Bioestatística: conceitos e modelos

Julio M. Singer

Departamento de Estatística
Universidade de São Paulo, Brazil
www.ime.usp.br/∼jmsinger

Etapas desejáveis para boa análise estatística

- Definição (operacional) do problema
- Planejamento do estudo
- Coleta e armazenamento de dados
- Análise descritiva / correção do conjunto de dados
- Análise inferencial
  - Modelo probabilístico
  - Tradução dos objetivos em termos do elementos do modelo probabilístico
  - Ajuste do modelo estatístico
  - Avaliação dos resultados
- Tradução dos resultados para aplicação prática

Planilhas de dados

i) Não utilizar limitadores de celas (*borders*) ou cores;

ii) Reservar apenas primeira linha para os rótulos das variáveis;

iii) Não usar acentos ou outros símbolos nos rótulos;

iv) Não esquecer uma coluna para a variável indicadora das unidades de investigação (evitar informações confidenciais como nomes de pacientes);

v) Escolher ponto ou vírgula para separação de casas decimais;

vi) Especificar o número de casas decimais;

vii) Codificação para dados abaixo do limite de detecção (*e.g.*, < 0.05)

viii) Incluir dicionário

Exemplo de planilha

ident	idade	nummal	parasita	numgest	idgest	sexorn	pesorn	estrn
1	25	0	0	3	38	2	3665	46
2	30	0	0	9	37	1	2880	44
3	40	0	2	1	41	1	2960	52
4	26	3	0	2	40	1	2740	47
5	.	0	0	1	38	1	2975	50
6	18	0	0	.	38	2	2770	48
7	20	0	0	1	41	1	2755	48
8	15	0	0	1	39	1	2860	49
9	.	0	0	.	42	2	3000	50
10	18	0	4	1	40	1	3515	51
11	17	2	0	2	40	1	3645	54
12	18	1	1	3	40	2	2665	48
13	30	0	0	6	40	2	2995	49
14	19	0	0	1	40	1	2972	46
15	32	0	0	5	41	2	3045	50
34	.	0	0	.	39	2	2950	47

Dicionário

Rótulos	**Variável**	**Unidade de medida**
idade	Idade da mãe	anos
nummal	Quantidade de malárias durante a gestação	número inteiro
parasita	Espécie do parasita da malária	0: não infectada
		1: P. vivax
		2: P. falciparum
		3: malária mista
		4: indeterminado
numgest	Paridade (quantidade de gestações)	Número inteiro
idgest	Idade gestacional no parto	semanas
sexorn	Sexo do recém-nascido	1: masculino
		2: feminino
pesorn	Peso do recém-nascido	g
estrn	Estatura do recém-nascido	cm
pcefal	Perímetro cefálico do recém-nascido	cm
Obs:	Observações omissas são representadas por um ponto	

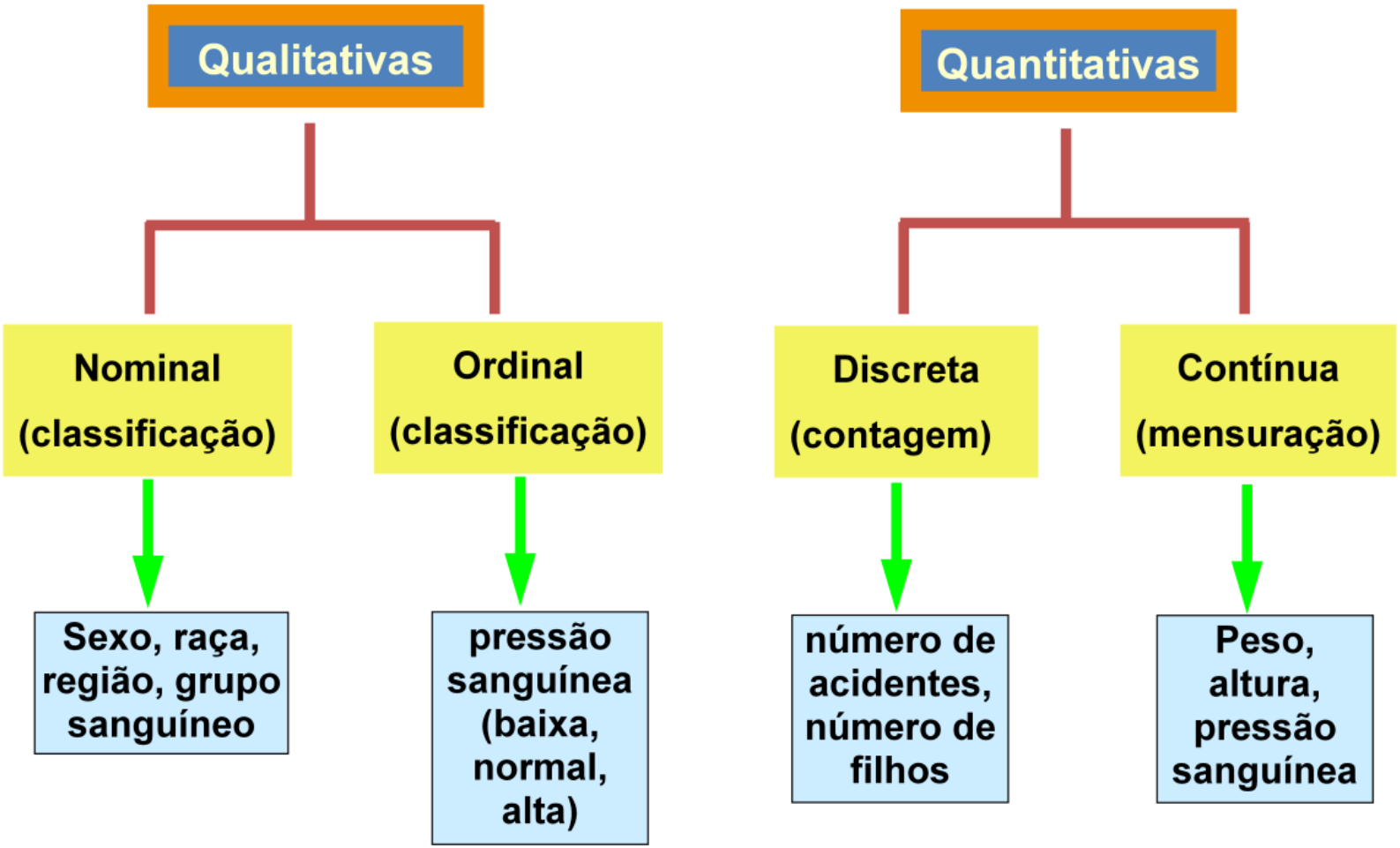
O pesadelo do estatístico

nome	idade	local	Data US	Cir-retosigmoide	US - retosigmoide	US-retosigmoide-LONG	Preparo US
Anne	23	MSA	16/04/2003	não	não		
Aparecida		HC	17.04.01 (verificar se tem US 08.2001)	não??	sim.Reto 3,3x1,5 cm	3,3	
Arlene		HC (Mauricio-Santa Joana)	24/07/2002	Verificar.	não		bom
Assunta	52	HC	22/07/2003	não	não		não
Beatriz		MSA	02/04/2004	não	não		bom
Bernadete	40	HC	15/04/2003	sim	sim.	2,6	bom
Bettina		MSA	26/05/2004	não	não		bom
Blanca		HC	23/07/2003	não	não		regular
Carla		MSA	23/10/2002	sim	não		
Carla		MSA	06/02/2004	sim	Sim,	4,5	
Carolina	24	MSA	25/07/2003	não	sim. Sigmoide: 1,0x0,4cm.	10	
Cassia kolaya		MSA	04/06/2004	sim	Sim 3,5x1,0x2,1	3,5	bom
Celia	31	HC	14/04/2003	não	não		ruim
					Sim		

Planilha para dados longitudinais

grupo	ident	sem	diam
AIG	2	30	7.7
AIG	2	31	8.0
AIG	2	32	8.2
AIG	2	34	9.1
AIG	2	35	9.4
AIG	2	36	9.8
AIG	12	28	7.1
AIG	12	29	7.1
AIG	12	37	7.3
AIG	12	39	9.0
AIG	12	30	9.4
⋮	⋮	⋮	⋮
PIG	17	33	7.5
PIG	17	34	7.7
PIG	17	36	8.2
PIG	29	26	6.3
PIG	29	27	6.5
PIG	29	28	6.6

Classificação de variáveis



Distribuição de frequências: var qualitativas

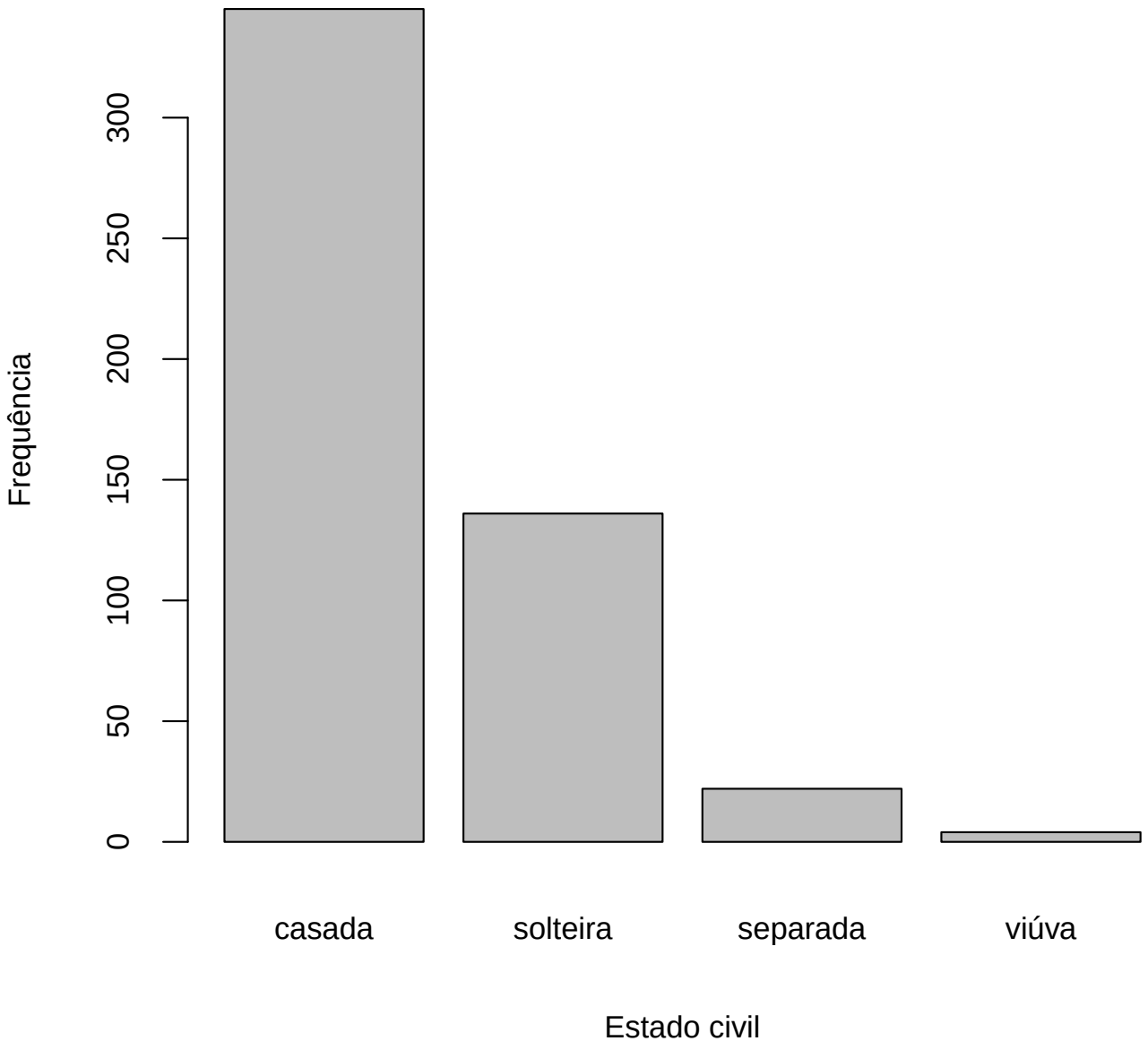
Estado civil	Frequência observada	Frequência relativa (%)
casada	345	68.0
separada	6	1.2
Separada	16	3.2
solteira	136	26.8
viúva	2	0.4
VIÚVA	2	0.4
Total	507	100.0

Estado civil	Frequência observada	Frequência relativa (%)
casada	345	68.0
solteira	136	26.8
separada	22	4.4
viúva	4	0.8
Total	507	100.0

