

Estatística Básica

PhD. Wagner Hugo Bonat

Laboratório de Estatística e Geoinformação-LEG
Universidade Federal do Paraná

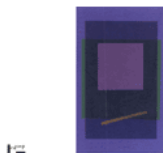
1/2017

Referência bibliográfica

- Livro-texto: Noções de Probabilidade e Estatística.
Autores: Marcos Nascimento Magalhães e
Antonio Carlos Pedroso de Lima.
Editora: EDUSP.

NOÇÕES DE PROBABILIDADE
E ESTATÍSTICA

Marcos Nascimento Magalhães
Antonio Carlos Pedroso de Lima



Laboratório de Estatística
e Geoinformação

Tópicos do curso

- Análise exploratória de dados.
- Probabilidades.
- Variáveis aleatórias discretas.
- Medidas resumo.
- Variáveis bidimensionais.
- Variáveis aleatórias contínuas.
- Inferência estatística - Estimação.
- Inferência estatística - Testes de hipóteses.
- Tópicos especiais.



O que é estatística?

- Estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.
- Estatística pode ser dividida em:
 - 1 Estatística descritiva.
 - 2 Probabilidade.
 - 3 Inferência estatística.
- Conceitos fundamentais:
 - 1 População: Conjunto de todos os elementos sob investigação.
 - 2 Amostra: Subconjunto da população.

Etapas da análise estatística

- Definição da população e característica de interesse.
- Amostragem: aleatória simples (com ou sem reposição), estratificada, por conglomerados, sistemática, etc.
- Estatística Descritiva (consistência e interpretações iniciais).
- Inferência estatística (baseado na amostra extrapolar para a população).
- Estimação de quantidades desconhecidas.
- Formulação e teste de hipóteses.

Organização de Dados

- Típica tabela de dados.

```
> # Carregando os dados
> dados <- read.table("questionario.txt", header = TRUE)
> head(dados)
```

	<i>Id</i>	<i>Turma</i>	<i>Sexo</i>	<i>Idade</i>	<i>Alt</i>	<i>Peso</i>	<i>Filhos</i>	<i>Fuma</i>	<i>Toler</i>	<i>Exerc</i>	<i>Cine</i>	<i>OpCine</i>	<i>TV</i>	<i>OpTV</i>
1	1	A	F	17	1.60	60.5	2	NAO	P	0	1	B	16	R
2	2	A	F	18	1.69	55.0	1	NAO	M	0	1	B	7	R
3	3	A	M	18	1.85	72.8	2	NAO	P	5	2	M	15	R
4	4	A	M	25	1.85	80.9	2	NAO	P	5	2	B	20	R
5	5	A	F	19	1.58	55.0	1	NAO	M	2	2	B	5	R
6	6	A	M	19	1.76	60.0	3	NAO	M	2	1	B	2	R

- Tipo de variáveis

- 1 Qualitativa nominal: Turma, Sexo, Fuma,
- 2 Qualitativa ordinal: Toler, OpCine, OpTV.
- 3 Quantitativa discreta: Idade, Filh, Exer, Cine, TV.
- 4 Quantitativa contínua: Alt, Peso.



Tabela de frequência

- Tabela de frequência: variáveis qualitativa nominais.

```
> # Sexo
> table(dados$Sexo)
```

```
  F  M
37 13
```

```
> prop.table(table(dados$Sexo))
```

```
  F  M
0.74 0.26
```

- Passo a passo.

Table 1: Tabela de frequência para a variável Sexo.

Sexo	n_i	f_i
F	37	0.74
M	13	0.26
Total	$n = 50$	1



Tabela de frequência

- Tabela de frequência: variáveis qualitativa ordinais e quantitativa discreta.

```
> # Idade
> table(dados$Idade)
```

```
17 18 19 20 21 23 24 25
 9 22 7 4 3 2 1 2
```

```
> prop.table(table(dados$Idade))
```

```
17 18 19 20 21 23 24 25
0.18 0.44 0.14 0.08 0.06 0.04 0.02 0.04
```

```
> cumsum(prop.table(table(dados$Idade)))
```

```
17 18 19 20 21 23 24 25
0.18 0.62 0.76 0.84 0.90 0.94 0.96 1.00
```



Tabela de frequência

- Passo a passo

Table 2: Tabela de frequência variável Idade.

