Relatório I - Estatistica

Artur Papa

1/25/2022

Relatório de Estatística Descritiva

Introdução

Estatística Descritiva é o ramo da estatística que visa sumarizar e descrever qualquer conjunto de dados, ou seja, é aquela que está preocupada em sintetizar os dados de maneira direta, preocupando-se menos com variações e intervalos de confiança dos dados. Desta maneira, vale ressaltar que neste trabalho iremos ver diversos gráficos e tabelas de um conjunto específico de dados e a relação entre cada um.

Materiais e métodos

O universo de estudo é composto pelos alunos matriculados na disciplina de estatística. Foram analisadas as seguintes variáveis: 1 - Idade 2 - Gênero 3 - CR (Coeficiente de rendimento) 4 - Trabalha 5 - Filhos 6 - Tipo de escola que concluiu o ensino médio 7 - Horas de estudo fora da sala de aula 8 - Número de livros lidos ano passado 9 - Frequência que utiliza a biblioteca 10 - Nível de fluência da língua inglesa Essas variáveis foram apresentadas através de tabelas e gráficos e avaliadas com relação às medidas descritivas, tais como média, mediana e desvio padrão com o objetivo de traçar o perfil dos alunos da disciplina. ## Análise descritiva dos dados Primeiramente temos a geração aleatória dos dados, feita usando as próprias bibliotecas fornecidas pela linguagem R.

```
set.seed(3886)

Idade <- sample(18:70, 50, replace = TRUE)

Genero <- sample(c("Masculino", "Feminino"), 50, replace = TRUE)
CR <- round(rnorm(50, mean = 70, sd = 2))

Trabalho <- sample(c("Sim", "Não"), 50, replace = TRUE)

Filhos <-sample(c("Sim", "Não"), 50, replace = TRUE)

EM <- sample(c("Pública", "Privada", "Pública/Privada"), 50, replace = TRUE)
Horas <- sample(0:20, 50, replace = TRUE)

Livros <- sample(0:5, 50, prob = c(0.5, 0.1, 0.1, 0.1, 0.1, 0.1), replace = TRUE)

BBT <- sample(c("Nunca usa", "Usa raramente",
                "Pouca frequência", "Usa frequentemente"), 50, replace = TRUE)

Ingles <- sample(c("Bom", "Regular",
                   "Ruim", "Péssimo"), 50, replace = TRUE)

Tab <- cbind(Idade, Genero, CR, Trabalho, Filhos, EM, Horas, Livros, BBT, Ingles)
```

```
Tab <- as.data.frame(Tab)
knitr::kable(head(Tab))
```

Idade	Genero	CR	Trabalho	Filhos	EM	Horas	Livros	BBT	Ingles
57	Feminino	71	Sim	Não	Pública/Privada	1	3	Nunca usa	Bom
49	Feminino	68	Sim	Não	Pública	8	1	Usa raramente	Regular
63	Feminino	71	Sim	Sim	Pública/Privada	6	5	Pouca frequência	Ruim
57	Masculino	69	Não	Sim	Pública	9	1	Usa frequentemente	Péssimo
68	Masculino	66	Não	Não	Privada	13	0	Usa frequentemente	Péssimo
37	Masculino	70	Sim	Sim	Pública	7	5	Usa frequentemente	Ruim

Antes de serem feitas as análises dos dados primeiro foi feito o "import" de algumas bibliotecas para auxiliarem na criação dos gráficos.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(RColorBrewer)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(tidyr)
library(forcats)
library(hrbrthemes)
library(viridis)
```

