Bioestatística

LCE0204

Cristian Villegas
Escola Superior de Agricultura Luiz de Queiroz
Universidade de São Paulo
clobos@usp.br

Segundo Semestre

Sumário

Parte I

Introdução à Bioestatística

Capítulo 1
Ideias da disciplina

Objetivos da aula

Responder as perguntas a seguir

1. O que é Estatística?
2. Para que serve a Estatística?
3. Quais são as três grandes áreas da estatística?
4. Quais são as Etapas numa Pesquisa Estatística?

1.1 Áreas da estatística

1. Estatística Descritiva
2. Probabilidades
3. Inferência Estatística

1.2 Etapas numa Pesquisa Estatística

1. Pensar a pergunta de pesquisa
2. Coleta de dados
3. Resumo dos dados
4. Responder a pergunta de pesquisa

1.2.1 Como responder a pergunta de pesquisa?

1.2.2 Como coletar dados?

1.2.3 Como resumir os dados?

1.2.4 Como inferir sobre os dados?

Parte II

Análise exploratória de dados

Capítulo 2
Tipos de variáveis

Objetivos da aula

Responder as perguntas a seguir

1. O que é uma variável NOMINAL ?
2. O que é uma variável ORDINAL ?
3. O que é uma variável DISCRETA ?
4. O que é uma variável CONTINUA ?

2.1 Tipos de variáveis

1. Qualitativas: Nominal e Ordinal.
2. Quantitativas: Discreta e Contínua.

Capítulo 3
Tabela de frequências e gráficos

Objetivos da aula

Responder as perguntas a seguir

1. Quais são os tipos de tabelas associados a cada tipo de variáveis que estudamos na disciplina de Estatística?
2. Quais são os tipos de gráficos associados a cada tipo de variáveis que estudamos na disciplina de Estatística?

3.1 Tabela de frequências para uma variável qualitativa nominal

Variável	n_i	f_i
C_1	n_1	$f_1 = \frac{n_1}{n}$
C_2	n_2	$f_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
C_k	n_k	$f_k = \frac{n_k}{n}$
Total	n	1

em que

- n_i é a frequência absoluta,
- $f_i = n_i/n$ é a frequência relativa,

Exemplo 1 *Foram entrevistados 250 brasileiros, com 18 anos ou mais, para saber a opinião deles sobre determinadas marcas de cervejas. Com base nos dados apresentados na seguinte tabela, calcule as frequências relativas*

Marcas de Cervejas	n_i
Itaipava	12
Skol	63
Bohemia	130
Antártica	45
Total	250

Tabela 3.1: Opinião dos brasileiros sobre determinadas marcas de cervejas

Resultado do exercício anterior

Interpretação?

Marcas de Cervejas	n_i	f_i
Itaipava	12	0.048
Skol	63	0.252
Bohemia	130	0.520
Antartica	45	0.180
Total	250	1

3.2 Gráficos associados a uma variável qualitativa nominal

- Gráfico de barras e
- Gráfico de setores ou de pizza.

Usando software livre R para gerar os gráficos

Site para fazer download do software `www.r-project.org`.

```
#-------------------------------------------------------------
# "Opinião dos brasileiros sobre marcas de cervejas"
#-------------------------------------------------------------
rm(list=ls(all=TRUE))
respostas <- c("Itaipava","Skol","Bohemia","Antártica")
frequencia<- c(12,63,130,45)
dados<- data.frame(respostas, ni=frequencia)
n<- sum(frequencia)
dados$fi<- dados$ni/n
```

Código R: Gráfico de barras

```
barplot(dados[,"ni"], legend = dados[,"respostas"],
col = c("blue", "red", "yellow","green"))
```

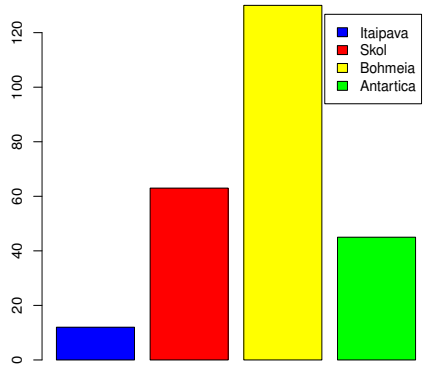

Figura 3.1: Opinião dos brasileiros sobre determinadas marcas de cervejas

Código R: Gráfico de setores ou de pizza

```
1 pie(dados$fi, col = c("blue", "red", "yellow","green"),labels=
2 c("Itaipava(4.8%)","Skol(25.2%)","Bohemia(52%)","Antartica(18%)"))
```

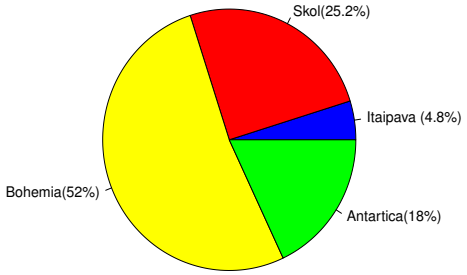

Figura 3.2: Opinião dos brasileiros sobre determinadas marcas de cervejas

3.3 Tabela de frequências para uma variável qualitativa ordinal

Variável	n_i	f_i	N_i	F_i
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

em que,

- n_i é a frequência absoluta,
- $f_i = n_i/n$ é a frequência relativa,
- $N_i = n_1 + n_2 + ... + n_i$ é a frequência absoluta acumulada e
- $F_i = f_1 + f_2 + ... + f_i$ é a frequência relativa acumulada.

Exemplo 2 *Foram entrevistados 2500 brasileiros, com 16 anos ou mais, para saber a opinião deles sobre determinado técnico de futebol. Com base nos dados da pesquisa apresentados na seguinte tabela, calcule as frequências relativas*

Opinião	n_i
Bom	1300
Regular	450
Ruim	125
Não sabe	625
Total	2500

Tabela 3.2: Opinião dos brasileiros sobre determinado técnico de futebol

Referência: Sônia Vieira (2008)

Resultado do exercício anterior

Respostas	n_i	f_i
Bom	1300	0.52
Regular	450	0.18
Ruim	125	0.05
Não sabe	625	0.25
Total	2500	1.00

Interpretação?

3.4 Gráficos associados a uma variável qualitativa ordinal

- Gráfico de barras e
- Gráfico de setores ou de pizza.

Usando software livre R para gerar os gráficos

```
#-------------------------------------------------------------
# "Opinião dos brasileiros sobre determinado técnico de futebol"
#  Fonte Viera(2008) Introdução à Bioestatística, página 29
#-------------------------------------------------------------
rm(list=ls(all=TRUE))
respostas <- c("Bom","Regular","Ruim","Não Sabe")
frequencia<- c(1300,450,125,625)
dados<- data.frame(respostas, ni=frequencia)
n<- sum(frequencia)
dados$fi<- dados$ni/n
```

Código R: Gráfico de barras

```
barplot(dados[,"ni"],legend = dados[,"respostas"],
col = c("blue", "red", "yellow","green"))
```

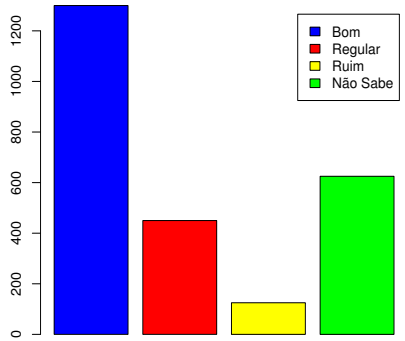

Figura 3.3: Opinião dos brasileiros sobre determinado técnico de futebol

Código R: Gráfico de setores ou de pizza

```
pie(dados$fi, col = c("blue", "red", "yellow","green"),
labels=c("Bom (52%)", "Regular(18%)", "Ruim(5%)", "Não sabe(25%)"))
```

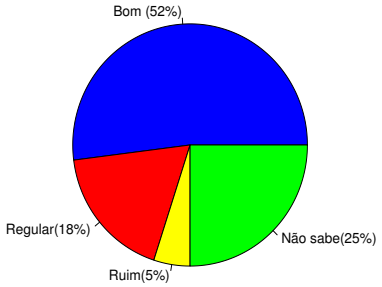

Figura 3.4: Opinião dos brasileiros sobre determinado técnico de futebol

3.5 Tabela de frequências para uma variável quantitativa discreta

Variável	n_i	f_i	N_i	F_i
C_1	n_1	f_1	N_1	F_1
C_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	n_k	f_k	$N_k = n$	$F_k = 1$
Total	n	1		

Exemplo 3 *As faltas ao trabalho de 30 empregados de uma clínica em determinado semestre estão na tabela a seguir. A partir dela, faça uma tabela de distribuição de frequências (absolutas, relativas e acumuladas).*

1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

Tabela 3.3: Número de faltas dadas por 30 empregados de uma clínica no semestre

Referência: Vieira (2008).

Resultado do exercício anterior

Número de faltas	n_i	f_i	N_i	F_i
0	9	0.300	9	0.300
1	10	0.333	19	0.633
2	5	0.167	24	0.800
3	3	0.100	27	0.900
4	2	0.067	29	0.967
6	1	0.033	30	1.000
Total	30	1		

Interpretação?

3.6 Gráficos associados a uma variável quantitativa discreta

- Gráfico de barras e
- Gráfico de frequências acumuladas (escada).

Usando software livre R para gerar os gráficos

```
#-------------------------------------------------------------------
#Núm. de faltas dadas por 30 empregados de uma clínica no semestre
#-------------------------------------------------------------------
faltas<- c(1 ,3 ,1 ,1 ,0 ,1 ,0 ,1 ,1 ,0,2 ,2 ,0 ,0 ,0 ,1 ,2 ,1 ,2,
0,0 ,1 ,6 ,4 ,3 ,3 ,1 ,2 ,4 ,0)

n<- length(faltas)
aux<- table(faltas)

dados1<- data.frame(aux)
dados2<- data.frame(aux/n)
final<- data.frame(faltas=dados1[,1], ni= dados1[,2],
fi= round(dados2[,2],3),Ni=cumsum(final$ni),Fi=cumsum(final$fi))
```

Código R: Gráfico de barras

```
barplot(final[,2], legend = final[,"faltas"],
xlab="Número de faltas", ylab="Frequência absoluta",
col = c("blue", "red", "yellow","green", "gray", "pink"))
```

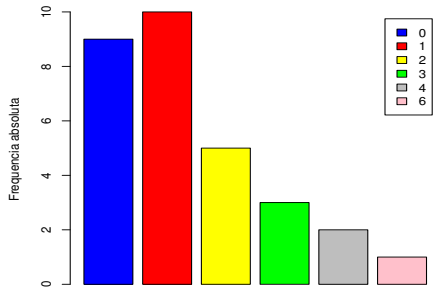

Figura 3.5: Número de faltas dadas por 30 empregados de uma clínica no semestre

Código R: Gráfico de frequências acumuladas (escada)

```
plot(c(0,1,2,3,4,6), final$Ni, xlab="Número de faltas",
ylab="Frequência absoluta acumulada",type="s", col="red")
```

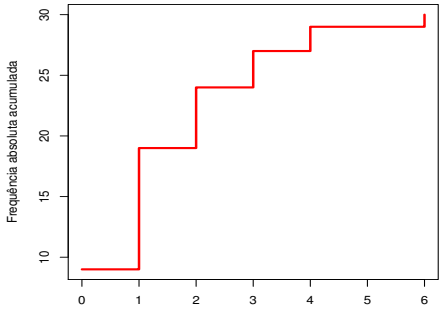

Figura 3.6: Número de faltas dadas por 30 empregados de uma clínica no semestre

3.7 Tabela de frequências para uma variável quantitativa contínua

em que X_i representa a marca de classe.

Intervalos	X_i	n_i	f_i	N_i	F_i
$[x_{11}, x_{12})$	$(x_{11} + x_{12})/2$	n_1	f_1	N_1	F_1
$[x_{21}, x_{22})$	$(x_{21} + x_{22})/2$	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_{k1}, x_{k2})$	$(x_{k1} + x_{k2})/2$	n_k	f_k	$N_k = n$	$F_k = 1$
Total		n	1		

Exemplo 4 *Os dados da tabela a seguir referem-se aos rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a Região Oeste Catarinense.*

3973	4660	4770	4980	5117	5540	6166	4500
4680	4778	4993	5166	5513	6388	4550	4685
4849	5056	5172	5823	4552	4760	4960	5063
5202	5889	4614	4769	4975	5110	5230	6047

Tabela 3.4: Rendimentos médios, em kg/ha, de 32 híbridos de milho, região Oeste, 1987/1988

Referência: Andrade e Ogliari (2007).

Quantas classes devemos considerar?

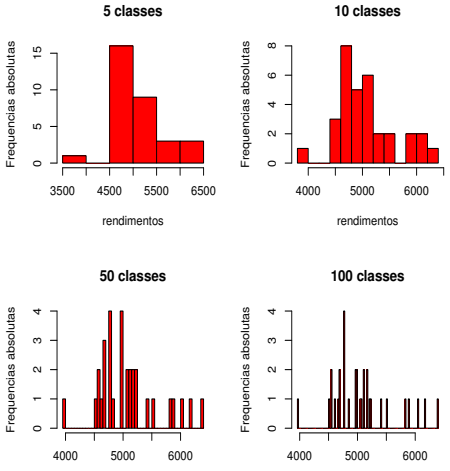

Figura 3.7: Histograma de Rendimentos médios considerando diferentes números de classes

Código R

```
par(mfrow=c(2,2))
hist(rendimentos, col="red",right=F, breaks=5, main="5 classes",
ylab="Frequencias absolutas")

hist(rendimentos, col="red",right=F, breaks=10, main="10 classes",
ylab="Frequencias absolutas")

hist(rendimentos, col="red",right=F, breaks=50, main="50 classes",
ylab="Frequencias absolutas")

hist(rendimentos, col="red",right=F, breaks=100, main="100 classes",
ylab="Frequencias absolutas")
```

Passos para construir uma tabela de frequências

- Determine o valor máximo e mínimo do conjunto de dados.
- Calcule a amplitude, que é a diferença entre o valor máximo e o valor mínimo.
- Determine o número de classes usando a regra de Sturges (1926), isto é, $k = 1 + 3.222 \log(n)$ em que n é o tamanho da amostra.
- Divida a amplitude dos dados pelo número de classes.
- O resultado da divisão é o intervalo de classe. É sempre melhor arredondar esse número para um valor mais alto, o que facilita o trabalho.
- Organize as classes de maneira que a primeira contenha o menor valor observado.

Passos para construir uma tabela de frequências (dados exemplo ??)

- Determine o valor máximo e mínimo do conjunto de dados.

```
> min(rendimentos)
[1] 3973

> max(rendimentos)
[1] 6388
```

- Calcule a amplitude, que é a diferença entre o valor máximo e o valor mínimo.

```
> (amplitude<- diff(range(rendimentos)))
[1] 2415
```

- Determine o número de classes usando a regra de Sturges(1926), isto é, $k = 1 + 3.222 \log(n)$ em que n é o tamanho da amostra.

```
> (k<- 1 + 3.222*log10(length(rendimentos)))#Regra de Sturges
[1] 5.849593
```

- Divida a amplitude dos dados pelo número de classes.

```
> amplitude/k
[1] 412.8492
```

- O resultado da divisão é o intervalo de classe. É sempre melhor arredondar esse número para um valor mais alto, o que facilita o trabalho.

```
Vamos aproximar para 500
```

- Organize as classes, de maneira que a primeira contenha o menor valor observado.

Resultado do exercício anterior

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

Interpretação?

3.8 Gráficos associados a uma variável quantitativa contínua

- Histograma.
- Polígono de Frequências.
  Gráfico de (X_i, n_i), $i = 1, ..., k$.
- Ogiva ou curva de frequências acumuladas.
  Gráfico de (Limite Superior$_i$, N_i) ou (Limite Superior$_i$, F_i), $i = 1, ..., k$.

