EPI5713 - Introdução ao R para a Análise de Dados

Análise bivariada

Gleice M S Conceição

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE SAÚDE PÚBLICA
DEPARTAMENTO DE EPIDEMIOLOGIA

Conteúdo

- ✓ Análise descritiva
- ✓ Teste t-Student para uma média
- ✓ Teste t-pareado para comparar duas médias, com observações dependentes
- ✓ Teste t-Student para comparar duas médias, com observações independentes

Conteúdo

- ✓ Coeficiente de correlação linear de Pearson
- ✓ Análise de regressão simples
- ✓ Análise de regressão múltipla
- ✓ Medidas de associação e medidas de risco
- ✓ Análise de regressão logística simples
- ✓ Análise de regressão logística múltipla

Tipos de Variáveis

O tipo da variável irá determinar a melhor forma de apresentar dos dados em tabelas e gráficos (análise de descritiva) e a técnica de análise inferencial mais adequada.

- **variáveis qualitativas ou categóricas**
  - **qualitativa nominal** *(sexo, tipo de patologia)*
  - **qualitativa ordinal** *(escolaridade)*
- **variáveis quantitativas ou numéricas**
  - **quantitativa discreta** *(número de cirurgias realizadas)*
  - **quantitativa contínua** *(peso, altura)*

Análise descritiva

Variáveis qualitativas

- tabelas de freqüências
- gráficos de barras
- gráfico em setores ou diagrama circular

Análise descritiva

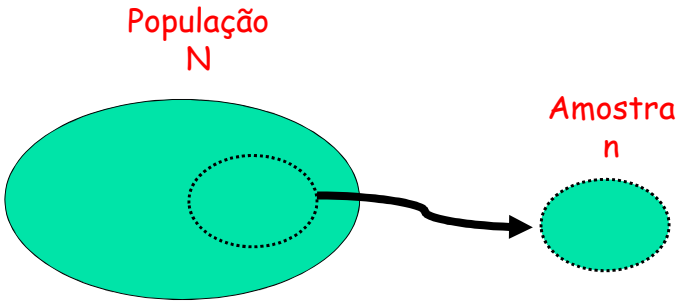
Variáveis quantitativas

- medidas de posição (média, mediana, moda)
- medidas de dispersão (variância, desvio-padrão, amplitude)
- intervalo de confiança
- box plot
- gráfico de dispersão unidimensional
- gráfico com média e intervalo de confiança
- histograma

Em particular, para as variáveis quantitativas discretas, as abordagens anteriores podem ser interessantes!

Uma única população
Teste t-Student para a média de uma população

Medidas populacionais são quase sempre desconhecidas.
Na maioria das vezes não é possível estudar a população toda.



A Estatística permite tirar conclusões sobre a população a partir do estudo de alguns de seus elementos (amostra).

Uma única população
Teste t-Student para a média de uma população

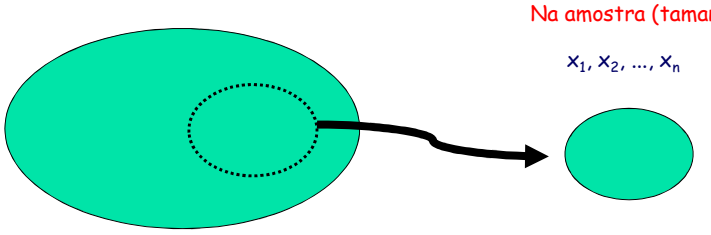
Seja X uma variável aleatória (quantitativa) de interesse

Na população (tamanho N):

$X_1, X_2, \ldots, X_N$

Na amostra (tamanho n):

$x_1, x_2, \ldots, x_n$



μ é a média de X na população = E(X)
σ^2 é a variância de X na população = Var(X)

$\overline{X}$ é a média de X na amostra
S^2 é a variância de X na amostra

Uma única população
Teste t-Student para a média de uma população

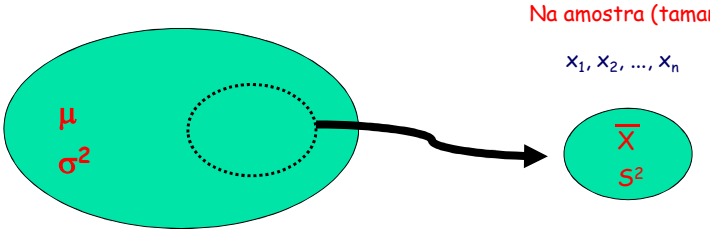
Seja X uma variável aleatória (quantitativa) de interesse

Na população (tamanho N):

$X_1, X_2, \ldots, X_N$

Na amostra (tamanho n):

$x_1, x_2, \ldots, x_n$



$$\mu = E(X) = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}$$

Teste t-Student para a média de uma população

Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica não é maior do que 23 μg por cigarro. Um laboratório realiza 32 análises desse índice, obtendo os dados abaixo.

28.80	16.87	19.39	19.69	20.30	20.67	20.84	21.48
21.57	21.79	22.19	22.95	22.97	23.05	23.23	23.50
23.63	23.76	23.78	23.82	25.57	25.64	25.71	27.91
28.48	28.61	28.64	28.69	28.75	29.33	29.38	29.90

Faça uma análise descritiva dos dados e tire conclusões iniciais.

Teste as hipóteses correspondentes.

Os dados estão armazenados em "Nicotina.xls".

Teste t-Student para a média de uma população

H_0: $\mu = 23$

H_a: $\mu > 23$

Medidas descritivas

- ✓ Média
- ✓ Desvio padrão
- ✓ Intervalo de confiança
- ✓ etc

Gráficos:

- ✓ Box plot
- ✓ Histograma
- ✓ etc

Teste t-Student para a média de uma população

H_0: $\mu = 23$

H_a: $\mu > 23$

Intervalo de confiança para μ (a concentração sérica média da substância na população)

$n = 32$

$\overline{X} = 24.40$

$S = 3.55$

$$IC(\mu;\gamma) = \overline{X} \pm t_{\gamma} S / \sqrt{n}$$

$IC(\mu, 0.95) = 24.40 \pm 2.0395 * 3.55/\sqrt{32}$

$IC(\mu, 0.95) = 24.40 \pm 1.28$

$IC(\mu, 0.95) = (23.12 : 25.68)$

Em qualquer teste de hipóteses, é importante:

- ✓ Definir a(s) variável(eis) de interesse e o(s) parâmetro(s) que a(s) caracteriza(m).
- ✓ Estabelecer as hipóteses nula e alternativa e interpretá-las.
- ✓ Identificar o parâmetro, o seu estimador e a distribuição do estimador; definir a estatística do teste (que deve conter parâmetro e estimador e ter uma distribuição conhecida) e sua distribuição.
- ✓ Especificar as suposições assumidas.
- ✓ Especificar o nível de significância (α)
- ✓ Apresentar o nível descritivo (p-valor)
- ✓ Apresentar a decisão (após comparar o p-valor com o valor de α)

Teste t-Student para a média de uma população

Variável(eis) de interesse e parâmetro(s) que a(s) caracteriza(m).

