

MAD-CB

Figure 1:

Estatística Descritiva – Entender as Variáveis

Variáveis Categóricas – Proporções, Taxas e Relações (ratios)

Livro Que Forneceu Alguns Exemplos Hoje

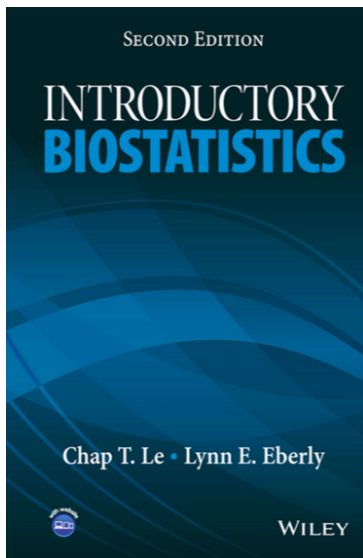


Figure 2:

- Estudo sobre diferenças raciais para incidência de cegueira por causa de glaucoma

```
str(cegGlauc)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2 obs. of  3 variables:
## $ raca : chr  "branco" "naobranco"
## $ pop  : num  32930233 3933333
## $ casos: num  2832 3227
```

```
kable(cegGlauc)
```

raca	pop	casos
branco	32930233	2832
naobranco	3933333	3227

- Contagem pura dos casos não ajuda muito para entender os dados
 - ▶ Diferença em tamanho das populações precisa ser tomado em conta

Proporção a Resgate

- Proporção dos casos a população por raça pode esclarecer o problema
- Proporção um número entre 0 e 1
 - ▶ Proporção acima de 1 não faz sentido
- Quando fala de incidência das doenças, esse proporção = **prevalência** da doença

$$\text{prevalência} = \frac{\text{número dos casos}}{\text{número total na população examinada}}$$

Calcular as Proporções

```
cegGlauc <- cegGlauc %>% mutate(prop = casos/pop)  
kable(cegGlauc)
```

raca	pop	casos	prop
branco	32930233	2832	0.0000860
naobranco	3933333	3227	0.0008204

- Proporção é um número muito pequeno
- Pode fazer ele mais significativo se multiplica por 10.000 ou 100.000
 - ▶ Coloca em termos de proporção por 10 mil ou 100 mil de população

Calcular Proporção por 100.000 Habitantes

```
cegGlauc <- cegGlauc %>% mutate(prop100mil = prop*100000)  
kable(cegGlauc)
```

raca	pop	casos	prop	prop100mil
branco	32930233	2832	0.0000860	8.60000
naobranco	3933333	3227	0.0008204	82.04238

- Agora, têm 2 números que podemos comparar
- Não brancos têm 82.0423798 casos por 100.000 pessoas
- Brancos só têm 8.5999999 casos por 100.000 pessoas
- Não brancos têm 9.54 o número de casos da doença que os brancos
- Doença desproporcionadamente afeta pessoas não brancos.

Valor das Proporções para os Dados Categóricos

- As proporções permitem que nós comparemos os grupos
 - ▶ Dá para os grupos uma base comum
- Proporção é a forma mais simples de normalização dos dados

- Como vimos semana passada, testes diagnósticos não são 100% exatos
 - ▶ Pessoas que têm doença com resultado negativo – **falso negativo**
 - ▶ Pessoas que não têm doença com resultado positivo – **falso positivo**
 - ▶ Classificação errada

Teste Diagnóstico para Câncer Cervical

- Teste com amostra de 24.103 mulheres para câncer cervical
- (Teste velho – só para demonstração)

Duas Proporções Que Nos Interessem

- Sensibilidade (Positivos Verdadeiros)

- ▶ Teste pode identificar essas pessoas que realmente são doentes?

$$\text{sensibilidade} = \frac{\text{número das pessoas doentes que testam positivo}}{\text{número total das pessoas doentes}}$$

- Especificidade (Negativos Verdadeiros)

- ▶ Teste pode só identificar pessoas doentes e não as saudáveis?

$$\text{especificidade} = \frac{\text{número das pessoas saudáveis que testam negativo}}{\text{número total das pessoas saudáveis}}$$

Dados de Amostra

```
kable(cervCan)
```

Estado	Neg	Pos	Tot
saudável	23362	362	23724
doente	225	154	379

```
sensib <- cervCan$Pos[Estado == "doente"] / cervCan$Tot[Estado == "doente"]  
paste("Sensibilidade = ", sensib * 100, "%")
```

```
## [1] "Sensibilidade = 40.6332453825858 %"
```

```
specif <- cervCan$Neg[Estado == "saudável"] / cervCan$Tot[Estado == "saudável"]  
paste("Especificidade = ", specif * 100, "%")
```

```
## [1] "Especificidade = 98.4741190355758 %"
```

