Estatística e Probabilidade Aula 2 - Representação gráfica de dados e medidas de centralidade;

Professor: Josceli Tenório

Data: 17/03/2022

Frequências e histograma

Alguns dados numéricos são obtidos por contagem para determinar o valor de uma variável, por exemplo, número de filhos e de TVs em uma residência,, enquanto outros dados são obtidos pela tomada de medidas como peso, altura e pressão arterial. A recomendação para plotagem de um histograma geralmente é diferente para esses dois casos.

Definições:

- Uma variável é discreta se o seu conjunto de valores possíveis é finito ou pode ser relacionado em uma sequência infinita.
- Uma variável é contínua se os seus valores possíveis consistem de um intervalo completo na reta de numeração.

Considere os dados constituídos de observações de uma variável discreta x. A frequência de qualquer valor particular de x é o número de vezes em que esse valor ocorre naquele conjunto.

A freqüência relativa de um valor é a fração ou proporção de vezes em que o valor ocorre:

$$freqüência\ relativa\ de\ um\ valor\ =\ número\ de\ ocorrências/número\ de\ observações\ do\ conjunto\ de\ dados$$

Suponha, por exemplo, que o nosso conjunto de dados consista em 200 observações de x = o número de defeitos graves em um novo carro de certo tipo. Se 70 desses valores x forem 1, então:

freqüência do valor x = 1: 70

freqüência relativa do valor x = 1: 70/200 = 0,35

Uma distribuição de frequência é uma tabulação das frequências e/ou frequências relativas.

Construção do histograma para dados discretos

Determine a frequência e a frequência relativa de cada valor de x. Depois, marque os valores possíveis de x em uma escala horizontal. Acima de cada valor, desenhe um retângulo cuja altura seja a frequência relativa (ou a frequência, como alternativa) daquele valor.

Exemplo:

Com que frequência uma equipe consegue atingir a bola num campeonato de beisebol mais de 10, 15 ou mesmo 20 vezes? A Tabela 1 mostra uma distribuição de frequência do número de acertos por equipe, por partida, para todos os jogos de nove séries entre 1989 e 1993.

Tabela 1. Distribuição de frequência de acertos em jogos de nove séries

Acertos/Jogo	Número de Jogos	Freqüência Relativa	Acertos/Jogo	Número de Jogos	Freqüência Relativa
0	20	0,0010	14	569	0,0294
1	72	0,0037	15	393	0,0203
2	209	0,0108	16	253	0,0131
3	527	0,0272	17	171	0,0088
4	1048	0,0541	18	97	0,0050
5	1457	0,0752	19	53	0,0027
6	1988	0,1026	20	31	0,0016
7	2256	0,1164	21	19	0,0010
8	2403	0,1240	22	13	0,0007
9	2256	0,1164	23	5	0,0003
10	1967	0,1015	24	1	0,0001
11	1509	0,0779	25	0	0,0000
12	1230	0,0635	26	1	0,0001
13	834	0,0430	27	1	0,0001
				19,383	1,0005

Exemplo:

Frequência relativa = 20/19383 = 0,0010

A Figura 1 mostra o histograma obtido:

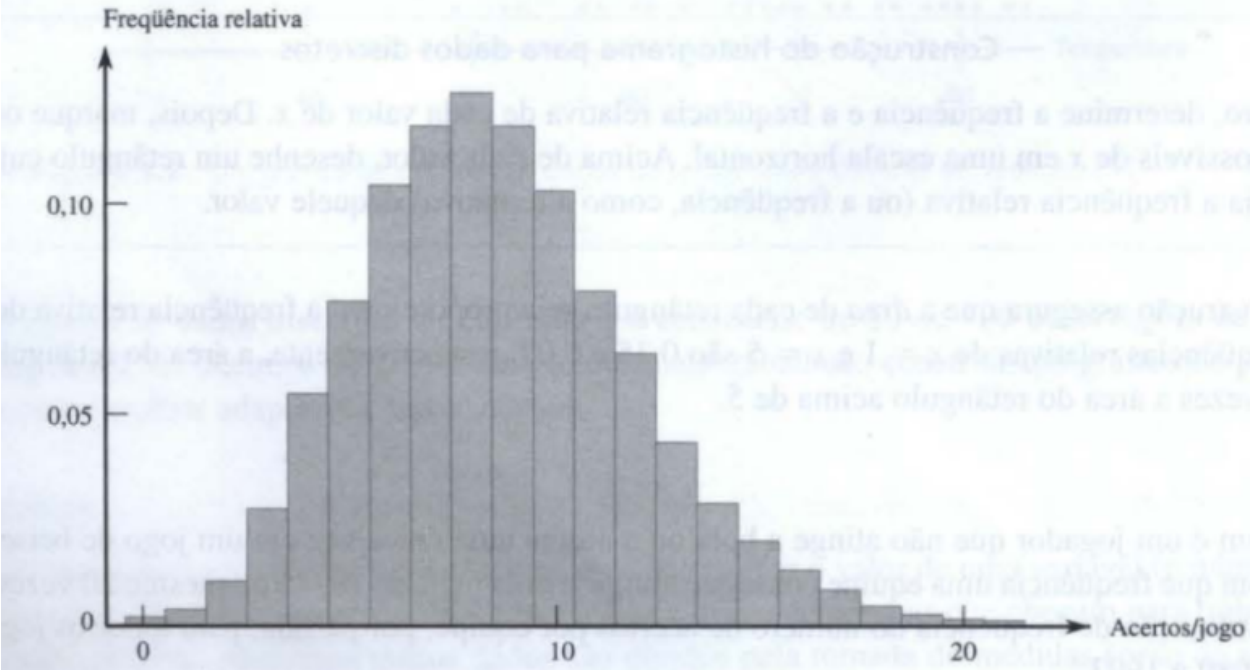

Figura 1. Histograma do número de acertos por jogo de nove séries.

Que informações podemos obter a partir da tabela e/ou do histograma?

Exemplo: Qual a proporção de jogos com no máximo dois acertos?

Como calcular?

Proporção (x <=2) = fr (x=0) + fr (x=1) + fr (x=2)

Proporção (x <=2)= 0,0010 + 0,0037 + 0,0108 = 0,0155

Construção de histograma para dados contínuos: classes de larguras iguais

Determinar a frequência e a frequência relativa de cada classe. Marcar os limites de classe em um eixo de medida horizontal. Acima de cada intervalo de classe, desenhe um retângulo cuja altura seja a frequência relativa correspondente (ou a frequência).

Exemplo:

As empresas de energia necessitam de informações sobre o consumo de seus clientes para obterem previsões precisas da demanda. O valor do consumo ajustado, em KWh, é mostrado na Tabela 2.

Tabela 2. 90 valores de consumo ajustado em KWh.

2,97	4,00	5,20	5,56	5,94	5,98	6,35	6,62	6,72	6,78
6,80	6,85	6,94	7,15	7,16	7,23	7,29	7,62	7,62	7,69
7,73	7,87	7,93	8,00	8,26	8,29	8,37	8,47	8,54	8,58
8,61	8,67	8,69	8,81	9,07	9,27	9,37	9,43	9,52	9,58
9,60	9,76	9,82	9,83	9,83	9,84	9,96	10,04	10,21	10,28
10,28	10,30	10,35	10,36	10,40	10,49	10,50	10,64	10,95	11,09
11,12	11,21	11,29	11,43	11,62	11,70	11,70	12,16	12,19	12,28
12,31	12,62	12,69	12,71	12,91	12,92	13,11	13,38	13,42	13,43
13,47	13,60	13,96	14,24	14,35	15,12	15,24	16,06	16,90	18,26

Para obter o histograma é necessário estabelecer classes. Não há uma regra única. Uma boa possibilidade é calcular a raiz quadrada do número de observações:

número de classes = raiz quadrada (número de observações)

Para o histograma é necessário contar o número de ocorrências em cada classe:

Classe	1–<3	3–<5	5–<7	7–<9	9–<11	11–<13	13–<15	15–<17	17–<19
Freqüência	1	1	11	21	25	17	9	4	1
Freqüência relativa	0,011	0,011	0,122	0,233	0,278	0,189	0,100	0,044	0,011

A Figura 2 mostra o histograma obtido;

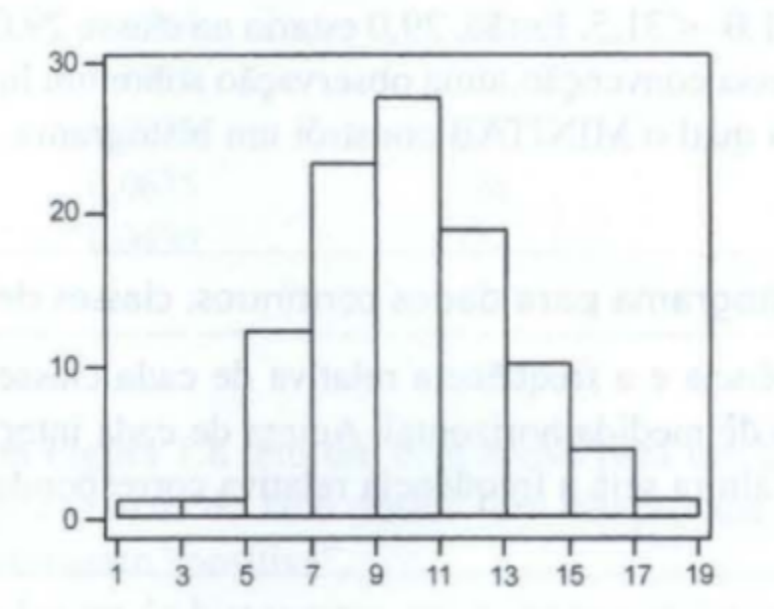

Figura 2. Histograma dos dados do consumo de energia.