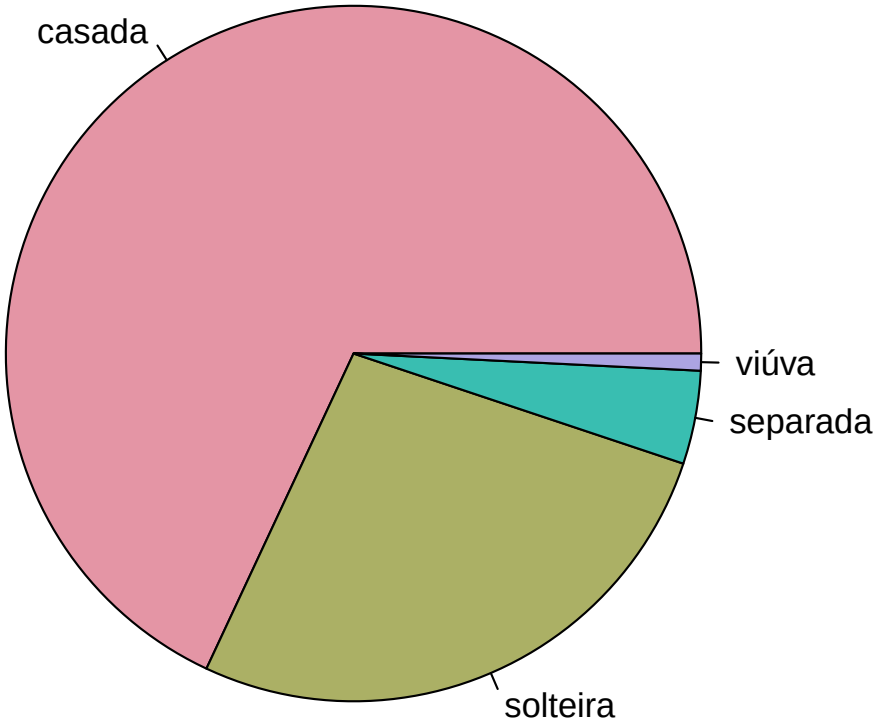
Distribuição de frequências: var qualitativas

Dismenorreia	Frequência observada	Frequência relativa (%)	Frequência relativa acumulada (%)
não	69	13.6	13.6
leve	46	9.1	22.7
moderada	116	22.9	45.6
intensa	209	41.3	86.9
incapacitante	68	13.1	100.0
Total	508	100.0	100.0

Gráficos de barras



Gráficos tipo torta

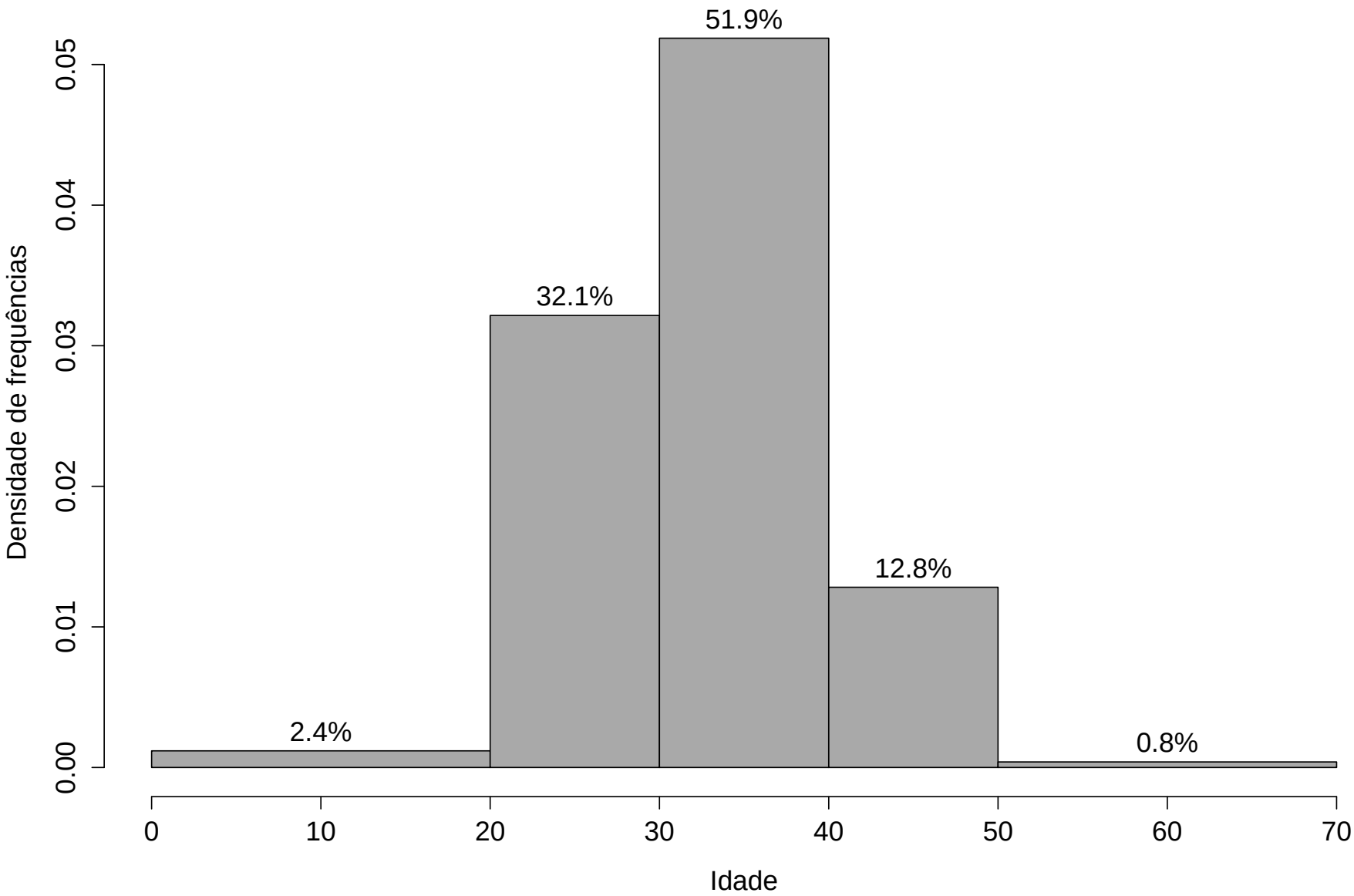


Estado civil

Distribuição de frequências: var contínuas

Idade	Frequência observada	Frequência relativa (%)	Frequência relativa acumulada (%)
$0 - 20$	12	2.4	2.4
$20 - 30$	163	32.1	34.5
$30 - 40$	263	51.9	86.4
$40 - 50$	65	12.8	99.2
$50 - 70$	4	0.8	100.0
Total	507	100.0	100.0

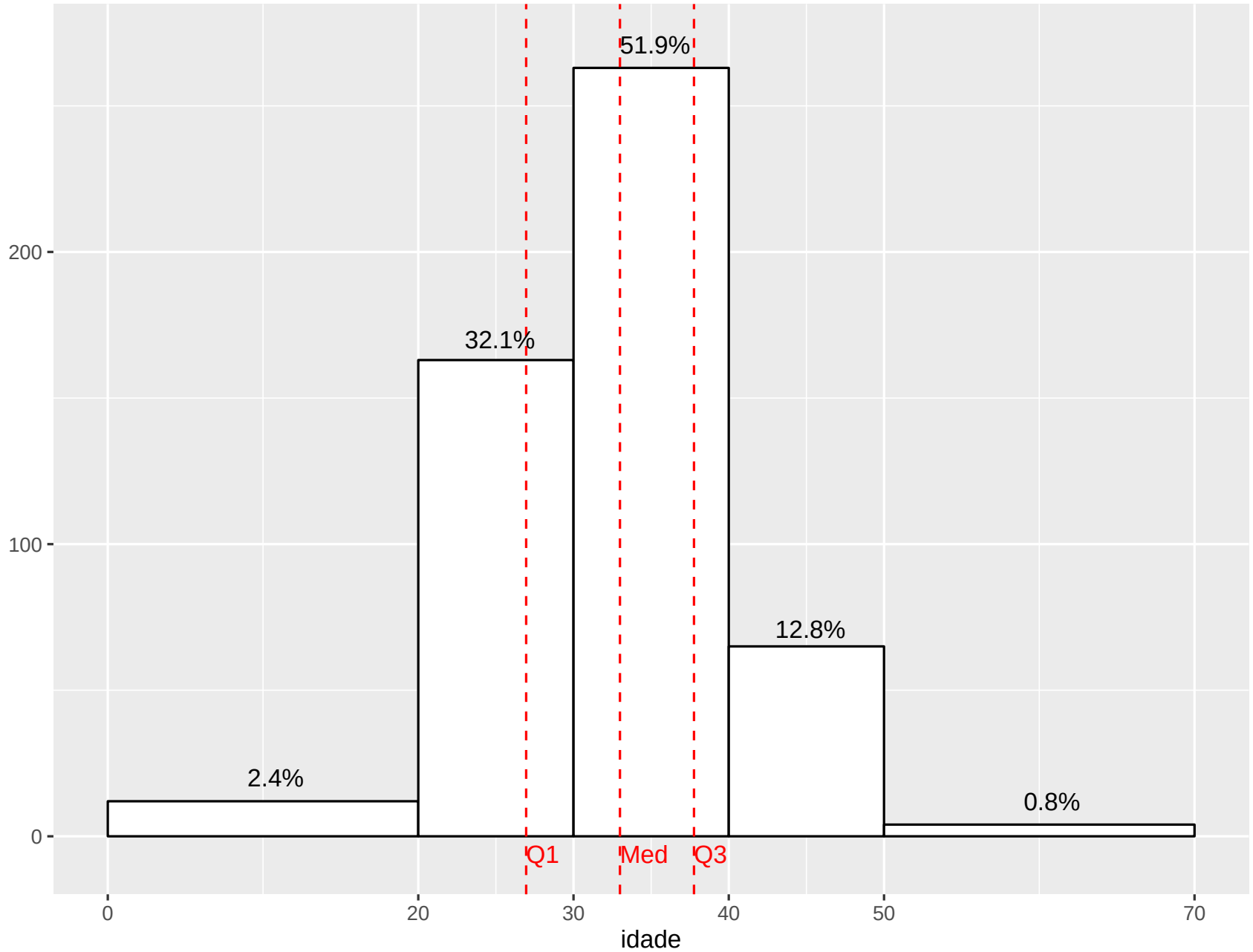
Histograma


51.9%
32.1%
12.8%
2.4%
0.8%
Densidade de frequências
0.00
0.01
0.02
0.03
0.04
0.05
0
10
20
30
40
50
60
70
Idade

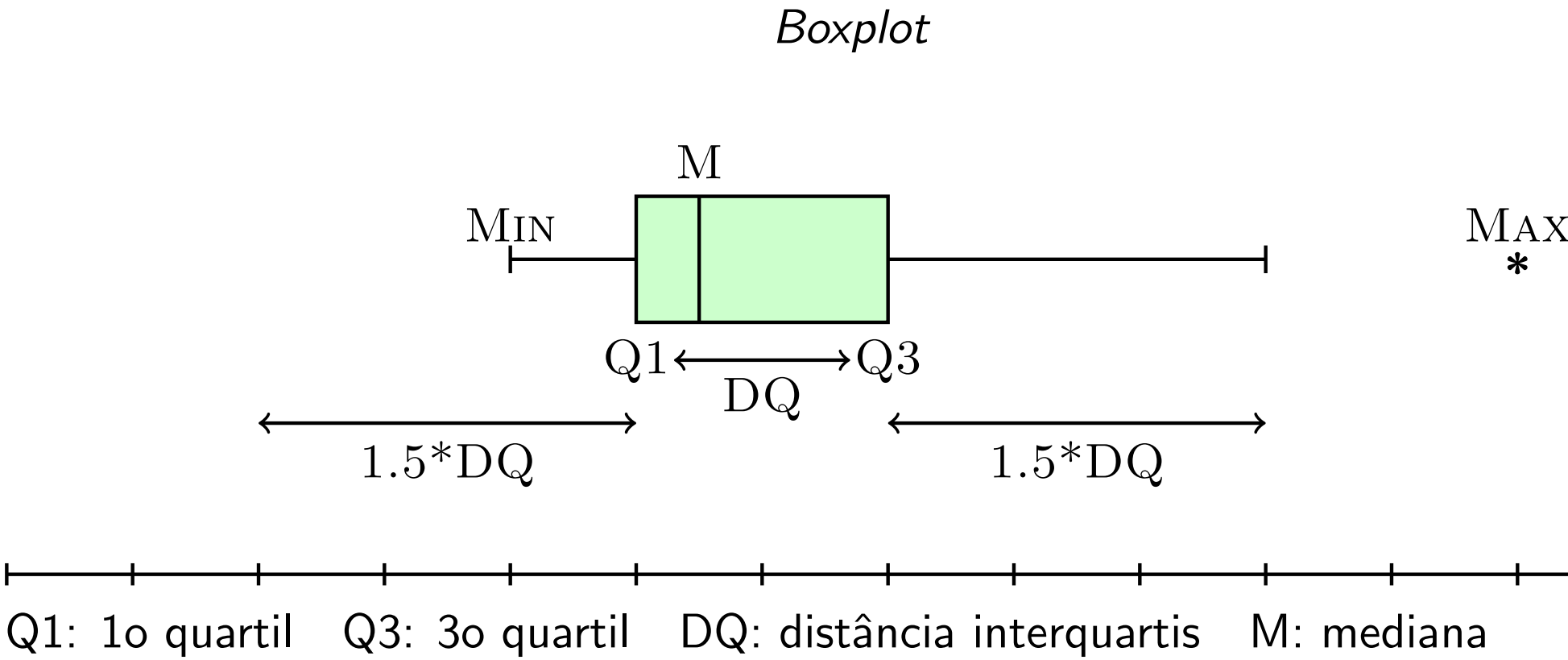
Medidas resumo (posição)