X – índice de nicotina do cigarro (μg),

com $\mathrm{E(X)} = \mu$ e $\mathrm{VAR(X)} = \sigma^2$

Estabelecer as hipóteses nula e alternativa e interpretá-las.

H_0: $\mu = 23$ - o fabricante tem razão, o índice médio de nicotina é o preconizado

H_a: $\mu > 23$ - o fabricante não tem razão, o índice médio de nicotina é maior do que o preconizado

Teste t-Student para a média de uma população

- ✓ Parâmetro: μ
- ✓ Estimador: $\bar{X}$
- ✓ Distribuição do estimador: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- ✓ Estatística do teste e sua distribuição: $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$
- ✓ Suposições assumidas: $X \sim N$

Teste t-Student para a média de uma população

$$p - valor = P(\bar{X} \geq \bar{X}_{obs} \mid H_o \text{ verdadeira}) =$$

$$= P(\bar{X} \geq 24.4 \mid \mu = 23) =$$

$$= P\left(\frac{\bar{X} - 23}{3.55/\sqrt{32}} \geq \frac{24.40 - 23}{3.55/\sqrt{32}}\right) =$$

$$= P(t_{31} \geq 2.235) = 0.0164$$

Teste t-Student para a média de uma população

α =0,05

Decisão:

Como p-valor = 0,0164, p-valor < 0,05, então rejeito H_0 e decido que o fabricante não tem razão, o índice médio de nicotina é maior do que o preconizado, isto é, $\mu > 23$ µg.

Análise Bidimensional

- Descrever o comportamento conjunto de duas variáveis
- Avaliar a existência de associação entre elas

Análise Bidimensional

1. Uma variável é qualitativa e a outra é quantitativa
2. As duas variáveis são quantitativas
3. As duas variáveis são qualitativas

Quando uma variável é qualitativa e a outra é quantitativa

Análise descritiva

- Descrever a variável quantitativa para cada categoria da variável qualitativa
- Usar tudo que aprendemos para descrever variáveis quantitativas (medidas descritivas, box-plot, histograma, etc.)

Análise inferencial

- Teste t-Student
- Teste t-pareado
- ANOVA

Quando as duas variáveis são quantitativas

Análise descritiva

- Diagrama de dispersão
- Coeficiente de correlação

Análise inferencial

- Regressão linear simples

Quando as duas variáveis são qualitativas

Análise descritiva

- Técnicas similares àquelas aprendidas na descrição de variáveis qualitativas:
- Tabelas de frequência conjunta
- Percentuais na linha e/ou coluna
- Gráfico de barras

Análise inferencial

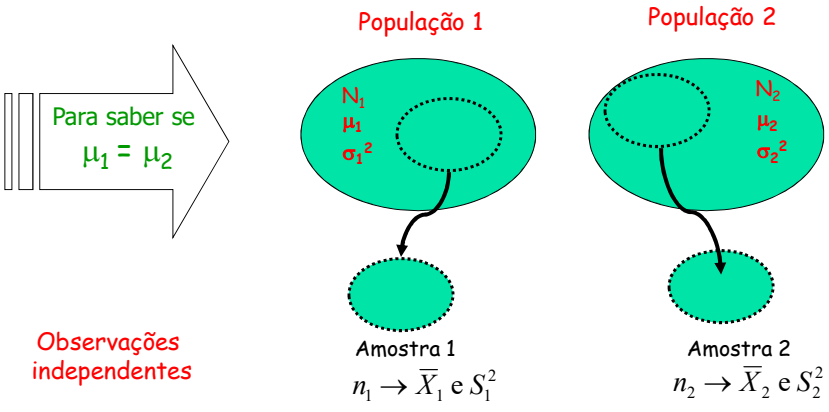
- Teste Qui-quadrado
- Medidas de risco ou de magnitude de efeito (RR, RA, RP, OR)

Quando uma variável é qualitativa e a outra é quantitativa

Comparando as médias de duas populações Teste t-Student

X_1 : v.a. de interesse na população 1, com $E(X_1) = \mu_1$ e $Var(X_1) = \sigma_1^2$

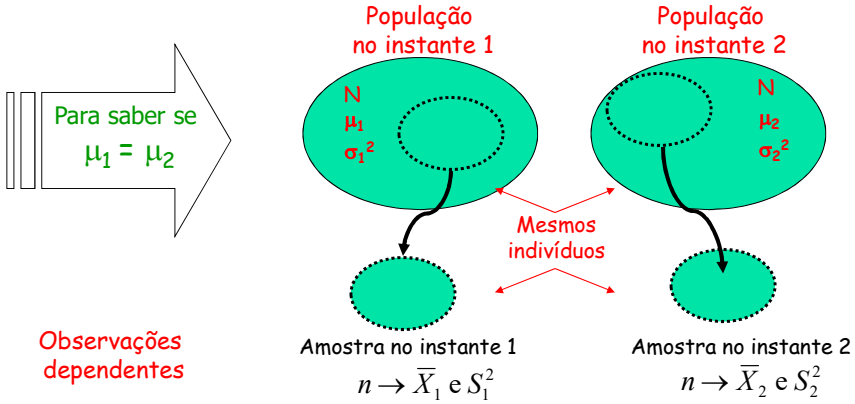
X_2 : v.a. de interesse na população 2, com $E(X_2) = \mu_2$ e $Var(X_2) = \sigma_2^2$



Comparando as médias de duas populações
Teste t-pareado

X_1 : v.a. de interesse na população, **no instante 1**, com $E(X_1) = \mu_1$ e $Var(X_1) = \sigma_1^2$

X_2 : v.a. de interesse na população, **no instante 2**, com $E(X_2) = \mu_2$ e $Var(X_2) = \sigma_2^2$



