Curtose

Origem dos termos

* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the



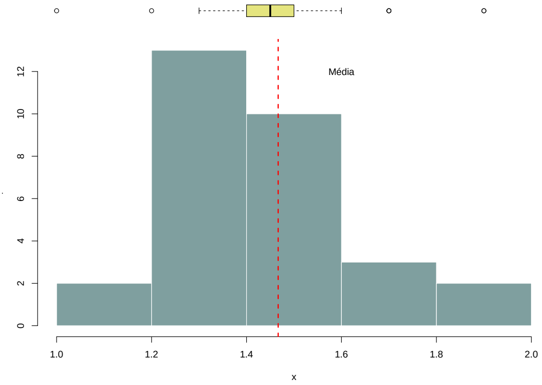
normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Student (em verdade, William Sealy Gosset) em seu artigo "*Errors of Routine Analysis*" de 1927, apresenta em uma nota de rodapé a explicação para os termos **platicúrtica** e **leptocúrtica**.

- Livre tradução: *Eu mesmo tenho em mente o significado das palavras da **memoria technica** acima, onde a primeira figura representa o ornitorrinco (**platypus**), e a segunda cangurus, conhecidos por "lebrar" ("**lepping**"), e portanto, talvez, com igual razão, devam ser lebres!*

Curtose

- **Exemplo:** novamente considere os dados de 30 flores de iris. A variável x agora representa o comprimento de pétala de 30 flores do tipo Íris. Observando o histograma de x, a distribuição parece ser **leptocúrtica**.



Curtose

- Vamos utilizar uma planilha para auxiliar na organização das operações necessárias para o cálculo do coeficiente de curtose b_2.

-	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^4$
	1,5	0,033	0,001	0,0000
	1,6	0,133	0,018	0,0003
	1,4	-0,067	0,004	0,0000
	1,7	0,233	0,054	0,0030
	1,3	-0,167	0,028	0,0008
	1,5	0,033	0,001	0,0000
	1,3	-0,167	0,028	0,0008
	1,5	0,033	0,001	0,0000
	1,7	0,233	0,054	0,0030
	1,7	0,233	0,054	0,0030
	1,4	-0,067	0,004	0,0000
	1,4	-0,067	0,004	0,0000
	1,5	0,033	0,001	0,0000
	1,3	-0,167	0,028	0,0008
	1,4	-0,067	0,004	0,0000
	1,3	-0,167	0,028	0,0008
	1,5	0,033	0,001	0,0000

Curtose

	1,4	-0,067	0,004	0,0000
	1,4	-0,067	0,004	0,0000
	1,5	0,033	0,001	0,0000
	1,9	0,433	0,188	0,0353
	1,9	0,433	0,188	0,0353
	1,4	-0,067	0,004	0,0000
	1,4	-0,067	0,004	0,0000
	1,3	-0,167	0,028	0,0008
	1,5	0,033	0,001	0,0000
	1,2	-0,267	0,071	0,0051
	1,6	0,133	0,018	0,0003
	1,0	-0,467	0,218	0,0474
	1,5	0,033	0,001	0,0000
Soma	44,0	0,000	1,047	0,1366

Curtose

Assim, temos

- $\bar{x} = 44/30 = 1,47$;
- $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1,047}{29} = 0,0361$;
- $s = \sqrt{s^2} = 0,19$;
- $b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} = \frac{0,1366}{30 \times (0,19)^3} = 3,49 > 3$, confirmando o que o gráfico da distribuição já indicava (distribuição **leptocúrtica**).

Propriedades da média e variância

Propriedades da média e variância

- Sejam $x_1, x_2, \ldots, x_n$ valores observados da variável x, tal que

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

são a média, e a variância da variável x.

Propriedades da média e variância

- **Exemplo (relembrando):** considere o peso em quilos de cinco ovelhas.

Ovelha 1	Ovelha 2	Ovelha 3	Ovelha 4	Ovelha 5
24	22,5	22,5	24	24,5

Propriedades da média e variância

- Temos que o peso médio e a variância são

$$\bar{peso} = 23,5; \quad s^2_{peso} = 0,875.s^2_{peso} = 0,875.$$

Propriedades da média e variância

- Defina a transformação linear y_i de x_i

$$y_i = ax_i + b, i = 1, 2, \ldots, n,$$

em que a e b **são constantes**.