- Dados: $X_1, \ldots, X_n$
- Medidas de posição
  - Moda: valor mais frequente
  - Média: $\overline{X} = \sum_{i=1}^{n} X_i/n$
  - Mediana: 50% dos dados $< Med$ e 50% dos dados $> Med$
  - Primeiro quartil: 25% dos dados $< Q_1$; 75% dos dados $> Q_1$
  - Terceiro quartil: 75% dos dados $< Q_3$; 25% dos dados $> Q_3$
- Rendimento por pessoa ocupada no Brasil em 2012
  - Média = R$ 1.497
  - Mediana = R$ 900
  - Percentil 99 = R$ 10.000

Histograma com quartis

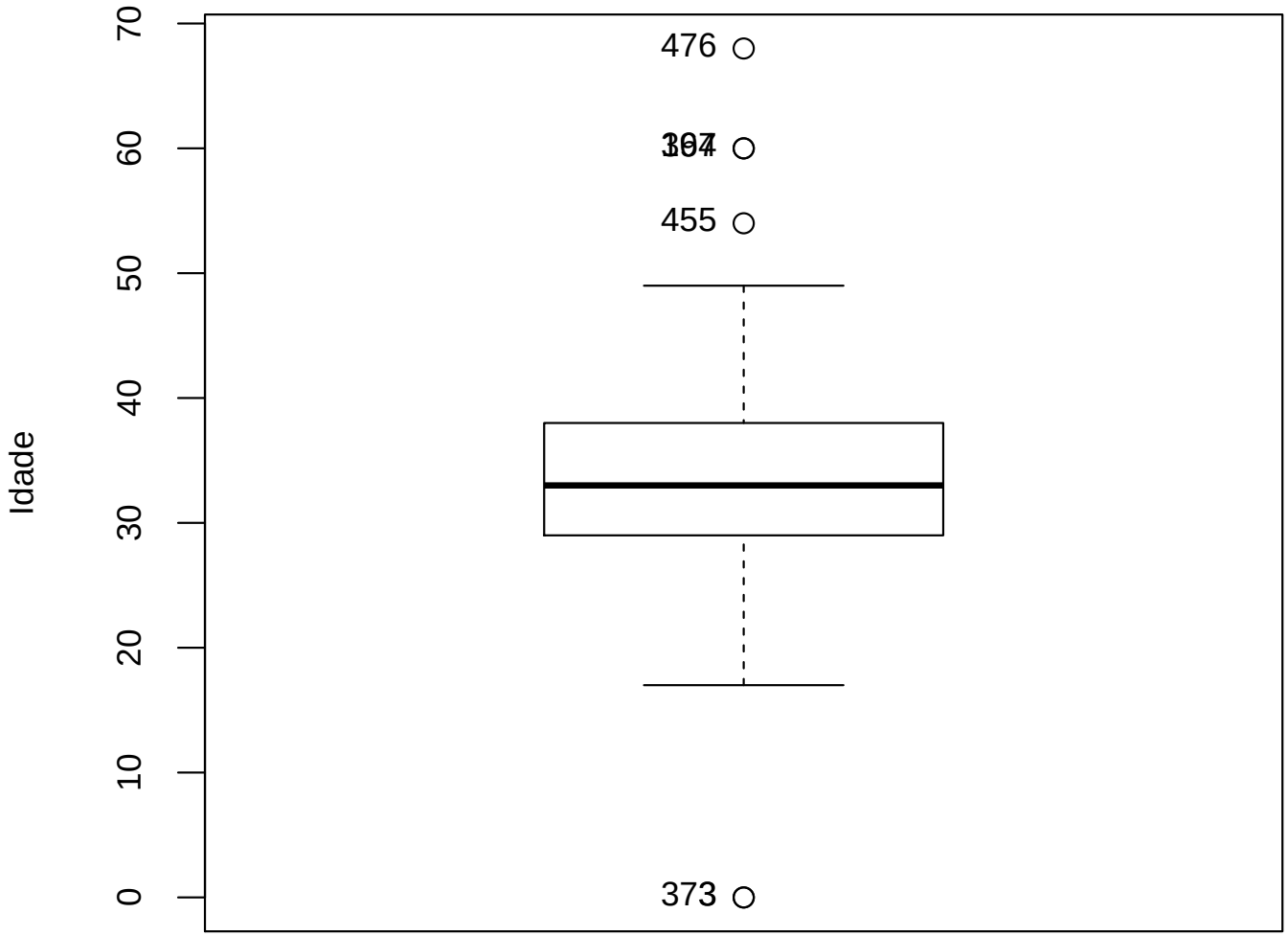


Boxplots


Boxplot
M
Min
Max
*
Q1
Q3
DQ
1.5*DQ
1.5*DQ
Q1: 1o quartil
Q3: 3o quartil
DQ: distância interquartis
M: mediana

Medidas resumo e boxplot (Idade)

Média	DP	min	Q1	Mediana	Q2	max	n
33.33	7.0	0	29	33	38	68	507



Medidas resumo (dispersão)

- Dados: $X_1, \ldots, X_n$
- Medidas de dispersão
  - Amplitude: Máximo$(X_1, \ldots, X_n)$ - Mínimo$(X_1, \ldots, X_n)$
  - Variância: $\text{Var}(X) = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$
  - Desvio padrão: $DP = \sqrt{Var(X)}$
  - Distância interquartis: $DQ = Q_3 - Q_1$
- A $= \{X_1 = 10 \text{ kg}, X_2 = 10 \text{ kg}, X_3 = 10 \text{ kg}\}$

  $\overline{X} = 10 \text{ kg}, Var(X) = 0 \text{ kg}^2, DP(X) = \sqrt{0} = 0 \text{ kg}$
- B $= \{X_1 = 5 \text{ kg}, X_2 = 10 \text{ kg}, X_3 = 15 \text{ kg}\}$

  $\overline{X} = 10 \text{ kg}, Var(X) = 16.7 \text{ kg}^2, DP(X) = \sqrt{16.70} = 4.1 \text{ kg}$
- C $= \{X_1 = 0 \text{ kg}, X_2 = 10 \text{ kg}, X_3 = 20 \text{ kg}\}$

  $\overline{X} = 10 \text{ kg}, Var(X) = 66.7 \text{ kg}^2, DP(X) = \sqrt{66.7} = 8.2 \text{ kg}$

Associação entre 2 variáveis qualitativas

Distribuição conjunta das variáveis $X=$ hipertensão arterial e $Y=$ insuficiência cardíaca

Insuficiência cardíaca	Hipertensão arterial		Total
	Tem	Não tem	
Tem	12	4	16
Não tem	20	14	34
Total	32	18	50

Porcentagens em relação ao total

Insuficiência cardíaca	Hipertensão		Total
	Tem	Não tem	
Tem	24%	8%	32%
Não tem	40%	28%	68%
Total	64%	36%	100%

Associação entre 2 variáveis qualitativas

Distribuição conjunta das variáveis $X=$ hipertensão arterial e $Y=$ insuficiência cardíaca

Insuficiência cardíaca	Hipertensão arterial		Total
	Tem	Não tem	
Tem	12	4	16
Não tem	20	14	34
Total	32	18	50

Porcentagens em relação aos totais das linhas

Insuficiência cardíaca	Hipertensão		Total
	Tem	Não tem	
Tem	75%	25%	100%
Não tem	40%	60%	100%
Total	64%	36%	100%

Associação entre 2 variáveis qualitativas

Valores esperados das frequências sob hipótese de inexistência de associação entre hipertensão e insuficiência cardíaca

Insuficiência Cardíaca	Hipertensão		Total
	Tem	Não Tem	
Tem	10.2	5.8	16
Não Tem	21.8	12.2	34
Total	32 (64%)	18 (36%)	50 (100%)

Valores obervados

Insuficiência cardíaca	Hipertensão arterial		Total
	Tem	Não tem	
Tem	12	4	16
Não tem	20	14	34
Total	32	18	50

$$\chi^2 = \sum_{i=1}^{4} (o_i - e_i)^2 / e_i$$

Associação entre fatores de risco e doença

Frequências de doentes observados num estudo prospectivo

Hábito tabagista	Câncer pulmonar		Total
	sem	com	
não fumante	80	20	100
fumante	35	15	50

Proporções de ocorrência de doença

Hábito tabagista	Câncer pulmonar		Total
	sem	com	
não fumante	$1-\pi_0$	π_0	1
fumante	$1-\pi_1$	π_1	1

π_0: proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser não fumantes

π_1: proporção (ou probabilidade) de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser não fumantes.

Associação entre fatores de risco e doença

Medidas de associação entre fatores de risco e doença

- Risco atribuível: $d = \pi_1 - \pi_0 \iff \pi_1 = d + \pi_0$
  aumento de d na proporção de doentes atribuível à exposição ao fator de risco
- Risco relativo: $r = \pi_1/\pi_0 \iff \pi_1 = r\pi_0 \iff \pi_1 = \pi_0 + (r-1)\pi_0$
  proporção de doentes entre os expostos é r vezes proporção de doentes entre os não expostos
- $\pi_0/(1-\pi_0)$: chance de indivíduo ser doente *vs.* não doente quando não exposto ao fator de risco
  $\pi_1/(1-\pi_1)$: chance de indivíduo ser doente *vs.* não doente quando exposto ao fator de risco
- Razão de chances: $\omega = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \iff \dfrac{\pi_1}{1-\pi_1} = \omega\dfrac{\pi_0}{1-\pi_0}$

Associação entre fatores de risco e doença

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

- Exemplo 1: $\pi_0 = 0.42$, $\pi_1 = 0.44$ então $d = 0.02$, $r \cong 1.05$, $\omega = 1.09$
- Exemplo 2: $\pi_0 = 0.02$, $\pi_1 = 0.04$ então $d = 0.02$, $r = 2.00$, $\omega \cong 2.04$

- $\omega = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = r\dfrac{1-\pi_0}{1-\pi_1} \approx r$, quando π_0 e π_1 são pequenos

Em geral, se trabalha com linearizações de r e ω (modelos log-lineares)

- $\log r = \log \pi_1 - \log \pi_0$
- $\log \omega = \log \pi_1 - \log \pi_0 - \log(1-\pi_1) + \log(1-\pi_0)$

Estudos prospectivo *vs.* retrospectivo

Estudo prospectivo

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

Estudo retrospectivo / caso-controle

Fator de risco	Estado do paciente	
	sem doença	doente
não exposto	$1-p_0$	$1-p_1$
exposto	p_0	p_1
Total	1	1

π_0: proporção de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser não fumantes

π_1: proporção de indivíduos que contraem câncer pulmonar dentre os que SABEMOS ser fumantes

p_0: proporção de não fumantes dentre aqueles que SABEMOS não ter câncer pulmonar

p_1: proporção de fumantes dentre aqueles SABEMOS ter câncer pulmonar

Estudos prospectivo *vs.* retrospectivo

Estudo prospectivo

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

Estudo retrospectivo / caso-controle

Fator de risco	Estado do paciente	
	sem doença	doente
não exposto	$1-p_0$	$1-p_1$
exposto	p_0	p_1
Total	1	1

Teorema de Bayes

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

Razões de chances: exemplos

Frequências de doentes observados num estudo prospectivo

Hábito tabagista	Câncer pulmonar		Total
	sem	com	
não fumante	80	20	100
fumante	35	15	50

- Risco atribuível: $d = 0.30 - 0.20 = 0.10$ (risco de câncer pulmonar aumenta de 10% para fumantes relativamente aos não fumantes)
- Risco relativo: $r = 0.30/0.20 = 1.50$ (risco de câncer pulmonar para fumantes é 1.5 vezes o risco correspondente para não fumantes)
- Chances: chance de câncer pulmonar para fumantes é $0.429 = 0.30/0.70$; chance de câncer pulmonar para não fumantes é $0.25 = 0.20/0.80$
- Razão de chances: $\omega = 0.43/0.25 = 1.72$ (chance de câncer pulmonar para fumantes é 1.72 vezes a chance correspondente para não fumantes)

Razões de chances: exemplos

Frequências de fumantes observados num estudo retrospectivo

Hábito tabagista	Câncer pulmonar	
	sem	com
não fumante	80	20
fumante	35	15
Total	115	35

- Chance de um indivíduo ser fumante SABENDO que tem câncer pulmonar é $0.751 = 0.429/0.571$
- Chance de um indivíduo ser fumante SABENDO que não tem câncer pulmonar é $0.437 = 0.304/0.696$
- Razão de chances correspondente é $\omega = 0.751/0.437 = 1.72$
- Estudo retrospectivo não permite calcular chances de câncer pulmonar para fumantes e não fumantes mas razão de chances é igual àquela calculada por meio de um estudo prospectivo
- Concluímos que a chance de câncer pulmonar para fumantes é 1.72 vezes a chance de câncer pulmonar para não fumantes apesar de não saber quanto vale cada uma delas.

