Métodos Estatísticos Básicos

Aula 2 - Distribuição de Frequências

Regis A. Ely

Departamento de Economia
Universidade Federal de Pelotas

16 de maio de 2020

Conteúdo

O que é frequência na estatística?

Conceitos preliminares

Distribuição de frequência simples

Distribuição de frequência com intervalos de classe
- Elementos da frequência com intervalos de classe
- Construção de intervalos de classe

Tipos de distribuição de frequências

Exemplo no R
- Carros, cavalos e cilindradas
- Gráfico de barras, histograma e polígono de frequência

O que é frequência na estatística?

A frequência de uma variável é o número de ocorrências em uma determinada categoria ou período de tempo

- A distribuição de frequência sintetiza informação sobre os dados
- Existem diversas maneiras de agruparmos dados
- A distribuição de frequência é normalmente apresentada em tabelas ou gráficos

Conceitos preliminares

Alguns conceitos que utilizaremos nesta aula

- **Dados brutos**: tabela ou arquivo com os dados originais, sem qualquer tipo de organização numérica
- **Rol**: tabela obtida após a ordenação dos dados de forma crescente ou decrescente
- **Intervalo de classe**: intervalos de valores igualmente espaçados para agrupar dados

Distribuição de frequência simples

A distribuição de frequência simples é uma tabela com os dados condensados de acordo com a repetição de seus valores

- Utilizada para variáveis qualitativas ou quantitativas discretas

Idade	Frequência
21	1
22	4
23	2
24	5
25	2
26	3

Distribuição de frequência com intervalos de classe

Podemos agrupar os dados em intervalos de classe

- Essencial quando trabalhamos com variáveis contínuas

Idade	Frequência
$21 \mid\!- 23$	5
$23 \mid\!- 25$	7
$25 \mid\!- 27$	5

O símbolo $\mid\!-$ significa que incluímos o valor à esquerda e excluímos o valor à direita, outra notação possível é $[21, 23)$

Elementos da frequência com intervalos de classe

A partir de uma frequência com intervalo de classe podemos definir:

- **Classe**: intervalos que utilizamos para agrupar os dados, sendo K o número total de classes e i o número da classe (*ex: na tabela anterior,* $K = 3$ *e* $K_2 = 23 \vdash 25$)
- **Limites de classe**: são os extremos de cada classe, sendo l_i o limite inferior da classe e L_i o limite superior (*ex:* $l_2 = 23$ *e* $L_2 = 25$)
- **Amplitude do intervalo de classe**: diferença entre o limite superior e inferior da classe, $h_i = L_i - l_i$ (*ex:* $h_2 = 2$)
  - A amplitude deve ser sempre igual entre os intervalos de classe

Elementos da frequência com intervalos de classe

- **Amplitude total da distribuição**: diferença entre o limite superior da última classe e o limite inferior da primeira classe, $AT = L(max) - l(min)$ ($ex:\ AT = 27 - 21 = 6$)
- **Amplitude total da amostra**: diferença entre o valor máximo e o valor mínimo da amostra, $AA = X_{max} - X_{min}$ ($ex:$ $AA = 26 - 21 = 5$)
  - AT será sempre maior ou igual a AA
- **Ponto médio da classe**: ponto que divide o intervalo de classe em duas partes iguais, $X_i = \frac{l_i + L_i}{2}$ ($ex:\ X_2 = \frac{23+25}{2} = 24$)

Construção de intervalos de classe

Como dividir um conjunto de dados em intervalos de classe?

1. Calculamos a amplitude da amostra, AA
2. Definimos o número de classes, K
   - **Regra de Sturges**: número inteiro imediatamente superior à $K = 1 + log_2 n$, sendo n o total de observações *(Sturges, 1926)*
3. Definimos a amplitude dos intervalos de classe: $h > AA/K$
4. Ajustamos o limite inferior da primeira classe e a amplitude dos intervalos de classe conforme desejarmos[1]

[1] Deve-se construir intervalos que não tenham frequência zero e que respeitam o limite inferior da amplitude (*regra do item 3*)

Tipos de distribuição de frequências

Existem quatro tipos de distribuição de frequências

- **Frequência simples absoluta**: número de observações de cada classe, f_i, sendo a soma sempre igual ao total da amostra, $\sum_i f_i = n$
- **Frequência relativa**: razão entre a frequência absoluta de cada classe e a frequência total da distribuição, fr_i, sendo a soma sempre igual a um, $\sum_i fr_i = 1$
- **Frequência simples acumulada**: soma acumulada das frequências simples absolutas, F_i
- **Frequência relativa acumulada**: soma acumulada das frequências relativas, Fr_i.