```
## Loading required package: viridisLite
```

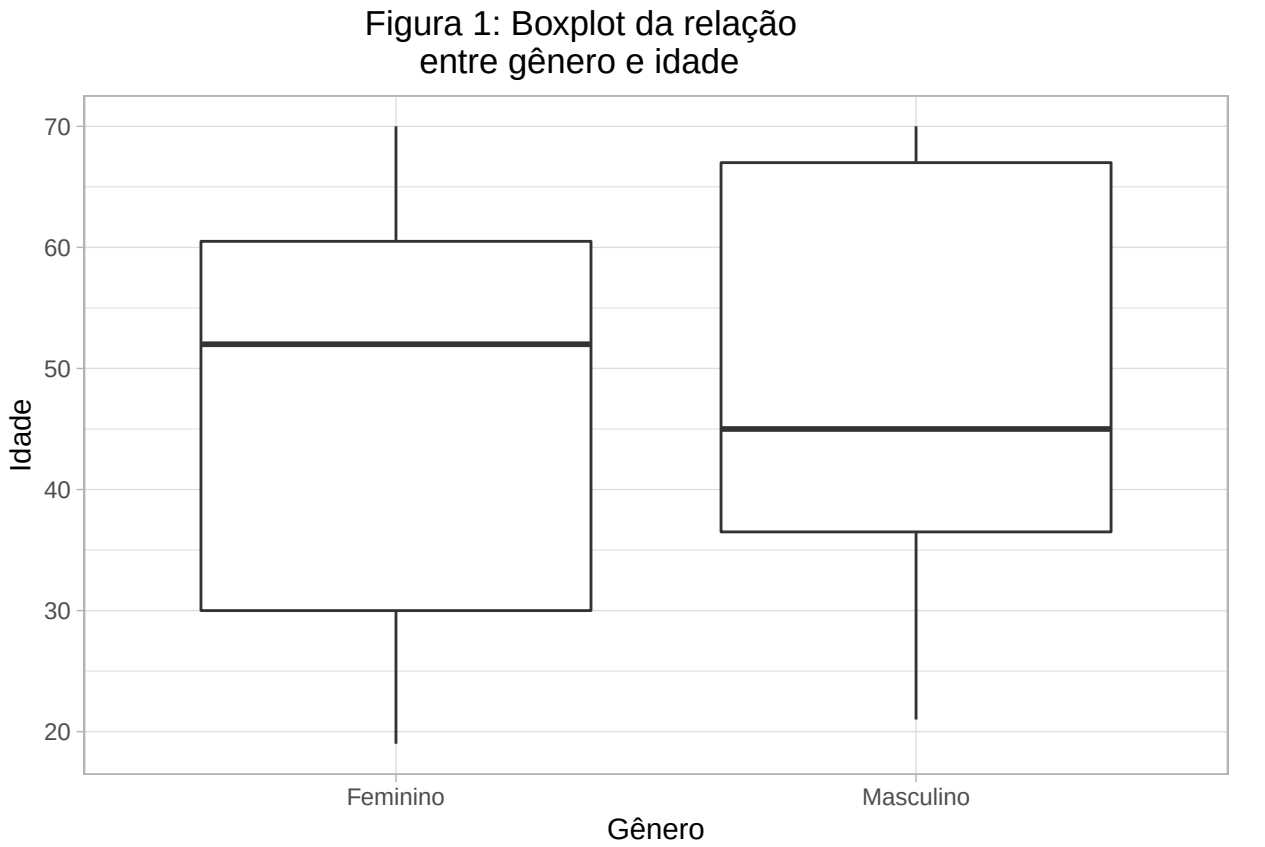
```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Na primeira análise, pode-se ver a relação entre gênero e idade, e observando o "boxplot" percebe-se que há homens mais velhos do que mulheres, embora a média de idade do sexo masculino seja menor que a do sexo feminino.

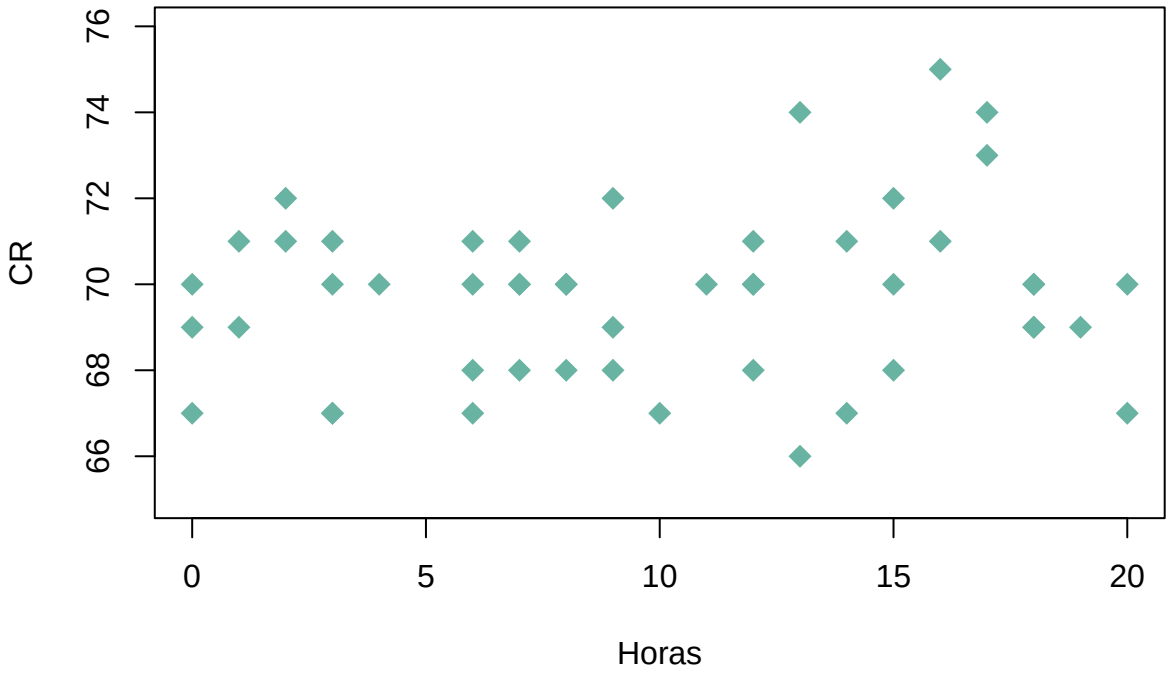
```
Tab$Idade <- as.numeric(Tab$Idade)
ggplot(Tab, aes(x = Genero, y = Idade)) +
geom_boxplot() + theme_light() +
ggtitle("Figura 1: Boxplot da relação \n entre gênero e idade") +
xlab("Gênero") + theme(plot.title = element_text(hjust = 0.4))
```