- Então, temos como **propriedades da média e variância**:

$$\bar{y} = a\bar{x} + b, \quad \text{e} \quad s_y^2 = a^2 s_x^2.$$

Propriedades da média e variância

- Logo, $s_y = \sqrt{s_y^2} = \sqrt{a^2 s_x^2} = as_x$.
  - Ou seja, temos que a média e variância da variável transformada pode ser expressa em função da média, variância e as constantes a e b da variável original.
- Note também que a variância **não é influenciada pela constante** b.

Propriedades da média e variância

- É fácil demonstrar tais propriedades, partindo da definição da média e variância e aplicando a transformação.

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (ax_i + b) \\
&= \frac{1}{n}\left(a\sum_{i=1}^{n} x_i + n \times b\right) \\
&= a\frac{\sum_{i=1}^{n} x_i}{n} + n \times \frac{b}{n} \\
&= a\bar{x} + b,
\end{aligned}
$$

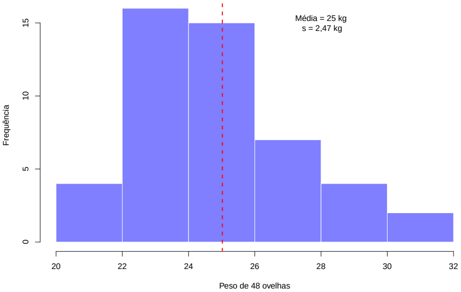
e

Propriedades da média e variância

$$
\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[ax_i+b-(a\bar{x}+b)]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[ax_i-a\bar{x}+b-b)]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[a(x_i-\bar{x})]^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left[a^2(x_i-\bar{x})^2\right] \\
&= a^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \\
&= a^2 s_x^2.
\end{aligned}
$$

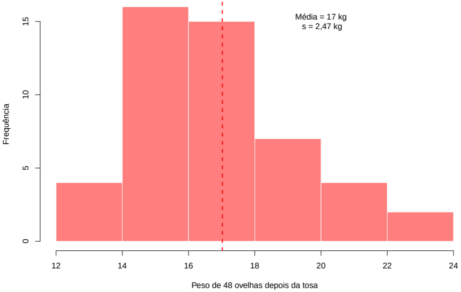
Propriedades da média e variância

- **Exemplos (1):** considere um conjunto de 48 ovelhas. A distribuição do peso deste grupo de animais é apresentada a seguir.



Propriedades da média e variância

- Suponha agora, que após a tosa, cada ovelha "perde" **exatamente** 8 kg. Após a tosa, a distribuição do peso é mais uma vez apresentada.



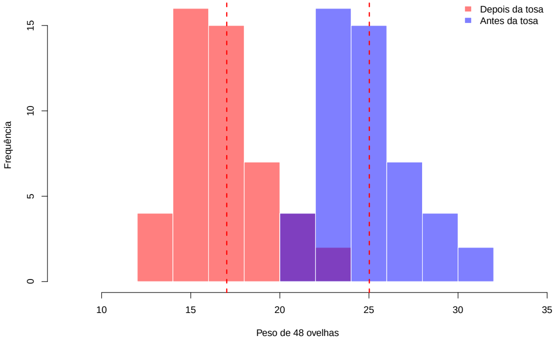
Propriedades da média e variância

- Note que, se o peso antes da tosa é representado por x, então podemos escrever o peso após a tosa como $y = ax + b$, em que $a = 1$ e $b = -8$.
- Percebemos que a média diminui **exatamente** 8 kg em relação a média do peso antes da tosa[5].
  - Já a variância após a tosa é a mesma que a variância do peso antes da tosa[6].

[5] $\bar{y} = (1)\bar{x} - 8 = 25 - 8 = 17$.
[6] $s_y^2 = (1)^2 s_x^2 = (1)(2,47)^2 = (2,47)^2$.