Código R: Histograma

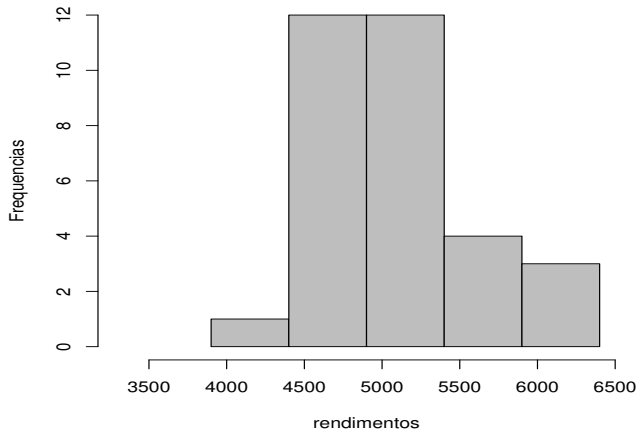

Figura 3.8: Histograma de Rendimentos médios

Código R: Polígono de frequências

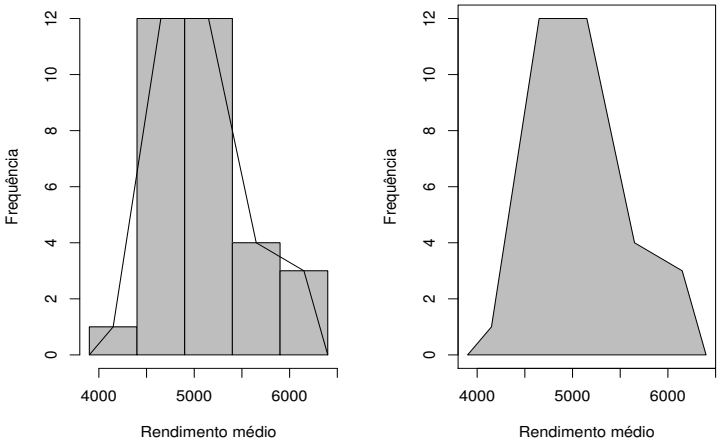

Figura 3.9: Polígono de Frequências dos Rendimentos médios

Código R: Ogiva (Curva de frequências acumuladas)

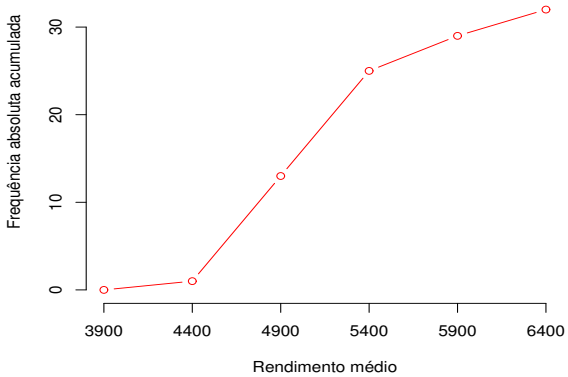

Figura 3.10: Ogiva dos Rendimentos médios

Código R: dados e histograma usando a regra de Sturges

Rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a Região Oeste Catarinense.

```
rendimentos<- c(3973 ,4660 ,4770 ,4980 ,5117 ,5403 ,6166,4500,
4680 ,4778 ,4993 ,5166 ,5513 ,6388 ,4550,4685,4849 ,5056 ,5172,
5823 ,4552 ,4760 ,4960,5063,5202 ,5889 ,4614 ,4769 ,4975 ,5110 ,
5230,6047)

hist(rendimentos, breaks=c(3900 ,4400 ,4900 ,5400 ,5900 ,6400),
ylab="Frequencias absolutas", main="", xlim=c(3300,6500),
col="gray")
```

Código R: histograma e polígono de frequências

```
par(mfrow=c(1,2))
h=hist(rendimentos,breaks=c(3900 ,4400 ,4900 ,5400 ,5900 ,6400),
main="",col="gray",xlab="Rendimento médio",ylab="Frequência")
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
type = "l")

plot(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
type = "n",main="",xlab="Rendimento médio",ylab="Frequência")
polygon(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0),
col="gray", border="black")
```

Código R: ogiva

```
1 library(fdth)
2 aux100=fdt(rendimentos, start=3900,h=500,end=6400)
3 plot(aux100,type='cfp', xlab="Rendimento médio",
4 ylab="Frequência absoluta acumulada")
```

Parte III

Medidas de tendência central

Capítulo 4
Introdução

- Média
- Moda
- Mediana

Conceitos básicos de somatório

Definição 1 *O somatório de $x_1, ..., x_n$ variáveis é definido por*

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n.$$

Propriedades

Sejam k, a e b constantes

1. $\sum_{i=1}^{n} k = nk$
2. $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$
3. $\sum_{i=1}^{n} (x_i \pm k) = \sum_{i=1}^{n} x_i \pm nk$
4. $\sum_{i=1}^{n} (a \pm bx_i) = na \pm b\sum_{i=1}^{n} x_i$
5. $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$
6. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

7. $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Definição 2 *O somatório que depende de* $x_1, ..., x_n$ *e* $y_1, ..., y_n$ *variáveis é definido por*

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + ... + x_n y_n.$$

Propriedades para duas variáveis

Sejam k, a e b constantes

1. $\sum_{i=1}^{n} kx_i y_i = k \sum_{i=1}^{n} x_i y_i$
2. $\sum_{i=1}^{n} (x_i y_i \pm k) = \sum_{i=1}^{n} x_i y_i \pm nk$
3. $\sum_{i=1}^{n} (ax_i \pm by_i) = a \sum_{i=1}^{n} x_i \pm b \sum_{i=1}^{n} y_i$

4.1 Média para dados não agrupados

A medida de tendência central mais conhecida e mais utilizada é a média aritmética, ou simplesmente média. Como se calcula a média?

Definição 3 *A média aritmética de um conjunto de dados numéricos é obtida somando todos os dados e dividindo o resultado pelo número deles. A média, que denotamos por $\bar{x}$ (lê-se x-barra), é definida por*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + ... + x_n}{n}.$$

Exemplo 5 *Um professor de Educação Física mediu a circunferência abdominal de 10 homens que se apresentaram em uma academia. Obteve os valores , em centímetros: 88, 83, 79, 76, 78, 70, 80, 82, 86 e 105. Calcule a média*

Solução

$$\bar{x} = \frac{88 + 83 + ... + 105}{10} = \frac{827}{10} = 82.7cm$$

Interpretação?: Os homens mediram, em média 82.7 cm de circunferência abdominal.

4.2 Mediana para dados não agrupados

Definição 4 *A mediana (M_e) é o valor que ocupa a posição central do conjunto dos dados ordenados.*

- A mediana divide a amostra em duas partes: uma com números menores ou iguais à mediana, outra com números maiores ou iguais à mediana.
- Quando o número de dados é ímpar, existe um único valor na posição central.
- Quando o número de dados é par, existem dois valores na posição central. A mediana é a média desses dois valores. Em resumo,

$$M_e = \begin{cases} x_{[\frac{n+1}{2}]} & \text{n ímpar} \\ \dfrac{x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}}{2} & \text{n par} \end{cases}$$

Exemplo 6 *Calcule a mediana do peso, em quilogramas, de cinco bebês nascidos em um hospital: 3.500, 2.850, 3.370, 2.250 e 3.970.*

- Coloque os dados em ordem crescente como segue 2.250, 2.850, 3.370, 3.500, 3.970. A mediana é o valor que está na posição central, ou seja, 3.370 kg. A mediana usando a fórmula anterior fica

dada por

$$M_e = x_{[\frac{5+1}{2}]} = x[3] = 3.370\text{kg}.$$

- Se no exemplo **??** os dados tivessem sido 3.500, 2.850, 3.370, 2.250, então a mediana seria

$$M_e = \frac{x_{[\frac{4}{2}]} + x_{[\frac{4}{2}+1]}}{2} = \frac{x[2] + x[3]}{2} = \frac{2.850 + 3.370}{2} = 3.110\text{kg}.$$

4.3 Moda para dados não agrupados

Definição 5 *A moda é o valor que ocorre com maior frequência.*

Exemplo 7 *Determine a moda dos dados: 0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6.*

A moda é 7, porque é o valor que ocorre com o maior número de vezes.

- Um conjunto de dados pode não ter moda porque nenhum valor se repete maior número de vezes, ou ter duas ou mais modas.
- O conjunto de dados

  $$0, 2, 4, 6, 8, 10$$

  não tem moda.
- O conjunto de dados

  $$1, 2, 2, 3, 4, 4, 5, 6, 7$$

  tem duas modas: 2 e 4.

4.4 Média para uma variável quantitativa discreta(dados agrupados)

Definição 6 *A média aritmética de dados agrupados em uma tabela de distribuição de frequências, isto é, de* $x_1, ...x_k$ *que se repetem* $n_1, ..., n_k$ *vezes na amostra, é*

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n},$$

em que $n = \sum_{i=1}^{k} n_i$.

Exemplo 8 *Para calcular a média do número de filhos em idade escolar que têm os funcionários de uma empresa, a psicóloga que trabalha em Recursos Humanos obteve uma amostra de 20 funcionários. Os dados estão apresentados em seguida. Como se calcula a média?.*

Referência: Vieira (2008)

1	0	1	0	2	1	2	1	2	2
1	5	0	1	1	1	3	0	0	0

Tabela 4.1: Número de filhos em idade escolar de 20 funcionários

4.5 Média para uma variável quantitativa contínua (dados agrupados)

Definição 7 *A média aritmética de dados agrupados em uma tabela de distribuição de frequências é dada por*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\, X_i = \frac{n_1 X_1 + ... + n_k X_k}{n}$$

em que k é o número de classes e X_i é a marca de classe.

Exemplo 9 *Calcule a média para os dados do exemplo* **??**.

Número de filhos em idade escolar	n_i	$x_i n_i$
0	6	0
1	8	8
2	4	8
3	1	3
4	0	0
5	1	5
Total	20	24

$$\bar{x} = \frac{0 \times 6 + ... + 5 \times 1}{20} = \frac{24}{20} = 1.2 \quad \text{filhos.}$$

Comentário: O número médio de filhos em idade escolar é 1.

4.6 Mediana para dados agrupados

Definição 8 *A mediana para dados agrupados é calculada da seguinte forma*

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e - 1}}{n_{M_e}} \right) \times a_{M_e}$$

em que

- LI_{Me}*: Limite inferior da classe mediana.*

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$\bar{x} = \frac{(4150 \times 1 + ... + 6150 \times 3)}{32} = 5087.5\text{kg/ha}.$$

- n: *Tamanho da amostra.*
- N_{M_e-1}: *Frequência absoluta acumulada anterior à classe* M_e.
- n_{M_e}: *Frequência absoluta da classe* M_e.
- a_{M_e}: *Amplitude da classe* M_e.

Exemplo 10 *Calcule a mediana para os dados do exemplo* **??**.

Exemplo 11 *Calcule a mediana para os dados do exemplo* **??**.

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e - 1}}{n_{M_e}} \right) \times a_{M_e} = ????????.$$

4.7 Moda para dados agrupados

Definição 9 *A moda para dados agrupados é calculada da seguinte forma.*

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times a_{M_o}$$

em que,

- LI_{Mo}*: Limite inferior da classe modal.*
- $\Delta_1 = n_{(Mo)} - n_{(Mo-1)}$ *e* $\Delta_2 = n_{(Mo)} - n_{(Mo+1)}$.

Rendimentos Médios	X_i	n_i	f_i	N_i	F_i
$[3900 - 4400)$	4150	1	0.031	1	0.031
$[4400 - 4900)$	4650	12	0.375	13	0.406
$[4900 - 5400)$	5150	12	0.375	25	0.781
$[5400 - 5900)$	5650	4	0.125	29	0.906
$[5900 - 6400)$	6150	3	0.094	32	1.000
Total		32	1		

$$M_e = LI_{M_e} + \left(\frac{\frac{n}{2} - N_{M_e-1}}{n_{M_e}}\right) \times a_{M_e} = 4900 + \left(\frac{32/2 - 13}{12}\right) \times 500 = 5025\,\text{kg/ha}.$$

- $n_{(M_o)}$: *Frequência absoluta da classe modal.*
- $n_{(Mo-1)}$: *Frequência absoluta anterior à classe modal.*
- $n_{(Mo+1)}$: *Frequência absoluta posterior à classe modal.*
- a_{M_o}: *Amplitude da classe* M_o.

Exemplo 12 *Calcule a moda para os dados, apresentados a seguir, de produção de resina(kg) de 40 arvores de Pinus elliotti.*

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times a_{M_o} = ??????????????????$$

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 4.2: Produção de resina (kg) de 40 arvores de Pinus elliotti

Resposta do exercício anterior

$$M_o = LI_{M_o} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times a_{M_o} = 2.01 + \left(\frac{12-6}{12-6+12-9}\right) \times 0.70 = 2.477\text{kg}.$$

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 4.3: Produção de resina (kg) de 40 arvores de Pinus elliotti

Parte IV

Medidas de dispersão

Introdução

Exemplo 13 *Considere as notas de uma prova de estatística aplicada a três turmas*

- *Grupo 1: 3, 4, 5, 6, 7.*
- *Grupo 2: 1, 3, 5, 7, 9.*
- *Grupo 3: 5, 5, 5, 5, 5. Calcule a média e a mediana de cada grupo.*

Comentários? Precisamos de uma medida de variabilidade.

Código R: Gráfico para estudar dispersão

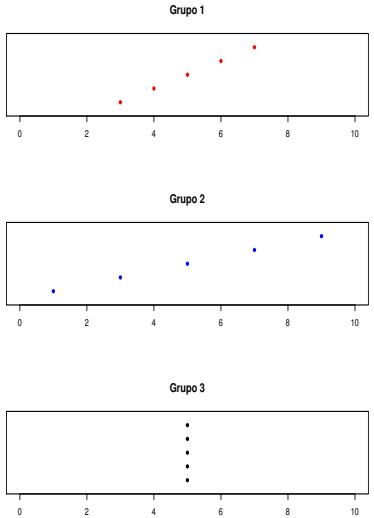

Figura 4.1: Notas de uma prova de estatística aplicada a três turmas

Capítulo 5
Medidas de dispersão para dados não agrupados

5.1 Amplitude

Definição 10 *Uma medida da variabilidade é a amplitude, que é obtida subtraindo o valor mais baixo de um conjunto de observações do valor mais alto, isto é,*

$$\textit{Amplitude} = \textit{máximo - mínimo}$$

Alguns comentários da amplitude

- é fácil de ser calculada e suas unidades são as mesmas que as da variável,
- não utiliza todas as observações (só duas delas) e
- pode ser muito afetada por alguma observação extrema.

5.2 Variância e desvío padrão

Definição 11 *A variância s^2 é definida como a média das diferenças quadráticas de n valores em relação à sua média aritmética, ou seja,*

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

Essa medida é sempre uma quantidade positiva. Como suas unidades são as do quadrado da variável, é mais fácil usar sua raiz quadrada.

Definição 12 *O desvio padrão ou desvio típico é definido como a raiz quadrada de s^2, isto é,*

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

O desvio padrão é uma medida de variabilidade ou dispersão e é medida na mesma dimensão que as das observações.

Exemplo 14 *Calcule a amplitude, variância e desvio padrão das seguintes quantidades medidas em metros: 3, 3, 4, 4, 5.*

Solução

- A amplitude dessas obervações é 5-3=2 metros.
- $\bar{x} = (3 + 3 + 4 + 4 + 5)/5 = 3.8$ metros.
- $s^2 = 0.70$ metros^2.
- $s = \sqrt{0.70 \text{metros}^2} = 0.84$ metros.

Capítulo 6
Medidas de dispersão para dados agrupados

6.1 Variáveis discretas

Seja s^2 e $s = \sqrt{s^2}$, a variância e o desvio padrão respectivamente, então para dados agrupados temos que

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,(x_i - \bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,x_i^2 - n\,\bar{x}^2\right)$$

Exemplo 15 *Calcular a variância, o desvio padrão para o conjunto de dados amostrais apresentados na tabela abaixo.*

x_i	n_i
1	2
3	4
5	2

Tabela 6.1: Distribuição do número de irmãos dos professores do LES

Resposta do exercício anterior

$$\bar{x} = \frac{1 \times 2 + 3 \times 4 + 5 \times 2}{8} = 3 \text{ irmãos}$$

$$s^2 = \frac{(1-3)^2 \times 2 + (3-3)^2 \times 4 + (5-3)^2 \times 2}{8-1} = 2.29 \text{ irmãos}^2$$

$$s = \sqrt{2.29 \text{ irmãos}^2} = 1.51 \text{ irmãos}$$

6.2 Variáveis continuas

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,(X_i - \bar{x})^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\,X_i^2 - n\,\bar{x}^2\right)$$

Exemplo 16 *Veja exemplo* ??.

Produção de resina (kg)	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 6.2: Produção de resina (kg) de 40 arvores de Pinus elliotti

Resposta do exercício anterior

Temos que

$$s^2 = \frac{1}{40-1}\left(\sum_{i=1}^{7} n_i\, X_i^2 - 40 \times \bar{x}^2\right)$$

em que,

$$\bar{x} = \frac{1}{40}(0.96 \times 3 + ... + 5.16 \times 1) = 2.6925 \text{ kg}.$$

Logo,

$$s^2 = \frac{1}{39}\left(3 \times 0.96^2 + ... + 1 \times 5.16^2 - 40 \times 2.6925^2\right) = 0.8791\,\text{kg}^2.$$

Assim, $s = 0.9376$kg.

6.3 Coeficiente de variação

Definição 13 *O coeficiente de variação se define por*

$$CV = \frac{s}{\bar{x}} \times 100\%$$

em que s é o desvio padrão e $\bar{x}$ é a média.

O coeficiente de variação

- é uma medida de dispersão relativa
- elimina o efeito da magnitude dos dados
- exprime a variabilidade em relação à média

Exemplo 17 *Os dados estudados neste exemplo correspondem às idades e alturas da turma de Cálculo Conclusão: Os alunos são, mais dispersos quanto a idade do que quanto à altura.*

Variáveis	Média	Desvio Padrão	CV
Altura	171.33	11.10	6.4 %
Idade	19	1.62	8.5 %

Tabela 6.3: Altura e Idade dos alunos.

Parte V

Medidas de posição

Capítulo 7
Introdução

- Quartis
- Decis
- Percentis

7.1 Quartis, Decis e Percentis

Definição 14 *Os quartis dividem os dados em 4 conjuntos iguais* (Q_1, Q_2, Q_3). Q_2 *representa a mediana.*

Definição 15 *Os decis dividem os dados em 10 conjuntos iguais* $(D_1, ..., D_9)$. D_5 *representa a mediana.*

Definição 16 *Os percentis dividem os dados em 100 conjuntos iguais* $(P_1, ..., P_{99})$. P_{50} *representa a mediana.*

- Podemos observar que a mediana coincide com o quartil 2 (Q_2), decil 5 (D_5) e percentil 50 (P_{50}).

7.2 Percentis para dados não agrupados

Definição 17 *O percentil P_j para dados não agrupados é definido como*

$$P_j = \begin{cases} x_{[i+1]} & f > 0 \\ \dfrac{x_{[i]} + x_{[i+1]}}{2} & f = 0 \end{cases}$$

$j = 1, ..., 99$. *A forma de calcular percentil é a seguinte* $n \times p = i + f$, *em que* i *parte representa a parte inteira e* f *parte decimal do produto* $n \times p$, $0 < p < 1$.

Exemplo 18 *Veja exemplo ?? e calcule o percentil 25, 33, 50, 63 e 75.*

- $40 \times 0.25 = 10 + 0$, logo $P_{25} = \dfrac{x_{[10]} + x_{[11]}}{2} = 2.05$kg.
- $40 \times 0.33 = 13 + 0.2$, logo $P_{33} = x_{[14]} = 2.16$kg.
- $40 \times 0.50 = 20 + 0$, logo $P_{50} = \dfrac{x_{[20]} + x_{[21]}}{2} = 2.65$kg.
- $40 \times 0.63 = 25 + 0.2$, logo $P_{63} = x_{[26]} = 3.09$kg.
- $40 \times 0.75 = 30 + 0$, logo $P_{75} = \dfrac{x_{[30]} + x_{[31]}}{2} = 3.46$kg.