Idade	n_i	f_i	f_{ac}
17	9	0.18	0.18
18	22	0.44	0.62
19	7	0.14	0.76
20	4	0.08	0.84
21	3	0.06	0.90
22	0	0.00	0.90
23	2	0.04	0.94
24	1	0.02	0.96
25	2	0.04	1.00
Total	$n = 50$	1	

Tabela de frequência

- Tabela de frequência: variáveis quantitativa contínuas ou discretas com muitos valores.

```
> # Peso
> # Criando as categorias
> dados$Peso_Cat <- NA
> dados[which(dados$Peso >= 40 & dados$Peso < 50),]$Peso_Cat <- '40 |- 50'
> dados[which(dados$Peso >= 50 & dados$Peso < 60),]$Peso_Cat <- '50 |- 60'
> dados[which(dados$Peso >= 60 & dados$Peso < 70),]$Peso_Cat <- '60 |- 70'
> dados[which(dados$Peso >= 70 & dados$Peso < 80),]$Peso_Cat <- '70 |- 80'
> dados[which(dados$Peso >= 80 & dados$Peso < 90),]$Peso_Cat <- '80 |- 90'
> dados[which(dados$Peso >= 90 & dados$Peso < 100),]$Peso_Cat <- '90 |- 100'
> cbind("ni" = table(dados$Peso_Cat),
+       "fi" = prop.table(table(dados$Peso_Cat)),
+       "fac" = cumsum(prop.table(table(dados$Peso_Cat))) ) )
```

```
      ni    fi    fac
40 |- 50    8 0.16 0.16
50 |- 60   22 0.44 0.60
60 |- 70    8 0.16 0.76
70 |- 80    6 0.12 0.88
80 |- 90    5 0.10 0.98
90 |- 100    1 0.02 1.00
```

```
>
```

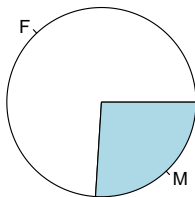
- Sugestão: Faça uma tabela de frequência pra a variável TV.

Representação gráfica

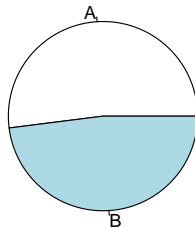
- Diagrama circular: qualitativa nominal e ordinal.

```
> par(mfrow = c(1,2), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))  
> pie(table(dados$Sexo), main = "Sexo")  
> pie(table(dados$Turma), main = "Turma")
```

Sexo



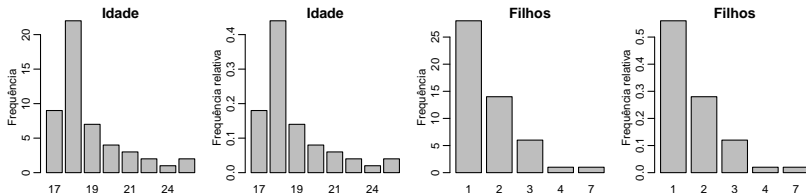
Turma



Representação gráfica

- Gráfico de barras: qualitativa ordinal e quantitativa discreta.

```
> par(mfrow = c(1,4), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> barplot(table(dados$Idade), main = "Idade", ylab = "Frequência")
> barplot(prop.table(table(dados$Idade)), main = "Idade", ylab = "Frequência relativa")
> barplot(table(dados$Filh), main = "Filhos", ylab = "Frequência")
> barplot(prop.table(table(dados$Filh)), main = "Filhos", ylab = "Frequência relativa")
```

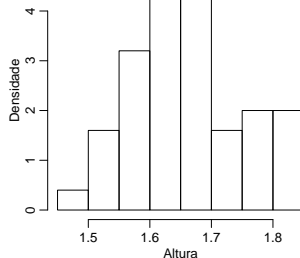
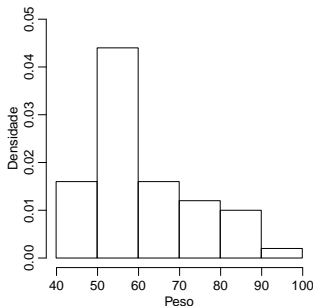


- Sugestão: Faça um gráfico de barras para as outras variáveis do conjunto de dados que este gráfico é adequado.