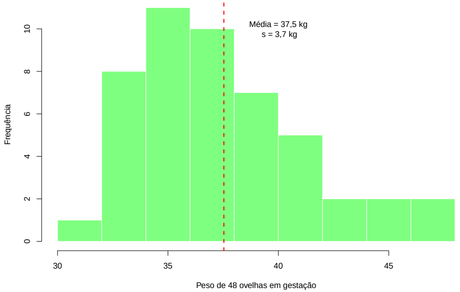
Propriedades da média e variância

- Comparando as distribuições de x (peso antes da tosa) e y (peso após a tosa) percebemos que a forma da distribuição não muda, mas a distribuição é deslocada (em oito unidades) para a esquerda (constante b negativa).



Propriedades da média e variância

- **Exemplos (2):** considere mais uma vez o conjunto de 48 ovelhas (com o seu peso original; antes da tosa). Suponha que quando as ovelhas estão em gestação[7], o seu peso é **exatamente** 1,5 vezes o seu peso original.



[7] Considere um certo momento da gestação das ovelhas (o final da gestação, por exemplo).

Propriedades da média e variância

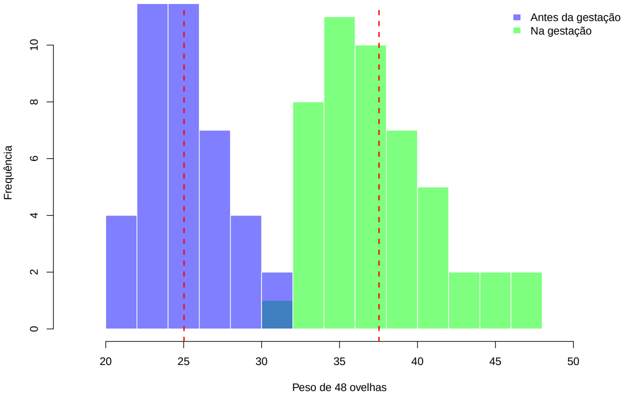
- Mais uma vez, representando o peso antes da tosa por x, podemos escrever o peso na gestação como $y = ax + b$, em que $a = 1,5$ e $b = 0$.
- Percebemos que a média é **exatamente** 1,5 a média do peso antes da gestação[8].
- A variância do peso na gestação é a variância do peso original multiplicado por $1,5^2$ (a^2)[9].

[8] $\bar{y} = (1,5)\bar{x} - 0 = (1,5)25 = 37,5.$
[9] $s_y^2 = (1,5)^2 s_x^2 = (1,5)^2(2,47)^2 = 13,73.$

Propriedades da média e variância

- Comparando as distribuições de x (peso original) e y (peso na gestação) percebemos que a distribuição do peso é deslocada para a direita ($a > 0$) e a dispersão aumenta ($|a| > 1$).

Propriedades da média e variância



Padronização

Padronização

- Uma transformação linear do tipo $y_i = ax_i + b$ merece destaque.
  - Se especificarmos $a = \frac{1}{s_x}$ e $b = -\frac{\bar{x}}{s_x}$ temos a seguinte transformação

$$z_i = \left(\frac{1}{s_x}\right)x_i + \left(-\frac{\bar{x}}{s_x}\right) = \frac{x_i - \bar{x}}{s_x}, i = 1, \ldots, n.$$

Padronização

- Como se compartam $\bar{z}$ e s_z?

$$\bar{z} = \frac{1}{s_x}\bar{x} - \frac{\bar{x}}{s_x} = 0, \quad \text{e} \quad s_z^2 = \left(\frac{1}{s_x}\right)^2 s_x^2 = 1.$$

Padronização

- Esta transformação é também chamada de **padronização** e a variável resultante z é chamada de **escore padronizado** de x.
- Como visto, esta transformação tem como propriedades média zero e variância um.
- O escore padronizado pode ser interpretado como o **número de desvios padrões** que uma observação está da média.
  - Os dados abaixo da média (na escala original) têm escores z padronizados negativos;
  - Os dados acima da média possuem z-escores positivos.