Que informações podemos obter da tabela e/ou do histograma?

Exemplo: Qual a proporção de observações inferior a 9?

Pelo histograma, proporção de observações inferior a 9 será:

Proporção (n<=9) = 0,01 + 0,01 + 0,12 + 0,23 = 0,37

Medidas de centralidade

Os resumos visuais de dados são excelentes ferramentas para obter impressões e ideias iniciais. Uma análise mais formal de dados frequentemente exige o cálculo e a

interpretação de medidas-resumo numéricas simples. Uma característica importante de um conjunto de números é sua localização e, em particular, seu centro.

Trataremos inicialmente os dados numéricos e após os dados qualitativos.

Média

Para um determinado conjunto de números $x_1, x_2, \ldots, x_n$, a medida mais familiar e útil do centro é a média do conjunto. Como quase sempre temos os vários x; constituindo uma amostra, frequentemente chamaremos a média aritmética de média amostral.

Definição

A **média amostral** $\bar{x}$ das observações $x_1, x_2, \ldots, x_n$, é dada por

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

O numerador de $\bar{x}$ pode ser escrito mais informalmente como $\sum x_i$, onde a soma se dá sobre todas as observações da amostra.

Mediana

A palavra mediana é sinônimo de "metade" e a mediana amostral é o valor do meio quando as observações são ordenadas da menor para a maior. Quando as observações estiverem representadas por $x_1, \ldots, x_n$.

Definição:

A **mediana amostral** é obtida pela ordenação das n observações da menor para a maior (com os valores repetidos incluídos, de forma que cada observação da amostra seja exibida na lista ordenada). Assim,

$$\tilde{x} = \begin{cases} \text{O único valor médio se } n \text{ for ímpar} & = \left(\frac{n+1}{2}\right) \text{enésimo valor ordenado} \\ \text{A média dos dois valores médios se } n \text{ for par} & = \text{média dos valores ordenados } \left(\frac{n}{2}\right) \text{ e } \left(\frac{n}{2}+1\right) \end{cases}$$

Exemplo: considere os dados a seguir sobre a concentração do receptor de transferrina de uma amostra de mulheres com evidências laboratoriais de uma visível anemia por deficiência de ferro.

$$x_1 = 15{,}2 \quad x_2 = 9{,}3 \quad x_3 = 7{,}6 \quad x_4 = 11{,}9 \quad x_5 = 10{,}4 \quad x_6 = 9{,}7$$
$$x_7 = 20{,}4 \quad x_8 = 9{,}4 \quad x_9 = 11{,}5 \quad x_{10} = 16{,}2 \quad x_{11} = 9{,}4 \quad x_{12} = 8{,}3$$

Após a ordenação:

7,6 8,3 9,3 9,4 9,4 9,7 10,4 11 ,5 11 ,9 15,2 16,2 20,4

Corno n = 12 é par, tiramos a média n/2 =do sexto e sétimo valores ordenados:

$$\text{mediana amostral} = \frac{9{,}7 + 10{,}4}{2} = 10{,}05$$

Dados categorizados

Quando os dados são categorizados, uma distribuição de frequência ou distribuição de frequência relativa fornece um resumo tabular eficiente dos dados. Entretanto, é possível calcular uma proporção amostral ou média amostral considerando uma categoria. Se fizermos x representar o número da amostra na categoria 1, o número na categoria 2 será n - x.

Exemplo: Considere uma amostra de consumidores que possuem ou não smartphone (1 para quem possui e 0 para quem não possui). Uma amostra de tamanho n = 10 pode então resultar em 1, 1 O, 1, 1, 1, O, O, 1, 1. A média dessa amostra numérica é:

$$\frac{x_1 + \ldots + x_n}{n} = \frac{1+1+0+\cdots+1+1}{10} = \frac{7}{10} = \frac{x}{n} = \text{proporção amostral}$$

Se em uma situação de dados categorizados focarmos a atenção em urna determinada categoria e codificarmos os resultados da amostra de forma que 1 seja registrado como um indivíduo da categoria e O para um indivíduo fora dela, a proporção amostral de indivíduos da categoria será a média amostra[da seqüência de 1 s e Os. Assim, uma média amostral pode ser usada para resumir os resultados de uma amostra categorizada.