Comparando as médias de duas populações
Teste t-pareado

Deseja-se testar a hipótese de a quantidade de proteínas totais no plasma, depois de determinada operação em portadores de esquistossomose mansônica, ser diferente da quantidade antes da operação. Foi utilizada uma amostra de 17 pacientes, cujos resultados estão na tabela ao lado e no arquivo "Esquistossomose.xls"

Paciente	Antes	Depois
1	6,9	6,9
2	7,8	8,6
3	6,6	8,7
4	5,9	7,3
5	7,8	7,8
6	6,4	8,2
7	8,8	9,3
8	7,3	7,3
9	8	7,6
10	8,6	7,8
11	7,7	7,6
12	7,9	7,8
13	8,7	8,1
14	5,8	6,8
15	9,2	8,3
16	9,3	10,2
17	8,9	9,1

Comparando as médias de duas populações
Teste t-pareado

D – Diferença entre as quantidades de proteína antes e depois da operação (D= *Antes - Depois*), com $E(D) = \mu_D$ e $Var(D) = \sigma_D^2$

$$H_0: \mu_D = 0$$

$$H_a: \mu_D \neq 0$$

Média e DP de D na amostra:

$$\bar{D} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad S_D = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}$$

Paciente	Antes	Depois	D
1	6,9	6,9	0,0
2	7,8	8,6	-0,8
3	6,6	8,7	-2,1
4	5,9	7,3	-1,4
5	7,8	7,8	0,0
6	6,4	8,2	-1,8
7	8,8	9,3	-0,5
8	7,3	7,3	0,0
9	8	7,6	0,4
10	8,6	7,8	0,8
11	7,7	7,6	0,1
12	7,9	7,8	0,1
13	8,7	8,1	0,6
14	5,8	6,8	-1,0
15	9,2	8,3	0,9
16	9,3	10,2	-0,9
17	8,9	9,1	-0,2

Comparando as médias de duas populações
Teste t-pareado

Variável de interesse e o(s) parâmetro(s) que a(s) caracteriza(m).

D – Diferença entre as quantidades de proteína antes e depois da operação

(D= *Antes - Depois*),

D= $X_1 - X_2$, com $E(D) = \mu_D$ e $Var(D) = \sigma_D^2$

Hipóteses nula e alternativa.

$$H_0: \mu_1 = \mu_2 \Rightarrow H_0: \mu_1 - \mu_2 = 0 \Rightarrow H_0: \mu_D = 0$$
$$H_a: \mu_1 \neq \mu_2 \quad H_a: \mu_1 - \mu_2 \neq 0 \quad H_a: \mu_D \neq 0$$

onde $\mu_D = \mu_1 - \mu_2$

$H_0: \mu_D = 0$ A quantidade média de proteína é a mesma antes e depois ...

$H_a: \mu_D \neq 0$ A quantidade média de proteína não é a mesma antes e depois ...

Comparando as médias de duas populações
Teste t-pareado

Parâmetro: μ_D

Estimador: $\overline{D}$

Distribuição do estimador: $\overline{D} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$

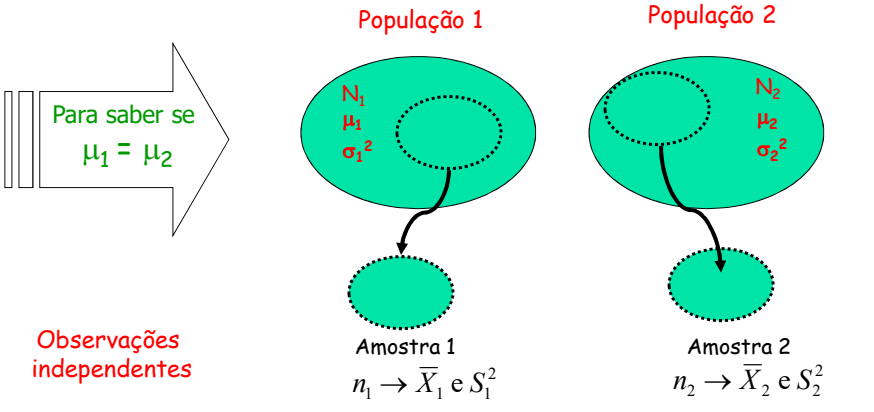
Estatística do teste e sua distribuição: $T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$

Suposições assumidas: X_1 e $X_2 \sim N$

Comparando as médias de duas populações
Teste t-Student

X_1 : v.a. de interesse na população 1, com $E(X_1) = \mu_1$ e $Var(X_1) = \sigma_1^2$

X_2 : v.a. de interesse na população 2, com $E(X_2) = \mu_2$ e $Var(X_2) = \sigma_2^2$



Comparando as médias de duas populações
Teste t-student

"Dando Branco" em Testes

Muitos estudantes já passaram pela desagradável experiência de entrar em pânico em um teste porque a primeira questão era excepcionalmente difícil. A ordem dos itens foi estudada em relação a seu efeito sobre a ansiedade. Os escores seguintes são medidas de "ansiedade de teste debilitadora", que muitos de nós chamamos de pânico ou "dar branco" (com base em dados de "Item Arrangement, Cognitive Entry Characteristics, Sex and Test Anxiety as Predictors of Achievement in Examination Performance", de Klimko, *Journal of Experimental Education, Vol. 52, No. 4*).

Questões ordenadas de Fácil para Difícil					Questões ordenadas de Difícil para Fácil			
24,64	39,29	16,32	32,83	28,02	33,62	34,02	26,63	30,26
33,31	20,60	21 13	26,69	28,90	35,91	26,68	29,49	35,32
26,43	24,23	7,10	32,86	21,06	27,24	32,34	29,34	33,35
28,89	28,71	31 73	30,02	21,96	27,62	42,91	30,20	32,54
25,49	38,81	27,85	30,29	30,72				

Comparando as médias de duas populações
Teste t-student

Variável(eis) de interesse e parâmetro(s) que a(s) caracteriza(m).