Estatísticas para testes diagnósticos

Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	n_{11}	n_{12}	n_{1+}
não doente (ND)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- n_{ij}: frequência de indivíduos com o i-ésimo *status* da doença ($i = 1$ para doentes e $i = 2$ para não doentes) e j-ésimo resultado do teste ($j = 1$ para resultado positivo e $j = 2$ para resultado negativo)
- $n_{1+} = n_{11} + n_{12}$ e $n_{+1} = n_{11} + n_{21}$

Estatísticas para testes diagnósticos

Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	n_{11}	n_{12}	n_{1+}
não doente (ND)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- Sensibilidade: probabilidade de resultado positivo para pacientes doentes $[S = P(T+|D)]$ estimada por $s = n_{11}/n_{1+}$;
- Especificidade: probabilidade de resultado negativo para pacientes não doentes $[E = P(T-|ND)]$ estimada por $e = n_{22}/n_{2+}$;
- Falso positivo: probabilidade de resultado positivo para pacientes não doentes $[FP = P(T+|ND)]$ estimada por $fp = n_{21}/n_{2+}$;
- Falso negativo: probabilidade de resultado negativo para pacientes doentes $[FN = P(T-|D)]$ estimada por $fn = n_{12}/n_{1+}$;

Estatísticas para testes diagnósticos

Frequência de pacientes submetidos a um teste diagnóstico

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	n_{11}	n_{12}	n_{1+}
não doente (ND)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- Valor preditivo positivo: probabilidade de que o paciente seja doente SABENDO que o resultado do teste é positivo $[VPP = P(D|T+)]$ e pode ser estimada por $vpp = n_{11}/n_{+1}$;
- Valor preditivo negativo: probabilidade de que o paciente não seja doente SABENDO que o resultado do teste é negativo $[VPN = P(ND|T-)]$ estimada por $vpn = n_{22}/n_{+2}$;
- Acurácia probabilidade de resultados corretos $[AC = P\{(D \cap T+) \cup (ND \cap T-)\}]$ estimada por $ac = (n_{11} + n_{22})/n$.

Testes diagnósticos e prevalência

Número de pacientes submetidos a teste diagnóstico (prevalência = 15%)

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	20	10	30
não doente (ND)	80	90	170
Total	100	100	200

Número de pacientes submetidos a teste diagnóstico (prevalência = 30%)

Verdadeiro status	Resultado do teste		Total
	positivo (T+)	negativo (T-)	
doente (D)	40	20	60
não doente (ND)	66	74	140
Total	106	94	200

Testes diagnósticos e prevalência

Características do teste aplicado

	População com doença	
Característica	menos prevalente	mais prevalente
Sensibilidade	67%	67%
Especifidade	53%	53%
VPP	20%	38%
VPN	90%	79%
Acurácia	55%	57%

- Sensibilidade, especificidade: características do teste
- VPP, VPN, Acurácia: dependem da prevalência

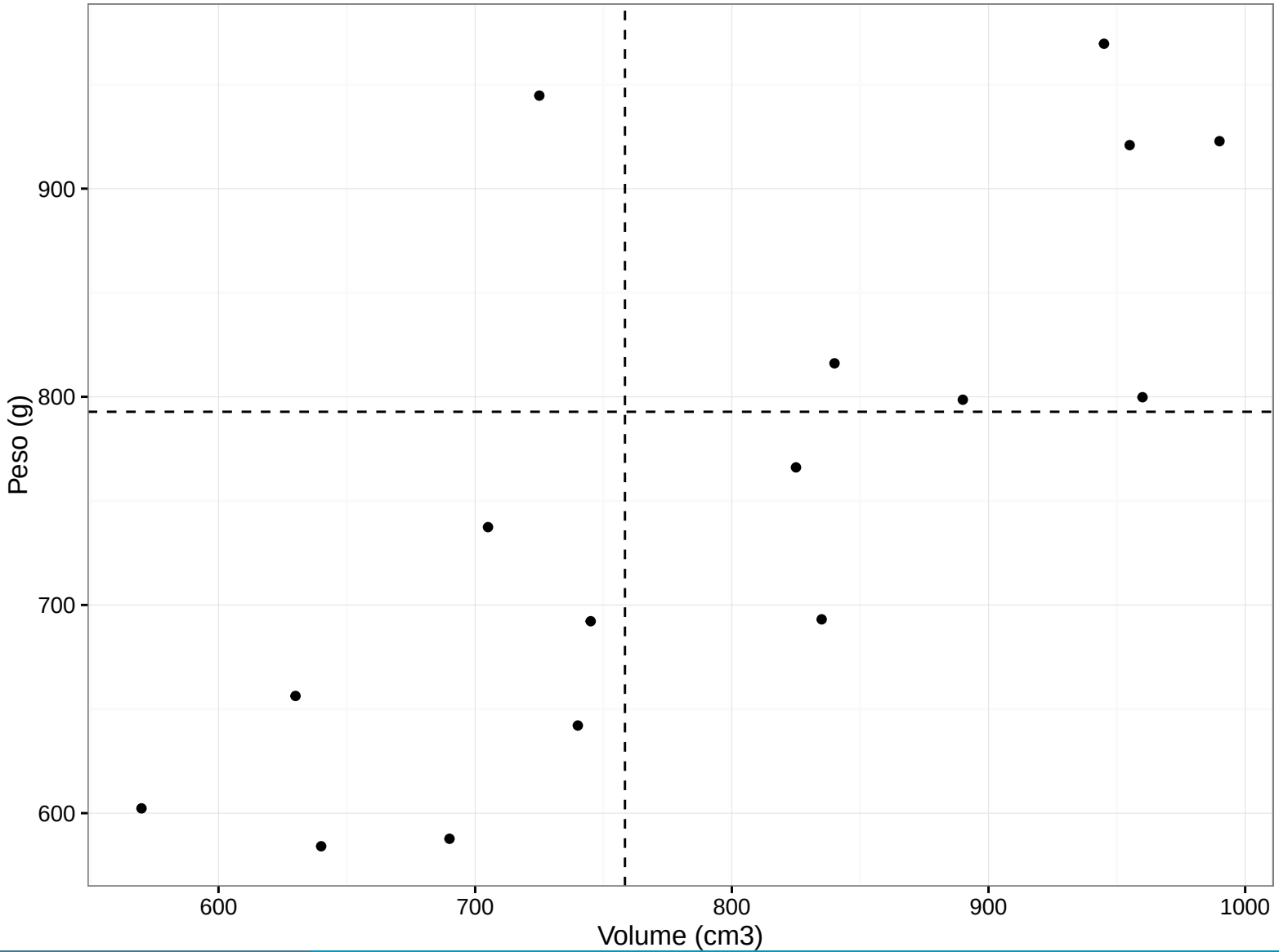
Associação entre duas variáveis quantitativas

Peso e volume do lobo direito de enxertos de fígado

Volume1 (cm^3)	Volume2 (cm^3)	Volume (cm^3)	Peso (g)
672.3	640.4	656.3	630
686.6	697.8	692.2	745
583.1	592.4	587.7	690
850.1	747.1	798.6	890
729.2	803.0	766.1	825
776.3	823.3	799.8	960
715.1	671.1	693.1	835
634.5	570.2	602.3	570
773.8	701.0	737.4	705
928.3	913.6	920.9	955
916.1	929.5	922.8	990
983.2	906.2	944.7	725
750.5	881.7	816.1	840
571.3	596.9	584.1	640
646.8	637.4	642.1	740
1021.6	917.5	969.6	945

Gráfico de dispersão

Gráfico de dispersão: peso e volume do lobo direito de enxertos de fígado



Coeficiente de correlação

- Pearson
  - Mede correlação linear
  - Varia entre -1 e +1

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{[\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2]^{1/2}}$$

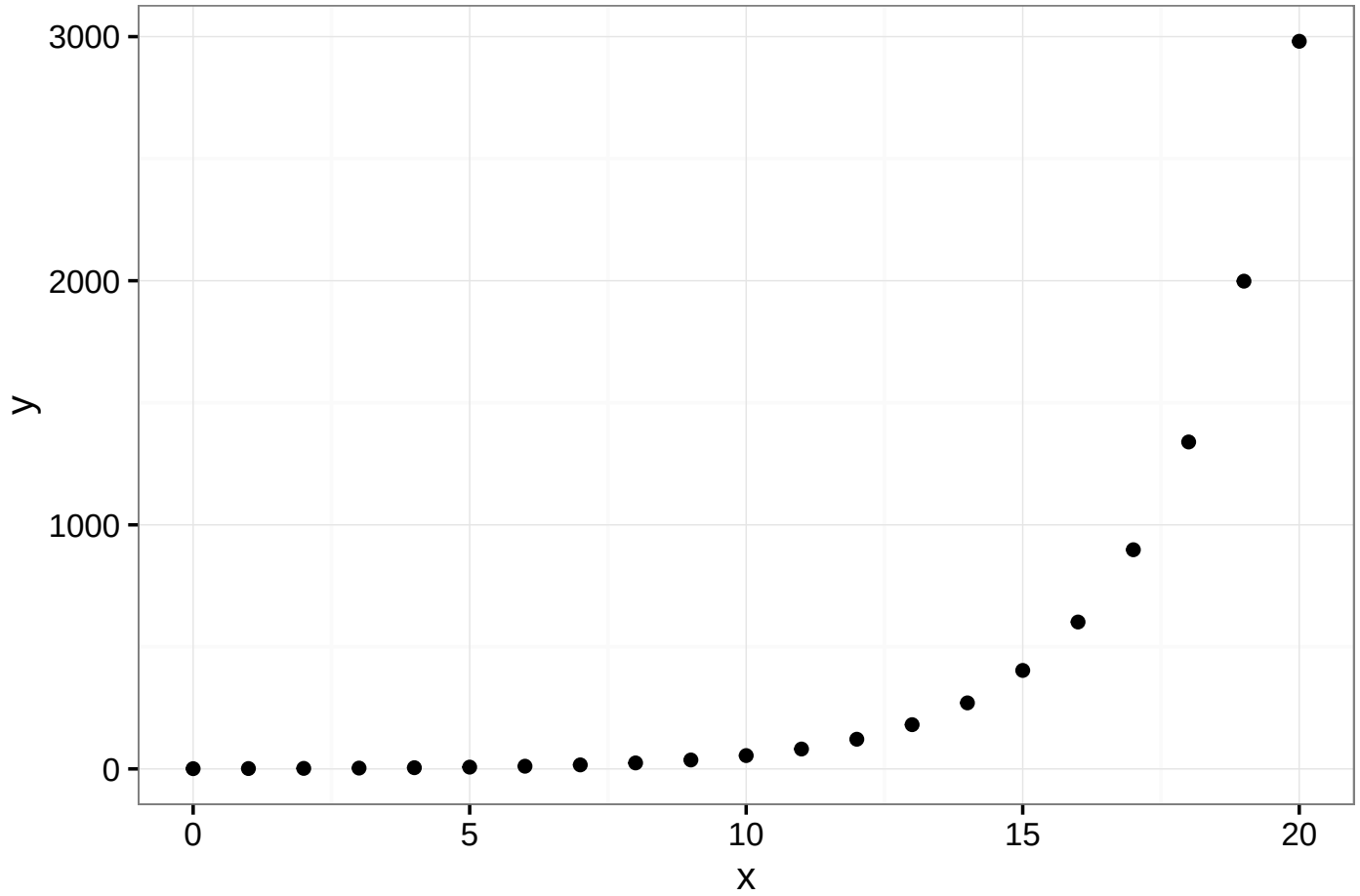
  - Exemplo fígado: $r_P = 0.76$
- Spearman
  - Similar ao coeficiente de Pearson com observações substituídas por seus postos (índice correspondente à sua posição no conjunto ordenado)
  - Varia entre -1 e +1
  - Mede correlação monotônica

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{[\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2]^{1/2}}$$