Para o segundo gráfico, vemos um "scatterplot" comparando as horas de estudo e o coeficiente de rendimento, analisando a imagem percebemos que aqueles que possuem uma média de horas de estudo por volta de 15 tendem a ter um coeficiente de rendimento maior. Dito isso, podemos inferir que esse valor seria um valor ótimo para essa relação, visto que aqueles que estudam pouco não conseguem obter uma nota alta, porém aqueles que estudam demais também não conseguem ter um bom rendimento.

```
plot(Tab$Horas, Tab$CR,
     xlim=c(0,20) , ylim=c(65, 76),
     pch=18,
     cex=1.6,
     col="#69b3a2",
     xlab="Horas", ylab="CR",
     main="Figura 2: Horas x CR"
     )
```

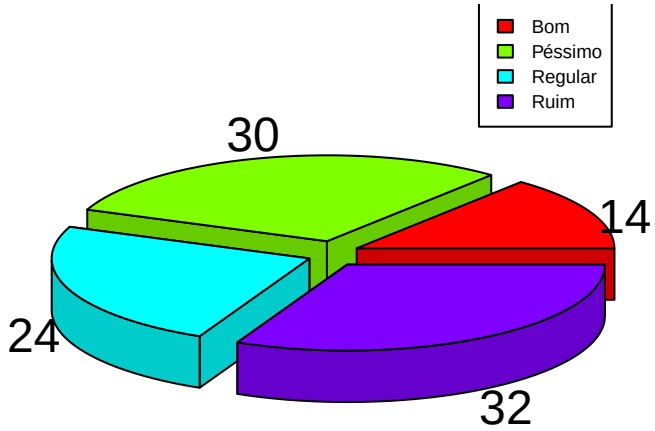
Figura 2: Horas x CR



Para a terceira figura temos a proporção da fluência de inglês entre os estudantes, vendo o gráfico de pizza podemos notar que poucos sabem falar inglês, já que temos 62% dos alunos que possuem uma fluência "Péssima/Ruim" e apenas 14% com um compreendimento bom da língua inglesa.

```
library(plotrix)
tab_en <- table(Tab$Ingles)
p1 <- prop.table(tab_en)
percent <- round(100*(tab_en/sum(tab_en)), digits = 2)
pie3D(p1, labels = percent, explode = 0.1, main = "Figura 3: Fluência dos estudantes", col = rainbow(le
legend("topright", c("Bom", "Péssimo", "Regular", "Ruim"), cex = 0.6, fill = rainbow(length(p1)))
```

Figura 3: Fluência dos estudantes



Para o quarto gráfico, podemos ver que há uma frequência maior para as pessoas que usam a biblioteca do que para aqueles que não usam.

```
tab_lib <- table(Tab$BBT)
tab_lib <- as.data.frame(tab_lib)

ggplot(tab_lib, aes(x = Var1, y=Freq))+
  geom_bar(stat="identity", fill="#4B0082", alpha=.6, width=.4) +
    ggtitle("Figura 4: Frequência do uso \n da biblioteca") +
    coord_flip() +
    xlab("") +
    theme_bw()
```