X_1 : escore de ansiedade na população 1 (questões FD), com $E(X_1) = \mu_1$ e $Var(X_1) = \sigma_1^2$

X_2 : escore de ansiedade na população 2 (questões DF) , com $E(X_2) = \mu_2$ e $Var(X_2) = \sigma_2^2$

Hipóteses nula e alternativa.

$$\begin{matrix} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 < \mu_2 \end{matrix} \Rightarrow \begin{matrix} H_0: \mu_1 - \mu_2 = 0 \\ H_a: \mu_1 - \mu_2 < 0 \end{matrix} \Rightarrow \begin{matrix} H_0: \mu_D = 0 \\ H_a: \mu_D < 0 \end{matrix} \quad \text{onde } \mu_D = \mu_1 - \mu_2$$

Comparando as médias de duas populações
Teste t-student

Parâmetro: $\mu_D = \mu_1 - \mu_2$

Estimador: $\bar{X}_D = \bar{X}_1 - \bar{X}_2$

Distribuição do estimador: $\bar{X}_D \sim N\left(\mu_D, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

Estatística do teste e sua distribuição:

✓ Se as variâncias populacionais (σ_1^2 e σ_2^2) fossem conhecidas (pouco provável!!):

$$Z = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Comparando as médias de duas populações
Teste t-student

Estatística do teste e sua distribuição:

✓ Se as variâncias populacionais (σ_1^2 e σ_2^2) forem desconhecidas, precisamos substituí-las por seus estimadores (S_1^2 e S_2^2) na expressão abaixo, chegando a uma distribuição t-Student.

$$Z = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Mas temos que considerar duas situações possíveis para as variâncias populacionais:

a) as variâncias populacionais são diferentes, isto é, $\sigma_1^2 \neq \sigma_2^2$

b) as variâncias populacionais são iguais, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Comparando as médias de duas populações
Teste t-student

Estatística do teste e sua distribuição:

a) as variâncias populacionais são diferentes, isto é, $\sigma_1^2 \neq \sigma_2^2$

Fácil! Substituímos cada variância pelo seu respectivo estimador:

$$Z = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \qquad \Rightarrow T = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\vartheta$$

$$\text{onde } \vartheta = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Comparando as médias de duas populações
Teste t-student

Estatística do teste e sua distribuição:

b) as variâncias populacionais são iguais, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Para estimar a variância única σ^2, utilizamos uma média ponderada de S_1^2 e S_2^2 :

$$S_{comb}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

E substituímos cada variância por S_{comb}^2:

$$Z = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \quad \Rightarrow T = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{S_{comb}^2}{n_1} + \frac{S_{comb}^2}{n_2}}} \sim t_{n_1+n_2-2}$$

Comparando as médias de duas populações
Teste t-student

Estatística do teste e sua distribuição:

a) variâncias diferentes $(\sigma_1^2 \neq \sigma_2^2)$

$$T = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\vartheta$$

onde $\vartheta = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

b) variâncias iguais $(\sigma_1^2 = \sigma_2^2)$

$$T = \frac{\bar{X}_D - \mu_D}{\sqrt{\frac{S_{comb}^2}{n_1} + \frac{S_{comb}^2}{n_2}}} \sim t_{n_1+n_2-2}$$

onde $S_{comb}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

Comparando as médias de duas populações
Teste t-student

Para decidir se as variâncias populacionais são iguais ou diferentes:

Teste F para comparação de variâncias de duas populações:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Comparando as médias de duas populações
Teste t-student

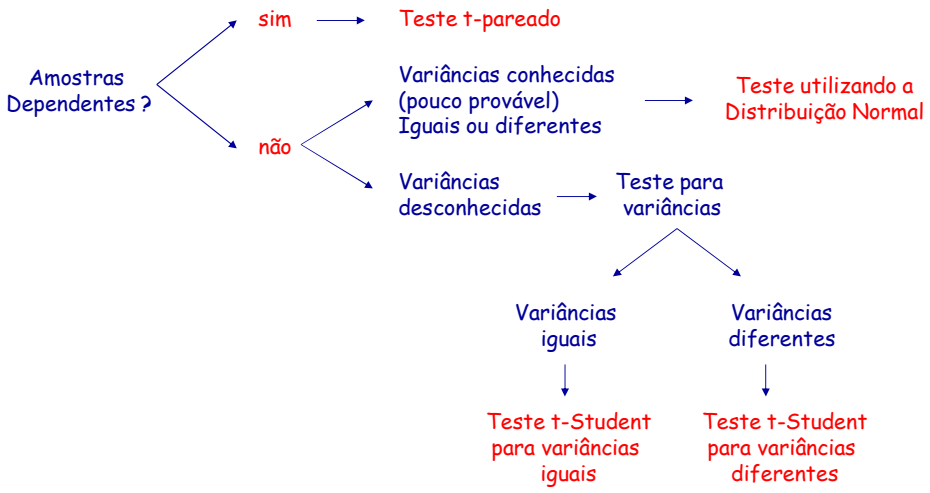
Suposições assumidas

- ✓ X_1 e X_2 têm distribuição *Normal*.
  Isto garante que $\bar{X}_1$ e $\bar{X}_2$ terão distribuição *Normal* e, consequentemente, $\bar{X}_D$ terá distribuição *Normal*.
- ✓ Se o tamanho da amostra for grande, podemos usar o TLC.

Para avaliar se X_1 e X_2 têm distribuição *Normal*:

- *Histogramas*
- *QQ-plots*
- *Testes de Normalidade*

Comparando duas médias



Comparando as médias de duas populações
Teste t-student

"Dando Branco" em Testes

Muitos estudantes já passaram pela desagradável experiência de entrar em pânico em um teste porque a primeira questão era excepcionalmente difícil. A ordem dos itens foi estudada em relação a seu efeito sobre a ansiedade. Os escores seguintes são medidas de "ansiedade de teste debilitadora", que muitos de nós chamamos de pânico ou "dar branco" (com base em dados de "Item Arrangement, Cognitive Entry Characteristics, Sex and Test Anxiety as Predictors of Achievement in Examination Performance", de Klimko, *Journal of Experimental Education, Vol. 52, No. 4*).

Questões ordenadas de Fácil para Difícil					Questões ordenadas de Difícil para Fácil			
24,64	39,29	16,32	32,83	28,02	33,62	34,02	26,63	30,26
33,31	20,60	21 13	26,69	28,90	35,91	26,68	29,49	35,32
26,43	24,23	7,10	32,86	21,06	27,24	32,34	29,34	33,35
28,89	28,71	31 73	30,02	21,96	27,62	42,91	30,20	32,54
25,49	38,81	27,85	30,29	30,72				

Quando as duas variáveis são quantitativas

Duas variáveis quantitativas

Análise de Regressão

- Duas variáveis quantitativas
- Descrever a relação entre elas
- Eventualmente, prever o valor de uma delas para um determinado indivíduo quando só conhecemos o valor da outra

Exemplos

- Volume do plasma sanguíneo e peso
- Tempo de reação a um estímulo e idade
- Perda de peso e concentração de uma determinada substância
- Número de óbitos e concentração de um determinado poluente

Análise de Regressão

✓Variável resposta, dependente ou preditiva

A variável que está sendo afetada pela outra ou outras, que acreditamos depender das outras, que pode ser explicada ou prevista pelas outras.