  - Exemplo fígado: $r_S = 0.75$

Coeficiente de correlação

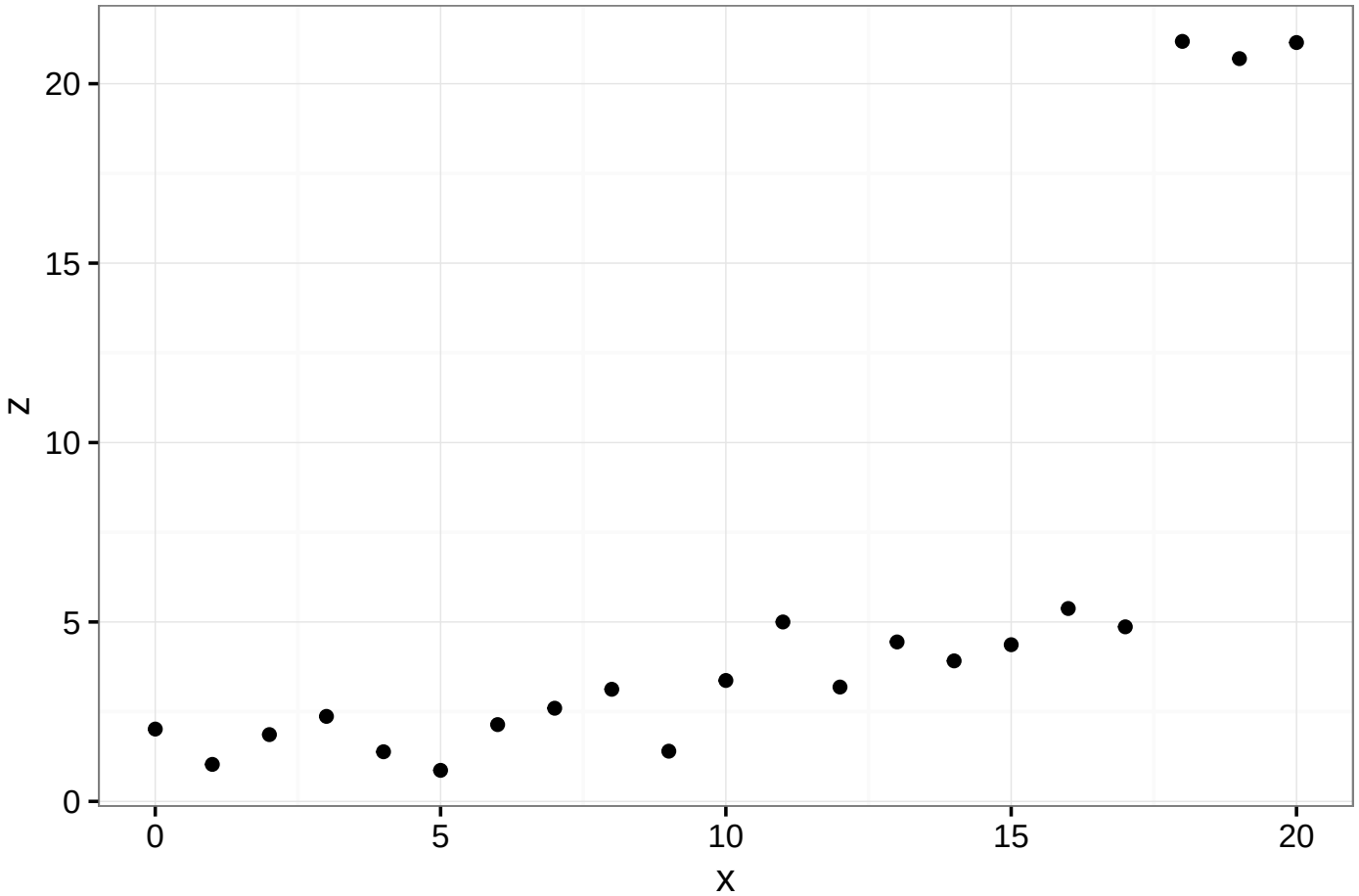
Gráfico de dispersão entre valores de duas variáveis X e Y



- Coeficiente de correlação (linear) de Pearson: $r_P = 0.75$
- Coeficiente de correlação de Spearman: $r_S = 1$

Coeficiente de correlação

Gráfico de dispersão entre valores de duas variáveis X e Z



- Coeficiente de correlação (linear) de Pearson: $r_P = 0.73$
- Coeficiente de correlação de Spearman: $r_S = 0.90$

Associação entre variáveis quantitativas e qualitativas

Medidas resumo para a concentração de Zn (ppm) em cascas de tipuanas

Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	199.4	110.9	29.2	122.1	187.1	232.8	595.8	59
Coletora	139.7	90.7	35.2	74.4	127.4	164.7	385.5	52
Local I	100.6	73.4	20.1	41.9	73.0	139.4	297.7	48
Local II	59.1	42.1	11.0	31.7	45.7	79.0	206.4	34

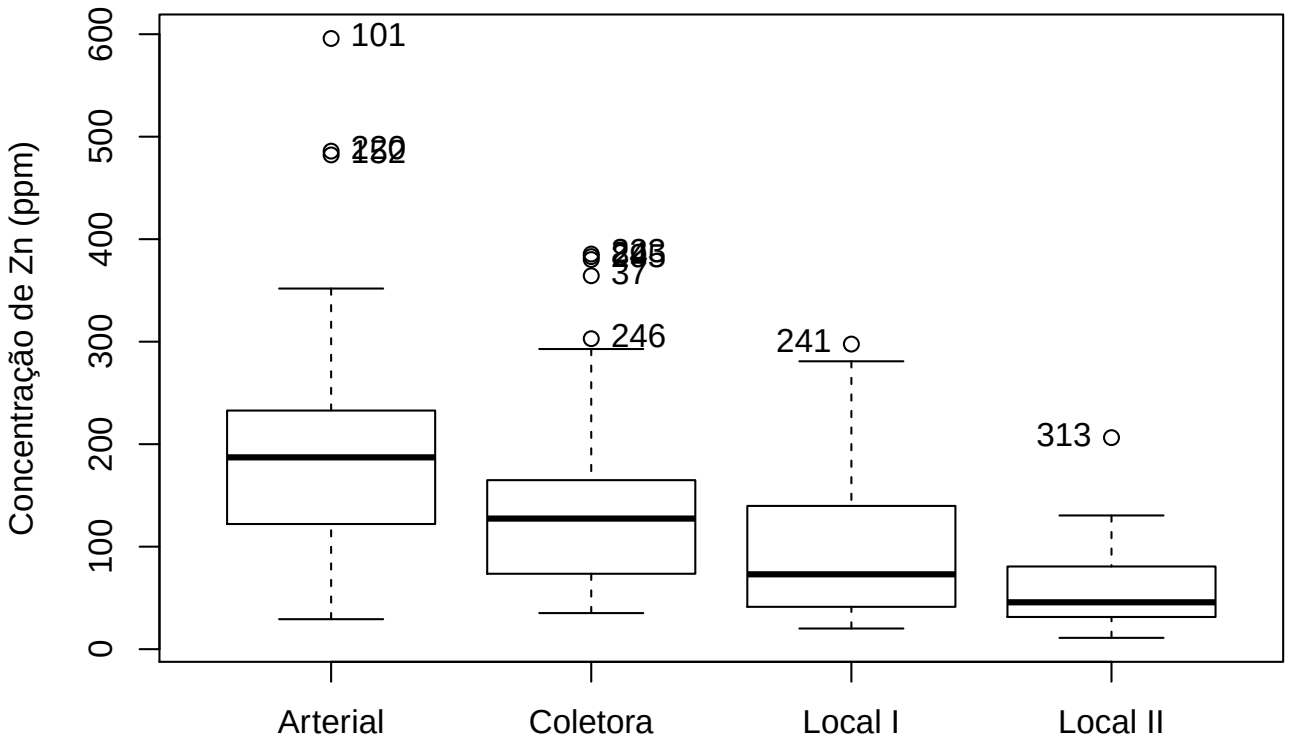
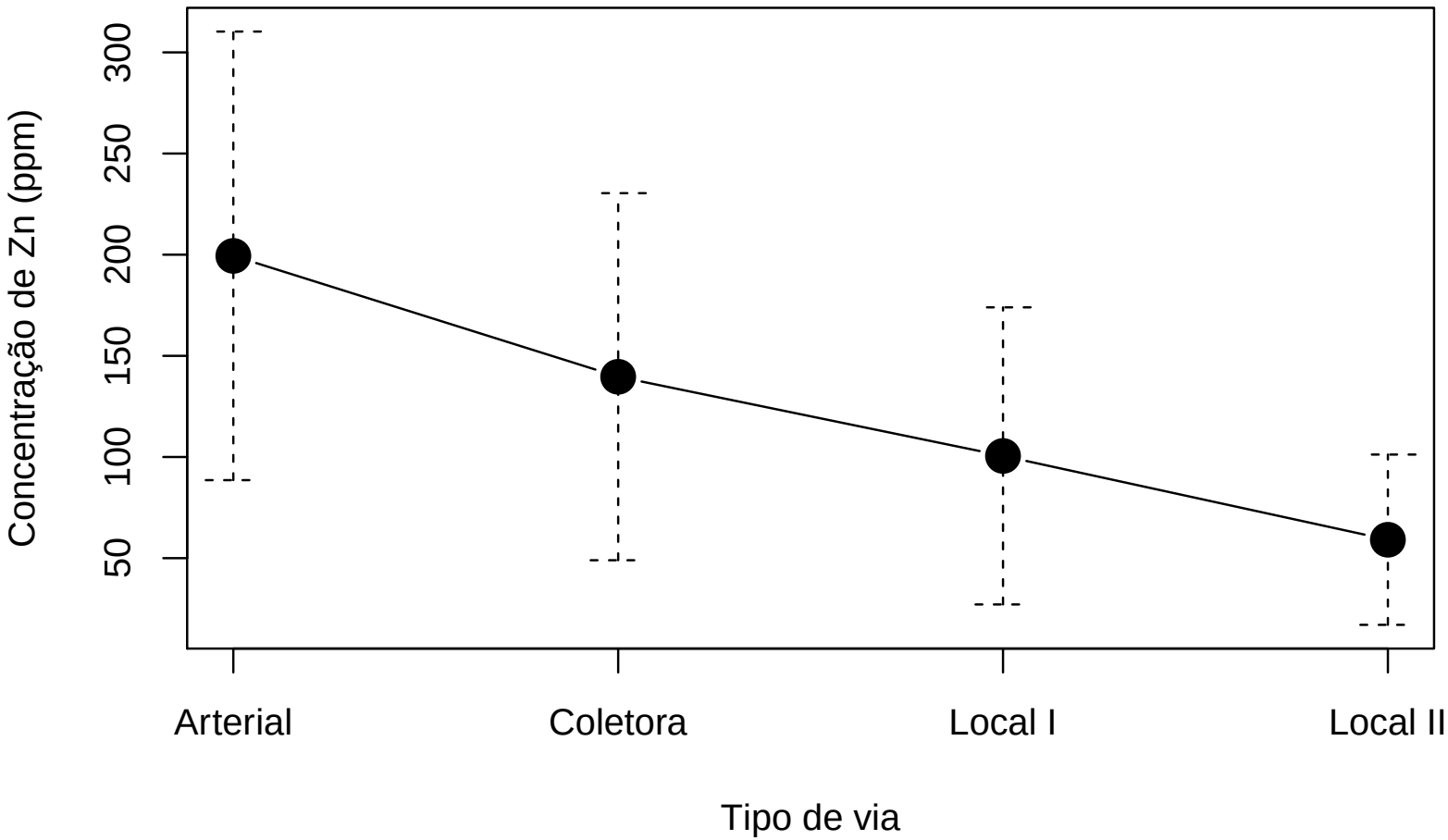