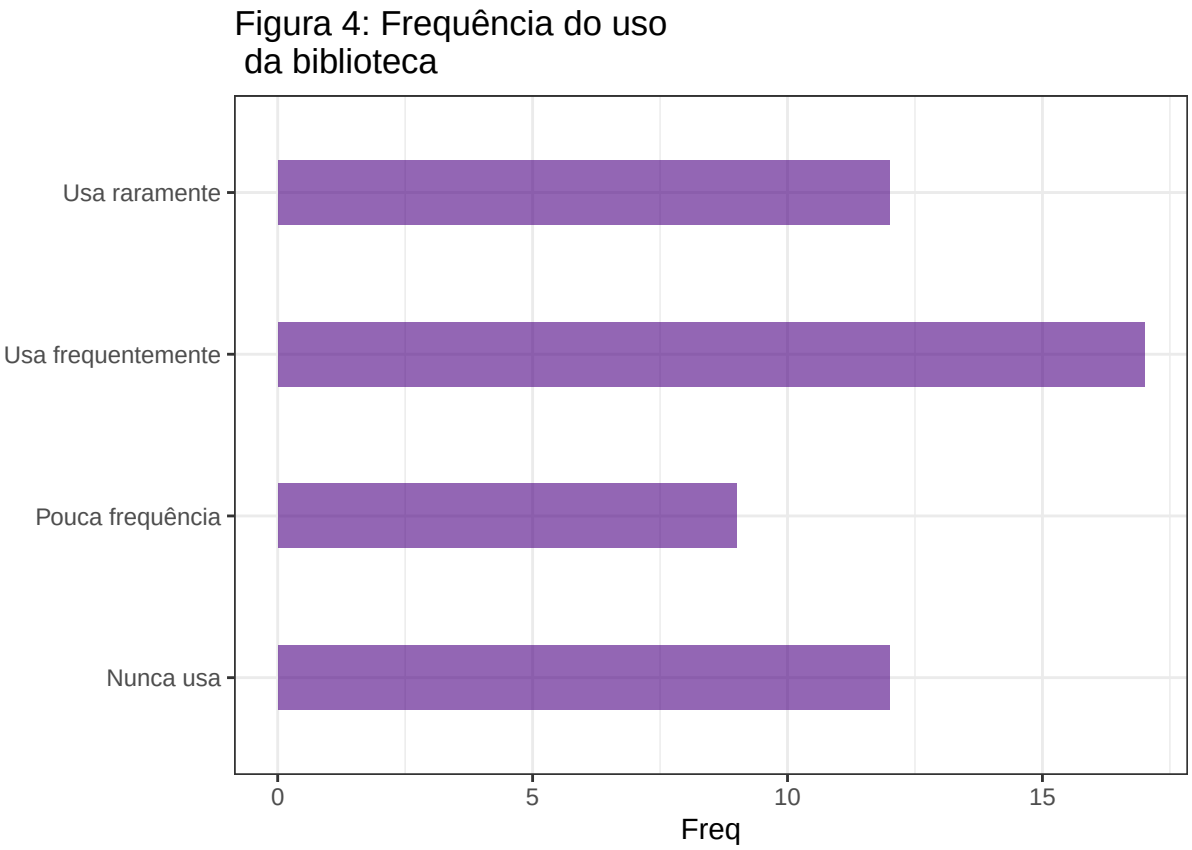


Tabela de frequência para a quantidade de pessoas que têm filhos e que não têm e nota-se que não há uma discrepância alta entre as duas categorias.

```
tab_sons <- table(Tab$Filhos)
tab_sons <- prop.table(tab_sons)
sons <- data.frame(tab_sons)
knitr::kable(head(rename(sons, c("Tem Filho(a)" = "Var1"))))
```

Tem Filho(a)	Freq
Não	0.54
Sim	0.46

Na tabela de frequência do ensino médio nota-se que mais da metade dos alunos estudaram em uma escola privada, já que o valor chega a 82% contando as variáveis “Privada” e “Privada/Pública”.