7.3 Percentis para dados agrupados

Definição 18 *O percentil P_j para dados agrupados é definido como*

$$P_j = LI_k + \left(\frac{n \times \frac{j}{100} - N_{k-1}}{n_k} \right) \times a_k \quad j = 1, ..., 99.$$

Observação 1 *A seguir alguns casos particulares de percentis*

$$\begin{aligned} P_{25} &= LI_k + \left(\frac{n \times \frac{25}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_1 \\ P_{50} &= LI_k + \left(\frac{n \times \frac{50}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_2 \\ P_{75} &= LI_k + \left(\frac{n \times \frac{75}{100} - N_{k-1}}{n_k} \right) \times a_k = Q_3 \end{aligned}$$

Exemplo 19 *Veja o exemplo ?? (produção de resina(kg) de 40 arvores de Pinus elliotti) e calcule o percentil 25, 50 e 75.*

Classes	X_i	n_i	f_i	N_i	F_i
$[0.61; 1.31)$	0.96	3	0.075	3	0.075
$[1.31; 2.01)$	1.66	6	0.150	9	0.225
$[2.01; 2.71)$	2.36	12	0.350	21	0.525
$[2.71; 3.41)$	3.06	9	0.225	30	0.750
$[3.41; 4.11)$	3.76	9	0.225	39	0.975
$[4.11; 4.81)$	4.46	0	0.000	39	0.975
$[4.81; 5.51)$	5.16	1	0.025	40	1.000

Tabela 7.1: Produção de resina(kg) de 40 arvores de Pinus elliotti.

Resultado do exercício anterior

A seguir calculamos o percentil 25, 50 e 75, respectivamente

$$P_{25} = LI_k + \left(\frac{n \times \frac{25}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.01 + \left(\frac{40 \times 1/4 - 9}{12} \right) \times 0.70 = 2.068$$

$$P_{50} = LI_k + \left(\frac{n \times \frac{50}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.01 + \left(\frac{40 \times 1/2 - 9}{12} \right) \times 0.70 = 2.652$$

$$P_{75} = LI_k + \left(\frac{n \times \frac{75}{100} - N_{k-1}}{n_k} \right) \times a_k = 2.71 + \left(\frac{40 \times 3/4 - 21}{9} \right) \times 0.70 = 3.410$$

7.4 Gráfico de caixas-e-bigodes (boxplot)

- Determinar valor mínimo dos dados.
- Determinar valor máximo dos dados.
- Determinar Q_1, Q_2 e Q_3.
- Determinar se há pontos atípicos $Q_1 - 1.5IQR$ ou $Q_3 + 1.5IQR$, em que $IQR = Q_3 - Q_1$ é a amplitude interquatilica.

Código R: Quartis (dados brutos)

```
> Quartis<- boxplot(resina, plot=F)
> Quartis.novo<- data.frame(Quartis$stats)
> rownames(Quartis.novo)<- c("Minimo","Quar. 1","Quar. 2",
"Quar. 3","Maximo")
> Quartis.novo
        Quartis.stats
Minimo           0.71
Quar. 1          2.05
Quar. 2          2.65
Quar. 3          3.46
Maximo           5.41
```

Exemplo 20 *Com base no exemplo ?? (produção de resina(kg) de 40 arvores de Pinus elliotti) construir boxplot.*

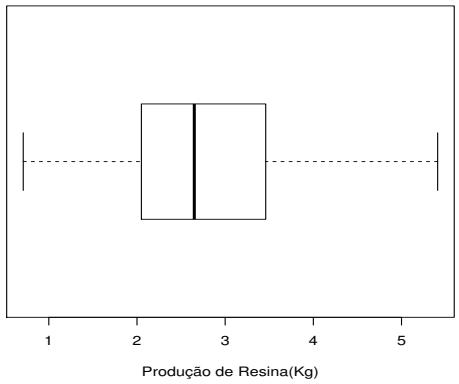

Figura 7.1: Gráfico Caixas-e-bigodes para dados de resina (Kg)

Exemplo 21 *Estatura de alunos da turma de Bioestatística por sexo.*

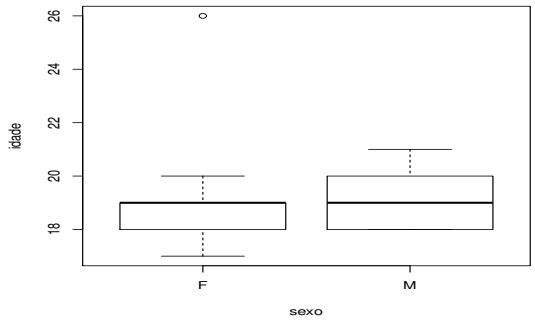

Figura 7.2: Gráfico Caixas-e-bigodes para dados de resina (Kg)

Parte VI

Medidas de simetria

7.5 Introdução

Tem por objetivo básico medir o quanto a distribuição de frequências do conjunto de valores observados se afasta da condição de simetria.

7.6 Distribuição simétrica

- $\bar{x} = M_e = M_o$.

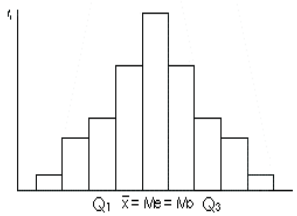

Figura 7.3: Distribuição simétrica

7.7 Distribuição assimétrica negativa ou assimétrica à esquerda

- $\bar{x} < M_e < M_o$

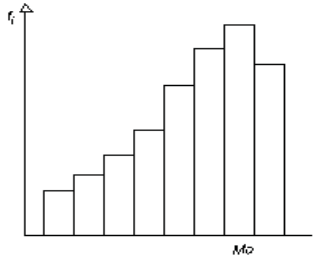

Figura 7.4: Distribuição assimétrica à esquerda

7.8 Distribuição assimétrica positiva ou assimétrica à direita

- $M_o < M_e < \bar{x}$

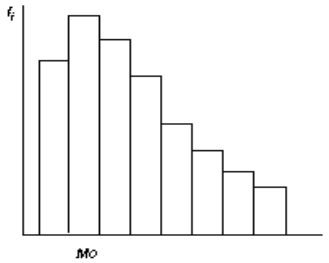

Figura 7.5: Distribuição assimétrica à direita

Referências

Andrade, Dalton F e Ogliari, Paulo J (2010). Estatística para as ciências agrárias e biológicas com noções de experimentação. Editora da UFSC.

Vieira, Sônia (2008). Introdução à Bioestatística. 4a edição: Elsevier.

Parte VII

Regressão e correlação

Capítulo 8
Introdução

8.1 Correlação

Seja r o coeficiente de correlação linear

$$r = \frac{Sxy}{\sqrt{SxxSyy}} \quad \text{em que,}$$

$$Sxy = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}, \quad Sxx = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \quad Syy = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Conjunto A

Para o conjunto A, temos que $Sxy = 84, Sxx = 82.5, Syy = 133.6, \bar{x} = 5.5, \bar{x} = 6.2, n = 10$. Logo, $r = 0.80$ correlação positiva

Conjunto B

Para o conjunto B, temos que $Sxy = -82.5, Sxx = 82.5, Syy = 133, \bar{x} = 5.5, \bar{x} = 6.2, n = 10$. Logo, $r = -0.78$ correlação negativa

Código R para calcular correlação

```
> (conjuntoA<- data.frame(xA, yA))
xA yA
1   1  0
2   2  2
3   3  6
4   4  3
5   5  9
6   6  4
7   7 10
8   8  8
9   9 12
10 10  8

> (conjuntoB<- data.frame(xB, yB))
xB yB
1   1  8
2   2 12
```

```
3   3  8
4   4 10
5   5  4
6   6  9
7   7  3
8   8  6
9   9  0
10 10  2
```

```
> correlacao<- function(x, y){
  + n<- length(x)
  + Sxy<- sum(x*y)-n*mean(x)*mean(y)
  + Sxx<- sum(x*x)-n*mean(x)*mean(x)
  + Syy<- sum(y*y)-n*mean(y)*mean(y)
  + r<- Sxy/sqrt(Sxx*Syy)
  +
    + saidas<- list()
    + saidas$Sxy<- Sxy
    + saidas$Sxx<- Sxx
    + saidas$Syy<- Syy
```

```
    + saidas$mediaX<- mean(x)
    + saidas$mediaY<- mean(y)
    + saidas$n<- length(x)
    + saidas$r<- r
    + return(saidas)
    + }

> correlacao(xA, yA)
$Sxy
[1] 84

$Sxx
[1] 82.5

$Syy
[1] 133.6

$mediaX
[1] 5.5
```

```
$mediaY
[1] 6.2

$n
[1] 10

$r
[1] 0.8001089

> correlacao(xA, yB)
$Sxy
[1] -82

$Sxx
[1] 82.5

$Syy
[1] 133.6

$mediaX
```

```
[1] 5.5

$mediaY
[1] 6.2

$n
[1] 10

$r
[1] -0.7810587

> cor(conjuntoA)
xA        yA
xA 1.0000000 0.8001089
yA 0.8001089 1.0000000

> cor(conjuntoB)
xB         yB
xB  1.0000000 -0.7810587
yB -0.7810587  1.0000000
```

```
> data.frame(conjuntoA, conjuntoB)
xA yA xB yB
1   1  0  1  8
2   2  2  2 12
3   3  6  3  8
4   4  3  4 10
5   5  9  5  4
6   6  4  6  9
7   7 10  7  3
8   8  8  8  6
9   9 12  9  0
10 10  8 10  2

> par(mfrow=c(1,2))
> plot(conjuntoA, pch=20, lwd=3, main="ConjuntoA")
> plot(conjuntoB, pch=20, lwd=3, main="ConjuntoB")
```

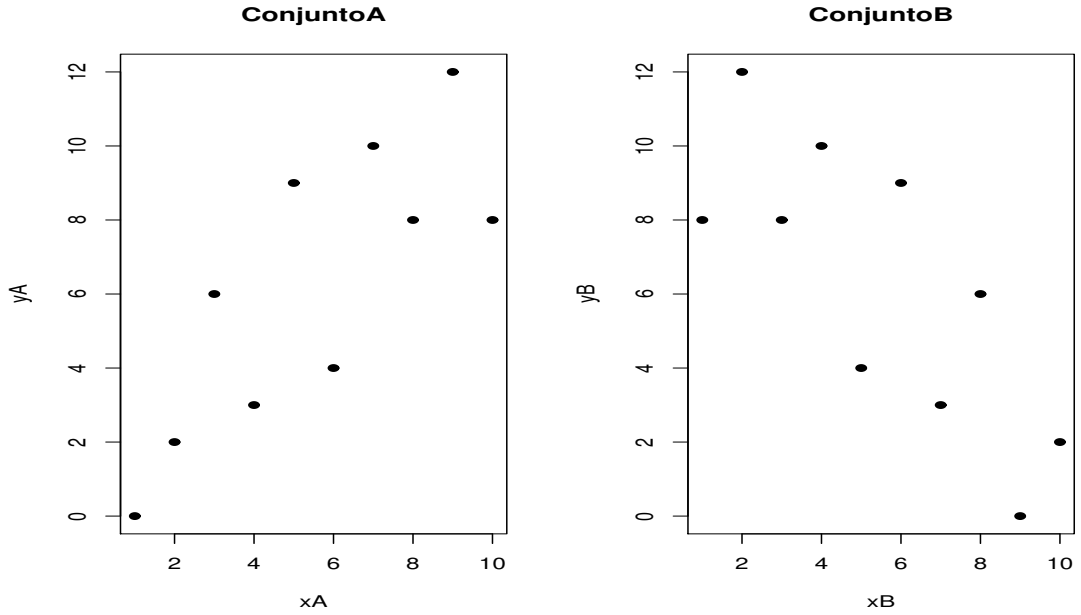

Figura 8.1: Gráfico de dispersão

8.2 Regressão

A equação da reta é dada por

$$y_i = a + b * x_i$$

Equação da reta estimada (com base nos dados (x,y))

$$\hat{y}_i = \hat{a} + \hat{b} * x_i, \qquad \text{em que}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \hat{b} = \frac{Sxy}{Sxx}$$

$$Sxy = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}, \quad Sxx = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

A equação da reta estimada, fica dada por

$Sxy = 371.35, Sxx = 171.875, \bar{x} = 8.625, \bar{y} = 17.65, n = 8, a = -0.985, b = 2.161$

$$\hat{y}_i = -0.985 + 2.161 * x_i$$

Código R

```
#funcao que calcula regressao
rm(list=ls(all=TRUE))
regressao<- function(x, y){
  n<- length(x)
  Sxy<- sum(x*y)-n*mean(x)*mean(y)
  Sxx<- sum(x*x)-n*mean(x)*mean(x)
  b<- Sxy/Sxx
  a<- mean(y)-b*mean(x)
  saidas<- list()
  saidas$Sxy<- Sxy
  saidas$Sxx<- Sxx
  saidas$mediaX<- mean(x)
  saidas$mediaY<- mean(y)
  saidas$n<- length(x)
  saidas$a<- a
  saidas$b<- b
  return(saidas)
```

```
}

#Exemplo
> tempo<- c(2 ,3 ,5 ,8 ,10, 12, 14, 15)
> quantidade<- c( 3.5, 5.7, 9.9, 16.3, 19.3, 25.7 ,28.2, 32.6)

> regressao(tempo,quantidade)
$Sxy
[1] 371.35

$Sxx
[1] 171.875

$mediaX
[1] 8.625

$mediaY
[1] 17.65

$n
```

```
[1] 8

$a
[1] -0.9850182

$b
[1] 2.160582

> plot(tempo, quantidade, pch=20, lwd=3, main="")
> abline(lm(quantidade~tempo)$coef, col="red",lwd=2)
```

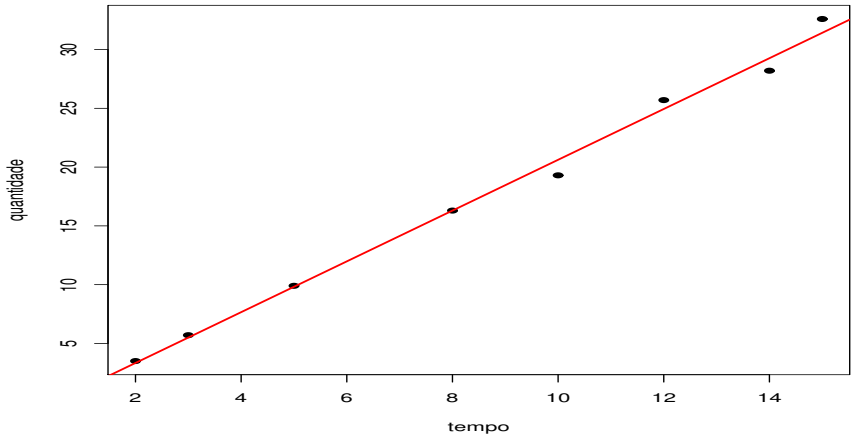

Figura 8.2: Gráfico de dispersão

Parte VIII

Teoria de probabilidades

Capítulo 9
Introdução

Foi no século XVII, com os chamados jogos de azar que surgiram os primeiros estudos de probabilidades. Grandes nomes da história da matemática são responsáveis pelo corpo de conhecimentos que constitui hoje a teoria das probabilidades:

1. Pascal (1623-1662),
2. Pierre de Fermat (1601-1665),
3. Huygens (1629-1695),
4. Isaac Newton (1642-1727),
5. Jacob Bernoulli (1654-1705),
6. Laplace (1749-1827),
7. Bayes (1702-1761),
8. Kolmogorov (1903-1987) entre outros.

Comecemos examinando as seguintes afirmações

1. É provável que João vá ao teatro amanhã
2. É provável que Adão e Eva tenham existido

Em ambas estão presentes as ideias de

1. Incerteza
2. Grau de confiança que depositamos naquilo que afirmamos

Note que a palavra provável também dá a ideia de futuro, mas na afirmação 2 estamos falando de algo que deve ter ocorrido no passado, se é que ocorreu. Isto porque na afirmação 2 a probabilidade não está ligada ao tempo, mas sim à eventual veracidade da própria afirmação.

Capítulo 10
Conceitos básicos

Antes de definirmos probabilidades vamos introduzir alguns conceitos básicos

10.1 Experimento aleatório

Definição 19 *É aquele que pode ser repetido nas mesmas condições indefinidamente sem que saibamos um resultado, de um evento de interesse, a priori, isto é, antes de sua realização, mas conhecemos todos os possíveis resultados.*

Notação ε

Exemplo 22 *A seguir alguns exemplos*

- *Lançamento de um dado.*
- *Tempo de duração de uma lâmpada.*
- *Número de veículos que passam por uma praça de pedágio durante um certo intervalo.*

10.2 Espaço amostral

Definição 20 *Conjunto de todos os possíveis resultados de um experimento aleatório.*

Notação Ω.

Exemplo 23 *A seguir alguns exemplos*

- *Lançamento de um dado* $\Omega = \{1, ..., 6\}$
- *Tempo de duração de uma lâmpada* $\Omega = (0, \infty)$
- *Número de veículos que passam por uma praça de pedágio durante um certo intervalo* $\Omega = \{0, 1, 2, ...\}$

10.3 Evento

Definição 21 *Subconjunto do espaço amostral*

Notação A, B, C, ...

Exemplo 24 *Lançamento de um dado* $\Omega = \{1, ..., 6\}$.

- *Evento A: Resultado é par* $A = \{2, 4, 6\}$ *(evento composto).*
- *Evento B: Resultado é maior do que 6* $B = \phi$ *(evento impossível).*
- *Evento C: Resultado menor do que 7* $C = \Omega$ *(evento certo).*
- *Evento D: Resultado igual a 1* $D = \{1\}$ *(evento simples).*

Parte IX

Teoria de conjuntos

[Teoria de conjuntos]

10.4 União

$A \cup B$ é quando A ou B ou ambos ocorrem.

10.5 Intersecção

$A \cap B$ é quando ocorrem A e B.

10.6 Eventos disjuntos ou mutuamente exclusivos

Quando dois eventos A e B não podem ocorrer simultaneamente, isto é, $A \cap B = \phi$

10.7 Evento complementar

A^c ou $\bar{A}$ é quando não ocorre A.