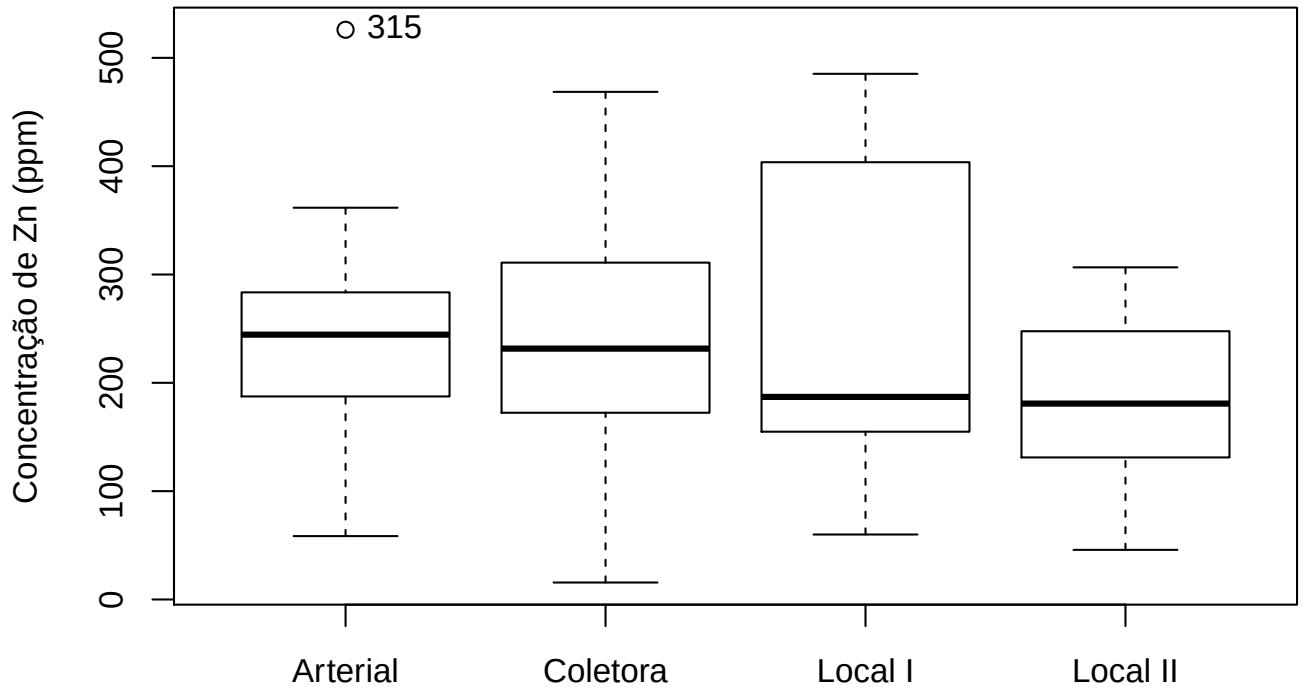
Gráfico de perfis médios (com barras de desvios padrões) para comparação das distribuições da concentração de Zn nas cascas de tipuanas



Associação entre variáveis quantitativas e qualitativas

Medidas resumo para a concentração de Zn (ppm) em cascas de alfeneiros

Tipo de via	Média	Desvio padrão	Min	Q1	Mediana	Q3	Max	n
Arterial	244.2	102.4	58.5	187.4	244.5	283.5	526.0	19
Coletora	234.8	102.7	15.6	172.4	231.6	311.0	468.6	31
Local I	256.3	142.4	60.0	154.9	187.0	403.7	485.3	19
Local II	184.4	96.4	45.8	131.1	180.8	247.6	306.6	7

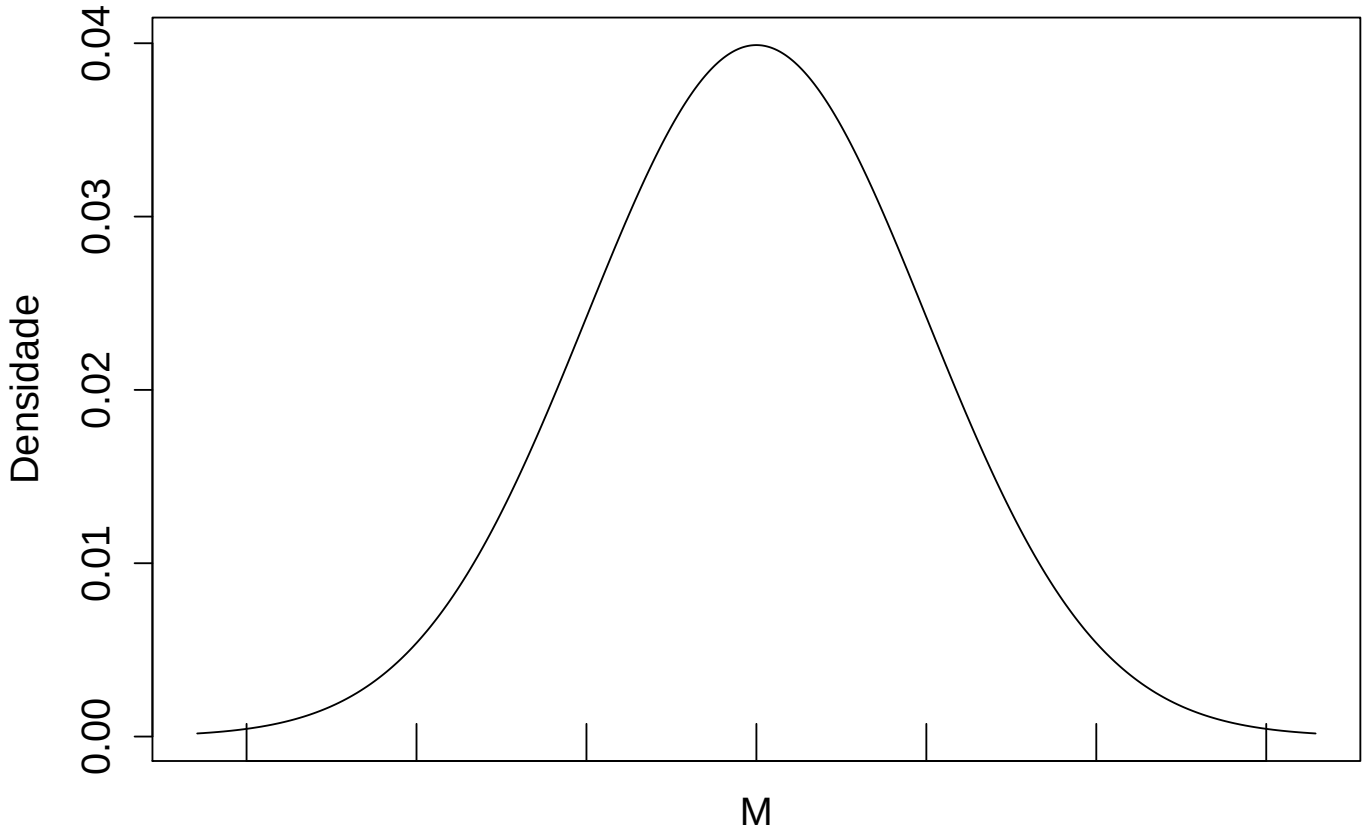


Princípios de inferência estatística

- Objetivo: generalizar conclusões obtidas numa amostra para a população (geralmente conceitual) de onde ela foi obtida
- Processo depende de modelos: única maneira de associar características da amostra àquelas da população
- Processo envolve a possibilidade de se cometerem erros (erros amostrais)
- Estatística permite a quantificação da probabilidade de se cometerem erros
- Estatística frequentista e bayesiana

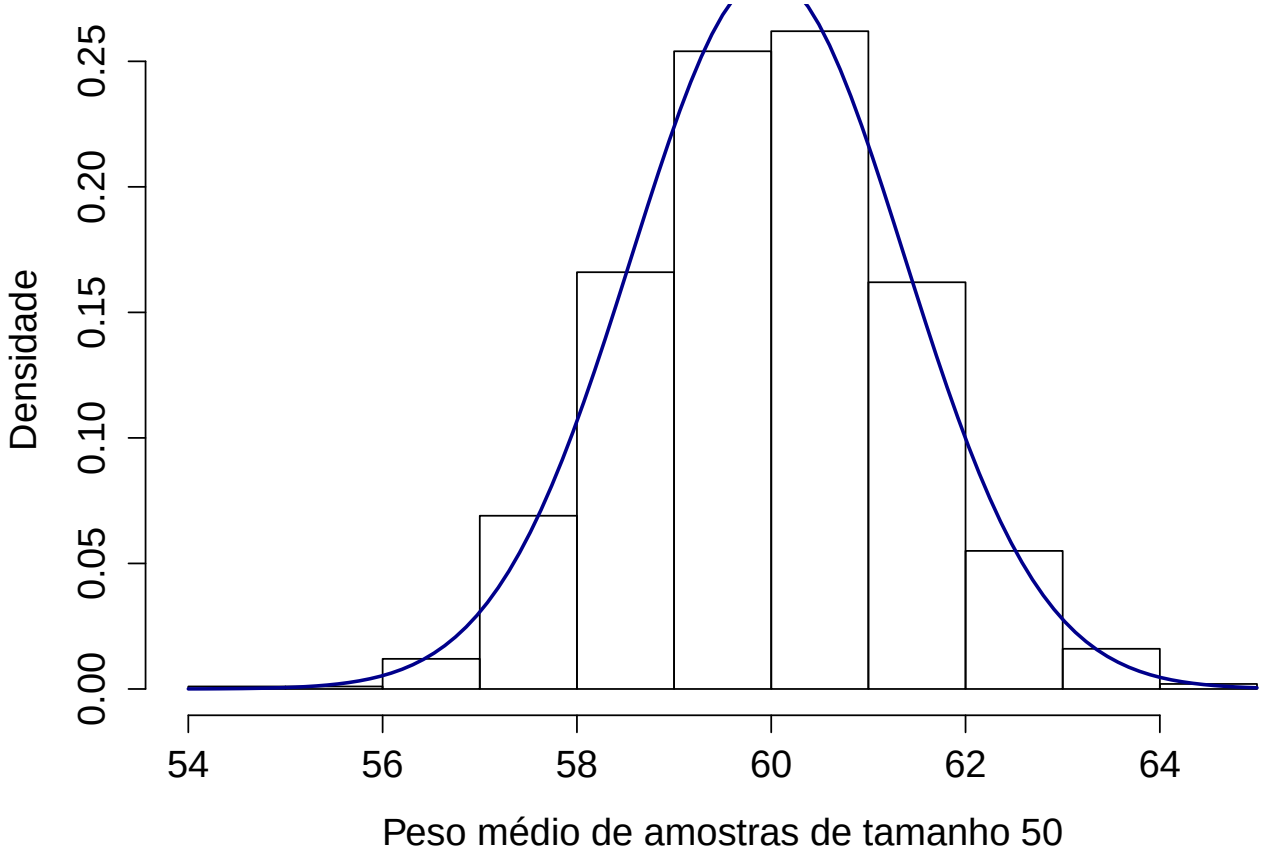
Inferência frequentista

- População (conjunto de pesos, por exemplo): $X_1, \ldots, X_N$
  - Média $= M$ (desconhecida) Pretende-se estimar por meio de amostra
  - Desvio padrão (DP) $= S$ (desconhecido)
  - Modelo probabilístico (histograma conceitual): Normal



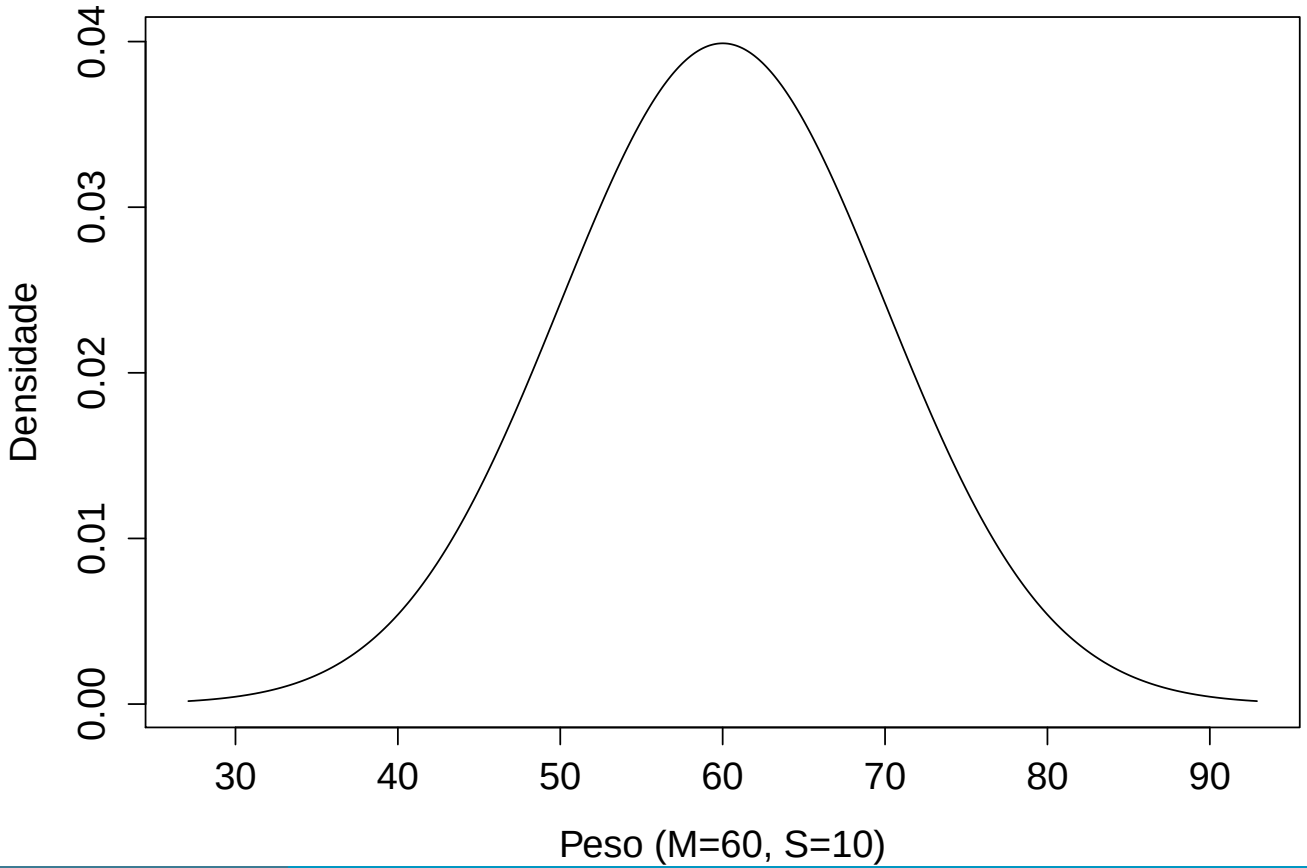
Estimação da média M

- Amostra 1 (n=50): $X_1, \ldots, X_n$ Média=61.1, DP = 11.2
- Amostra 2 (n=50): $X_1, \ldots, X_n$ Média=58.7, DP = 11.1
- ...
- Amostra 1000 (n=50): $X_1, \ldots, X_n$ Média=59.4, DP = 9.38