Representação gráfica

● Histogramas: quantitativa contínua.

```
> par(mfrow = c(1,2), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> hist(dados$Peso, prob = TRUE, ylab = "Densidade", xlab = "Peso",
+      ylim = c(0, 0.05), breaks = c(39.9, 49.9, 59.9, 69.9, 79.9, 89.9, 99.9), main = "")
> hist(dados$Alt, prob = TRUE, ylab = "Densidade", xlab = "Altura", main = "")
```



- Altura de cada retângulo é a densidade definida pelo quociente da área pela amplitude da faixa.

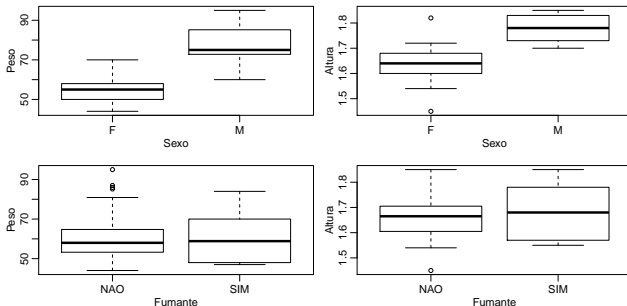
Mediana e quartis

- Mediana: valor da variável que divide o conjunto de dados ordenados em dois subgrupos de mesmo tamanho.
- Quartis: valores da variável que divide o conjunto de dados ordenados em quatro subgrupos de mesmo tamanho.
- Posição dos quartis:
 - 1 $Q_1 = 0.25 * (N + 1)$ arredonde.
 - 2 $Q_2 =$ média dos valores nas posições $(N/2)$ e $(N/2) + 1$ se N par e $Q_2 = (N + 1)/2$ se N ímpar.
 - 3 $Q_3 = 0.75 * (N + 1)$ arredonde.
- Example: Considere o conjunto de dados: 8.43(1), 8.65(2), 9.96(3), 10.91(4), 10.46(5) e 10.83(6).
- $Q_1 = 0.25 * 7 = 1.75 \approx 2$, ou seja 8.65.
- $Q_2 =$ média dos valores nas posições 3 e 4, ou seja, $(9.96 + 10.91)/2 = 10.43$.
- $Q_3 = 0.75 * 7 = 5.25 \approx 5$, ou seja, 10.46.

Representação gráfica

Boxplots: quantitativa contínua.

```
> par(mfrow = c(2,2), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> boxplot(dados$Peso ~ dados$Sexo, ylab = "Peso", xlab = "Sexo")
> boxplot(dados$Alt ~ dados$Sexo, ylab = "Altura", xlab = "Sexo")
> boxplot(dados$Peso ~ dados$Fum, ylab = "Peso", xlab = "Fumante")
> boxplot(dados$Alt ~ dados$Fum, ylab = "Altura", xlab = "Fumante")
```

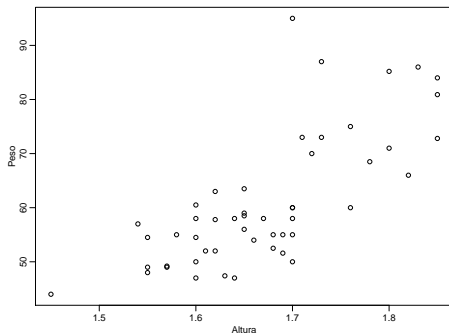


- Excelente para explorar relações entre variáveis quantitativas e qualitativas.

Representação gráfica

- Diagrama de dispersão: relação entre quantitativas.

```
> par(mfrow = c(1,1), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> plot(dados$Peso ~ dados$Alt, ylab = "Peso", xlab = "Altura")
```



- Excelente para explorar relação entre variáveis quantitativas.

Representação gráfica

- Gráfico de Mosaico: relação entre qualitativas.

```
> par(mfrow = c(1,1), mar=c(2.6, 2.8, 1.2, 0.5), mgp = c(1.6, 0.6, 0))
> mosaicplot(table(dados$Sexo, dados$Fuma), main = "")
```



- Excelente para explorar relação entre variáveis qualitativas (nominais ou ordinais).

Exercícios recomendados

- Seção 1.1 Ex. 1, 2 e 3.
- Seção 1.2 Ex. 1 e 4.
- Seção 1.4 Ex. 1, 3, 5, 8, 9, 12, 18 e 20.
- Avaliação próxima aula 5 questões.
- Próximo assunto: Capítulo 2 - Probabilidade.