Padronização

Observações

1. Uma **regra geral** para detecção de observações atípicas em uma distribuição é, após padronizar a variável, avaliar valores z acima de 1,96 e valores de z abaixo de -1,96 ($|z| > 1,96$)[10].
2. Com a padronização é possível produzir **uma escala comum** para variáveis que foram medidas em escalas diferentes (por exemplo, peso em quilos e altura em metros). Isso permite que possamos combinar diversas variáveis, como é o caso de alguns indicadores.

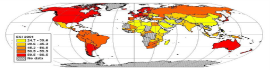
---

[10] Esta regra geral tem sua justificativa nas propriedades da distribuição de probabilidades **normal padrão** (média zero e variância um). Uma regra mais prática é considerar valores atípicos como $|z| > 2$, pois $1,96 \approx 2$.

Padronização

- O Índice de Sustentabilidade Ambiental (*Environmental Sustainability Index* - ESI) é resultado da colaboração entre:
  - *World Economic Forums Global Leaders for Tomorrow* (GLT) *Environment Task Force*
  - *Yale Center for Environmental Law and Policy* (YCELP)
  - *Columbia University Center for International Earth Science Information Network* (CIESIN)

- O ESI é uma medida de progresso global em direção à sustentabilidade ambiental desenvolvido para 122 países (ESI 2001).
- A alta classificação do ESI indica que um país tem alcançado um nível mais alto de sustentabilidade ambiental que a maioria dos outros países.
- Uma baixa posição do ESI sinaliza que um país enfrenta sérios problemas para a sustentabilidade ambiental ao longo de múltiplas dimensões.
- As pontuações ESI são baseadas em um conjunto de 22 indicadores centrais, cada um destes combina de duas a seis variáveis para um total de 67 variáveis subjacentes.
- Os indicadores e as variáveis foram escolhidos através de uma revisão cuidadosa da literatura ambiental e dados disponíveis, combinados com ampla consulta e análise.



(1) Padronização das variáveis: as 67 variáveis são padronizadas, ou seja, diminui-se a média e divide-se pelo respectivo desvio padrão.
- **Exemplo:** variável SO2 padronizada é obtida da seguinte forma:

$$Z_{SO2} = \frac{SO2 - \overline{SO2}}{DP(SO2)}.$$

  Desta forma, temos que a média de Z_{SO2} é zero e seu desvio padrão é um.
- **Por que padronizar?** A padronização permite que as 67 variáveis variem na mesma escala.

(2) Cálculo dos indicadores: cada indicador é a média das variáveis padronizadas que compõem o indicador.
- **Exemplo:** o indicador *Air Quality* é composto pelas variáveis Z_{SO2}, Z_{NO2} e Z_{TSP}. O indicador é portanto

$$\text{SYS\_AIR} = \frac{Z_{SO2} + Z_{NO2} + Z_{TSP}}{3}.$$

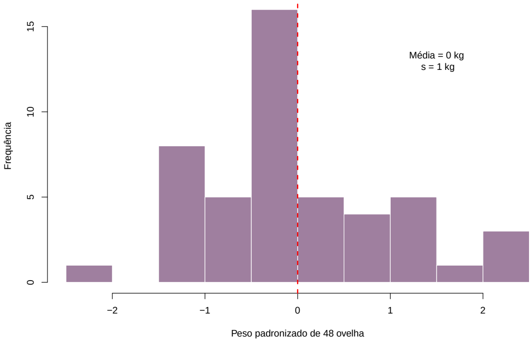
  Para o Brasil, temos $Z_{SO2} = -2,42$; $Z_{NO2} = -0,13$; $Z_{TSP} = 0,53$, e assim

$$\text{SYS\_AIR} = \frac{-2,42 - 0,13 + 0,53}{3} = \frac{-2,02}{3} = -0,67.$$

- Aqui se nota a importância das variáveis estarem na mesma escala, assim o cálculo dos indicadores não é afetado pela escala das variáveis.

Padronização

- **Exemplos (3):** veja como fica a distribuição do peso das 48 ovelhas quando padronizado.



Para casa

Para casa

A tabela a seguir apresenta as **frequências absolutas** dos 48 valores observados da variável **peso de porcos**:

Figure 1: Eram porcos, e não ovelhas!