- Taxa é semelhante a uma proporção normalmente com referência a tempo
 - ▶ Taxa de mudança; Ex: taxa de crescimento de um tumor
- Taxas podem exceder 1
 - ▶ Se algo dobre em tamanho num período, taxa de crescimento seria 100% ou 1

$$taxa = \frac{\text{valor novo} - \text{valor velho}}{\text{valor velho}} = \frac{\text{mudança}}{\text{valor velho}}$$

Exemplo: Crescimento em Número de Genotipagens para Pacientes com HIV

```
kable(genotip, captions = "Genotipagens por Ano")
```

yr	n
2010	3444
2011	3639
2012	5102
2013	5945
2014	6856
2015	6304

- Qual é a taxa de crescimento entre 2010 e 2015 em porcentagem?

```
taxacres <- 100 * (genotip$n[genotip$yr == 2015] -  
                  genotip$n[genotip$yr == 2010]) /  
                  genotip$n[genotip$yr == 2010]
```

- Taxa de crescimento = 83.0429733 %

Relações (Ratios)

- Relação das duas quantidades semelhantes medidas em grupos diferentes ou sob condições diferentes
- Ex: relação de homens e mulheres que fumam
- Se temos 200 homens na amostra que fumam e 150 mulheres,
 - ▶ Relação seria $200/150$ ou 1,33 homens que fumam por cada mulher

- Relação dos riscos de uma doença entre dois grupos diferentes

$$\text{risco relativo (RR)} = \frac{\text{incidência da doença em grupo 1}}{\text{incidência da doença em grupo 2}}$$

Doença Cardíaca Coronária (CHD) e Fumar

- Amostra de Americanos do Estudo Framingham
- Estudo case-control
 - ▶ Case - tem CHD nos últimos 10 anos
 - ▶ Control - saudável
- Fumar – Fator Confounding
 - ▶ Quanto fumar muda os resultados de CHD

```
suppressWarnings(tabCHD)
```

```
##           CHD10Anos
## fumante  Nao  Sim
##       Nao 1834  311
##       Sim 1762  333
```

```
print(paste("Fumantes Total:", "Não = ", fumtots[1], "| Sim = ", fumtots[2]))
```

```
## [1] "Fumantes Total: Não =  2145 | Sim =  2095"
```

```
print(paste("CHD 10 Anos Total:", "Não = ", CHDtots[1], "| Sim = ", CHDtots[2]))
```

```
## [1] "CHD 10 Anos Total: Não =  3596 | Sim =  644"
```

Calcular Incidência nos Dois Grupos

- Fumantes(Sim)

fumantes e doentes / todos os fumantes = $333 / 2095 = 0.1589499$

- Não-fumantes

não-fumantes e doentes / todos os não-fumantes = $311 / 2145 = 0.1449883$

- Risco Relativo (RR)

$$RR = \frac{\textit{incidência fumantes}}{\textit{incidência nãofumantes}}$$

```
rr <- fumincid / naofumincid
```

- $RR = 1.0962942$

Resumo de Métodos Descritivos para os Dados Categricos

- Proporções
 - ▶ Prevalência
 - ▶ Proporção por Unidade de População (Normalização)
 - ▶ Testes Diagnósticos
 - ▶ Sensibilidade e Especificidade
- Taxas
 - ▶ Medir mudanças
- Relações
 - ▶ Risco Relativo

Variáveis Contínuas

Distribuição de Frequência

- Exemplo simples: pesos de 57 crianças numa creche
- O que podemos aprender sobre a distribuição de frequência dos pesos?

Medidas de Tendência Central

- Tem um ponto que melhor descreve a distribuição?
- Qual número ocorre mais frequentemente? (**Modo**)
- Qual número sente do meio da distribuição? (**Mediana**)
- Qual número melhor representa o centro da distribuição? (**Média**)

- Mais simples
- O valor mais frequente da distribuição
- Entre os pesos, o modo é 27 lbs.
- É o mais frequente, ocorre 4 vezes
- R não tem uma função por modo.
 - ▶ Compartilhei uma função `modex` nas versões `.RMD` e `.R` destes slides
 - ▶ Outra função usa o nome `mode`