✓Variável explicativa, independente ou preditora

A variável que afeta a outra, que pode ajudar a explicar a variabilidade da outra e a prever a outra.

Regressão Linear Simples

Se o estudo envolver apenas uma variável explicativa, o método será chamado de Regressão Linear Simples.

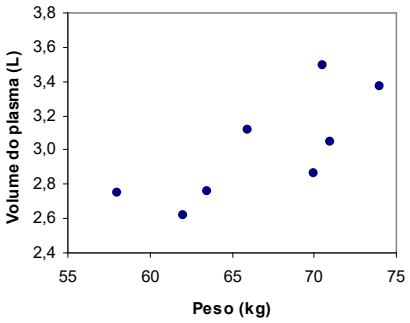
Exemplo

Volume do plasma sanguíneo (L) e peso (Kg)

Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	**3,0**	**66,9**

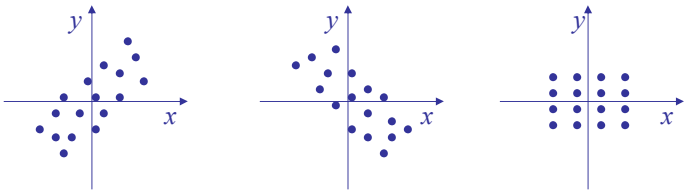
Peso e volume do plasma em 8 indivíduos saudáveis

	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	**3,0**	**66,9**



Tipos de associação entre variáveis

(a) (b) (c)



Por exemplo, $\frac{x \cdot y}{n}$

Coeficiente de correlação

$$r = corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{X})}{S_X}\frac{(y_i - \bar{Y})}{S_Y}$$

isto é, a média dos produtos dos valores padronizados das variáveis X e Y.

Coeficiente de correlação

ou

$$corr(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

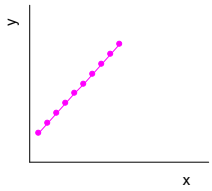
Coeficiente de correlação

- $-1 \leq corr(X,Y) \leq 1$
- Valores próximos de 1 ou –1 indicam uma associação forte
- Valores próximos de zero quando não existe associação

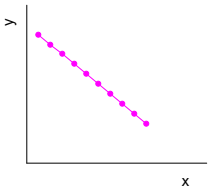
O coeficiente de correlação mede:

- Presença de associação linear
- Força de uma associação linear

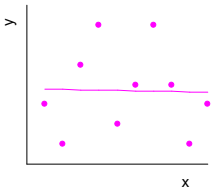
Coeficiente de correlação



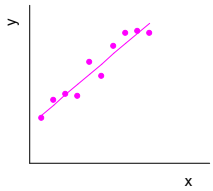
r = 1



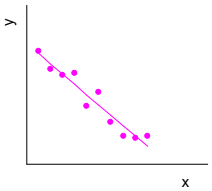
r = -1



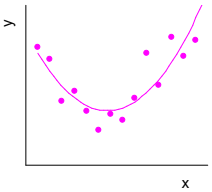
r próximo de 0



r próximo de 1



r próximo de -1



r próximo de 0

Coeficiente de correlação

Alguns autores sugerem avaliar a presença de associação linear a partir do coeficiente de correlação do seguinte modo:

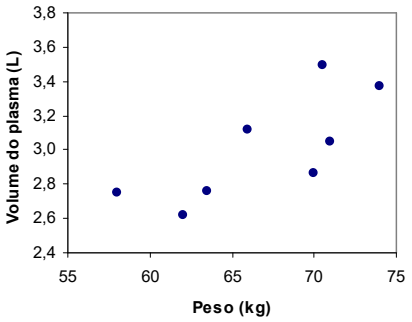
- de 0,10 a 0,39 - fraca
- de 0,40 a 0,69 - moderada
- de 0,70 até 1 - forte

Mas não há, de fato, uma norma rígida sobre isto.

Deve-se levar em conta o contexto e sempre avaliar a associação observando conjuntamente o coeficiente de correlação e o diagrama de dispersão.

Peso e volume do plasma em 8 indivíduos saudáveis

	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	**3,0**	**66,9**

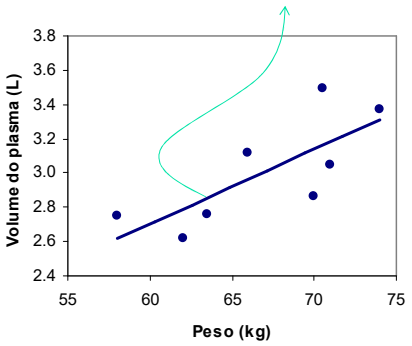


corr(X,Y) = 0,76

Peso e volume do plasma em 8 indivíduos saudáveis

Reta de Regressão

	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	**3,0**	**66,9**



corr(X,Y) = 0,76

Regressão Linear Simples

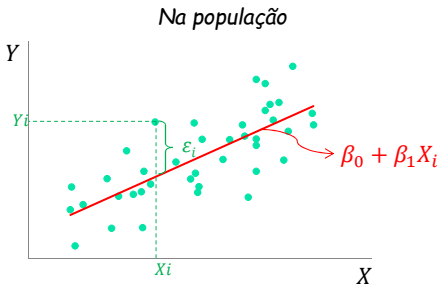
O modelo de regressão linear simples pode ser escrito como:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

onde

$\varepsilon_i \sim N(0, \sigma^2)$

$(\varepsilon_i, \varepsilon_i)$ independentes para todo *i, j*.