Exemplo 25 *Seja ε lançamento de um dado e $\Omega = \{1, ..., 6\}$. Seja $A = \{2, 4, 6\}$ e $B = \{1\}$. Determine $A \cup B$, $A \cap B$ e A^c.*

Parte X

Conceitos de Probabilidade

[Conceitos de Probabilidade]

Capítulo 11
Introdução

- Definição clássica,
- Definição frequentista e
- Definição axiomática.

11.1 Definição clássica de Probabilidade ou a priori

Seja ε um experimento aleatorio e Ω um espaço amostral associado formado por n resultados igualmente prováveis. Seja $A \subset \Omega$ um evento com m elementos. A probabilidade de A, denotada por $P(A)$, lê-se pe de A, é definida como sendo

$$P(A) = \frac{m}{n}.$$

Isto é, a probabilidade do evento A é o quociente entre o número de m casos favoráveis e o número n de casos possíveis.

Exemplo 26 *Calcular a probabilidade de no lançamento de um dado equilibrado obter-se*

- *Um resultado igual a 4.*
- *Um resultado ímpar.*

11.2 Definição frequentista de Probabilidade ou a posteriori

Definição 22 *Seja ε um experimento e A um evento de um espaço amostral associado ao experimento ε. Suponha-se que ε seja repetido n vezes e seja m o número de vezes que A ocorre nas n repetições de ε. Então, a frequência relativa do evento A, denotada por f_r, é o quociente*

$$f_r = \frac{m}{n} = \frac{\textit{número de vezes que A ocorre}}{\textit{número de vezes que } \varepsilon \textit{ é repetido}}$$

Exemplo 27 *A seguir dois exemplos,*

- *Uma moeda foi lançada 200 vezes e forneceu 102 caras. Então, a frequência relativa de caras é*

$$f_r = 102/200 = 0,51.$$

- *Um dado foi lançado 100 vezes e a face 6 apareceu 18 vezes. Então, a frequência relativa do evento $A =$ face 6 é*

$$f_r = 18/100 = 0,18.$$

Definição 23 *Seja ε um experimento e A um evento de um espaço amostral associado Ω.*

Suponhamos que ε é repetido n vezes e seja $f_r(A)$ a frequência relativa do evento. Então, a probabilidade de A é definida como sendo o limite de $f_r(A)$ quando n tende ao infinito. Ou seja

$$P(A) = \lim_{n \to \infty} f_r(A).$$

Deve-se notar que a frequência relativa do evento A é uma aproximação da probabilidade de A. As duas se igualam apenas no limite de infinitos experimentos. Em geral, para um valor de n, razoavelmente grande $f_r(A)$ é uma boa aproximação de $P(A)$.

11.3 Definição axiomática de Probabilidade

Definição 24 *Seja ε um experimento aleatório com um espaço amostral associado Ω. A cada evento $A \subset \Omega$ associa-se um número real, representado por $P(A)$ e denominado probabilidade de A, que satisfaz as seguintes propriedades (axiomas)*

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. *Se $A_1, A_2, ..., A_n$ forem, dois a dois, eventos disjuntos, então*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

Propriedades Como consequência dos axiomas estabelecidos acima, podemos ainda verificar outras propriedades das probabilidades de um evento

1. $P(\phi) = 0$.
2. Se A e A^c são eventos complementares, então

$$P(A^c) = 1 - P(A).$$

3. Se A e B são dois eventos quaisquer, então

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

4. Se A e B são eventos disjuntos, então

$$P(A \cup B) = P(A) + P(B).$$

Observação 2 *Quando estamos resolvendo um problema de probabilidade toda vez que for* ***ou*** *implica em soma e quando for* ***e*** *em produto.*

11.4 Probabilidade condicional e independência

Definição 25 *Sejam A e B dois eventos de um espaço amostral Ω, associado a um experimento ε, em que $P(A) > 0$. A probabilidade de B ocorrer condicionada a A ter ocorrido, será representada por $P(B\backslash A)$, e lida como probabilidade de B dado A ou probabilidade de B condicionada a A, e calculada por*

$$P(B\backslash A) = \frac{P(A \cap B)}{P(A)}.$$

Note que,

$$P(A\backslash B) = \frac{P(A \cap B)}{P(B)}.$$

Exemplo 28 *Suponha que se quer extrair duas peças ao acaso sem reposição de um lote que contém 100 peças das quais 80 peças são boas e 20 defeituosas. Defina-se os seguintes eventos*

- $A = \{$ *A primeira peça é defeituosa*$\}$ *e*
- $B = \{$ *A segunda peça é defeituosa*$\}$.

Determine $P(B\backslash A)$.

Tarefa

Exemplo 29 *Seleccionamos dois itens, ao acaso, um a um e sem reposição, de um lote que contém 10 intens do tipo A e 5 do tipo B. Qual é a probabilidade de que*

- *o primeiro item seja do tipo A?*
- *o segundo item seja do tipo B se o primeiro item foi do tipo A?*

11.5 Teorema da multiplicação (Regra do Produto)

Com o conceito de probabilidade condicional é possível apresentar uma maneira de se calcular a probabilidade da interseção de dois eventos A e B em função destes eventos. Esta expressão é denominada de teorema da multiplicação

$$P(AB) = P(A \cap B) = P(B \backslash A)P(A) = P(A \backslash B)P(B). \tag{11.1}$$

Exemplo 30 *Considere uma urna com 3 bolas brancas e 7 bolas verdes. Duas bolas são retiradas da urna, uma depois da outra sem reposição. Determine Ω e as probabilidades associadas com cada elemento do espaço amostral.*

Observação 3 *A regra do produto geralmente é util para encontrar probabilidades, quando a amostragem é sem reposição. A equação (??) pode ser generalizada à intersecção de "n eventos"$A_1, \ldots, A_n$ por medio das probabilidades condicionais sucessivas.*

Lema 1 *Sejam $A_1, \ldots, A_n$ eventos do espaço amostral Ω, então:*

$$P(A_1, \ldots, A_n) = P(A_1) \times P(A_2 \backslash A_1) \ \times \ P(A_3 \backslash A_1 A_2) \times \ldots \times P(A_n \backslash A_1 A_2 \ldots A_{n-1})$$

Tarefa

Exemplo 31 *Com base no exemplo anterior, calculemos a probabilidade do seguinte resultado* $B_1B_2V_3V_4B_5$ *em 5 retiradas de bolas de uma urna sem reposição.*

11.6 Eventos Independentes

Definição 26 *Dois eventos A e B são independentes se*

$$P(B \backslash A) = P(B).$$

Alguns comentários da definição anterior

- Se a probabilidade condicional é igual à probabilidade não condicional, então conhecer a ocorrência de A não muda a ocorrência de B.
- Essencialmente é equivalente a $P(A \cap B) = P(A)P(B)$, a regra multiplicativa para a probabilidade de uma intersecção se e somente se os eventos são independentes.
- Finalmente, tem uma relação com a falta de influência física dos eventos em cada um dos outros se não há influência na situação modelada, então assumimos independência no modelo.

Definição 27 *Dois eventos A e B são independentes se*

$$P(A \cap B) = P(A)P(B). \tag{11.2}$$

Alguns comentários da definição anterior

1. Essa definição tem vantagens sobre $P(B\backslash A) = P(B)$. Por um lado é simétrica e não atribui valores desiguais para A e B. Além disso, $P(B\backslash A) = P(B)$ não existe quando $P(A)$ é zero, enquanto que $P(A \cap B) = P(A)P(B)$ faz sentido para qualquer evento.
2. Pode ser verdadeiro ou falso, dependendo dos eventos, mas pode ser verificado, ainda se as probabilidades são zero ou não. Eventos independentes não é o mesmo que eventos disjuntos.
3. Se A e B são disjuntos, então $A \cap B$ é o conjunto vazio, cuja probabilidade é zero, enquanto que se são independentes então $A \cap B$ tem probabilidade igual ao produto $P(A)$ e $P(B)$.

Exemplo 32 *Lançam-se três moedas. Verificar se são independentes os eventos:*

1. *A: saída de cara na primeira moeda e*
2. *B: saída de coroa na segunda e terceira moedas.*

Tarefa

Exemplo 33 *Uma urna contem 6 bolas azuis e 4 bolas brancas. 2 bolas são extraídas, uma depois a outra. São os eventos A_1 :a primeira bola é azul e B_2: a segunda bola é branca independentes?.*

11.7 Independência de mais de dois eventos

Definição 28 *3 eventos A, B e C são independentes se satisfazem o seguinte:*

- $P(A \cap B) = P(A)\,P(B)$
- $P(A \cap C) = P(A)\,P(C)$
- $P(B \cap C) = P(B)\,P(C)$
- $P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$

Assim n eventos $A_1, \ldots, A_n$ são independentes se:

$$\begin{aligned} P(A_i \cap A_j) &= P(A_i)\,P(A_j) \qquad \forall\, i \neq j \\ P(A_i \cap A_j \cap A_k) &= P(A_i)\,P(A_j)\,P(A_k) \qquad \forall\, i \neq j \neq k \\ &\vdots \\ P(A_1 \cap A_2 \cap \ldots \cap A_n) &= \prod_{i=1}^{n} P(A_i) \end{aligned}$$

Exemplo 34 *Jogamos um dado duas vezes. Se definimos os seguintes eventos:* A=*{o primeiro dado*

mostra um numero par}, B=*{o segundo dado mostra um numero ímpar}*, C=*{Ambos os dados mostram um numero par ou ímpar}. Os eventos* A, B *e* C *são independentes?*

Observação 4 *O teorema de Bayes proporciona uma regra para calcular a probabilidade condicional de cada evento* A_i *dado* B *a partir das probabilidades condicionais de* B *dado cada um dos eventos* A_i *e a probabilidade não condicional de cada* A_i*.*

Resposta do exercício anterior

- $A \cup B = \{1, 2, 4, 6\}$,
- $A \cap B = \phi$ e
- $A^c = \{1, 3, 5\}$.

Resposta do exercício anterior $\Omega = \{1, 2, 3, 4, 5, 6\}$

- Um resultado igual a 4, $A = \{4\}$, então $P(A) = 1/6$.
- Um resultado ímpar, $B = \{1, 3, 5\}$, então $P(B) = 3/6 = 1/2$.

Parte XI

Teorema de Bayes

Capítulo 12
Introdução

As duas formulas desta seção, a lei de probabilidade total e o teorema de Bayes, se aplicam quando Ω pode ser particionado em n eventos $A_1, A_2, A_3, \ldots A_n$, disjuntos cuja união é Ω.

12.1 Regra da Probabilidade Total

Se uma coleção de n eventos $A_1, A_2, A_3, \ldots, A_n$ formam uma partição de Ω, e se $P(A_i) > 0$, $i = 1, \ldots, n$, então para um evento B,

$$\begin{aligned} P(B) &= \sum_{i=1}^{n} P(B \cap A_i)\,, \qquad \text{Regra do Produto} \\ &= \sum_{i=1}^{n} P(B \backslash A_i)\, P(A_i). \end{aligned}$$

Exemplo 35 *3 urnas contêm bolas azuis e bolas brancas. A urna um contem 1 bola azul e 3 brancas, a urna dois contem 3 bolas azuis e 7 brancas, e a urna 3 contem 80 bolas azules e 20 brancas. Uma urna é escolhida ao acaso (cada elecção tem a mesma probabilidade de ser seleccionada) e uma bola é escolhida desde a urna com igual probabilidade. Qual a probabilidade de que a bola seja azul?*

12.2 Teorema de Bayes

Se uma coleção finita de eventos $A_1, A_2, A_3, \ldots, A_n$ forma uma partição de Ω, e se $P(A_i) > 0$ $\forall i = 1, \ldots, n$, então para algum evento B e alguma partição A_i, então:

$$P(A_i \backslash B) = \frac{P(B \backslash A_i)P(A_i)}{\sum_{i=1}^{n} P(B \backslash A_i)P(A_i)} \tag{12.1}$$

em que,

- $P(A_i)$, é uma probabilidade a priori, isto é, antes realizar o experimento.
- $P(B \backslash A_i)$, é uma probabilidade condicional.
- $P(A_i \backslash B)$, é uma probabilidade a posteriori, isto é, quando o experimento já foi realizado.

Exemplo 36 *Suponha que um frabricante de sorvetes recebe 20% de todo o leite que utiliza de uma fazenda F_1, 30% de uma fazenda F_2 e 50% de uma fazenda F_3. Um órgão de fiscalização inspecionou as fazendas e observou que 20% do leite produzido por F_1 estava adulterado por adição de água, enquanto que para F_2 e F_3 essa proporção era de 5% e 2%, respectivamente. Na indústria de sorvetes os galões de leite são armazenados em um refrigerador sem identificação das fazendas. Para um galão escolhido ao acaso, qual é a probabilidade de que a amostra adulterada tenha sido obtida do leito fornecido pela fazenda F_1?*

12.3 Exercício adicional

Exemplo 37 *Seja o experimento aleatório lançar um dado e observar a face voltada para cima.*

1. *Determine* Ω
2. *Sejam os eventos A sair um número par, B sair um número ímpar, C sair o número 1 ou 3, D sair um número maior que 2. Determine* $A \cap B$, $A \cup B$, $A \cap C$, $A \cap D$, A^c.

Exemplo 38 *Um casal possui 2 filhos. Qual a probabilidade de ambos serem do sexo masculino?*

Exemplo 39 *Suponha que peças saiam de uma linha de produção e sejam classificadas como defituosas (D) ou como não defeituosas (D^c), isto é, perfeitas. Admita que três dessas peças, da produção de um dia, sejam escolhidas ao acaso e classificadas de acordo com esse esquema. Suponhamos que seja 0,2 a probabilidade de uma peça ser defeituosa e 0,8 a de ser não defeituosa. Admitimos que essas probabilidades sejam as mesmas para cada peça, ao menos enquanto durar o nosso estudo. Finalmente, admita-se que a classificação de qualquer peça em particular, sewja independente da classificação de qualquer outra peça. Empregando essas suposições, determine o espaço amostral para esse experimento e as probabilidades associadas a cada evento.*

Parte XII

Variáveis Aleatórias

Capítulo 13
Introdução

Na prática é, muitas vezes, mais interessante associarmos um número a um evento aleatório e

calcularmos a probabilidade da ocorrência desse número do que a probabilidade do evento.
Introduziremos a seguir o conceito de variáveis aleatórias.

13.1 Definição de variável aleatória

Definição 29 *Seja ε um experimento aleatório e Ω o espaço amostral associado com ε. Uma função X que associa a cada um dos elementos de $\omega \in \Omega$, um número real $X(\omega)$, se denomina variável aleatória. Isto, pode ser representado da seguinte forma*

$$\begin{aligned} X \ : \ & \Omega \to \mathbb{R} \\ & \omega \rightsquigarrow X(\omega) \end{aligned}$$

Exemplo 40 *Se lança uma moeda duas vezes e se define a variável aleatória X como o número de caras obtido nos dois lançamentos. Defina ε, Ω e os possíveis valores da variável aleatória X.*

Observação 5 *Uma variável aleatória pode ser classificada em*

1. *variável aleatória discreta ou*
2. *variável aleatória continua.*

13.2 Variável aleatória discreta

Definição 30 *Uma variável aleatória é discreta quando os possíveis valores da variável aleatória assumem valores em um conjunto enumerável.*

Exemplo 41 *A seguir alguns exemplos,*

- *número de sementes que germinam.*
- *número de chamadas telefônicas numa central da TIM em 30 minutos.*
- *número de acidentes na rua XV de novembro.*
- *número de mulheres na ESALQ.*

13.3 Variável aleatória continua

Definição 31 *Uma variável aleatória é continua quando os possíveis valores da variável aleatória não assumem valores em um conjunto enumerável.*

Exemplo 42 *A seguir alguns exemplos,*

- *rendimento de milho (kg/ha),*
- *diâmetro de uma árvore,*
- *ângulo entre o norte e a direção tomada por um pássaro no sentido horário,*
- *altura de plantas.*

Teorema 1 *O caso mais simples de variável aleatória é a função indicadora que definimos a seguir. Seja $A \subset \Omega$. Então, a função indicadora de A, I_A é definida por*

$$I_A(\omega) = \begin{cases} 1 & \text{se } \omega \in A; \\ 0 & \text{se } \omega \in A^c. \end{cases}$$

Exemplo 43 *A seguir alguns exemplos,*

- *para uma variável aleatória discreta*

$$I_{\{0,1,2,3\}}(x) = \begin{cases} 1 & \textit{se } x \in \{0,1,2,3\}; \\ 0 & \textit{se } x \notin \{0,1,2,3\}. \end{cases}$$

- *para uma variável aleatória continua*

$$I_{\mathbb{R}^+}(x) = \begin{cases} 1 & \textit{se } x \in \mathbb{R}^+; \\ 0 & \textit{se } x \notin \mathbb{R}^+. \end{cases}$$

13.4 Função de probabilidades

Definição 32 *Uma função $P(X = x)$ de uma variável aleatória discreta se denomina função de probabilidades se satisfaz as seguintes duas condições*

$$\begin{aligned} P(X = x) &\geq 0 \qquad x \in R_x \qquad \text{e} \\ \sum_{x \in R_x} P(X = x) &= 1, \end{aligned}$$

em que, R_x denota os possíveis valores da variável aleatória X. A distribuição de probabilidades de X é o conjunto de pares ordenados $(x_i, P(X = x_i))$, em que x_i representa os diferentes valores da variável aleatória X e $P(X = x_i)$ a probabilidade de ocorrência de x_i.

Exemplo 44 *Seja X uma variável aleatória com função de probabilidades*

$$P(X = x) = \frac{1}{6} \qquad \text{para } x = 1, 2, 3, 4, 5, 6$$

Determine se $P(X = x)$ é uma função de probabilidades.

13.5 Função densidade de probabilidades

Definição 33 *Uma função $f(x)$ de uma variável aleatória continua se denomina função densidade de probabilidades se satisfaz as seguintes duas condições:*

$$\begin{aligned} f(x) &\geq 0 \qquad x \in R_x \\ \int_{x \in R_x} f(x)dx &= 1, \end{aligned}$$

em que, R_x denota os possíveis valores da variável aleatória X.