- Representa o valor diretamente no meio da distribuição
- Barra central num boxplot
- Para calcular, primeiro pôr os itens em ordem de menor até maior
- Se tiver um número impar dos itens, selecione aquele no meio
- Se tiver um número par dos itens, selecione os dois no meio e divide a diferença em dois
 - ▶ Mesma coisa que a média dos 2 itens no meio
- R tem a função `median` que faz todo o trabalho
- Muito útil quando tem distribuições bastante skewed
- A mediana de nossos pesos é 32

Média

- Estatística mais usada (e abusada) hoje
- Mede o centro de gravidade de todos os itens
- É a soma de todos os valores dos itens dividido por o número dos itens
- População

$$\mu = \sum_{i=1}^n X_i$$

- Amostra

$$\bar{x} = \sum_{i=1}^n x_i$$

- A média de nossos pesos é 36.7192982

- Mais simples: quais são os valores máximos e mínimos? **range**
 - ▶ Para nossos dados: 12, 79
- Diferença entre 1º e 3º Quartéis: Inter-Quartile Range **IQR**
 - ▶ Mais restrito para 50% da distribuição em volta da mediana
 - ▶ para nossos dados: 21
- Medir a dispersão em volta da média **desvio padrão**
 - ▶ Falamos semana passada sobre ele
 - ▶ Média das divergências entre os dados individuais e a média
 - ▶ É o raiz quadrado da variância $\sqrt{\sigma^2}$
 - ▶ Com a média, constitui os 2 parâmetros que definam a distribuição normal

Coeficiente de Variação

- Mede quanto dispersão tem in termos de uma relação
- Definido como $c_v = \frac{\sigma}{\mu}$
- Quantas média é o desvio padrão
- Não é padronizado
- Mas indica se o desvio padrão é alto demais para fazer análises muito subtis

descTools:Desc Mostra Medidas Descritivas Bem

```
> Desc(peso, plotit = TRUE)
```

peso (numeric)

length	n	NAs	unique	0s	mean	meanCI
57	57	0	31	0	36.72	32.51
	100.0%	0.0%		0.0%		40.93
.05	.10	.25	median	.75	.90	.95
15.20	21.60	25.00	32.00	46.00	59.40	68.20
range	sd	vcoef	mad	IQR	skew	kurt
67.00	15.87	0.43	14.23	21.00	0.72	-0.09

lowest : 12.0 (3), 16.0, 19.0, 21.0, 22.0 (2)

highest: 65.0, 68.0, 69.0, 74.0, 79.0

Figure 3:

Inclui Gráficos Úteis

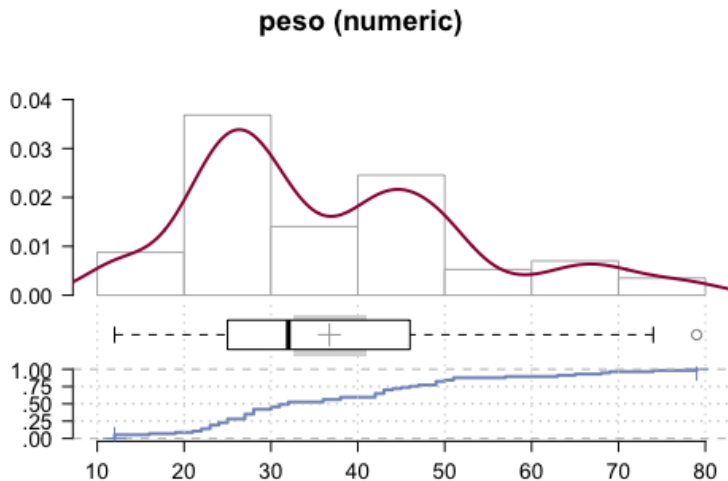
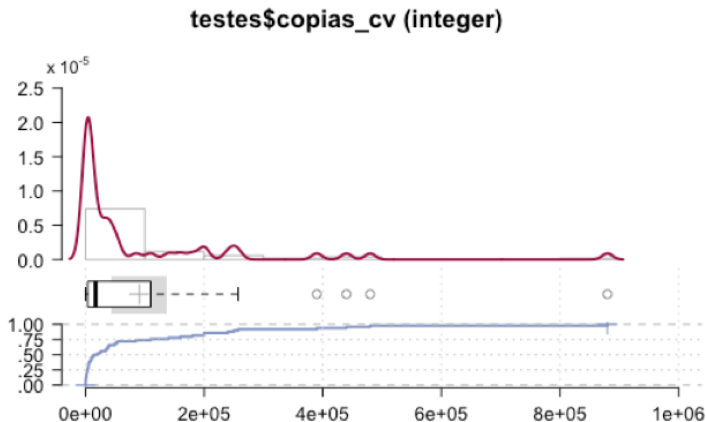


