Estatística Descritiva com R

Rômulo Coutinho Araújo

1. RESUMO DE DADOS

TIPOS DE VARIÁVEIS

Basicamente, temos dois tipos principais de variáveis: as *qualitativas* e as *quantitativas*.

As variáveis qualitativas subdividem-se em **nominal**, que somente identifica uma das realizações de uma variável, como ocorre no nome de uma pessoa, CPF ou número de telefone, por exemplo; **ordinal**, quando, além da identificação das realizações de uma variável, é possível ordená-las, como na variável escolaridade, por exemplo.

As variáveis quantitativas são classificadas em **discretas**, cujos possíveis valores formam um conjunto finito ou enumerável de números, resultado geralmente de uma contagem, como ocorre na variável 'anos completos de experiência', por exemplo; e também é possível classificar as variáveis quantitativas em **contínuas**, as quais possuem realizações a partir dos números reais, o que ocorre gereamente a partir de medições, como altura, peso, etc.

Tais tipos de variáveis definem técnicas próprias para resumo da informação, e algumas vezes tais técnicas são adaptáveis para diferentes tipos de variáveis.

DISTRIBUIÇÕES DE FREQUÊNCIAS

A distribuição de frequências é o primeiro panorama sobre a(s) varável(eis) analisada(s), e mostra o comportamento de suas possíveis realizações.

Exemplo de tabela de frequências utilizando os dados nativos do R, arquivo *mtcars*, do qual extraímos a informação do número de cilindros de cada modelo de veículo:

```
library(knitr)
library(pander)
pander(mtcars, caption =
        "Dados sobre Modelos de Veículos (mtcars)",split.table = Inf)
```

Table 1: Dados sobre Modelos de Veículos (mtcars)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Fleetwood Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

```
dados <- as.data.frame(table(mtcars$cyl))
names(dados) <- c("Cilindros", "Frequência")
dados$Proporção <- round(dados$Frequência/sum(dados$Frequência),2)
dados$Porcentagem <- round(dados$Proporção*100,2)
dados2 <- data.frame(Cilindros="Total", Frequência=sum(dados$Frequência), Proporção=sum(dados$Proporção)
dados <- merge(dados,dados2, all=T)
pander(dados, caption =
        "Frequência para o Número de Cilindros (mtcars)",split.table = Inf)
```

Table 2: Frequência para o Número de Cilindros (mtcars)

Cilindros	Frequência	Proporção	Porcentagem
4	11	0.34	34
6	7	0.22	22
8	14	0.44	44
Total	32	1	100

Os dados de proporção, ou porcentagem, são bastante úteis na comparação com pesquisas distintas, o que pode ser inviável apenas com o frequência absoluta.

No caso de contruir tabela de distribuição para variáveis contínuas (ou discreta com grande variabilidade), usam-se intervalos para que os dados sejam agrupados, observando sempre que o número de intervalos define o quanto de informação será perdida nesse processo, ou seja, quanto menos intervalos, mais informação perdida, porém, um alto número de intervalos implica na redução da clareza na informação; portanto, é preciso encontrar um equilíbrio quanto à quantidade de intervalos usados.

É possível criar uma tabela de frequência usando algumas regras para determinação do número de classes (o que não invalida o uso do bom senso):

a) Regra de Sturges (Regra do Logaritmo): $k = 1 + 3.3log(n)$

b) Regra da Potência de 2: $k =$ menor valor inteiro tal que $2^k \geq n$

c) Regra da Raíz Quadrada: $k = \sqrt{n}$

A partir do conjunto de dados *mtcars*, podemos contruir uma tabela de frequência da variável *potência* (*hp*). De início, vamos precisar da amplitude total do conjunto de dados:

```
(amp <- max(mtcars$hp)-min(mtcars$hp)) # Amplitude
```

```
## [1] 283
```

Para o cálculo da amplitude de cada classe, usamos o valor de k (número de classes) a partir da Regra de Sturges:

```r
(k <- 1+3.3*log(length(mtcars$hp)))
```

```
## [1] 12.43693
```

O valor da amplitude de cada classe é então dado por (arredondado):

```r
(h <- ceiling(amp/k))
```

```
## [1] 23
```

São criadas abaixo as classes que dividirão a tabela e sua apresentação:

```r
classe <- seq(min(mtcars$hp),max(mtcars$hp)+16, by=h) #Soma-se 16 ao máximo para
# termos um valor divisível por 23.
classe
```

```
##  [1]  52  75  98 121 144 167 190 213 236 259 282 305 328 351
```

```r
potencia_classes <- cut(mtcars$hp, classe, right = FALSE)
potencia_tabela <- as.data.frame(table(potencia_classes))
names(potencia_tabela) <- c("Potência", "Frequência")
potencia_tabela$Proporção <- round(potencia_tabela$Frequência/sum(potencia_tabela$Frequência),4)
potencia_tabela$Porcentagem <- round(potencia_tabela$Proporção*100,4)
potencia_tabela2 <- data.frame(Potência="Total", Frequência=sum(potencia_tabela$Frequência), Proporçã
Tabela_HP <- merge(potencia_tabela,potencia_tabela2, all=T)
pander(Tabela_HP, caption =
        "Frequência da Potência dos Veículos (em HP)")
```

Table 3: Frequência da Potência dos Veículos (em HP)

Potência	Frequência	Proporção	Porcentagem
[52,75)	5	0.1562	15.62
[75,98)	4	0.125	12.5
[98,121)	6	0.1875	18.75
[121,144)	2	0.0625	6.25
[144,167)	2	0.0625	6.25
[167,190)	6	0.1875	18.75
[190,213)	1	0.0312	3.12
[213,236)	2	0.0625	6.25
[236,259)	2	0.0625	6.25
[259,282)	1	0.0312	3.12
[282,305)	0	0	0
[305,328)	0	0	0
[328,351)	1	0.0312	3.12
Total	32	1	100