Para casa

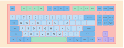
Peso (x_i)	Frequência (n_i)
20,0	1
21,5	2
22,0	1
22,5	5
23,0	1
23,5	4
24,0	6
24,5	8
25,0	2
25,5	3
26,0	2
26,5	2
27,0	2
27,5	3
28,5	2
29,5	1
30,0	1
31,0	2

Exercício

- Quem são os valores de mínimo e máximo da variável peso?
  - Qual é a amplitude de variação?
- Qual é a média da variável peso?
- Qual é o desvio padrão da variável peso?
- Quem é a mediana da variável peso?
- Considerando que o porco é classificado como **baixo peso** se este possui peso **menor ou igual a 23**, qual é o percentual de porcos com baixo peso?
- Utilizando gráficos ou medidas resumo, avalie a assimetria e a curtose da distribuição do peso.

ComplementaR

ComplementaR



Esta seção é complementar. São apresentadas algumas poucas funções em R relacionadas a discussão da aula. Para tal, vamos utilizar o exemplo original de (BUSSAB; MORETTIN, 2017) sobre os dados dos empregados da seção de orçamentos da Companhia MB. A planilha eletrônica correspondente encontra-se no arquivo `companhia_mb.xlsx`. Vamos começar carregando os dados para o R. Existem várias formas de se carregar **arquivos de dados** em diferentes no R. Como arquivo de interesse encontra-se no formato do Excel (xlsx), vamos utilizar a função `read_excel` do pacote `readxl`[11].

[11] Caso você não tenha o pacote, instale-o:`install.packages("readxl")`.

ComplementaR

```r
# install.packages("readxl")
library(readxl)

dados <- read_excel(path = "companhia_mb.xlsx")
```

```r
class(dados) # classe do objeto dados
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```r
dim(dados) # dimensão do objeto dados
```

```
## [1] 36  7
```

ComplementaR

- Note que o objeto dados é uma tabela de dados bruto.

```
head(dados) # apresenta as primeiras linhas do objeto dados
```

```
## # A tibble: 6 x 7
##       N `Estado Civil` `Grau de Instruç~ `N de Filhos` `Salario (x Sal M~ Id
##   <dbl> <chr>          <chr>                     <dbl>              <dbl> <d
## 1     1 solteiro       ensino fundament~            NA               4
## 2     2 casado         ensino fundament~             1               4.56
## 3     3 casado         ensino fundament~             2               5.25
## 4     4 solteiro       ensino médio                 NA               5.73
## 5     5 solteiro       ensino fundament~            NA               6.26
## 6     6 casado         ensino fundament~             0               6.66
```

ComplementaR

- As funções `skewness` e `kurtosis` do pacote `e1071` calcula, respectivamente, o coeficiente de assimetria e curtose, sendo que este último retorna $b_2 - 3$.
- Assim, o ponto de referência para classificar uma distribuição como mesocúrtica $(b_2 - 3 = 0)$, platicúrtica $(b_2 - 3 < 0)$, ou leptocúrtica $(b_2 - 3 > 0)$ é o zero.

```
# install.packages("e1071")
library(e1071)

skewness(dados$Idade)
```

```
## [1] -0.06162251
```

```
kurtosis(dados$Idade)
```

```
## [1] -0.7619338
```

```
skewness(dados$`Salario (x Sal Min)`)
```

```
## [1] 0.5997938
```

ComplementaR

```
kurtosis(dados$`Salario (x Sal Min)`)
```

```
## [1] -0.3291263
```

```
skewness(dados$`N de Filhos`, na.rm = TRUE)
```

```
## [1] 0.6360186
```

```
kurtosis(dados$`N de Filhos`, na.rm = TRUE)
```

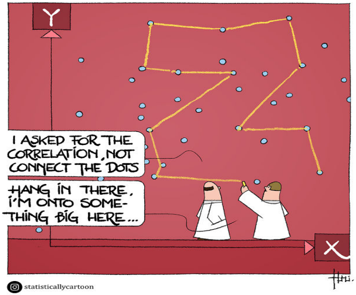
```
## [1] 0.2211539
```

Próxima aula

- Distribuições bivariadas.

Por hoje é só!

Bons estudos!



BUSSAB, W. de O.; MORETTIN, P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.