Regressão Linear Simples

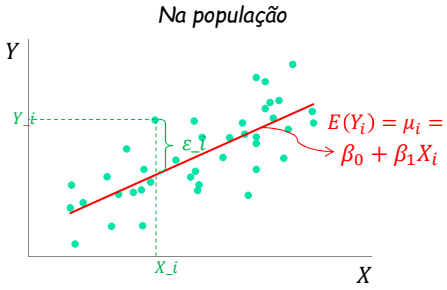
Isso equivale a:

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$$

onde

$Y_i \sim N(\mu_i; \sigma^2)$

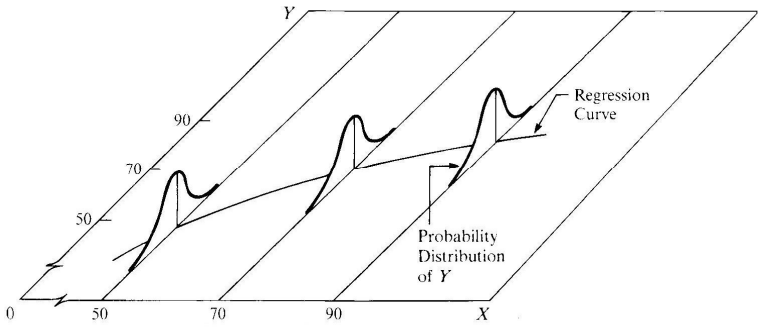
(Y_i, Y_i) independentes para todo *i, j*.



Ou seja,

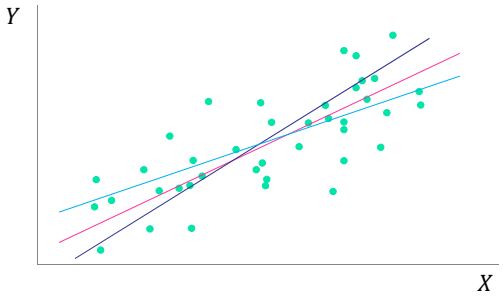
- ✓ *O valor médio ou esperado de Y (μ_i) é uma função de linha reta sobre os Xi.*
- ✓ *Para cada valor de X, Y é uma v.a. com distribuição normal* $\rightarrow Y_i \sim N(\mu_i\,;\,\sigma^2)$
- ✓ *A variância de Y é a mesma, qualquer que seja X.* $\rightarrow Var(Y_i) = \sigma^2$, *constante*

Regressão Linear Simples



Representação gráfica do modelo de regressão linear simples

Regressão Linear Simples



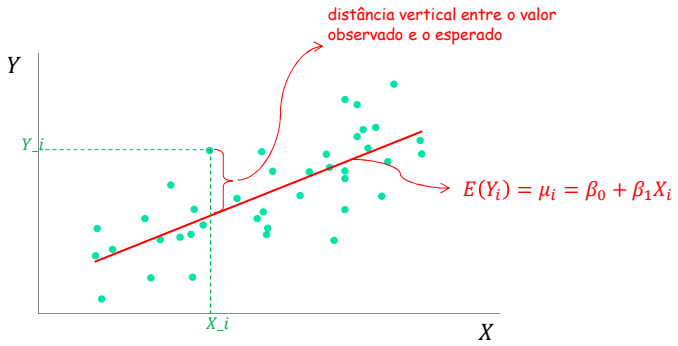
$\beta_0 = ?$

$\beta_1 = ?$

Como escolher a melhor reta ?

Estimação dos parâmetros

Método de Mínimos Quadrados



Método de Mínimos Quadrados

Para cada observação, consideramos a distância entre o valor de Y que foi observado e o valor de Y esperado (ou médio, aquele que é previsto pela reta)

$$Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

E procuramos valores de β_0 e β_1 que minimizam essa distância:

$$Q = \sum[Y_i - (\beta_0 + \beta_1 X_i)]^2 = \sum[Y_i - \beta_0 - \beta_1 X_i]^2$$

Método de Mínimos Quadrados

Os estimadores de β_0 e β_1 serão os valores b_0 e b_1 para os quais Q é a menor possível na amostra que está sendo considerada:

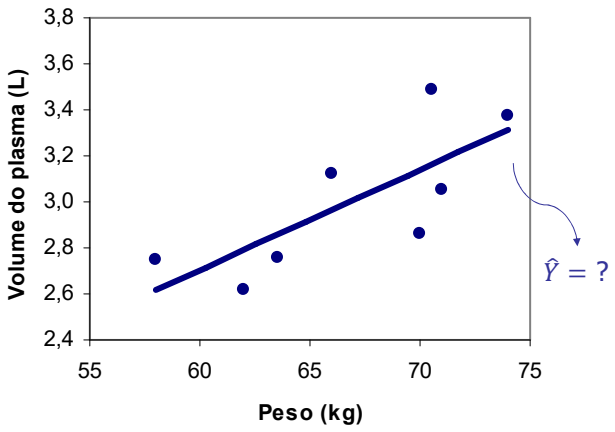
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

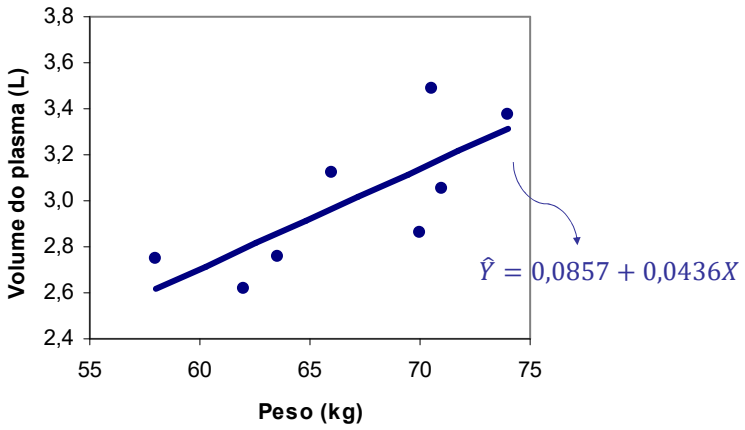
A reta estimada será

$$\hat{Y}_i = b_0 + b_1 X_i$$

Peso e volume do plasma em 8 indivíduos saudáveis



Peso e volume do plasma em 8 indivíduos saudáveis



Interpretação dos coeficientes

- ✓ b_0

$\hat{Y} = b_0 + b_1 X$

Se $X = 0 \Rightarrow \hat{Y} = b_0$

Então, b_0 é o valor esperado (ou médio) de Y quando X=0, é o intercepto da reta ajustada.

Interpretação dos coeficientes

✓ b_1

$\hat{Y} = b_0 + b_1 X$

Se aumentarmos X em uma unidade

$\hat{Y}_{novo} = b_0 + b_1(X + 1)$

$\hat{Y}_{novo} = b_0 + b_1 X + b_1$

$\hat{Y}_{novo} = \hat{y} + b_1$

Então, b_1 é o aumento esperado *(ou médio)* em *y* quando aumentamos *X* de 1 unidade, é o "efeito" de *X* em *Y*.