Exemplo 45 *Se X é uma variável aleatória continua com função*

$$f(x) = 1 \qquad \textit{para}\, x \in [0, 1].$$

$f(x)$ é uma função densidade de probabilidades?

Tarefa

Exemplo 46 *Seja X uma variável aleatória continua*

$$f(x) = \frac{1}{b-a} \qquad \textit{para}\, x \in [a, b].$$

$f(x)$ é uma função densidade de probabilidades?

Exemplo 47 *Seja X uma variável aleatória continua*

$$f(x) = \lambda e^{-\lambda x} \qquad \textit{para}\, x \in (0, \infty), \qquad \lambda > 0$$

$f(x)$ *é uma função densidade de probabilidades?*

Exemplo 48 *Seja X uma variável aleatória continua*

$$f(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda} x} \qquad \textit{para}\, x \in (0, \infty), \qquad \lambda > 0$$

$f(x)$ *é uma função densidade de probabilidades?*

13.6 Função de distribuição acumulada

Definição 34 *Dada a variável aleatória X, chamaremos de função de distribuição acumulada a função* $F(x)$ *definida por:*

$$\begin{aligned} F \ : \ & \mathbb{R} \to [0,1] \\ & x \rightsquigarrow F(x) = P(X \leq x) \end{aligned}$$

13.7 Função de distribuição acumulada para uma variável aleatória discreta

Definição 35 *Seja uma variável aleatória discreta X, então a função de distribuição acumulada se defino como*

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X \leq x_i).$$

Exemplo 49 *Seja X uma variável aleatória discreta com função de probabilidades dada por*

$$P(X = x) = \frac{3!}{(3-x)!x!}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{3-x} I_{\{0,1,2,3\}}(x)$$

Determine e faça o gráfico de $F(x)$.

13.8 Função de distribuição acumulada para uma variável aleatória continua

Definição 36 *Seja uma variável aleatória continua X, então a função de distribuição acumulada se defino como*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Exemplo 50 *Seja X uma variável aleatória continua com função densidade de probabilidades dada por*

$$f(x) = e^{-x} \qquad \text{para } x \in (0, \infty).$$

Determine $F(x)$.

Relação entre $f(x)$ e $F(x)$ para uma variável aleatória continua

Seja $f(x)$ uma função densidade de probabilidades, isto é, uma função não negativa que integra 1. Qual é a relação entre $F(x)$ e $f(x)$?

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, d\, t. \tag{13.1}$$

Note da equação (??) que com base no teorema teorema fundamental do cálculo integral

$$f(x) = \frac{d\, F(x)}{d\, x}.$$

Observação 6 *Para uma variável aleatória continua*

$$\begin{aligned} P(X = x) &= 0 \quad x \in \mathbb{R} \\ P(a < X < b) &= P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) \\ &= \int_{a}^{b} f(x)\, d\, x = F(b) - F(a). \end{aligned}$$

13.9 Esperança de uma variável aleatória discreta

Definição 37 *A esperança de uma variável aleatória discreta X, é definida por*

$$E(X) = \sum_{x \in R_x} x\, P(X = x).$$

Exemplo 51 *Determine $E(X)$ para a seguinte variável aleatória discreta*

$$P(X = x) = p^x\,(1-p)^{1-x} \qquad \textit{para}\; x \in \{0, 1\}.$$

13.10 Esperança de uma variável aleatória continua

Definição 38 *A esperança de uma variável aleatória continua X, é definida por*

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, d\, x.$$

Exemplo 52 *Determine $E(X)$ para a seguinte variável aleatória continua*

$$f(x) = \frac{1}{b-a} \qquad \textit{para}\; x \in [a, b],$$

Tarefa

Exemplo 53 *Determine* $E(I_A(x))$, *em que*

$$I_A(x) = \begin{cases} 1, & \text{se } x \in A; \\ 0, & \text{se } x \notin A. \end{cases}$$

Exemplo 54 *Determine* $E(X)$ *para a seguinte variável aleatória continua*

$$f(x) = \frac{1}{\lambda} e^{-x/\lambda} \qquad \text{para } x \in (0, \infty).$$

Exemplo 55 *Determine* $E(X)$ *para a seguinte variável aleatória continua*

$$f(x) = \lambda e^{-\lambda x} \qquad \text{para } x \in (0, \infty).$$

Propriedades da esperança

Sejam X e Y duas variáveis aleatórias, a, $b \in \mathbb{R}$ (constantes), então

1. $E(a) = a$.
2. $E(aX \pm bY) = aE(X) \pm bE(Y)$.
3. $E(aX) = aE(X)$.
4. $E(aX \pm b) = aE(X) \pm b$.
5. $E[(X - a)^2] = E(X^2) - 2aE(X) + a^2$.
6. $E(XY) = E(X)E(Y)$, se X e Y são variáveis aleatórias independentes.

Exemplo 56 *Seja X uma variável aleatória discreta com*

$$P(X = x) = p^x (1 - p)^{1-x} \qquad \textit{para } x \in \{0, 1\}.$$

Determine $E(2X + 1)$.

13.11 Variância para uma variável aleatória

Definição 39 *Seja X uma variável aleatória e* $\mu = E(X)$. *A variância de X é definida por*

$$\begin{aligned} V(X) &= E(X-\mu)^2 \\ &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2, \qquad \text{usando propriedades de esperança} \\ &= E(X^2) - 2\mu\mu + \mu^2, \qquad \mu = E(X) \\ &= E(X^2) - \mu^2, \qquad \mu = E(X) \\ &= E(X^2) - E(X)^2. \end{aligned}$$

Geralmente usamos a seguinte definição de variância

$$V(X) = E(X^2) - E(X)^2.$$

Note que $V(X) = E(X-\mu)^2 \geq 0$.

13.12 Variância para uma variável aleatória discreta

Definição 40 *A variância para uma variável aleatória discreta é dada por*

$$V(X) = \sum_{x \in R_x} x^2 \, P(X = x) - \left[\sum_{x \in R_x} x \, P(X = x) \right]^2 .$$

Exemplo 57 *Sejam X_1 e X_2 duas variáveis aleatórias. Com base na seguinte tabela calcule $V(X_1)$ e $V(X_2)$ e faça alguns comentários.*

x	1	2	3	4	5
$P(X_1 = x_1)$	0.1	0.2	0.4	0.2	0.1
$P(X_2 = x_2)$	0.3	0.1	0.2	0.1	0.3

13.13 Variância para uma variável aleatória continua

Definição 41 *A variância para uma variável aleatória continua é dada por*

$$V(X) = \int_{x \in R_x} x^2 \, f(x) \, d\, x - \left[\int_{x \in R_x} x \, f(x) \, d\, x \right]^2 .$$

Exemplo 58 *Determine* $V(X)$*, com base em*

$$f(x) = 1 \qquad \textit{para } x \in [0, 1].$$

Tarefa

Exemplo 59 *Determine* $V(X)$*, com base em*

$$f(x) = \frac{1}{b-a} \qquad \textit{para}\, x \in [a, b].$$

Exemplo 60 *Determine* $V(X)$*, com base em*

$$f(x) = \lambda\, e^{-\lambda\, x} \qquad \textit{para}\, x \in (0, \infty) \qquad \lambda > 0.$$

Propriedades da variância

Sejam X e Y variáveis aleatórias, a e b constantes, então

1. $V(aX + b)$
2. $V(a) = 0$
3. $V(aX) = a^2V(X)$
4. $V(-X) = V(X)$
5. $V(X \pm Y) = V(X) \pm V(Y)$, se X e Y são variáveis aleatórias independentes.

Tarefa

Exemplo 61 *Seja X uma variável aleatória discreta com*

$$P(X = x) = p^x \, (1 - p)^{1-x} \qquad \textit{para } x \in \{0, 1\}.$$

Determine $V(2X + 1)$.

Parte XIII

Modelos probabilísticos discretos

Capítulo 14
Distriuição Binomial e Poisson

A seguir estudamos três modelos probabilísticos discretos

- Modelo Bernoulli
- Modelo Binomial
- Modelo Poisson

14.1 Distribuição Bernoulli

Se um experimento possui dois possíveis resultados, sucesso e fracasso. Seja p a probabilidade de sucesso e $1-p$ a probabilidade de fracasso. A variável aleatória Bernoulli denota o número de sucessos em uma única tentativa do experimento aleatório, assim $R_x = \{0, 1\}$. A função de probabilidades está dada por

$$P(X = x) = p^x(1-p)^{1-x} \qquad \text{para } x \in \{0, 1\}, \qquad p \in (0, 1). \tag{14.1}$$

Notação $X \sim Ber(p)$.

Tarefa

Observação 7 *A esperança e variância de uma variável aleatória* $X \sim Ber(p)$ *são, respectivamente*

$$E(X) = p \qquad \text{e} \qquad V(X) = p\,(1-p).$$

14.2 Distribuição Binomial

Uma variável aleatória X que conta o número total de sucessos em n ensaios (tentativas) independentes de Bernoulli de um mesmo experimento aleatório é uma variável aleatória Binomial com parâmetros n e p, em que p denota a probabilidade constante de sucesso em cada ensaio Bernoulli, assim $R_x = \{0, 1, \ldots, n\}$. A função de probabilidades de X é dada por

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \qquad \text{para } x \in \{0, 1, 2, \ldots, n\}. \qquad p \in (0, 1) \qquad (14.2)$$

Notação $X \sim Bin(n, p)$.

Observação 8 *A esperança e variância de uma variável aleatória $X \sim Bin(n, p)$ são, respectivamente*

$$E(X) = n\,p \qquad \text{e} \qquad V(X) = n\,p\,(1-p)$$

Exemplos de distribuição Binomial

Exemplo 62 *A probabilidade de que um paciente se recupere de uma doença rara do sangue é 0.4. Sabemos que 15 pessoas tem a doença.*

1. *Qual é a probabilidade de que pelo menos 10 pessoas sobrevivam?*
2. *Qual é a probabilidade de que sobrevivam entre 3 e 8 pessoas?*
3. *Qual é a probabilidade de que sobrevivam exatamente 5 pessoas?*
4. *Calcular* $E(X)$.
5. *Calcular* $V(X)$.

Exemplo 63 *Numa criação de coelhos, 40*$\%$ *são machos. Qual a probabilidade de que nasçam pelo menos 2 coelhos machos num dia em que nasceram 20 coelhos?*

14.3 Distribuição de Poisson

Consideremos a probabilidade de ocorrência de sucessos em um determinado intervalo ou uma região específica, assim $R_x = \{0, 1, 2, ...\}$. A função de probabilidades de X é dada por

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{para} \in R_x = \{0, 1, 2 \ldots\} \qquad \lambda > 0. \tag{14.3}$$

Notação $X \sim P(\lambda)$.

Observação 9 *Podemos provar que se $X \sim P(\lambda)$, então*

$$E(X) = \lambda \qquad e \qquad V(X) = \lambda$$

Exemplos de distribuição Poisson

1. Número de carros que passam por um cruzamento por minuto, durante uma certa hora do dia
2. número de erros tipográficos por página, em um material impresso.
3. número de colônias de bactérias numa dada cultura por 0,01 mm^2, numa plaqueta de microscópio.
4. número de mortes por ataque de coração por ano, numa cidade.

Exemplos de distribuição Poisson

Exemplo 64 *O número médio de partículas radioativas que pasam por um contador durante um milisegundo num experimento de laboratório é 4. Qual a probabilidade de que entrem 6 partículas ao contador num milisegundo determinado?*

Exemplo 65 *Num livro de 800 páginas há 800 erros de impressão. Qual a probabilidade de que uma página contenha pelo menos 3 erros?*

Exemplo 66 *Numa central telefônica chegam 300 telefonemas por hora. Qual a probabilidade de que*

1. *num minuto não haja nenhum chamado?*
2. *em 2 minutos haja 2 chamados?*
3. *em t minutos não haja chamados?*

Teorema 2 *Se* $X \sim B(n,p)$ *e supondo* n *grande* $(n \to \infty)$ *e* p *pequeno (p* $\to$ *0), então* $\lambda = np$*, isto*

é,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!} \quad \textit{isto é,} \tag{14.4}$$

$$\lim_{p \to 0\, n \to \infty} P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{14.5}$$

Este teorema essencialmente diz que podemos aproximar a distribuição Binomial pela distribuição Poisson sempre que n seja grande e p pequeno.

Exemplo 67 *Uma companhia de seguros afirma que 0.1% da população tem certo tipo de acidentes cada ano. Se os 10000 segurados da companhia foram selecionados aleatoriamente desde a população. Qual será a probabilidade de que no máximo de 5 de estos clientes, tenham um acidente o proximo ano?*

Solução do exercício anterior

A : Pessoa segurada pela companhia sofre um acidente.

$X \sim B(10000, 0.001)$, logo

$$P(X \leq 5) = \sum_{x=0}^{5} \binom{10000}{x} (0.001)^x \, (0.999)^{(10000-x)}$$

Como n é grande e p é pequeno, calcularemos esta probabilidade usando a aproximação da distribuição Binomial pela distribuição Poisson, isto é, $\lambda = 10000 \times 0.001 = 10$. Por tanto,

$$P(X \leq 5) = \sum_{x=0}^{5} \frac{e^{-\lambda}\lambda^x}{x!} = 0.0671 \qquad \text{Conferir!!!!.}$$

Exercícios adicionais

1. Considere a variável aleatória número de vezes que ocorre face cara em 5 lançamentos de uma moeda equilibrada, determine a probabilidade de ocorrer
   - Duas caras.
   - Quatro caras.
   - No máximo duas caras.
2. Num rebanho bovino 30% dos animais estão atacados de febre aftosa. Retira-se ao acaso, uma amostra de 10 animais.
   - Verifique se a variável *número de animais doentes* pode ser estudada pelo modelo binomial. Justifique.
   - Qual a probabilidade de se encontrar 6 animais doentes?.
   - Qual a probabilidade de se encontrar pelo menos 4 animais doentes?.
3. Numa criação de coelhos, a taxa de nascimento de machos é de 40%. Qual a probabilidade de que nasçam pelo menos 2 coelhos machos, num dia em que nasceram 19 coelhos? (Mostrar gráfico no R).
4. Num certo ano, o IBAMA registrou no litoral catarinense (área de reserva), 18 mortes de golfinhos.

- Qual é a probabilidade de, num determinado mês do próximo ano, ocorrerem menos de 2 mortes?
- Qual é a probabilidade de ocorrerem 2 mortes no próximo semestre?

5. Em um certo tipo de fabricação de fita magnética, ocorrem defeitos a uma taxa de 1 a cada 2000 metros. Qual a probabilidade de que um rolo com 2000 metros de fita magnética
   - não tenha defeitos?
   - tenha no máximo dois defeitos?
   - tenha pelo menos dois defeitos?

Parte XIV

Modelos probabilísticos contínuos

Capítulo 15
Introdução

O modelo normal ocupa uma posição de grande destaque tanto a nível teórico como prático, isso porque o modelo normal representa com boa aproximação muitos fenômenos da natureza como, por exemplo, a característica altura de plantas de Amaranthus, cuja distribuição de frequência é dada na figura ??. Observe que existe uma tendência das observações se concentrarem próximo do valor central, ou seja, da média da distribuição, e esta concentração vai diminuindo a medida que os valores de altura vão aumentando e diminuindo, ou seja, existe baixa concentração de plantas baixas, assim como de plantas altas. A distribuição é aproximadamente simétrica, isto é, tomando a média como ponto central, a lado esquerdo é aproximadamente igual ao lado direito.

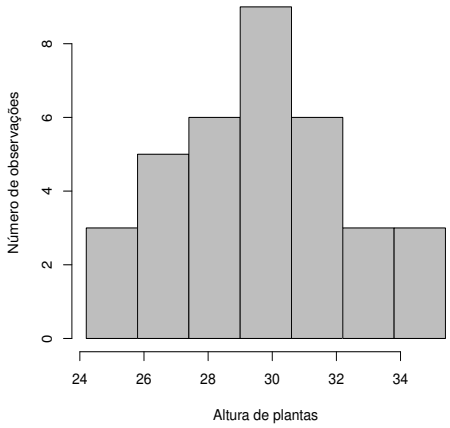

Figura 15.1: Distribuição de frequência da altura de plantas de Amaranthus (cm)

15.1 Distribuição Normal

Definição 42 *Dizemos que uma variável aleatória contínua X tem distribuição normal, com parâmetros μ e σ^2, em que $\mu \in (-\infty, +\infty)$ e $\sigma^2 \in (0, +\infty)$, representam a média e a variância da população X , respectivamente, se a sua função densidade de probabilidade for dada por:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad x \in (-\infty, +\infty)$$

em que $\exp$ *representa a base dos logaritmos naturais e vale aproximadamente* $2,7182$, $\pi = 3,1416$ *e* σ *é o desvio padrão. Notação* $X \sim N(\mu, \sigma^2)$.

