Atividade 01: Análise exploratoria de dados

Thalles Cotta Fontainha

2023-10-19

Table of Contents

#### Disciplina: Estatística Aplicada à Projetos de Pesquisa

### (Exercício 1): Abra o arquivo ebmt3.RData no R Commander.

> load("/home/sony/Documentos/Mestrado/Conjuntos\_de\_dados\_v2\_Thalles/ebmt3.RData")

> setwd("/home/sony/Documentos/Mestrado/Conjuntos\_de\_dados\_v2\_Thalles")

### (Exercício 2) Faça uma tabela de frequências para cada uma das variáveis: age e rfsstat. Quais os percentuais de cada faixa etária e de mortalidade/recidiva no estudo?

### Frequencies: age, rfsstat

> local({  
+ .Table <- with(ebmt3, table(age))  
+ cat("\ncounts:\n")  
+ print(.Table)  
+ cat("\npercentages:\n")  
+ print(round(100\*.Table/sum(.Table), 2))  
+ })

counts:  
age  
 <=20 20-40 >40   
 419 1057 728   
  
percentages:  
age  
 <=20 20-40 >40   
19.01 47.96 33.03

> local({  
+ .Table <- with(ebmt3, table(rfsstat))  
+ cat("\ncounts:\n")  
+ print(.Table)  
+ cat("\npercentages:\n")  
+ print(round(100\*.Table/sum(.Table), 2))  
+ })

counts:  
rfsstat  
 censura morte/recidiva   
 1363 841   
  
percentages:  
rfsstat  
 censura morte/recidiva   
 61.84 38.16

#### (Resposta 2): No arquivo ebmt3.RData a distribuição percentual da população em diferentes faixas etárias é a seguinte: 19,01% têm idade igual ou inferior a 20 anos, 47,96% têm idade entre 20 e 40 anos, e 33,03% têm idade superior a 40 anos.

### (Exercício 3) Faça uma tabela de dupla entrada com as variáveis age (linhas) e rfsstat (colunas), solicitando que sejam mostrados os percentuais nas linhas. Quais os percentuais de pacientes que sobreviveram em cada faixa etária?

> library(abind, pos=16)

### Two-Way Contingency Table: age, rfsstat

> local({  
+ .Table <- xtabs(~age+rfsstat, data=ebmt3)  
+ cat("\nFrequency table:\n")  
+ print(.Table)  
+ cat("\nRow percentages:\n")  
+ print(rowPercents(.Table))  
+ })

Frequency table:  
 rfsstat  
age censura morte/recidiva  
 <=20 295 124  
 20-40 686 371  
 >40 382 346  
  
Row percentages:  
 rfsstat  
age censura morte/recidiva Total Count  
 <=20 70.4 29.6 100 419  
 20-40 64.9 35.1 100 1057  
 >40 52.5 47.5 100 728

#### (Resposta 3): Sabendo que “rfstime” é Tempo em dias desde o transplante até a recidiva ou morte ou último acompanhamento (tempo de sobrevivência livre de recidiva).

#### Pacientes com idade <=20 anos: 70.4% sobreviveram (censura), 29.6% enfrentaram morte ou recidiva.

#### Pacientes com idade entre 20-40 anos: 64.9% sobreviveram, 35.1% enfrentaram morte ou recidiva.

#### Pacientes com idade >40 anos: 52.5% sobreviveram, enquanto 47.5% enfrentaram morte ou recidiva.

### (Exercício 4) Obtenha a média, mediana, P25, P75 das variáveis prtime e rfstime para cada faixa etária.

> library(e1071, pos=17)

### Resumos Numéricos: ebmt3

> numSummary(ebmt3[,c("prtime", "rfstime"), drop=FALSE], groups=ebmt3$age, statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(.25,.75))

Variable: prtime   
 mean sd IQR 25% 75% n  
<=20 389.2912 633.2222 397.5 26.00 423.50 419  
20-40 466.8714 690.9332 708.0 26.00 734.00 1057  
>40 426.0810 649.4042 574.0 24.75 598.75 728  
  
Variable: rfstime   
 mean sd IQR 25% 75% n  
<=20 992.4535 728.4596 1368.50 262.50 1631.0 419  
20-40 959.9716 746.0292 1283.00 241.00 1524.0 1057  
>40 850.8558 720.6033 1268.25 172.25 1440.5 728