Tipos de distribuição de frequências

Resumindo todos os tipos de distribuição de frequências em uma tabela

Classes	f_i	fr_i	F_i	Fr_i	X_i
$21 \mid\!- 23$	5	$5/17$	5	$4/17$	22
$23 \mid\!- 25$	7	$7/17$	12	$12/17$	24
$25 \mid\!- 27$	5	$5/17$	17	1	26
Total	17	1	-	-	-

Carros, cavalos e cilindradas

Vamos trabalhar com a base de dados `mtcars`, que contém dados de automóveis produzidos em 1973-74

```
head(mtcars[, 1:8], 5)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0
```

Carros, cavalos e cilindradas

Vamos acessar as dez primeiras observações dos dados brutos da potência dos automóveis

```
mtcars$hp[1:10]
```

```
 [1] 110 110  93 110 175 105 245  62  95 123
```

Se quisermos ordenar os dados em rol podemos usar a função `sort`

```
sort(mtcars$hp)[1:10]
```

```
 [1]  52  62  65  66  66  91  93  95  97 105
```

Carros, cavalos e cilindradas

Para calcularmos uma distribuição de frequência simples do número de carburadores dos automóveis utilizamos a função `table`

```
table(mtcars$carb)
```

```
 1  2  3  4  6  8
 7 10  3 10  1  1
```

São 7 automóveis com 1 carburador, 10 com 2, 3 com 3, e assim por diante

Carros, cavalos e cilindradas

O **R** automatiza o processo de criação dos intervalos de classe:

- A função `range` calcula a amplitude da amostra
- A função `nclass.Sturges` calcula o número de classes de acordo com a regra de Sturges
- A função `pretty` constrói intervalos ajustados ("bonitos")
- A função `cut` associa um intervalo para cada observação
- A função `cbind` transpõe o resultado da frequência com intervalos de classe gerada pela função `table` para melhor visualização

Carros, cavalos e cilindradas

Distribuição com intervalos de classe da potência dos automóveis:

```
quebras <- pretty(range(mtcars$hp),
                  n = nclass.Sturges(mtcars$hp))
classes <- cut(mtcars$hp, quebras, right = FALSE)
cbind(table(classes))
```

```
          [,1]
[50,100)     9
[100,150)    8
[150,200)    8
[200,250)    5
[250,300)    1
[300,350)    1
```

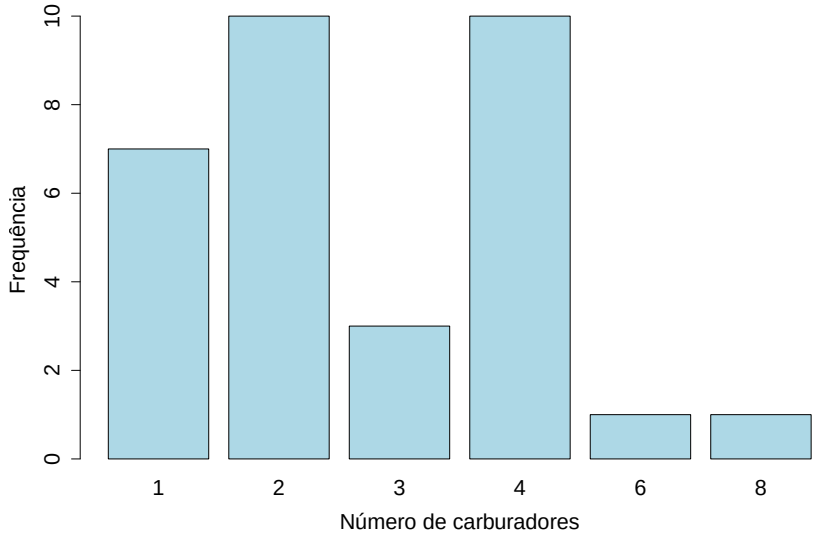
Gráfico de barras

Gráficos de barras são úteis para representar distribuições de frequência simples