Interpretação dos coeficientes

$$\hat{Y} = 0{,}0857 + 0{,}0436X$$

Um aumento de um 1kg no peso corresponderá a um aumento de 0,04 litros de plasma, em média.

Partição da Soma de Quadrados

A Soma de Quadrados Total (SQT) pode ser particionada da seguinte forma

$$SQT = SQ\,\mathrm{Re}\,g + SQR$$

onde:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{com (n-1) g.l.}$$

$$SQ\,\mathrm{Re}\,g = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{com 1 g.l.}$$

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{com (n-2) g.l.}$$

Coeficiente de determinação ou explicação

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQT}$$

O R^2 Mede a proporção da variabilidade total que é explicada pelo modelo adotado.

No modelo de Regressão Linear Simples, o R^2 é igual ao coeficiente de correlação (r) ao quadrado.

Quadro de Análise de Variância

Fonte de variação	g.l.	SQ	QM	$E(QM)$	F_0	$p-valor$
Regressão	1	SQReg	QMReg	$\sigma^2+\beta_1^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	$\frac{QM\,\mathrm{Re}\,g}{QMR}\sim F_{(1,n-2)}$	
Resíduo	n - 2	SQR	QMR	σ^2		
Total	n - 1	SQT				

Estabelecendo as hipóteses

$$H_o:\beta_1=0$$

$$H_a:\beta_1\neq 0$$

A estatística do teste é

$$T=\frac{b_1-\beta_1}{\sqrt{\frac{QMR}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}}\sim t_{n-2}$$

Verificando se o modelo está bem ajustado

- Análise de resíduos
- Resíduo: a distância entre o valor de Y observado e o valor de Y ajustado pelo modelo.
- $e_i = Y_i - \hat{Y}_i$
- Os resíduos do modelo devem estar aleatoriamente dispersos em torno de zero
- O modelo supõe que os resíduos têm distribuição Normal, com média zero e variância constante
- Para checar essas suposições: Gráficos

Regressão Linear Múltipla

- ✓ Similar ao modelo simples
- ✓ Interpretação dos parâmetros é semelhante
- ✓ Somas de quadrados e R2 obtido da mesma forma
- ✓ Etc...

Regressão Linear Múltipla

Exemplo

A depuração da creatinina (Y) é uma medida importante da função renal, mas é difícil de obtê-la em um consultório clínico, pois requer coleta da urina de 24 horas. Para determinar se esta medida pode ser prevista a partir de alguns dados mais facilmente disponíveis, um especialista em rins avaliou 33 indivíduos do sexo masculino, obtendo informações sobre a concentração de creatinina (X1), idade (X2) e peso (X3). Os dados estão no arquivo "Rim.xls".

Quando as duas variáveis são qualitativas

Duas variáveis qualitativas

Exposição	Doença: Sim	Doença: Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

- Medidas de associação (χ^2)
- Medidas de risco ou magnitude de efeito

Medida de associação

Qui-quadrado de Pearson

$$\chi^2 = \sum_{i=1}^{l \times c} \frac{(o_i - e_i)^2}{e_i}$$

onde o_i *é o valor observado na i-ésima casela*

e_i *é o valor esperado na i-ésima casela:*

$$e_i = \frac{\text{total da linha} \times \text{total da coluna}}{\text{total geral}}$$

l é o número de linhas

c é o número de colunas

Um valor grande de χ^2 indica associação entre as variáveis!

Duas variáveis qualitativas

Exposição	Doença: Sim	Doença: Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

a + b = nº total de indíviduos expostos
c + d = nº total de indíviduos não expostos
a + c = nº total de indíviduos com a doença
b + d = nº total de indíviduos sem a doença

Duas variáveis qualitativas

Exposição	Doença: Sim	Doença: Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

- Em estudos de coorte : a + b e c + d são fixados
- Em estudos de caso-controle: a + c e b + d são fixados
- Em estudos transversais: apenas N é fixado

Medidas de risco em estudos de Coorte

Exposição	Doença: Sim	Doença: Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

fixos

Sejam:

I_e - incidência da doença (ou risco absoluto) em expostos = a/(a+b)

I_0 - incidência (ou risco absoluto) em não expostos = c/(c+d)

Risco relativo (RR):

Razão entre a incidência da doença em expostos (I_e) e a incidência em não expostos (I_0):

$$RR = \frac{I_e}{I_0} = \frac{a/(a+b)}{c/(c+d)}$$

Medidas de risco em estudos de Coorte

Exposição	Doença: Sim	Doença: Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

fixos

Sejam:

I_e - incidência da doença (ou risco absoluto) em expostos = a/(a+b)

I_0 - incidência (ou risco absoluto) em não expostos = c/(c+d)

Risco Atribuível (RA):

Diferença entre a incidência da doença em expostos (I_e) e a incidência em não expostos (I_0):

$$RA = Ie - Ir = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

Medida de "risco" em estudos Caso-Controle

Doença

Exposição	Sim	Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

fixos

Chance: razão de duas probabilidades:

- Probabilidade de exposição entre casos = a/(a+c)
- Probabilidade de não exposição entre casos = c/(a+c)
- Chance de exposição entre casos =

$$= \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

Chance : $\dfrac{\text{probabilidade do evento}}{1 - \text{probabilidade do evento}}$

Medida de "risco" em estudos Caso-Controle

Doença

Exposição	Sim	Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N

fixos

Sejam:

O_{ca} – *odds* ou chance de exposição entre casos = a/c

O_0 – *odds* ou chance de exposição entre controles = b/d

Odds ratio (OR) :

Definida como a razão entre a chance de exposição entre casos (O_{ca}) e a chance de exposição entre controles (O_{co}):

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Medida de "risco" em estudos Transversais

Exposição	Doença: Sim	Não	Total
Sim	a	b	a+b
Não	c	d	c+d
Total	a+c	b+d	N (fixo)

Sejam:

p_e - prevalência da doença entre expostos = a/(a+b)

p_0 - prevalência entre não expostos = c/(c+d)

Razão de prevalências (RP):

Definido como a razão entre a prevalência da doença entre expostos (p_e) e a prevalência entre não expostos (p_0):

$$RP = \frac{p_e}{p_0} = \frac{a/(a+b)}{c/(c+d)}$$

Regressão Logística Simples

- Estudar a associação entre duas variáveis
- Variável resposta é qualitativa com duas categorias (dicotômica), também chamada de "desfecho".
- Variável explicativa pode ser qualitativa ou quantitativa

Ex:

- Ocorrência de infarto do miocárdio (sim ou não) e uso de contraceptivo oral
- Ocorrência de câncer de pulmão (sim ou não) e hábito de fumar
- Cura (sim ou não) e tipo de tratamento

Regressão Logística Simples

Lembrando...