Figure 4:

Transformações das Variáveis

- Os dados da carga viral de semana passada
 - ▶ Altamente skewed



Pode Mudar a Escala da Variável para Fazer ele Mais Útil para Análise

- Agora, tem range de 879916
- Coeficiente de variação é 1.809823
- Pode transformar a escala de variável
 - ▶ Raiz quadrado
 - ▶ Logaritmo (ou base 10 ou base e [logaritmo Neperiano])

Raiz Quadrado Primeiro

```
Desc(sqrt(testes$copias_cv), conf.level = NA,  
      main = "Raiz Quadrado de Cópias CV", plotit = FALSE)
```

```
## -----  
## Raiz Quadrado de Cópias CV  
##  
##      length      n      NAs      unique      Os  
##      5e+01      5e+01      0      = n      0  
##      100.0%      0.0%      0.0%  
##  
##      .05      .10      .25      median      .75  
## 1.740791e+01 2.689752e+01 6.082763e+01 1.298715e+02 3.216338e+02  
##  
##      range      sd      vcoef      mad      IQR  
## 9.289180e+02 2.139415e+02 9.992760e-01 1.332724e+02 2.608061e+02  
##  
##      mean      meanCI  
## 2.140965e+02 1.532950e+02  
##      2.748979e+02  
##  
##      .90      .95  
## 5.007260e+02 6.458536e+02  
##  
##      skew      kurt  
## 1.347936e+00 1.241788e+00  
##  
## lowest : 9.165151e+00, 1.549193e+01, 1.643168e+01, 1.860108e+01, 1.923538e+01  
## highest: 5.072603e+02, 6.244998e+02, 6.633250e+02, 6.928203e+02, 9.380832e+02
```

Gráfico de Transformação de Raiz Quadrado

Raiz Quadrado de Cópias CV

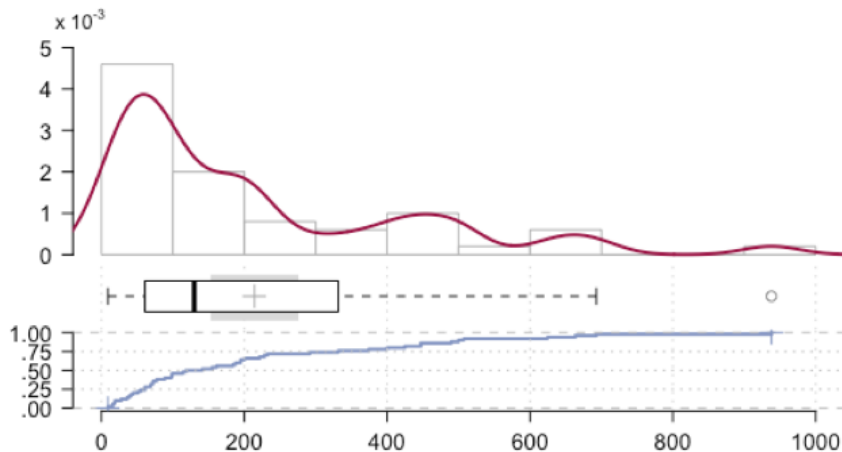


Figure 5:

Transformação Log Base 10 – Típico para CV

```
Desc(log10(testes$copias_cv), conf.level = NA,  
     main = "Log Base 10 de Cópias CV", plotit = FALSE)
```

```
## -----  
## Log Base 10 de Cópias CV  
##  
##      length      n      NAs      unique      Os  
##      5e+01      5e+01      0      = n      0  
##      100.0%      0.0%      0.0%  
##  
##      .05      .10      .25      median      .75  
## 2.479834e+00 2.854662e+00 3.568202e+00 4.223579e+00 5.013399e+00  
##  
##      range      sd      vcoef      mad      IQR  
## 4.020203e+00 9.890114e-01 2.359805e-01 1.076588e+00 1.445198e+00  
##  
##      mean      meanCI  
## 4.191073e+00 3.909999e+00  
##      4.472147e+00  
##  
##      .90      .95  
## 5.399192e+00 5.619878e+00  
##  
##      skew      kurt  
## -2.592686e-01 -7.936995e-01  
##  
## lowest : 1.924279e+00, 2.380211e+00, 2.431364e+00, 2.539076e+00, 2.568202e+00  
## highest: 5.410462e+00, 5.591065e+00, 5.643453e+00, 5.681241e+00, 5.944483e+00
```

Gráfico da Transformação Log Base 10

Log Base 10 de Cópias CV