Aplicações

1. Os transdutores de temperatura de um determinado tipo são enviados em lotes de 50. Uma amostra de 60 lotes foi selecionada e o número de transdutores fora das especificações em cada lote foi determinado, resultando nos dados a seguir:

2	1	2	4	0	1	3	2	0	5	3	3	1	3	2	4	7	0	2	3
0	4	2	1	3	1	1	3	4	1	2	3	2	2	8	4	5	1	3	1
5	0	2	3	2	1	0	6	4	2	1	6	0	3	3	3	6	1	2	3

   a) Determine as frequências e frequências relativas dos valores observados de x = número de transdutores fora das especificações em um lote.
   b) Que proporção de lotes na amostra possui no máximo cinco transdutores fora das especificações? Que proporção tem menos de cinco? Que proporção possui no mínimo cinco unidades fora das especificações?
   c) Desenhe um histograma dos dados, usando a frequência relativa na escala vertical e comente suas características.

2. Os dados seguintes são referentes à vida útil das brocas (número de furos que uma broca faz antes de quebrar), quando os furos são feitos em uma determinada liga de bronze.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

a) Por que uma distribuição de frequência não pode ter por base os intervalos de classe 0-50, 50-100, 100-150 e assim por diante?
b) Construa uma distribuição de frequência e um histograma dos dados usando limites de classes O, 50, 100, ... e então faça comentários sobre as características interessantes.
c) Que proporção das observações de vida útil dessa amostra é inferior a 100? Que proporção das observações é igual ou maior que 200?

3. Considere as observações a seguir sobre resistência ao cisalhamento (MPa) de uma junta soldada de uma determinada forma.

22,2	40,4	16,4	73,7	36,6	109,9
30,0	4,4	33,1	66,7	81,5	

a) Determine o valor da média amostral.
b) Determine o valor da mediana amostral. Por que esse valor é tão diferente da média?
c) Calcule a média aparada, excluindo a menor e a maior observações.

Referência

Devore J L. Probability and Statistics for Engineering and the Sciences. Thomson ed. 5a. ed. 284 p.

R programming - Aula 0

```
#Vetores - tipos de dados
numerico <- c(100, 10, 49)
caractere <- c("a", "b", "c")
string<-c("segunda","terça")
boolean <- c(TRUE, FALSE, TRUE)
numerico[1]
boolean[c(2,3)]

#Lista
a <- list(
  a = 1:3,
  b = "sextou",
  c = pi,
  d = list(-1, -5)
)
a

vet <- c(-24, -50, 100, -350, 10)
names(vet) <- c("segunda-feira", "terça-feira", "quarta-feira",
"quinta-feira", "sexta-feira")
vet[1]

#Converter uma variável para factor()
status <- c("estudante", "não estudante", "estudante", "não
estudante")
estudante <- factor(status)
estudante

temperatura <- c("alta", "baixa", "alta","baixa", "media")
factor_temperatura <- factor(temperatura, order = TRUE, levels =
c("baixa", "media", "alta"))
temperatura

#Estrutura condicional
x <- 30
y <- 50
if (x > y) {
  print("x é maior")
} else {
  print("y é maior")
```

```
}

#Estrutura de repetição
valor <- c(16, 9, 13, 5, 2, 17, 14)
for(i in 3:length(valor)) {
  print(valor[i])
}


#DATAFRAMES

#O efeito da vitamina C no crescimento dos dentes em
porquinhos-da-índia
data(ToothGrowth)
str(ToothGrowth)
View(ToothGrowth)
#calculando a media
mean(ToothGrowth$len)
sd(ToothGrowth$len)
head(ToothGrowth)
head(ToothGrowth, 10)
str(ToothGrowth)


#GRAFICO
library(ggplot2)

# scatter plot
ggplot(ToothGrowth, aes(x = len, y = dose, color = supp)) +
  geom_point(size = 4)

#histograma
ggplot(ToothGrowth, aes(x = len)) +
  geom_histogram()

#densidade
ggplot(ToothGrowth, aes(x = len)) +
  geom_density()

#violin
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=dose, y=len)) +
  geom_violin() +
coord_flip()
```

```
#violin separado por cores
ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) +
  geom_violin() +
  coord_flip()

#violin com pontos
ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) +
  geom_violin() +
  geom_dotplot(binaxis='y', stackdir='center', dotsize=1)
  coord_flip()
```