Pode-se demonstrar que:

- $E(X) = \mu$
- $V(X) = \sigma^2$
- $f(x)$ *é simetrica ao redor de* $x = \mu$,

Gráficos da distribuição normal

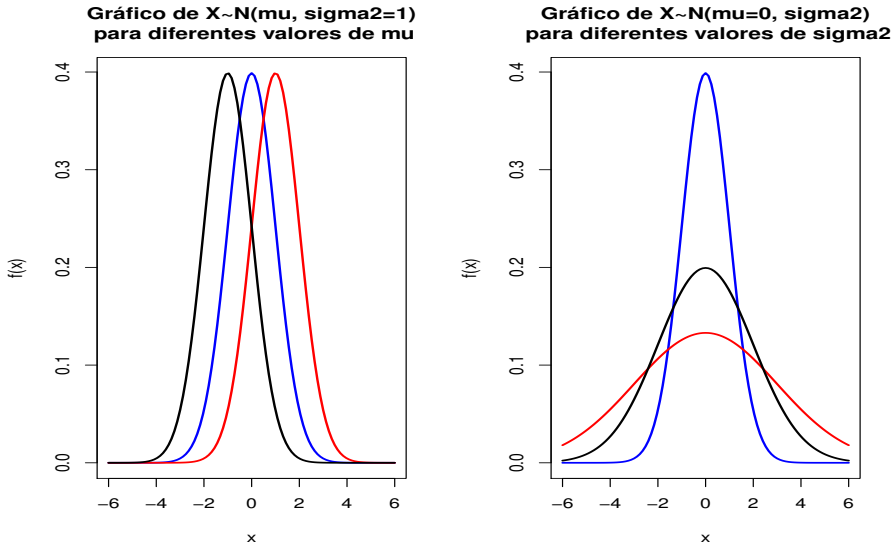

Figura 15.2: Gráficos da distribuição $\mathrm{N}(\mu, \sigma^2)$

Código R

```
#Média diferente igual variância
par(mfrow=c(1,2))
curve(dnorm(x,0,1),-6,6,lwd=2,col="blue",ylab="f(x)",
      main="Gráfico de X~N(mu, sigma2=1)\n para diferentes valores de mu")
curve(dnorm(x,1,1),-6,6,lwd=2,col="red", add=T)
curve(dnorm(x,-1,1),-6,6,lwd=2,col="black", add=T)

#média igual diferente variância
curve(dnorm(x,0,1),-6,6,lwd=2,col="blue",ylab="f(x)",
      main="Gráfico de X~N(mu=0, sigma2)\n para diferentes valores de sigma2")
curve(dnorm(x,0,3),-6,6,lwd=2,col="red", add=T)
curve(dnorm(x,0,2),-6,6,lwd=2,col="black", add=T)
```

15.2 Cálculos de probabilidades

A probabilidade de uma variável aleatória com distribuição normal tomar um valor entre dois pontos quaisquer, por exemplo, entre os pontos a e b é igual a área sob a curva normal compreendida entre aqueles dois pontos. Veja a figura 5.11. Suponha, então, que $X \sim N(\mu, \sigma^2)$ e queiramos determinar a probabilidade de X estar entre a e b, portanto, como estamos interessados em obter uma área, devemos realizar o seguinte cálculo:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

Acontece que essa integral não pode ser calculada exatamente, consequentemente, a probabilidade só pode ser obtida aproximadamente, e por métodos numéricos. Podemos obter estas probabilidades com o uso de programas computacionais estatísticos, entre os quais podemos citar o Statistica, Minitab, SAS e R.

Exemplo: Cálculos de probabilidades

Se $X \sim N(0,1)$, calcule $P(-3 \leq X \leq -1)$. Temos que,

Exemplo 68

$$P(-3 \leq X \leq -1) = \int_{-3}^{-1} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx$$

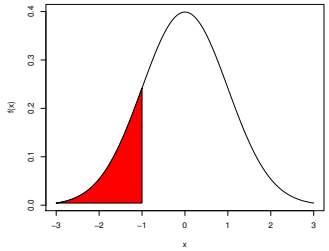

Figura 15.3: Cálculo de $P(-3 \leq X \leq -1)$, $X \sim N(0,1)$

15.3 A distribuição normal padrão

Definição 43 *Se $X \sim N(\mu, \sigma^2)$, então a variável aleatória Z definida por:*

$$Z = \frac{X - \mu}{\sigma}$$

tem uma distribuição $N(0, 1)$, isto é, tem distribuição normal com média $\mu = 0$ e variância $\sigma^2 = 1$, cuja função densidade de probabilidade é dada por:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}.$$

15.4 O uso da tabela da distribuição normal padrão

Exemplo 69 *Calcule as seguintes probabilidades, supondo que* $Z \sim N(0,1)$

- $P(Z \leq 2,10)$
- $P(Z \geq 2,10)$
- $P(Z \geq -2,10)$
- $P(Z \leq -2,10)$
- $P(-2,10 \leq Z \leq 2,10)$

Exemplo 70 *Calcule as seguintes probabilidades, supondo que* $X \sim N(3,16)$

- $P(X \leq 2)$
- $P(X \geq 5)$
- $P(X \geq -5)$
- $P(X \leq -2)$
- $P(2 \leq X \leq 5)$

15.5 Aplicação

Estudos meteorológicos indicam que a precipitação pluviométrica mensal em períodos de seca numa certa região pode ser considerada como seguindo a distribuição normal de média $30mm$ e variância $16mm^2$.

- Em um mês de seca qual a probabilidade de que chova mais de 34mm?
- Em um mês de seca qual a probabilidade de que chova menos de 42mm?
- Em um mês de seca qual a probabilidade de que chova entre 34mm e 42mm?

Parte XV

Inferência estatística

Capítulo 16
Conceitos Básicos

Agora, vamos ver como reunir a Análise Exploratória de Dados, Modelos Probabilísticos e

Amostragem, para podermos desenvolver um estudo importantíssimo dentro da estatística, conhecido pelo nome de Inferência Estatística, isto é, como tirar conclusões sobre parâmetros da *população* (por exemplo, sobre médias (μ), proporções (p), variâncias (σ^2)) com base no estudo de somente uma parte da população, ou seja, com base em uma *amostra*.

16.1 População

Definição 44 *Uma população, em estatística, é formada por todos os valores possíveis de uma característica desejável. Esses valores não precisam ser todos diferentes, nem um número finito.*

Exemplo 71 *Exemplos de populações*

1. *todos os valores possíveis da produção de milho em quilogramas por hectare (kg/ha);*
2. *todos os pesos ao nascer de coelhos da raça gigante, em gramas;*
3. *todos os valores de diâmetros de Biomphalarias do Poção do Córrego Grande;*
4. *todos os valores de micronúcleos de roedores de uma região poluída.*

16.2 Amostra

Definição 45 *Uma amostra, é uma parte (subconjunto) da população*

Exemplo 72 *Exemplos de amostras*

1. *os rendimentos de milho, em kg/ha, de uma amostra de 5 unidades experimentais (canteiros);*
2. *os pesos ao nascer de uma ninhada de coelhos da raça gigante;*
3. *os diâmetros de uma amostra de 30 Biomphalarias do Poção do Córrego Grande;*
4. *os valores de micronúcleos de uma amostra de 25 roedores.*

16.3 Estatística

Definição 46 *Uma estatística é uma medida usada para descrever uma característica da amostra.*

Exemplo 73 *Exemplos de estatísticas são*

1. $\bar{X}$ *a média da amostra;*
2. S *o desvio padrão da amostra e*
3. P *a proporção da amostra.*

16.4 Parâmetros

Definição 47 *Um parâmetro é uma medida usada para descrever uma característica da população.*

Geralmente são representados por letras gregas, assim, por exemplo, μ representa a média populacional; π representa a proporção populacional e σ representa o desvio padrão populacional.

16.5 Estimativa

Definição 48 *Quando uma estatística assume um determinado valor, temos o que denomina-se de estimativa. Temos os dados de uma particular amostra, calculamos o valor da estatística de interesse, este valor é a nossa estimativa.*

Exemplo 74 *Alguns exemplos de estimtiva são*

1. *a estimativa da produção média por planta da cultivar Gala é de* $\bar{x} = 84$ *kg/planta.*
2. *a estimativa da proporção de peixes com comprimento total menor do que 50 mm é* $p = 46\%$.

Observação 10 *Os dois problemas básicos da inferência estatística são:*

1. *Estimação e*
2. *Testes de Hipóteses.*

Vamos, através de um exemplo, ilustrar estas duas situações.

Exemplo de problema de estimação

Exemplo 75 *Um pesquisador está interessado em avaliar a produção média por planta, μ, da cultivar de maçã denominada Gala, para as seguintes condições: plantas com idade de aproximadamente 5 anos, em bom estado fitossanitário, cultivadas com alta tecnologia e para a região I do zoneamento agroclimático de Santa Catarina. A população é formada por todas as plantas da cultivar Gala nas condições citadas. Mais especificamente, a população é constituída por todos os valores de produção por planta. Para essa finalidade, o pesquisador vai coletar uma amostra aleatória de, por exemplo, 10 plantas, da referida cultivar nas condições descritas. Uma amostra de valores de produções por planta, em kg, foi:*

Plantas	1	2	3	4	5	6	7	8	9	10	$\bar{x}$	s
Produção	84	82	90	86	80	91	85	79	81	82	84	4,0552

Com os 10 valores de produção/planta podemos calcular uma estimativa da produção média verdadeira por planta, $\bar{x} = 84$ kg. Portanto, estamos usando a média da amostra, $\bar{X}$, como estimador da média verdadeira, μ. Essa estimativa é chamada de estimativa pontual, pois origina um único valor. Esse é um raciocínio tipicamente indutivo, onde se parte do particular (amostra) para o geral (população).

Observação 11 *Um fato importante que se observa quando trabalhamos com amostras, é que sempre*

vamos ter que a média verdadeira, μ é igual a média na amostra $\bar{X}$ mais um erro de amostragem. A representação disso é dada por:

$$\mu = \bar{X} + \textit{erro amostral},$$

em que o termo erro amostral é a diferença entre a estatística ($\bar{X}$) e o parâmetro (μ).

16.6 Precisão e confiança

Apesar do nome *erro*, isto não quer dizer que a amostragem foi feita de forma errada e, que, portanto, deve-se coletar uma nova amostra. Esse valor pode ser negativo ou positivo, pequeno, nulo ou grande. Em todas as pesquisas vamos estar envolvidos com o erro amostral. Dizemos que uma estimativa é *precisa*, se tivermos alto grau de *confiança* de que o erro amostral associado a estimativa em questão, é pequeno. A precisão e a confiança são dois conceitos chaves nesse estudo. A precisão pode ser entendida como a diferença máxima entre a estimativa e o parâmetro que o pesquisador deseja considerar no seu estudo. Voltaremos a tratar deste assunto posteriormente.

Ideia de intervalo de confiança

Uma outra forma de estimação é através da construção de intervalos de confiança. Nesse caso, temos uma estimativa intervalar, isto é, temos um intervalo, dentro do qual esperamos que o valor populacional se encontre. Por exemplo, para os dados de produção/planta da cultivar Gala ao invés de dizer que a estimativa é de 84 kg/planta, podemos dizer que a média está no intervalo de $81,10$ a $86,90$.

Observação 12 *Essa forma de estimação é muito mais informativa que a estimativa pontual. O pesquisador pode verificar se esse intervalo é curto (preciso, informativo) ou se é muito amplo (pouco informativo).*

Ideia sobre teste de hipóteses

O segundo problema é o de teste de hipóteses sobre os parâmetros. Por exemplo, um pesquisador deseja saber se a produção média/planta da cultivar Gala é a mesma da produção média/planta da cultivar Golden. Para isso, foi obtida uma outra amostra aleatória de 10 plantas da cultivar Golden sob as mesmas condições descritas para a cultivar Gala. Os dados das duas amostras aleatórias são apresentadas na tabela a seguir.

Tabela 16.1: Produção por planta, em Kg, de maçãs das cultivares Gala e Golden

Variedades	1	2	3	4	5	6	7	8	9	10	$\bar{x}$	s
Gala	84	82	90	86	80	91	85	79	81	82	84,0	4,06
Golden	95	102	85	93	104	89	98	99	107	106	97,8	7,32

As estimativas da produção média das duas cultivares, calculadas com os dados das duas amostras foram 84 Kg/planta e 97,8 kg/planta para as cultivares Gala e Golden, respectivamente. Portanto, a diferença verificada entre as duas cultivares, com essas duas amostras, foi de 13,8 kg/planta a favor da cultivar Golden.

Observando-se os dados individualmente, verificamos que para as plantas 3 e 6, as produções na cultivar Gala foram superiores a da Golden. Portanto, podemos pensar que é perfeitamente possível obtermos um par de amostras, dentre todas as amostras possíveis de serem sorteadas, no qual a produção média da cultivar Gala é superior a da Golden. Isso devido simplesmente a amostragem, ou seja, variações

devido a amostragem. Assim, o problema que se apresenta, é o de decidir o que é uma diferença real, isto é, devido à cultivar, ou uma diferença casual, isto é, devido a variação casual na amostra.

Logicamente, o pesquisador pretende generalizar os resultados obtidos na análise estatística, isto é, ele deseja saber se há diferença significativa entre as médias verdadeiras μ_{Gala} e μ_{Golden} (desconhecidas pelo pesquisador). Como ele está trabalhando com duas amostras aleatórias, dentre um grande número de possíveis amostras, ele não pode fazer afirmações com 100% de certeza, mas ele pode perfeitamente fazer uma afirmação probabilística, indicando a probabilidade de erro ao fazer uma afirmação sobre uma hipótese em teste. Para isso, utilizaremos as distribuições de probabilidades.

16.7 Amostra aleatória

Definição 49 *Uma amostra aleatória simples de tamanho n, de uma variável aleatória X , é aquela cujas n observações* $X_1, X_2, ..., X_n$ *são independentes e identicamente distribuídas.*

16.8 Distribuições Amostrais (Tarefa)

Observação 13 *Ler seção 6.3 (distribuições amostrais) do livro Estatística para as ciências agrárias e biológicas com noções de experimentação do Dalton F. Andrade e Paulo J. Ogliari.*

Parte XVI

Intervalo de confiança

Capítulo 17
Introdução

Estimação é o nome técnico para o processo que consiste em utilizar os dados de uma amostra para

avaliar parâmetros populacionais desconhecidos, ou, como o próprio nome indica, estimar os mesmos. Dentre as diversas características (parâmetros) de uma população que podem ser estimadas, vamos estudar as mais utilizadas, isto é,

1. a média μ,
2. a proporção π e
3. a variância σ^2.

Exemplo

Um pesquisador sempre está desenvolvendo um processo de estimação. Por exemplo, um Biólogo pode estar interessado na proporção de micronúcleos em 5000 células sanguíneas em peixes do genêro bagre; um Agrônomo pode estar interessado na produção média de uma cultura. Outros exemplos, os prejuízos causados pelo ataque de uma praga ou doença; o diâmetro de caramujos; o tamanho de Lulas encontradas no trato digestivo de Atuns; tamanho de crustáceos da classe Malacostraca e sub-classe Eumalacostraca, popularmente conhecida com o nome de Caprelas; parâmetros estatísticos genéticos (variância genética, ambiental, fenotípica).

Parte XVII

Conceitos Básicos

17.1 Estimador

Definição 50 *Um estimador é uma estatística que será usado para a estimação de um parâmetro populacional. Os estimadores mais frequentes são a média, a proporção e a variância amostral, representados por:* $\bar{X}$, P *e* S^2, *respectivamente, utilizados para estimar os parâmetros* μ, π *e* σ^2, *respectivamente.*

17.2 Métodos para encontrar estimadores

Os três métodos mais utilizados para encontrar estimadores (não serão estudados neste curso) são:

1. método da máxima verossimilhança,
2. método dos momentos e
3. método dos mínimos quadrados.

17.3 Estimativas Pontuais e Intervalares

De modo geral, vamos supor que os valores da população se distribuem segundo um dado modelo probabilístico, cujos parâmetros são desconhecidos e, portanto, precisam ser estimados. Lembramos que os estimadores possuem as suas correspondentes distribuições amostrais.

Na estimação por ponto, procede-se a estimação do parâmetro através de um único valor. A obtenção dos estimadores $\bar{X}$, P e S é feita de forma direta, aplicando as definições de média aritmética, proporção e desvio padrão aos dados da amostra, tomando-se o cuidado de que para o cálculo do desvio padrão usa-se $n-1$ no denominador.

Assim, uma estimativa pontual da média populacional μ é a média aritmética da amostra, $\bar{x}$. Uma estimativa da proporção populacional, π é obtida através do cálculo da proporção na amostra, dada por: $p = n_1/n$, onde n_1 é o número de elementos na amostra que possuem determinada característica desejada e n é o número total de elementos na amostra. Como estimativa do desvio padrão populacional, σ usa-se o desvio padrão da amostra, s, dado por:

$$s = \sqrt{\frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2}.$$

Por outro lado, na estimação por intervalo, encontramos um limite inferior e um limite superior, os quais vão formar um intervalo de valores, dentro do qual esperamos, com certo grau de confiança, que o verdadeiro valor do parâmetro esteja incluído. O intervalo de confiança é muito mais informativo do que uma estimativa através de um único valor. Pois, no intervalo, além de termos a informação pontual, também temos uma boa ideia da variabilidade do parâmetro.

Parte XVIII

Intervalos de Confiança baseados numa amostra

17.4 IC para μ quando σ^2 é conhecido

Suponha $X_1, \ldots, X_n$ uma å de tamanho n provenientes de uma população normal, com média μ e variância σ^2, isto é, $X \sim N(\mu, \sigma^2)$.

Um IC para μ do $100(1-\alpha)\%$ quando σ^2 é conhecido é dado por

$$IC_{100(1-\alpha)\%}(\mu) = \left[\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \tag{17.1}$$

em que $z_{1-\alpha/2}$ representa o percentil $1-\alpha/2$ de $Z \sim N(0,1)$, isto é $P(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$.

Exemplo 76 *A distribuição dos pesos de pacotes de sementes de milho, enchidos automaticamente por uma certa máquina, é normal, com desvio padrão (σ) conhecido e igual a 200g. Uma amostra de 15 pacotes retirada ao acaso apresentou os seguintes pesos, em kg, Construir e interpretar os intervalos*

20,05	20,10	20,25	19,78	19,69	19,90	20,20	19,89
19,70	20,30	19,93	20,25	20,18	20,01	20,09	

de 95% e 99% de confiança para o peso médio dos pacotes de sementes de milho.

17.5 IC para μ quando σ^2 é desconhecido

Um IC para μ do $100(1-\alpha)\%$ quando σ^2 é desconhecido é dado por

$$IC_{100(1-\alpha)\%}(\mu) = \left[\bar{x} \pm t^{(n-1)}_{1-\alpha/2}\frac{s}{\sqrt{n}}\right], \tag{17.2}$$

em que $t^{(n-1)}_{1-\alpha/2}$ representa o percentil $1-\alpha/2$ de $T \sim t(n-1)$, isto é $P(T \leq t^{(n-1)}_{1-\alpha/2}) = 1-\alpha/2$.