- No **R** podemos utilizar a função `barplot`

```
barplot(table(mtcars$carb))
```

Gráfico de barras



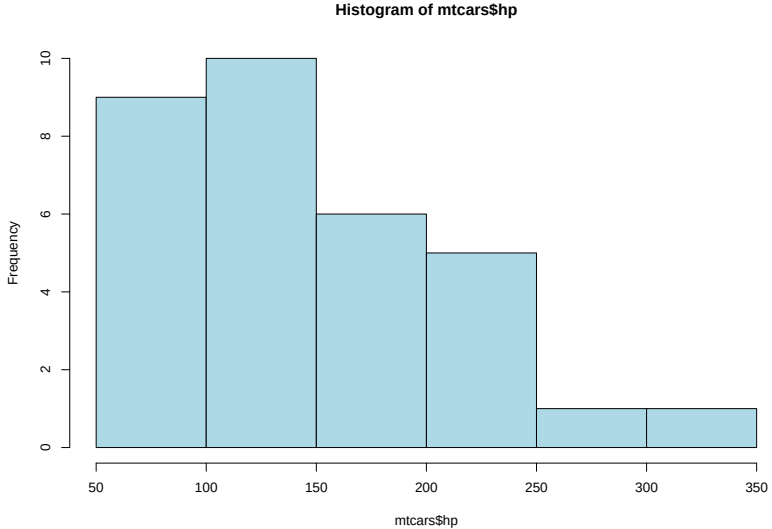
Histograma

Histogramas são úteis para representar distribuições com intervalos de classe

- No **R** podemos utilizar a função `hist`

```
hist(mtcars$hp)
```

Histograma


Histogram of mtcars$hp
Frequency
0
2
4
6
8
10
50
100
150
200
250
300
350
mtcars$hp

Histograma e polígono de frequência

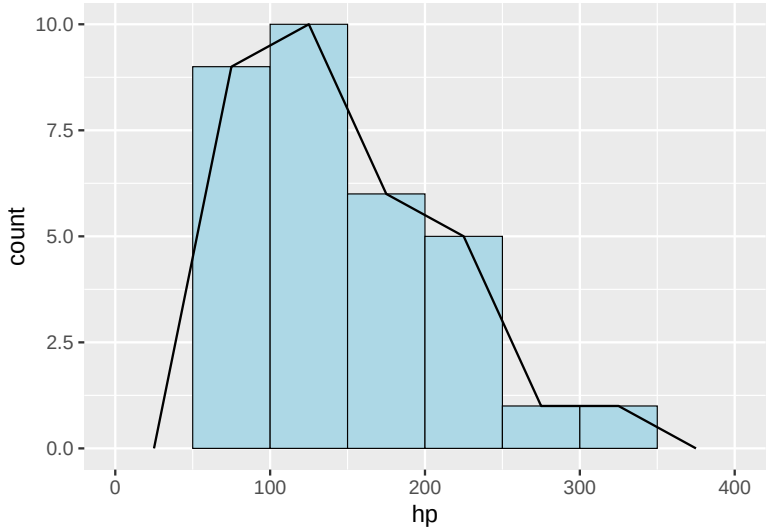
Polígonos de frequência são histogramas representados por linhas

- No **R** podemos utilizar a função `ggplot`[2] para plotar o histograma e o polígono de frequência em um mesmo gráfico

```
library(ggplot2)
ggplot(mtcars, aes(x = hp)) +
  geom_histogram(color = "black", fill = "lightblue",
                 breaks = quebras) +
  geom_freqpoly(color = "black", breaks = quebras,
                size = 1)
```

[2] A função `ggplot` possibilita a produção de gráficos mais complexos no R

Histograma e polígono de frequência



Referências

Sturges, H. *The choice of a class-interval*. Journal of the American Statistical Association, v. 21, n. 153, p. 65–66, 1926.