Erro padrão

- Média das médias das amostras (n=50) = 59.97 (M=60)
- DP das médias das amostras (n=50) = $1.41 \approx 10/\sqrt{50}$ (S=10)
- DP: característica do fenômeno (mede dispersão em torno da média)
- DP das médias (erro padrão) depende do tamanho da amostra (mede (precisão da média amostral)



Intervalos de confiança

- Intervalos de confiança: indicam a margem de erro que se pode cometer ao estimar a média (ou outra característica (populacional) por meio de uma amostra
- Formato: Estimativa $\pm$ Constante $\times$ Erro padrão da estimativa
- IC (95%) Para média M: $\overline{X} \pm 1.96 \times DP(X)/\sqrt{n}$
- $61.1 \pm 1.96 \times 11.2/\sqrt{50} = 61.1 \pm 3.1 = [58.0; 64.2]$
- Com base na média da amostra de tamanho 50 ($\overline{X} = 61.1$) concluímos que a média populacional M deve estar entre 58.0 e 64.2
- Coeficiente de confiança (95%): Repetindo esse procedimento 100 vezes, esperamos acertar em 95

Testes de hipóteses (valor p)

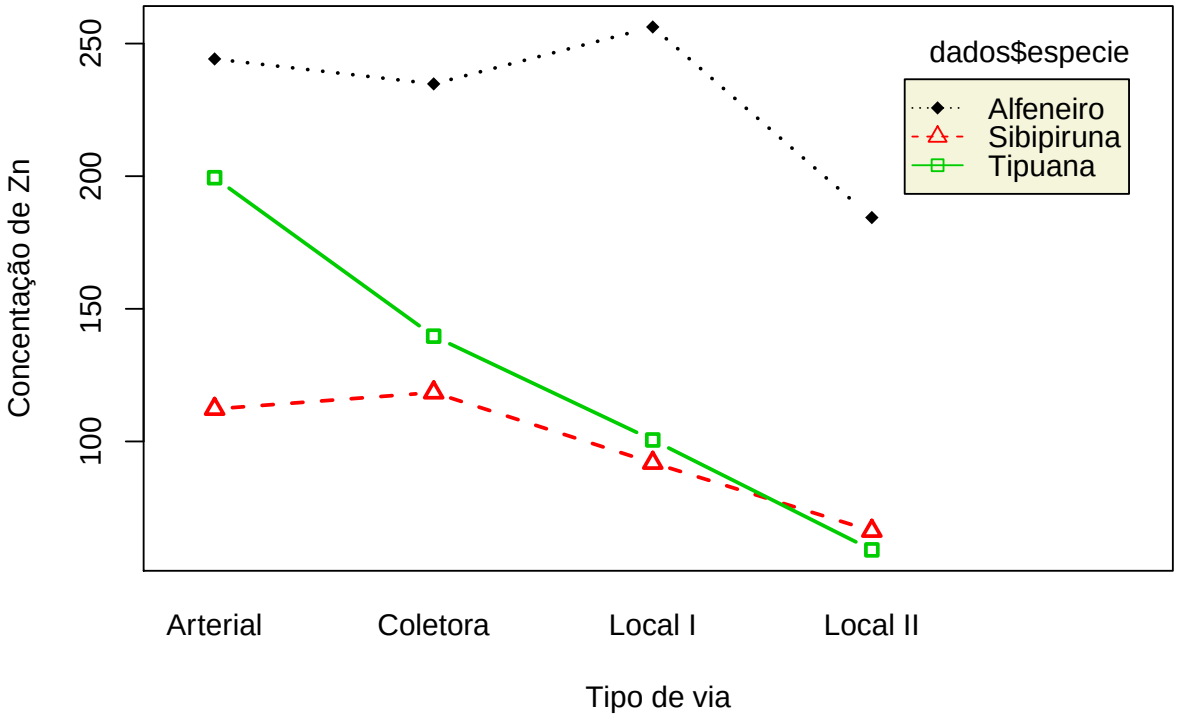
- Hipótese estatística: afirmação sobre alguma característica da população
- Exemplo: O peso médio da população é menor ou igual a 60
- Numa amostra (n=50) observamos $\overline{X} = 61.1$ (que é > 60)
- Dúvida: É possível obter esse valor numa amostra quando a média populacional é 60?
- Procedimento: Supor que a média populacional é 60 e calcular a probabilidade de observar $\overline{X} \geq 61.1$ numa amostra (n=50)
- Essa probabilidade é o valor p (p value)
- No exemplo, $p = 0.22$, ou seja, se a média populacional for 60, a probabilidade de observarmos uma média amostral (n=50) ≥ 61.1 é aproximadamente 22%
- Conclusão: com base nesses resultados não temos evidências de que a média populacional seja maior que 60

Comparação de populações (ANOVA)

- Objetivo: Comparar distribuições de frequências (populacionais)
- Suposições: distribuições normais com mesma variância (basta comparar médias)
- Hipótese: Existe "efeito" dos fatores (variáveis explicativas) nas médias populacionais da variável resposta?
- Exemplo: Avaliar a distribuição da concentração média de Zn em cascas de árvores de três espécies (alfeneiro, tipuana e sibipuruna) localizadas em vias com diferentes intensidades de tráfego (arterial, coletora, local 1 e local 2)

Fonte	GL	SQ	QM	F	$Pr(>F)$
Espécie	2	1093700	546850	68.440	$p < 0.001$
Tipo de via	3	419904	139968	17.517	$p < 0.001$
Interação	6	181407	30234	3.784	$p < 0.001$
Resíduo	485	3875272	7990		

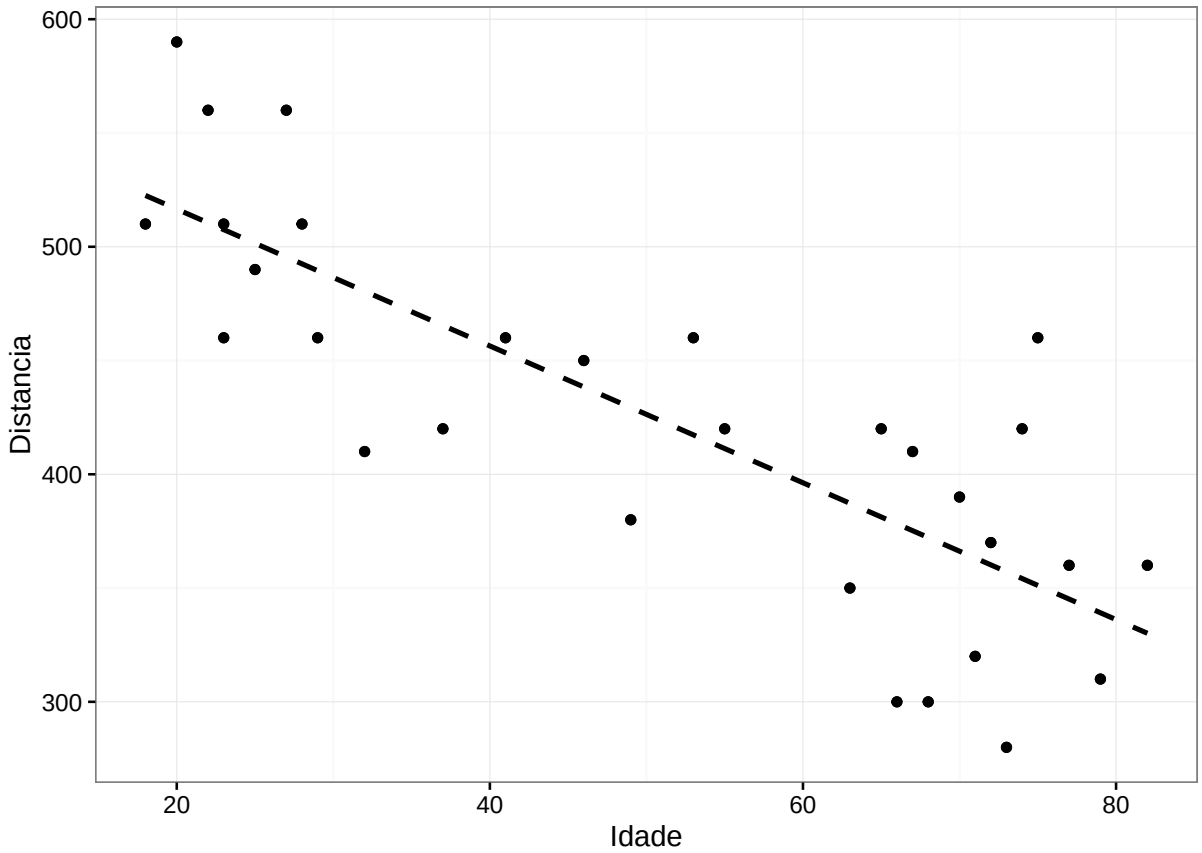
ANOVA com dois fatores (interação)



- Suposições avaliadas por meio de Estatística Descritiva (da amostra)
- Efeitos significativos: populações tem médias diferentes
- Questões subsequentes: quais são as diferenças? como quantificá-las?
- Diferenças estatísticas têm significado prático?

Regressão linear simples

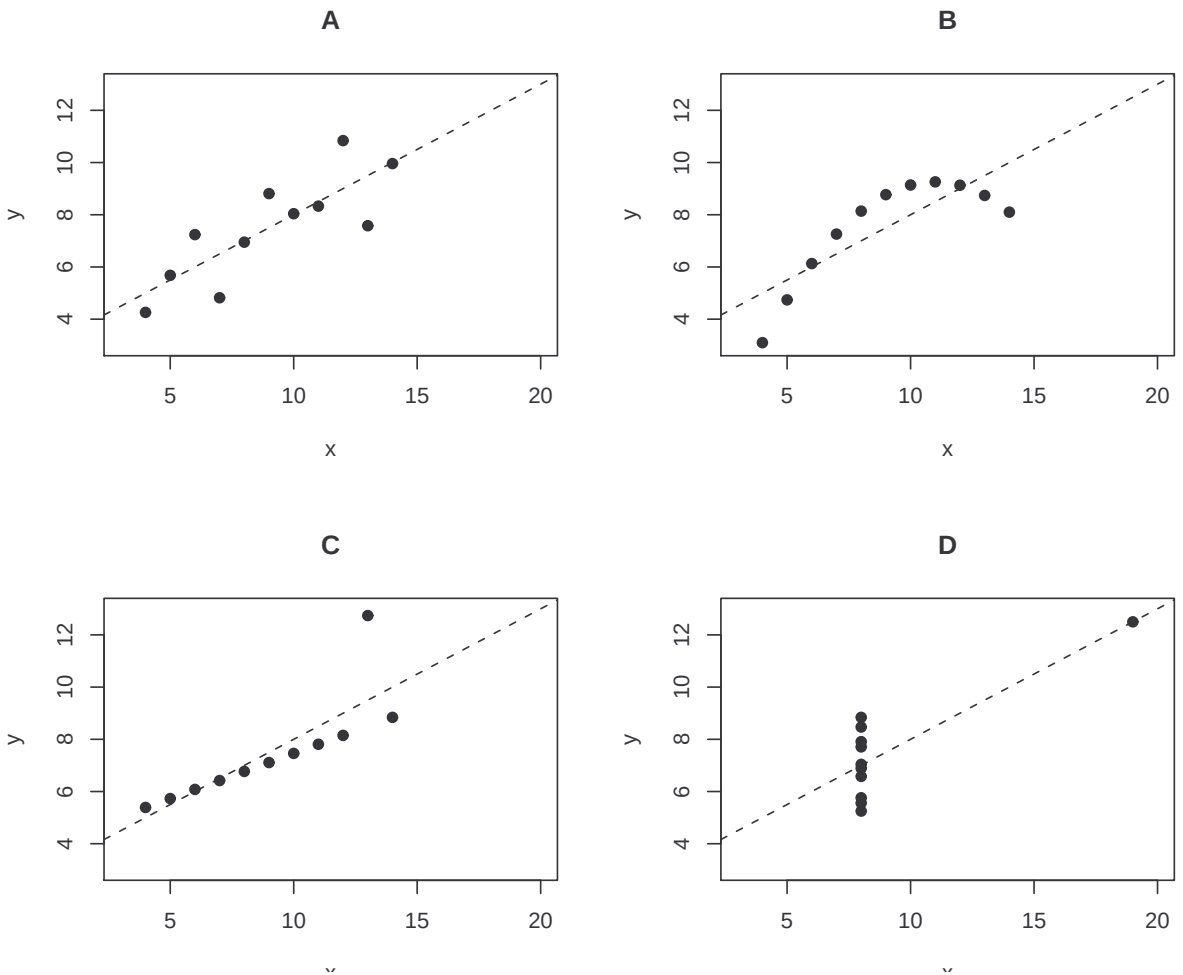
- Objetivo: Avaliar efeito de uma variável explicativa contínua na distribuição de uma variável resposta
- Exemplo: Avaliar o efeito da idade de motoristas na distância em que conseguem distinguir um objeto