Se Y é uma variável aleatória dicotômica, então pode ser escrita como

$$Y = \begin{cases} 0 \text{ (fracasso)}....1-p \\ 1 \text{ (sucesso)}.....p \end{cases}$$

E sua função de probabilidades será dada por:

y	0	1
P(Y=y)	1-p	p

ou, de modo resumido, $P(Y = y) = p^y(1-p)^{1-y}$, y = 0, 1

Além disso, E(Y) = p e Var(Y) = p(1-p)

Regressão Logística Simples

Se Y~Bernoulli, não é possível adotar o modelo de regressão linear simples para este caso, porque suas suposições não estariam satisfeitas.

Lembrando....

O modelo de RLS é dado por

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2), independentes$$

Regressão Logística Simples

Para resolver este problema, foram desenvolvidos vários modelos:

- Logístico
- Probito
- Complementar log-log
- e outros

Regressão Logística Simples

O modelo de regressão logística simples pode ser escrito como:

$$E(Y_i) = p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

onde

- β_0 e β_1 são parâmetros
- X_i é o valor da variável explicativa para o i-ésimo indivíduo
- Y_i é a o valor da variável resposta para o i-ésimo indivíduo
- $p_i = P(Y_i=1)$ é a probabilidade de ocorrência do evento de interesse para o i-ésimo indivíduo

Regressão Logística Simples

O modelo de regressão logística simples pode ser escrito como:

$$E(Y_i) = p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

Aplicando a transformação logito, teremos:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

Regressão Logística Simples

O modelo de regressão logística simples pode ser escrito como:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

A expressão do lado esquerdo é denominada logito ou log-odds.

Os parâmetros β_0 e β_1 são estimados pelo método de máxima verossimilhança.

Regressão Logística Simples

$$p_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

$$\frac{p_i}{1-p_i} = \frac{\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}}{1-\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}} = \frac{\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}}{\frac{1+e^{\beta_0+\beta_1 X_i}-e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}} = \frac{\frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}}{\frac{1}{1+e^{\beta_0+\beta_1 X_i}}} = e^{\beta_0+\beta_1 X_i}$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(e^{\beta_0+\beta_1 X_i}\right) = \beta_0 + \beta_1 X_i$$

Regressão Logística Simples

Interpretação dos parâmetros do modelo

1. Se a variável explicativa é contínua:

- Para $X = a$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 a \Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 a}$$

- Para $X = a+1$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 (a+1) \Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 a+\beta_1} = e^{\beta_0+\beta_1 a} e^{\beta_1}$$

Quando aumentamos X de uma unidade, a chance de ocorrer o desfecho fica multiplicada por $exp(\beta_1)$

Regressão Logística Simples

Interpretação dos parâmetros do modelo

1. Se a variável explicativa é contínua:

$$OR = \frac{\left(\frac{p}{1-p}\right)^{X=a+1}}{\left(\frac{p}{1-p}\right)^{X=a}} = \frac{e^{\beta_0+\beta_1 a+\beta_1}}{e^{\beta_0+\beta_1 a}} = e^{\beta_1}$$

A exponencial do coeficiente β_1 é a *OR*.

Regressão Logística Simples

Interpretação dos parâmetros do modelo

2. Se a variável explicativa é é categórica, com duas categorias (*X=0* ou *X=1*)

- Para *X=0* (ausência do atributo)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 \Rightarrow \frac{p}{1-p} = e^{\beta_0}$$

- Para *X=1* (presença do atributo)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1} = e^{\beta_0}e^{\beta_1}$$

A chance de desenvolver o desfecho entre os indivíduos que têm o atributo é $exp(\beta_1)$ vezes a chance indivíduos que não têm o atributo .

Regressão Logística Simples

Interpretação dos parâmetros do modelo

2. Se a variável explicativa é é categórica, com duas categorias (*X=0* ou *X=1*)

$$OR = \frac{\left(\frac{p}{1-p}\right)^{X=1}}{\left(\frac{p}{1-p}\right)^{X=0}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

A exponencial do coeficiente β_1 é a *OR*.

Regressão Logística Múltipla

Mesma interpretação dos parâmetros do modelo

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

Por exemplo:

- Se X_1 é categórica com duas categorias: A exponencial do coeficiente β_1 é a chance de indivíduos com a característica X_1 desenvolverem o desfecho quando comparados a indivíduos sem a característica X_1, independentemente das demais variáveis (ou mantendo as demais variáveis constantes).
- Se X_2 é contínua: Para cada aumento de uma unidade em X_2 , o risco de desenvolver o desfecho fica multiplicado pela exponencial de β_2, independentemente das demais variáveis (ou mantendo as demais variáveis constantes).
- E assim por diante...

Exemplo: Doença coronariana

Para avaliar fatores de risco para doença coronariana, foi realizado um estudo de coorte envolvendo 609 homens, brancos, sem histórico de doença coronariana, seguidos durante 7 anos. No início do estudo foram medidas as seguintes variáveis: nível de neurotransmissores, denominados catecolaminas, idade, nível de colesterol, hábito de fumar, perfil de eletrocardiograma, pressão arterial sistólica e diastólica. O desfecho de interesse foi a ocorrência de doença coronariana. Os dados do estudo estão no arquivo "Doenca Coronariana.xls", que contem as seguintes variáveis:

Exemplo: Doença coronariana

- indiv – variável que identifica cada indivíduo, assumindo valores de 1 a 609;
- dcor – ocorrência de doença coronariana durante o período de estudo (0=não, 1=sim);
- catecol - nível de catecolaminas (0=normal, 1=alto);
- idade - idade (em anos);
- colest – nível de colesterol;
- fumo – se o indivíduo alguma vez já fumou (0=nunca, 1=sim);
- eletc – eletrocardiograma (0=normal, 1= anormal);
- pad - pressão arterial diastólica;
- pas - pressão arterial sistólica;

Exemplo: Doença coronariana

Crie também as variáveis explicativas:

- faixaet - faixa etária (1: 40 a 49 anos; 2:50 a 59 anos, 3: 60 anos e mais)
- Pressão alta (pad >= 95 e pas >= 160)