Exemplo 77 *O peso médio, ao nascer, de bezerros da raça Ibagé examinada uma amostra de 20 partos, foi de 26 kg com um desvio padrão de 2kg. Construir e interpretar o intervalo de 95% de confiança para o peso médio de bezerros.*

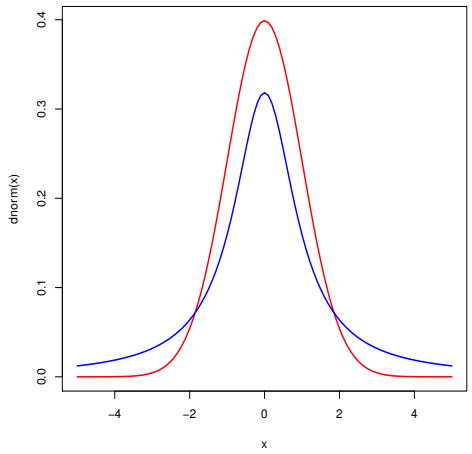

Figura 17.1: Comparação entre distribuição N(0,1) (vermelha) e t(1) (azul)

Exemplo

Exemplo 78 *Os resíduos industriais jogados nos rios, muitas vezes, absorvem o oxigênio necessário à respiração dos peixes e outras formas de vida aquática. Uma lei estadual exige um mínimo de 5 ppm de oxigênio dissolvido, a fim de que o conteúdo do mesmo seja suficiente para manter a vida aquática. Seis amostras de água retiradas de um rio revelaram os índices: 4.9, 5.1, 4.9, 5.0, 5.0 e 4.7 ppm de oxigênio dissolvido. Construir o intervalo com 95% de confiança para a verdadeira média de oxigênio, em ppm, e interpretar.*

17.6 IC para a proporção

O intervalo de confiança para uma proporção populacional (π), é muito semelhante ao intervalo de confiança para uma média populacional com σ conhecido. A principal diferença está no desvio padrão da distribuição amostral das proporções, que é dado por $s_P = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, assim um IC para π do $100(1-\alpha)\%$ é dado por

$$IC_{100(1-\alpha)\%}(\pi) = \left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{17.3}$$

em que $\hat{p}$ é a proporção estimada de π.

Exemplo 79 *Em certo lago, uma amostra de 1000 peixes acusou 290 tilápias. Construa um intervalo de 95% de confiança para a verdadeira proporção de tilápias na população piscosa do lago.*

Tarefa

Exemplo 80 *Uma amostra de 35 peixes da espécie Xenomelaniris brasiliensis coletada na localidade Praia da Barra da Lagoa, Florianópolis, SC, apresentou 46% de peixes com comprimento total acima de 50 mm. Encontre um intervalo, com 99% de confiança, dentro do qual deve estar a verdadeira proporção de peixes com comprimento acima de 50 mm.*

17.7 Erro de Estimação ou de Amostragem

Ao coletarmos uma amostra e calcularmos a média dos valores desta amostra $(\bar{X})$, dificilmente ela vai ser igual a média verdadeira (μ), apesar de estarem próximas, para amostras suficientemente grandes. Como a amostra é uma parte da população, é lógico pensar que os dois valores dificilmente vão coincidir. Lembre-se do estudo da distribuição amostral da média. Portanto, quando vamos estimar um parâmetro, sempre estamos sujeitos a cometer um erro, denominado erro de estimação ou de amostragem, que é a diferença entre o parâmetro e a estatística amostral, isto é,

$$e = \text{erro de estimação} = \mu - \bar{X}.$$

O erro de estimação associado ao

1. IC para μ quando σ^2 é conhecido é dado por $e = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$,
2. IC para μ quando σ^2 é desconhecido é dado por $e = t^{(n-1)}_{1-\alpha/2}\frac{s}{\sqrt{n}}$,
3. IC para π é dado por $e = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

17.8 Determinação do tamanho da amostra para μ para σ conhecido

Uma das perguntas mais frequentes em estatística é qual é o tamanho da amostra necessário para estimar a média? A resposta a esta pergunta, só é possível de ser dada, após o pesquisador da área de interesse, fornecer algumas informações, como veremos a seguir. Podemos determinar o tamanho da amostra (n), através da fórmula do erro de estimação associado a um intervalo de confiança,

$$e = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Para uma amostra aleatória simples, quando o desvio padrão populacional (σ) é conhecido, ou temos alguma informação sobre o mesmo, determinamos o tamanho da amostra pela expressão

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2.$$

Exemplo 81 *A distribuição dos pesos de pacotes de sementes de milho, enchidos automaticamente por uma certa máquina, é normal, com desvio padrão (σ) conhecido e igual a 200g. Uma amostra de 15 pacotes retirada ao acaso apresentou os seguintes pesos, em kg, Que tamanho de amostra será*

20,05	20,10	20,25	19,78	19,69	19,90	20,20	19,89
19,70	20,30	19,93	20,25	20,18	20,01	20,09	

necessário coletar para produzir um intervalo de 95% de confiança para a verdadeira média, com um erro de estimação de 50 gramas?

17.9 Determinação do tamanho da amostra para μ para σ desconhecido

Na prática, geralmente o desvio padrão populacional (σ) é desconhecido, ou não temos conhecimento de um limite superior para o mesmo. Nesse caso, deveríamos usar o desvio padrão da amostra (s), e a distribuição t de Student. Acontece que a amostra ainda não foi coletada para que possamos conhecer o valor de s, desvio padrão da amostra, então, uma solução é coletar uma amostra piloto de n' elementos para, com base nela, obtermos uma estimativa de s, empregando-se a seguir a expressão

$$n = \left(\frac{t_{1-\alpha/2} s}{e} \right)^2 .$$

Onde t é o valor de tabela, com $n' - 1$ graus de liberdade (tamanho da amostra piloto menos um), e probabilidade de erro igual a α. Se $n \leq n'$ implica que a amostra piloto já é suficiente para a estimação da média, caso contrário, devemos retirar mais elementos da população para completar o tamanho mínimo da amostra.

Exemplo 82 *O peso médio, ao nascer, de bezerros da raça Ibagé examinada uma amostra de 20 partos, foi de 26 kg com um desvio padrão de 2kg. Que tamanho de amostra será necessário para produzir um intervalo de confiança de 95% para a verdadeira média, com uma precisão de 5% da média da amostra preliminar?*

A amostra piloto de tamanho $n' = 20$*, nos forneceu* $\bar{x} = 26$ *kg e* $s = 2$ *kg. Temos ainda que a precisão desejada vale* $e = 0.05 * (26) = 1.3$ *kg e* $t(19, 0.975) = 2.093$*. Portanto, o tamanho da amostra vale*

$$n = \left(\frac{2.093 \times 2}{1.3}\right)^2 = 10.37 = 11.$$

Necessitamos de uma amostra de 11 bezerros para a precisão e confiança estipuladas pelo pesquisador. Como a amostra piloto tem tamanho $n' = 20$, maior que o tamanho da amostra necessário $n = 11$ bezerros, implica que a amostra piloto já é suficiente para o estudo.

17.10 Determinação do tamanho da amostra para π

Para encontrarmos o tamanho necessário de uma amostra para estimarmos uma proporção da população, procedemos de forma análoga ao que foi feito para o caso de estimação de uma média da população, isto é,

$$n = \left(\frac{z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{e} \right)^2 .$$

Exemplo 83 *Considere que uma amostra piloto de 35 peixes da espécie Xenomelaniris brasiliensis foi coletada na localidade da Praia da Barra da Lagoa, Florianópolis, SC, e apresentou 46% de peixes com comprimento total acima de 50 mm. Se quisermos estimar a proporção de peixes com comprimento acima de 50 mm, qual o tamanho da amostra necessário para que tenhamos 99% de confiança de que o erro de nossa estimativa não seja superior a 5%?*

Exercícios

Exemplo 84 *De uma amostra de 100 peixes da espécie Xenomelaniris brasiliensis, coletada na Armação do Pântano do Sul, Florianópolis, SC, verificou-se que 57 deles apresentavam comprimento total maior que 50 mm.*

1. *Com base nessa informação, determine o intervalo de confiança de 99% para a verdadeira proporção de peixes com comprimento total acima de 50 mm.*
2. *Qual o tamanho de amostra necessário para estimar a verdadeira proporção com precisão de 5%, usando uma confiança de 95%?.*

Exemplo 85 *O diâmetro médio de Biomphalaria tenagophila, examinada uma amostra de 35 animais, foi de 0,871 mm com um desvio padrão de 0,057 mm.*

1. *Dê a estimativa por intervalo do verdadeiro diâmetro médio utilizando um nível de confiança de 95%.*
2. *Que tamanho de amostra será necessário para produzir um intervalo de confiança de 95% para a verdadeira média, com uma precisão de 2% da média da amostra preliminar?*

Exemplo 86 *Em um experimento, 320 de 400 sementes germinaram. Determine o intervalo de confiança de 98% para a verdadeira proporção de sementes que germinam. Para realizar o teste de germinaçã, quantas sementes serão necessárias utilizar, se se deseja um intervalo de confiança de 95%, com precisão de 4%?*

Parte XIX

Intervalos de confiança duas amostras

Introdução

A comparação de duas populações ou dois tratamentos é uma das situações mais comuns encontradas na prática estatística. Chamamos essas situações de problemas de duas amostras.

Problema de duas amostras

1. O objetivo da inferência é a comparação das respostas a dois tratamentos ou a comparação das características de duas populações.
2. Temos uma amostra distinta de cada tratamento ou de cada população.

O calouro 15 é real, ou é um mito?

Há uma crença popular de que os estudantes universitários, tipicamente ganham 15 libras (6,8 kg) de peso durante seu primeiro ano de faculdade. Esse ganho de peso de 15 libras tem sido chamado *calouro 15*. Explicações razoáveis para esse fenômeno incluem o novo estresse da vida universitária (não incluindo um curso de estatística, que é pura alegria), novos hábitos de alimentação, níveis crescentes de consumo de álcool, menos tempo livre para atividades físicas, comida de lanchonete com abundância de gordura e carboidratos, a nova liberdade de escolher entre uma variedade de alimentos (incluindo pizzas suntuosas que demandam apenas um telefonema), e falta de sono suficiente que resulta em níveis mais baixos de leptina, que ajuda a regular o apetite e o metabolismo. Mas, o *calouro 15* é real, ou é um mito que tem sido perpetuado através de evidência anedótica e/ou dados falseados?.

O conjunto de dados inclui dois pesos para cada um dos 67 sujeitos do estudo. Cada sujeito foi pesado em setembro e novamente em abril de seu primeiro ano. Essas duas medidas foram feitas no inicio e no final dos sete meses de vida passada no campus entre essas duas datas. É importante reconhecer que cada par individual de medidas antes e depois se refere a um mesmo estudante, de modo que a lista de 67 medidas antes e as 67 medidas depois constituem dados emparelhados dos 67 sujeitos do estudo.

Exemplos

1. Um banco deseja saber qual de dois planos de incentivo aumentará mais o uso de seus cartões de crédito. Ele oferece cada incentivo a uma amostra aleatória de clientes de cartão de crédito e compara a quantidade debitada no cartão durante os seis meses seguintes.
2. Um psicólogo desenvolve um teste que mede o compromisso social. Ele compara o compromisso social de universitários com o de universitárias, aplicando o teste a duas amostras separadas de estudantes, uma de cada gênero.

Capítulo 18
IC para igualdade de variâncias

Capítulo 19
IC para a diferença de médias (amostras independentes)

19.1 Caso $\sigma_x^2 = \sigma_y^2 = \sigma^2$ desconhecidos

Seja $X_1, \ldots, X_{n_1}$ aa(n_1) em que cada $X \sim N(\mu_x, \sigma_x^2)$. Seja $Y_1, \ldots, Y_{n_2}$ aa(n_2) em que cada $Y \sim N(\mu_y, \sigma_y^2)$. Além disso as duas amostras são independentes.

$$IC_{100(1-\alpha)\%}(\mu_x - \mu_y) = \left[(\bar{x} - \bar{y}) \pm t_{1-\alpha/2}^{(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \tag{19.1}$$

em que $t_{1-\alpha/2}^{(n_1+n_2-2)}$ representa o percentil $1-\alpha/2$ de $T \sim t(n_1+n_2-2)$ e s_p^2 é definido por

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Exemplo 87 *Queremos estudar a influência que pode exercer o tabaco no peso das crianças ao nascer. Para isso, consideram-se dois grupos de mulheres grávidas (de fumantes e outro de não-fumantes) e obtém-se os seguintes dados sobre o peso de seus filhos:*

1. *Mãe fumantes,* $n_1 = 35$ *mulheres,* $\bar{x}_1 = 3.6$*kg e* $s_1 = 0,5$*kg*
2. *Mãe não-fumantes,* $n_2 = 27$ *mulheres,* $\bar{x}_2 = 3.2$*kg e* $s_2 = 0,8$*kg.*

Em ambos os grupos, os pesos dos recém nascidos provém de distribuições normais de médias desconhecidas e com variâncias que embora sejam desconhecidas podemos supor que sejam as mesmas. Construir um intervalo de confiança de 95% para a diferença das médias. Calcular o erro de estimação.

Exemplo 88 *As produções de duas variedades de milho, em toneladas por hectare, foram as seguintes Construir um IC de 95% para a diferença das médias entre as variedades A e B. Calcular o erro de*

Variedade A	1,3	1,4	1,1	1,4	1,5
Variedade B	1,8	1,6	1,9	1,9	1,8

estimação.

Exemplo 89 *Os tempos gastos na manobra dos arados Fuçador e Erechim, foram os seguintes*

Fuçador	0,20	0,22	0,18	0,23	0,12	0,20	0,13	0,12	0,13	0,22	0,17
Erechim	0,36	0,48	0,33	0,43	0,40	0,43	0,33	0,36	0,35	0,40	0,35

Construir um intervalo de confiança de 95% para a diferença das médias entre o tempo gasto na manobra dos arados Fuçador e Erechim . Calcular o erro de estimação.

19.2 IC para a diferença entre proporções

O IC $100(1-\alpha)\%$ para a diferença de proporções é dado a seguir

$$IC_{100(1-\alpha)\%}(\pi_1 - \pi_2) = \left[(\hat{p_1} - \hat{p_2}) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right] \tag{19.2}$$

em que $\hat{p}_1$ é a proporção estimada de π_1 e $\hat{p}_2$ é a proporção estimada de π_2.

Exemplo 90 *Acredita-se que a osteoporose está relacionada com o sexo. Para isso, escolhe-se uma amostra de 100 homens com mais de 50 anos e uma amostra de 200 mulheres mas mesmas condições. Obtém-se 10 homens e 40 mulheres com algum grau de osteoporose. O que podemos concluir com uma confiança de 95% ?*

Os airbags salvam vidas?

Exemplo 91 *A tabela a seguir lista resultados de uma amostra aleatória simples de ocupantes de bancos dianteiros envolvidos em acidentes de carro. Construir um intervalo de confiança de 95% para a*

	Airbag Disponível	Airbag não Disponível
Mortes de ocupantes	41	52
Número total de ocupantes	11541	9843

diferença de proporções. Calcular o erro de estimação.

Capítulo 20
IC para a diferença de médias (amostras dependentes)

Os dados de duas amostras constituem dados pareados quando estão relacionados dois a dois, segundo algum critério que introduz um influência marcante entre os diversos pares de valores. Também é importante observar que deve haver independência entre observações dentro de cada uma das amostras.

Exemplo 92 *Desejamos fazer um teste estatístico para verificar se existe diferença significativa entre as médias das notas obtidas na primeira avaliação e na segunda avaliação da disciplina de estatística. Então, para cada aluno, tomamos a sua nota na primeira avaliação e na segunda avaliação. Como existem diferenças entre os alunos (alguns estudam mais, outros tem mais facilidade com a disciplina, etc.), os pares de notas (cada aluno um par de notas) não são independentes. Existe o fator aluno introduzindo uma influência forte entre os pares de dados. Observe que para cada amostra, como os alunos são diferentes, as observações são independentes dentro delas.*

Um IC de $100(1-\alpha)\%$ para a diferença de médias (amostras dependentes) é dado por

$$IC_{100(1-\alpha)\%}(\mu_d) = \left[\bar{x}_d \pm t^{(n-1)}_{1-\alpha/2}\frac{s_d}{\sqrt{n}}\right] \tag{20.1}$$

em que

$$\bar{x}_d = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \text{é a média das diferenças,} \quad d_i = x_i - y_i$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{x}_d)^2 \qquad \text{é a variância das diferenças,}$$

e $t^{(n-1)}_{1-\alpha/2}$ representa o percentil $1-\alpha/2$ de $T \sim t(n-1)$.

Exemplo 93 *Foi conduzido um experimento para estudar o conteúdo de hemoglobina no sangue de suínos com deficiência de niacina. Aplicou-se 20 mg de niacina em 8 suínos. Encontre o intervalo de confiança com 95% para a verdadeira diferença entre as duas medias. Foram mensurados os níveis de hemoglobina no sangue antes e depois da aplicação da niacina, os resultados obtidos no experimento foram:*

Suínos	Antes (A)	Depois (B)
1	13,6	11,4
2	13,6	12,5
3	14,7	14,6
4	12,1	13,0
5	12,3	11,7
6	13,2	10,3
7	11,0	9,8
8	12,4	10,4

Exemplo 94 *Verificar com um 95% se o calibre da veia esplênica é, em média, o mesmo, antes e depois da oclusão da veia porta a partir dos seguintes dados de cães.*

Cão	1	2	3	4	5	6
Antes da oclusão	75	50	50	60	50	70
Depois da oclusão	85	75	70	65	60	90

Capítulo 21
Teste de Hipóteses

21.1 Introdução

Um problema que precisamos aprender a resolver é o de testar uma hipótese, isto é, feita uma determinada afirmação sobre um parâmetro populacional, por exemplo, sobre uma média populacional ou uma proporção populacional, será que os resultados de uma amostra contrariam ou não tal afirmação? Podemos estar interessados em verificar, por exemplo, se as seguintes afirmações são verdadeiras

- a produtividade do milho em Santa Catarina, é de 2300 kg/ha;
- os comprimentos médios dos ante-braços de duas espécies de morcêgos são iguais;
- a proporção de fixação de fitoplâncton em dois tipos de solos é a mesma;
- a produção média de duas cultivares de feijão é a mesma;
- épocas de plantio estão associadas com a sobrevivência das mudas.