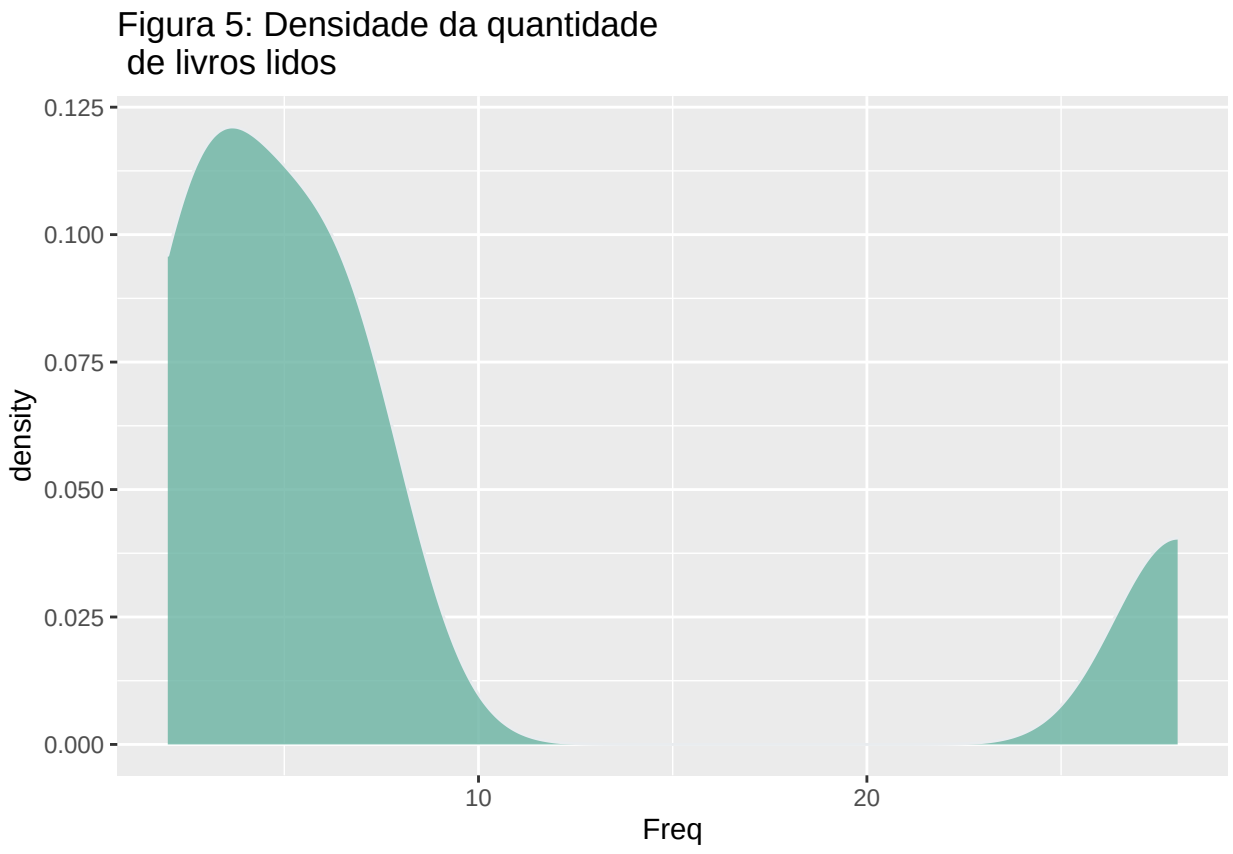
```
hs.tb <- table(Tab$EM)

hs1 <- cbind("Freq. Abs." = hs.tb)
hs2 <- cbind("Freq. Rel." = prop.table(hs.tb))
```

```
hs3 <- cbind("Freq. Abs." = addmargins(hs.tb))
hs4 <- cbind("Freq. Rel." = addmargins(prop.table(hs.tb)))

knitr::kable(cbind("Freq.Abs." = addmargins(hs.tb), "Freq.Rel." =
addmargins(prop.table(hs.tb))))
```

	Freq.Abs.	Freq.Rel.
Privada	22	0.44
Pública	9	0.18
Pública/Privada	19	0.38
Sum	50	1.00

Para o gráfico da densidade podemos observar o gap entre os valores de 11 até 25 aproximadamente, esse gap se dá pelo fato de não havermos uma frequencia com esse valor. Ademais, observa-se que a frequência é maior nos valores de 0 a 10 e menor para os valores de 26 a 30.

```
tab_books <- table(Tab$Livros)
tab_books <- as.data.frame(tab_books)

ggplot(tab_books, aes(x = Freq)) +
  ggtitle("Figura 5: Densidade da quantidade \n de livros lidos") +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)
```



Conclusão

Posto isso, podemos ver que nem sempre os dados irão ter relação entre si. Desse modo, se usarmos como exemplo se uma pessoa tem filhos com o nível de fluência em inglês, não poderíamos tirar nenhuma conclusão já que um não influencia no resultado do outro. Entretanto, caso comparemos as horas de estudo com o CR,

veremos que há uma relação direta entre eles. Por fim, é válido ressaltar a acertividade ao se criar os gráficos e as tabelas de seus respectivos dados, além disso é observado a notória importância de se coletar, filtrar e analisar os dados em geral.