Regressão linear simples

- Modelo: Distância esperada $= A + B \times$ IDADE
- Ajuste por mínimos quadrados: $R^2 = 0.63$
- Distância esperada $= 576.7\ (\pm 23.5)$ - $3.0\ (\pm 0.4) \times$ IDADE
  - A: Distância esperada para RECÉM NASCIDOS $=$ 576.7
  - B: Diminuição da distância esperada para cada aumento de 1 ano na IDADE dos motoristas $=$ 3.0
- Alternativa: Distância esperada $= A + B \times$ (IDADE - 18)
- Distância esperada $= 522.6\ (\pm 16.7)$ - $3.0\ (\pm 0.4) \times$ (IDADE-18)
  - A: Distância esperada para motoristas com 18 anos $=$ 522.6
  - Diminuição da distância esperada para cada aumento de 1 ano na IDADE dos motoristas $=$ 3.0

Cuidado com R^2



- Nos quatro casos $R^2 = 0.67$
- Análise de resíduos para avaliação da qualidade do ajuste.

Regressão linear múltipla

- Objetivo: Avaliar efeito de duas ou mais variáveis explicativas na distribuição de uma variável resposta
- Exemplo: Avaliar o efeito da CARGA na esteira e IMC no VO2max
- Modelo: VO2 esperado $= A + B$ CARGA $+ C$ IMC
  - A: VO2 esperado para sujeito com CARGA=0 e IMC=0
  - B: Variação no VO2 esperado por aumento de uma unidade na CARGA para sujeitos com mesmo IMC
  - C: Variação no VO2 esperado por aumento de uma unidade no IMC para sujeitos avaliados com mesma CARGA
- Modelo ajustado por mínimos quadrados: VO2 esperado $=$ $16.4\ (\pm 1.8) + 0.12\ (\pm 0.1)$ CARGA - $0.42\ (\pm 0.07)$ IMC
- $R^2 = 0.78$

Regressão logística univariada

- Objetivo: avaliar relação entre chance de ocorrência de um evento (INFARTO, 0=não, 1=sim) e uma variável explicativa (HAS, 0=não, 1=sim)
- Modelo: $\log(\text{chance INFARTO} \mid \text{HAS}) = A + B \times \text{HAS}$
  $(\text{chance INFARTO} \mid \text{HAS}) = \exp(A + B \times \text{HAS})$
- Interpretação dos coeficientes
  - $(\text{chance INFARTO} \mid \text{HAS} = 0) = \exp(A)$
  - $(\text{chance INFARTO} \mid \text{HAS} = 1) = \exp(A + B)$

$$\frac{\text{chance INFARTO} \mid \text{HAS} = 1}{\text{chance INFARTO} \mid \text{HAS} = 0} = \exp(B)$$

  - Chance de INFARTO para sujeitos com HAS é $\exp(B)$ vezes a chance de INFARTO para sujeitos sem HAS
  - Razão de chances (*odds ratio*) é $\exp(B)$

Regressão logística multivariada

- Modelo:
  log(chance INFARTO | HAS, IDADE) $= A + B \times$ HAS $+ C \times$ IDADE
  (chance INFARTO | HAS, IDADE) $= \exp(A + B \times$ HAS $+ C \times$ IDADE$)$
- Interpretação dos coeficientes
  - (chance INFARTO | HAS $= 0$, IDADE=K) $= \exp(A + C \times$ K$)$
  - (chance INFARTO | HAS $= 1$, IDADE=K) $= \exp(A + B + C \times$ K$)$

$$\frac{\text{chance INFARTO} \mid \text{HAS} = 1, \text{IDADE=K}}{\text{chance INFARTO} \mid \text{HAS} = 0, \text{IDADE=K}} = \exp(B)$$

  - Chance de INFARTO para sujeitos com HAS e IDADE $=$ K é $\exp(B)$ vezes a chance de INFARTO para sujeitos sem HAS e IDADE $=$ K
  - Razão de chances de INFARTO para sujeitos de mesma idade com e sem HAS é $\exp(B)$

Regressão logística multivariada

Interpretação dos coeficientes (continuação)

- (chance INFARTO | HAS = 0, IDADE=K) = $\exp(A + C \times \text{K})$
- (chance INFARTO | HAS = 0, IDADE=K+1) = $\exp[A + C \times (\text{K}+1)]$

$$\frac{\text{chance INFARTO} \mid \text{HAS} = 0, \text{IDADE}=(\text{K}+1)}{\text{chance INFARTO} \mid \text{HAS} = 0, \text{IDADE}=\text{K}} = \exp(C)$$

- Chance INFARTO para sujeitos com HAS=0 e IDADE = K+1 é $\exp(C)$ vezes a chance de INFARTO para sujeitos com HAS=0 e IDADE = K
- Chance INFARTO para sujeitos com mesmo nível de HAS (com ou sem) fica multiplicada por $\exp(C)$ para cada ano de aumento na idade
- Chance INFARTO para sujeitos de 60 anos é $\exp(10C)$ vezes a chance de INFARTO para sujeitos com 50 anos e mesmo nível de HAS
- Chance INFARTO para sujeitos de 60 anos com HAS é $\exp(B + 10C)$ vezes a chance INFARTO para sujeitos com 50 anos sem HAS

Regressão logística multivariada: exemplo

- VDD: Deficiência de vitamina D (< 20 ng/mL): 1 = sim, 0=não
- GRUPO: Controles (0 = CONT, n=76) e Infectados por HIV ou HCV (1 = INF, n=204)
- HOMAref = HOMA - 2.7 (ponto de corte)
- Modelo:
  (chance VDD | GRUPO, HOMAref) $= \exp(A + B \times$ GRUPO $+$ $C \times$ HOMAref)
- $A = -0.48 \pm 0.24$, $B = 0.91 \pm 0.28$, $C = 0.12 \pm 0.06$
- $\exp(A) = 0.62$, $\exp(B) = 2.48$, $\exp(C) = 1.13$
- chance VDD | CONT, HOMA=2.7) = 0.62, IC95% = [0.39 - 0.98]
- chance VDD | INF, HOMA=2.7) = 2.48, IC95% = [1.43 - 4.31]
- Essas chances ficam multiplicadas por 1.13 (IC95% = [1.01 - 1.27]) para cada acréscimo de uma unidade no HOMA

Curva ROC

- Regressão logística: permite calcular probabilidade de VDD dados GRUPO (X) e HOMA (W)

$$P(VDD|X,W) = \frac{\exp[A + B\ X + C\ (W - 2.7)]}{1 + \exp[A + B\ X + C\ (W - 2.7)]}$$

- Para $X = 1$, $W = 8.0$ temos $P(VDD|X = 1, W = 8.0) = 0.74$
- Para $X = 0$, $W = 4.3$ temos $P(VDD|X = 1, W = 8.0) = 0.43$
- Utilização dessas probabilidades para classificar pacientes como VDD
- Ponto de corte: classificar como VDD quando Probabilidade $> 60\%$

Tabela de resultados

Situação real	Resultado da decisão		Total
	VDD	não VDD	
VDD	n_{11}	n_{12}	n_{1+}
não VDD	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

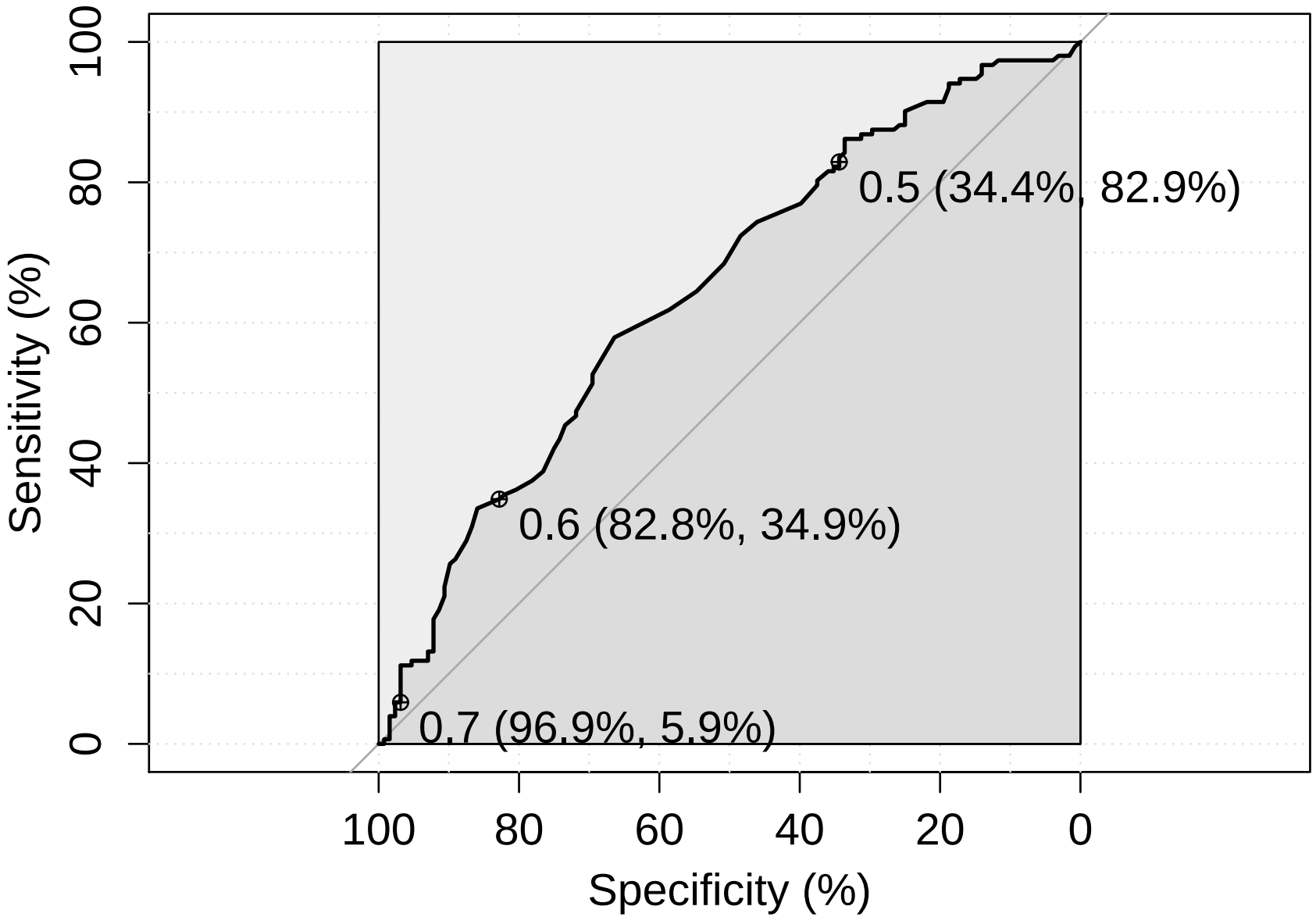
Curva ROC

Tabela de resultados

Situação real	Resultado da decisão		Total
	VDD	não VDD	
VDD	n_{11}	n_{12}	n_{1+}
não VDD	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- Sensibilidade = (1- taxa falsos negativos) = n_{11}/n_{1+}
- Especificidade = (1- taxa falsos positivos) = n_{22}/n_{2+}
- Curva ROC: gráfico de Sensibilidade *versus* Especificidade para diferentes pontos de corte
- Objetivo: selecionar ponto de corte que maximiza Sensibilidade e Especificidade

Curva ROC


Sensitivity (%)
100
80
60
40
20
0
100
80
60
40
20
0
Specificity (%)
0.5 (34.4%, 82.9%)
0.6 (82.8%, 34.9%)
0.7 (96.9%, 5.9%)