21.2 Objetivo

O objetivo de um teste estatístico de hipóteses é fornecer ferramentas que nos permitam aceitar ou rejeitar uma hipótese estatística através dos resultados de uma amostra.

21.3 Exemplo de proporção

Consideremos um teste de germinação de sementes, onde foram analisadas 400 sementes de milho, obtidas por um processo de amostragem aleatória, de um grande lote de sementes, encontrando-se, nesta amostra, um poder germinativo de 92,8%. Porém, a distribuidora afirma que não haverá menos de 94% de germinação no lote. O que devemos responder com o auxílio de um teste de hipóteses, é se podemos considerar a afirmação da distribuidora como sendo verdadeira ou não. Para todos os testes estatísticos, inicialmente devemos formular as hipóteses.

Sempre vamos ter duas hipóteses estatísticas, isto é,

1. Hipótese nula: é a hipótese que sugere que a afirmação que estamos fazendo sobre o parâmetro populacional é verdadeira. Essa hipótese é representada por H_0. No nosso exemplo, a hipótese nula é que a verdadeira porcentagem de germinação de sementes é de 94%, portanto, a distribuidora está certa, e a representamos por:

$$H_0 : \pi = 94\%$$

2. Hipótese alternativa: é a hipótese que sugere que a afirmação que estamos fazendo sobre o parâmetro populacional é falsa e a representamos por H_1. No nosso exemplo, a hipótese alternativa é que o poder germinativo do lote é menor que 94%, pois devemos nos precaver contra o lote ter menos do que 94% de germinação e, portanto, a distribuidora não está certa, e a

representamos por:

$$H_1 : \pi < 94\%.$$

Portanto, a construção da hipótese alternativa, depende do grau de conhecimento biológico ou agronômico sobre o fenômeno, ou das informações que se têm do problema em estudo.

21.4 Hipótese alternativa

Existem três afirmações que podemos fazer em uma hipótese alternativa

1. $H_1 : \pi \neq 94\%$ (temos um teste bilateral);
2. $H_1 : \pi > 94\%$ (temos um teste unilateral à direita);
3. $H_1 : \pi < 94\%$ (temos um teste unilateral à esquerda).

21.5 Conceitos Básicos

21.6 Erros Tipo I e Tipo II

Quando rejeitamos a hipótese nula, corremos o risco de estarmos tomando uma decisão errônea, ou seja, rejeitamos a hipótese nula quando na verdade deveríamos aceitá-la. Este risco é o nível de significância ou valor p do teste e é representado pela letra grega α. Esse nível de significância é também conhecido como erro tipo I e, a probabilidade de sua ocorrência vale α. Um segundo tipo de erro que podemos cometer, é aceitar a hipótese nula, quando ela é de fato falsa. Neste caso, temos o erro tipo II, o qual é representado pela letra grega β. Esquematicamente, temos

Decisão	H_0 é verdadeira	H_0 é falsa
não rejeitar H_0	decisão correta	erro tipo II
rejeitar H_0	erro tipo I	decisão correta

21.7 Teste de hipótese

Definição 51 *Um teste de hipótese estatística é uma regra ou procedimento para decidir se rejeitamos ou não* H_0.

21.8 Região Crítica

Definição 52 *É o conjunto de valores com os quais rejeitamos* H_0. *Notação* RC.

21.9 Nível de Significância

Definição 53 *O nível de significância de um teste é definido como*

$$\alpha \;=\; P(\textit{Erro tipo I}) = P(\textit{Rejeitar } H_0 \textit{ dado que } H_0 \textit{ é Verdadeiro}).$$

21.10 Testes de Médias Populacionais

O objetivo de testar-se hipóteses sobre médias verdadeiras é avaliar certas afirmações feitas sobre as mesmas. Por exemplo, podemos desejar verificar a afirmação de que as alturas médias de plantas de feijão, para sementes de alto e baixo vigor, são iguais.

Existem, basicamente, três tipos de afirmações que se podem fazer quando se estuda médias populacionais

1. a afirmação diz respeito a uma média populacional, então, temos o teste de uma média populacional. Exemplo, os pesos ao nascer de bezerros da raça Nelore, no planalto Catarinense, em agosto, é de 25,5 kg;
2. a afirmação diz que as médias de duas populações (dois tratamentos) são iguais, temos, então, o teste de comparação de duas médias. Exemplos
   (a) as produções médias de batatinhas de duas variedades são iguais e
   (b) as áreas foliares específicas médias da espécie Cecropia glaziovi, em amostras situadas na borda da mata e na mata fechada são iguais.
3. (**não será estudado neste curso**) a afirmação diz que as médias de mais de duas populações (mais do que dois tratamentos) são todas iguais, temos, então, o teste de comparação de k

médias, com k $>$ 2. Neste caso, devemos fazer uma Análise de Variância. Existem diversos livros especializados em planejamento e análise de experimentos, que tratam desse tipo de análise, por exemplo, os livros de Steel e Torrie (1960) e Vieira (1999). Por exemplo, desejamos saber se há diferenças entre três locais (Baía Norte, Baía Sul e Pântano do Sul), quanto ao número médio de micronúcleos por 5000 células sanguíneas de peixes do gênero bagre.

21.11 Teste para μ quando σ^2 é desconhecida (1 amostra)

Seja $X_1, \ldots, X_n$ uma aa(n) desde uma distribuição $N(\mu, \sigma^2)$ com σ^2 desconhecida.

Tabela 21.1: Hipóteses para μ quando σ^2 é desconhecida (1 amostra)

Hipóteses nula	**Estatística sob H_0**
$H_0 : \mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Hipóteses alternativa	**Região Crítica**
$H_1 : \mu \neq \mu_0$	$RC = \{t \leq t_{\alpha/2}(n-1) \quad \text{ou} \quad t \geq t_{1-\alpha/2}(n-1)\}$
$H_1 : \mu > \mu_0$	$RC = \{t \geq t_{1-\alpha}(n-1)\}$
$H_1 : \mu < \mu_0$	$RC = \{t \leq t_{\alpha}(n-1)\}$

Exemplo 95 *Supõe-se que a produtividade média de feijão da safra no Estado de Santa Catarina é de 800 kg/ha. Para investigar a veracidade dessa afirmação, consultou-se uma publicação do Instituto CEPA-SC, onde obteve-se os seguintes valores de produtividade média de feijão: Qual a conclusão ao*

Produtividade	1017	980	507	841	899	264	700	800	653

nível de significância de 5%?.

Exemplo 96 *Foi retirada uma amostra de tamanho 10, da população de pesos aos 210 dias de bezerros da raça Nelore. Os valores, em kg, foram os seguintes Teste as hipóteses* $H_0 : \mu = 186$ *vs*

pesos	178 199 182 186 188 191 189 185 174 158

$H_1 : \mu < 186$ *ao nível de significância de* 5%.

21.12 Teste para diferença de médias (caso independente)

Seja $X_1, \ldots, X_{n_1}$ uma aa(n_1) desde uma distribuição $N(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_{n_2}$ uma aa(n_2) desde uma distribuição $N(\mu_y, \sigma_y^2)$ com $\sigma_x^2 = \sigma_y^2 = \sigma^2$ desconhecidas.

Tabela 21.2: Hipóteses para diferença de médias (caso independente)

Hipóteses nula	**Estatística sob H_0**
$H_0 : \mu_x - \mu_y = \delta$	$t = \dfrac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $\quad s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
Hipóteses alternativa	**Região Crítica**
$H_1 : \mu_x - \mu_y \neq \delta$	$RC = \{t \leq t_{\alpha/2}(n_1+n_2-2) \text{ ou } t \geq t_{1-\alpha/2}(n_1+n_2-2)\}$
$H_1 : \mu_x - \mu_y > \delta$	$RC = \{t \geq t_{1-\alpha}(n_1+n_2-2)\}$
$H_1 : \mu_x - \mu_y < \delta$	$RC = \{t \leq t_{\alpha}(n_1+n_2-2)\}$

Exemplo 97 *As produções de duas variedades de milho, em toneladas por hectare, foram as seguintes Que podermos afirmar em relação às produções de duas variedades de milho. Use um nível de*

Variedade A	1,3	1,4	1,1	1,4	1,5
Variedade B	1,8	1,6	1,9	1,9	1,8

significância de 5%?.

Exemplo 98 *Os tempos gastos na manobra dos arados Fuçador e Erechim, foram os seguintes*

Fuçador	0,20	0,22	0,18	0,23	0,12	0,20	0,13	0,12	0,13	0,22	0,17
Erechim	0,36	0,48	0,33	0,43	0,40	0,43	0,33	0,36	0,35	0,40	0,35

Espera-se que o arado Fuçador produza melhores resultados (gaste menos tempo na manobra). Qual a conclusão ao nível de significância de 5%?.

21.13 Teste para diferença de médias (caso dependente)

Seja $X_1, \ldots, X_{n_1}$ uma aa(n_1) desde uma distribuição $N(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_{n_2}$ uma aa(n_2) desde uma distribuição $N(\mu_y, \sigma_y^2)$, $D_i = X_i - Y_i \sim N(\mu_d, \sigma_d^2)$.

Tabela 21.3: Hipóteses para diferença de médias (caso dependente)

Hipóteses nula	**Estatística sob H_0**
$H_0 : \mu_d = \delta$	$t = \dfrac{(\bar{x}_d - \delta)}{s_d/\sqrt{n}} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{x}_d)^2$
Hipóteses alternativa	**Região Crítica**
$H_1 : \mu_d \neq \delta$	$RC = \{t \leq t_{\alpha/2}(n-1) \quad \text{ou} \quad t \geq t_{1-\alpha/2}(n-1)\}$
$H_1 : \mu_d > \delta$	$RC = \{t \geq t_{1-\alpha}(n-1)\}$
$H_1 : \mu_d < \delta$	$RC = \{t \leq t_{\alpha}(n-1)\}$

Exemplo 99 *Foi conduzido um experimento para estudar o conteúdo de hemoglobina no sangue de suínos com deficiência de niacina. Aplicou-se 20 mg de niacina em 8 suínos. Podemos afirmar que o conteúdo de hemoglobina no sangue diminuiu, com a aplicação de niacina? (use nível de significância de 0.01). Foram mensurados os níveis de hemoglobina no sangue antes e depois da aplicação da niacina, os resultados obtidos no experimento foram:*

Suínos	Antes (A)	Depois (B)
1	13,6	11,4
2	13,6	12,5
3	14,7	14,6
4	12,1	13,0
5	12,3	11,7
6	13,2	10,3
7	11,0	9,8
8	12,4	10,4

21.14 Teste para proporção populacional

Feita uma afirmação sobre uma proporção, desejamos saber se os dados de uma amostra suportam ou não tal afirmação. Por exemplo, verificar se a afirmativa de que 20% dos indivíduos de uma comunidade apresentam certa característica genética.

Exemplo 100 *O rótulo de uma caixa de sementes informa que a porcentagem de germinação é de 90%. Entretanto, como a data limite de validade já foi ultrapassada, acredita-se que a porcentagem de germinação seja inferior a 90%. Faz-se um experimento e, de 400 sementes testadas, 350 germinaram. Ao nível de significância de 10%, rejeita-se a hipótese de que a porcentagem de germinação é de 90%?*

Tabela 21.4: Hipóteses para π (1 amostra)

Hipóteses nula	**Estatística sob H_0**
$H_0 : \pi = \pi_0$	$z = \dfrac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
Hipóteses alternativa	**Região Crítica**
$H_1 : \pi \neq \pi_0$	$RC = \{z \leq z_{\alpha/2} \quad \text{ou} \quad z \geq z_{1-\alpha/2}\}$
$H_1 : \pi > \pi_0$	$RC = \{z \geq z_{1-\alpha}\}$
$H_1 : \pi < \pi_0$	$RC = \{z \leq z_{\alpha}\}$

Exemplo 101 *Um Biólogo, com base em conhecimentos teóricos e práticos, afirma que a proporção de forófitos no estádio arbóreo pioneiro da Floresta Ombrófila na Ilha de Santa Catarina, apresenta 47% sem bromélias. Numa amostra de 35 forófitos, $p = 40\%$ não apresentaram bromélias. Teste a afirmativa do Biólogo ao nível de significância de 5%.*

Exemplo 102 *A proporção de analfabetos em um município era de 15% na gestão anterior. O prefeito atual implantou um programa de alfabetização desde o início de sua gestão e afirma que após 2 anos reduziu a proporção de analfabetos. Para verificar a afirmação do prefeito 60 cidadãos foram entrevistados. Se observamos 6 analfabetos entre os 60 entrevistados, qual é a conclusão ao nível de significância de 5%?*

Exemplo 103 *Suponha que um medicamento existente no mercado produza o efeito desejado em 60% dos casos nos quais é aplicado. Um laboratório produz um novo medicamento e afirma que ele é melhor do que o existente. Aplicou-se o medicamento em 10 pacientes. Se observamos que o medicamento novo produz o efeito desejado 8 pacientes, qual é a conclusão ao nível de significância de 5%?*

21.15 Teste para diferença de proporções populacionais

Suponha que você queira determinar se a proporção de estudantes universitários do sexo feminino que receberam diploma de bacharel em quatro anos é diferente da proporção de estudantes universitários do sexo masculino que receberam diploma de bacharel em quatro anos.

Tabela 21.5: Teste para diferença de proporções populacionais

Hipóteses nula	**Estatística sob H_0**
$H_0 : \pi_1 = \pi_2$	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}}$ $\quad \bar{p} = \frac{x_1+x_2}{n_1+n_2}$
Hipóteses alternativa	**Região Crítica**
$H_1 : \pi_1 \neq \pi_2$	$RC = \{z \leq z_{\alpha/2} \text{ ou } z \geq z_{1-\alpha/2}\}$
$H_1 : \pi_1 > \pi_2$	$RC = \{z \geq z_{1-\alpha}\}$
$H_1 : \pi_1 < \pi_2$	$RC = \{z \leq z_{\alpha}\}$

Exemplo 104 *Em um estudo de 200 mulheres adultas selecionadas aleatoriamente e 250 homens adultos, ambos usuários do facebook, 30% das mulheres e 38% dos homens diseram que planejam comprar on-line ao menos uma vez durante o mês seguinte. Usando um nível de significância de* $\alpha = 0.10$*, teste a afirmação de que há uma diferença entre a proporção de mulheres e a proporção de homens, usuários do facebook, que planejam comprar on-line.*

Exemplo 105 *Uma equipe de pesquisa médica conduziu um estudo para testar o efeito de um medicamento na redução do colesterol. Ao final do estudo, os pesquisadores descobriram que dos 4700 sujeitos selecionados aleatoriamente que tomaram o medicamento, 301 morreram de doenças do coração. Dos 4300 sujeitos selecionados aleatoriamente que tomaram um placebo, 357 morreram de doenças do coração. Usando um nível de significância de* $\alpha = 0.01$*, você pode concluir que a taxa de mortalidade por doenças do coração é menor para aqueles que tomaram a medicação do que para aqueles que tomaram o placebo?*

Exemplo 106 *Em um estudo que investiga morbilidade e mortalidade entre vítimas pediátricas de acidentes automobilísticos, a informação com relação à efetividade dos cintos de segurança foi coletada em um período de 18 meses. Duas amostras aleatórias foram selecionadas, uma da população de crianças que usavam cintos de segurança no momento do acidente e outra da população que não os usava. Na amostra de 123 crianças que usavam cinto de segurança no momento do acidente, três morreram. Na amostra de 290 crianças que não usavam cinto de segurança 13 morreram. Usando um nível de significância de* $\alpha = 0.05$*, que coisa você pode concluir?.*

Exemplo 107 *Em um estudo conduzido para investigar fatores não clínicos associados ao método de tratamento cirúrgico recebido para um câncer de mama em estágio inicial, algumas pacientes sofreram mastectomia radical e modificada, enquanto outras tiveram mastectomia parcial, acompanhada por terapia de radiação. Queremos determinar se a idade da paciente afeta o tipo de tratamento que*

receberam. Em particular, queremos saber se as proporções de mulheres abaixo de 55 anos são idênticas nos dois grupos de tratamento. Uma amostra aleatória de 658 mulheres que sofreram mastectomia parcial e subsequente terapia de radiação contém 292 mulheres abaixo de 55 anos; uma amostra de 1580 mulheres que receberam mastectomia radical modificada contém 397 mulheres abaixo dos 55 anos. Usando um nível de significância de $\alpha = 0.05$*, que coisa você pode concluir?.*

21.16 Nível descritivo: p (ou p-valor ou p-*value*)

http://soniavieira.blogspot.com.br/2012/09/o-que-e-p-valor.html

Essa probabilidade p mede a força da evidência contida nos dados, contra a hipótese nula H_0. Como saber se essa evidência é suficiente para rejeitar H_0?. Se o valor de p é **pequeno**, então é pouco provável observarmos valores iguais ou mais extremos que o da amostra, supondo a hipótese nula H_0 verdadeira. Logo, há indícios de que a hipótese nula não seja verdadeira e tendemos a rejeitá-la. Para valores **não tão pequenos** de p , não fica evidente que a hipótese nula H_0 seja falsa. Portanto, tendemos a não rejeitar H_0.

Assim,

- p **pequeno** $\Rightarrow$ rejeitamos H_0.
- p **não pequeno** $\Rightarrow$ não rejeitamos H_0.

21.17 Quão pequeno deve ser p para rejeitarmos H_0 ?

Lembrando que a idéia inicial de p era considerar um nível de significância associado à evidência amostral, então devemos compará-lo com o nível de significância α fixado, de modo que,

- $p \leq \alpha \Rightarrow$ rejeitamos H_0
- $p > \alpha \Rightarrow$ não rejeitamos H_0

Se $p \leq \alpha$, dizemos que a amostra forneceu evidência suficiente para rejeitar a hipótese nula H_0.

Observação 14 *Algumas observações*

- *Quanto menor o valor de p maior é a evidência contra a hipótese nula H_0 contida nos dados.*
- *Quanto menor o nível de significância α fixado, mais forte deve ser a evidência contra a hipótese nula para que ela seja rejeitada.*