#### (Resposta 4):

#### Em prtime; A média (mean) para idade <= 20 : 389.2912

#### Em prtime; O desvio padrão (sd) para idade <= 20 : 633.2222

#### Em prtime; A média (mean) para idade entre 20-40 : 466.8714

#### Em prtime; O desvio padrão (sd) para idade entre 20-40 : 690.9332

#### Em prtime; A média (mean) para idade entre > 40 : 426.0810

#### Em prtime; O desvio padrão (sd) para idade > 40 : 649.4042

#### e

#### Em rfstime; A média (mean) para idade <= 20 : 992.4535

#### Em rfstime; O desvio padrão (sd) para idade <= 20 : 728.4596

#### Em rfstime; A média (mean) para idade entre 20-40 : 959.9716

#### Em rfstime; O desvio padrão (sd) para idade entre 20-40 : 746.0292

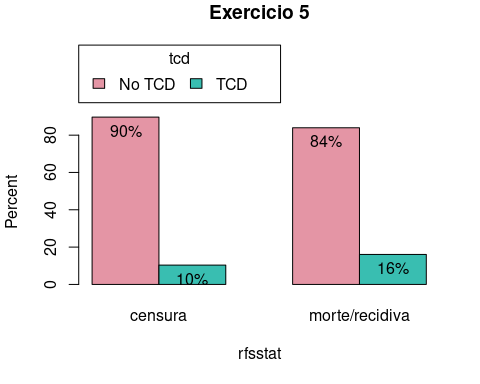
#### Em rfstime; A média (mean) para idade entre > 40 : 850.8558

#### Em rfstime; O desvio padrão (sd) para idade > 40 : 720.6033

### (Exercício 5): Faça um diagrama de barras lado a lado e condicional, com as percentagens da variável rfsstat para cada categoria da variável tcd. Comente o gráfico.

### Bar Plot: rfsstat

> with(ebmt3, Barplot(rfsstat, by=tcd, style="parallel", legend.pos="above", xlab="rfsstat", ylab="Percent", main="Exercicio 5", scale="percent", label.bars=TRUE))

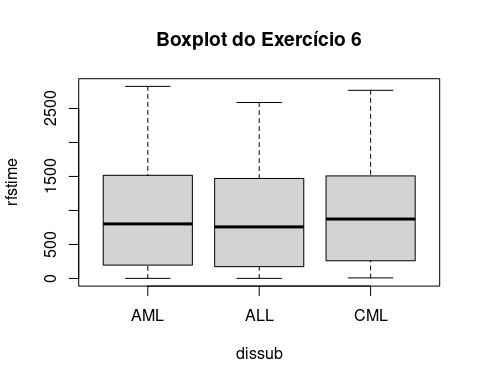


#### (Resposta 5): Sabendo que “tcd”: depleção de células T (“No TCD”, “TCD”). Entre o grupo “censura”, 90% tinham “No TCD” enquanto 10% tinham “TCD”. Já em “morte/recidiva” 84% tinham “No TCD” enquanto 16% tinham “TCD”.

### (Exercício 6): Faça um boxplot da variável rfstime para cada subclassificação da doença. Comente o gráfico.

### Boxplot: rfstime ~ dissub

> Boxplot(rfstime ~ dissub, data=ebmt3, id=list(method="none"), main="Boxplot do Exercício 6")



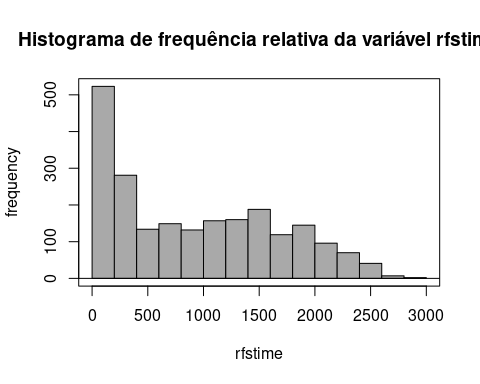
#### (Resposta 6): O boxplot é uma ferramenta visual útil para entender como o tempo de sobrevivência livre de recidiva (rfstime) varia entre diferentes subtipos de leucemia, como AML, ALL e CML. Ele mostra a mediana, quartis e valores extremos, oferecendo uma visão completa da distribuição desses dados. Ao observar o boxplot, os profissionais de saúde podem identificar facilmente casos incomuns em cada subtipo, fornecendo insights valiosos que podem exigir uma atenção especial na análise e tomada de decisões clínicas.

#### (Comentário do gráfico do exercício 6): A linha mais grossa é a mediana, o percentil 25 a parte de baixo, entre 0 e 500 no eixo y (rfstime) enquanto o percentil 75, valores proximos de 1500 no eixo y também. Ou seja, em geral os 3 tipos de subtipos de leucemia (AML, ALL, CML) possuem comportamento semelhante em termos de rfstime (tempo de sobrevivência livre de recidiva) com valores de mediana proximas também. E também não possem valores de “outlier” que é um valor que se afasta significativamente do padrão geral desse conjunto de dados expresso.

### (Exercício 7): Faça um histograma de frequência relativa da variável rfstime. Comente o gráfico.

### Histogram: rfstime

> with(ebmt3, Hist(rfstime, scale="frequency", breaks="Sturges", col="darkgray", main="Histograma de frequência relativa da variável rfstime"))



#### (Resposta 7): No histograma, nota-se que a maioria das ocorrências (frequência) está concentrada entre 0 e 500 dias no eixo do tempo de sobrevivência livre de recidiva (rfstime), representado no eixo horizontal (x). Este eixo mostra o tempo desde o transplante até a recidiva, morte ou último acompanhamento. O eixo vertical (y) reflete a quantidade de ocorrências dentro dessa faixa de tempo.