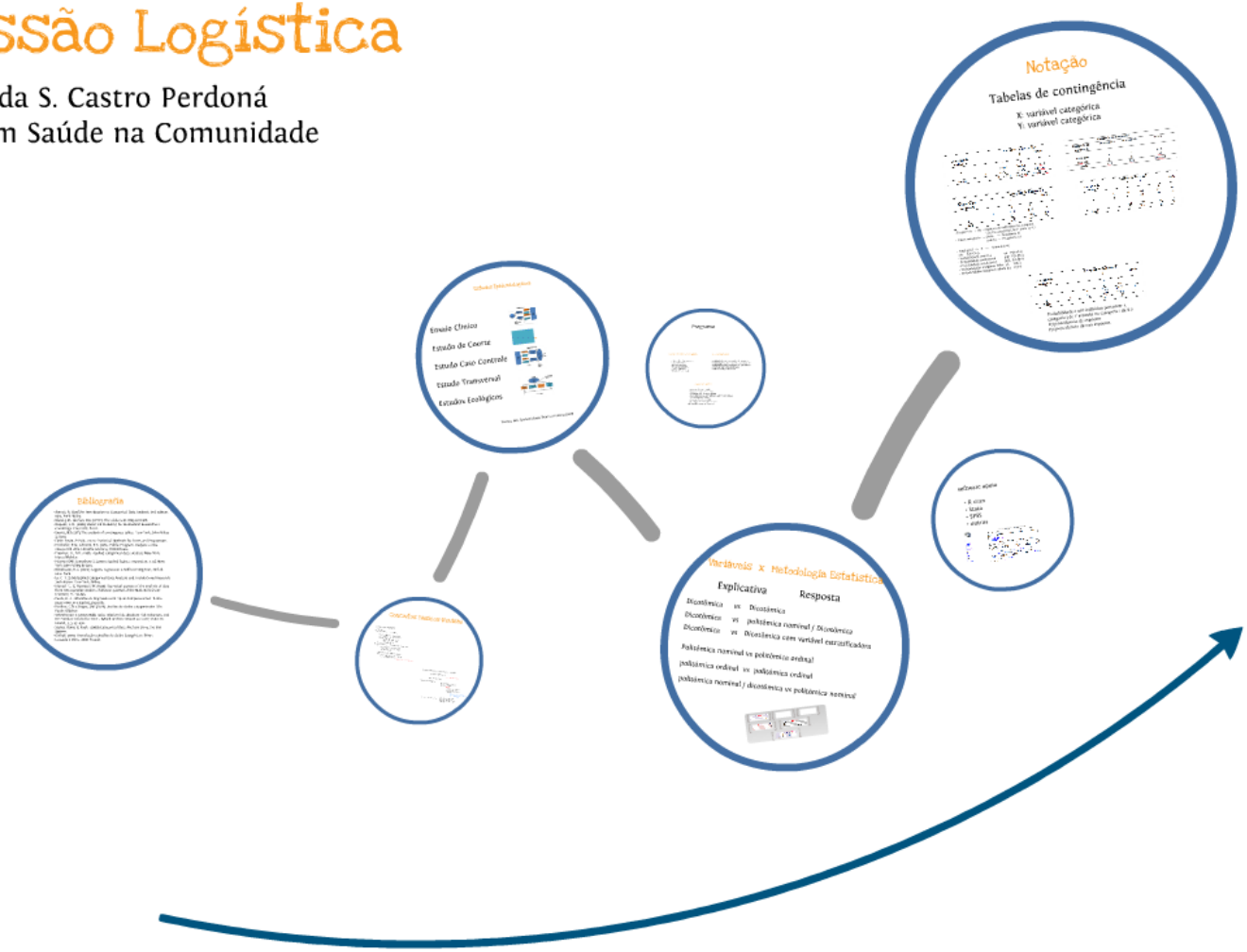
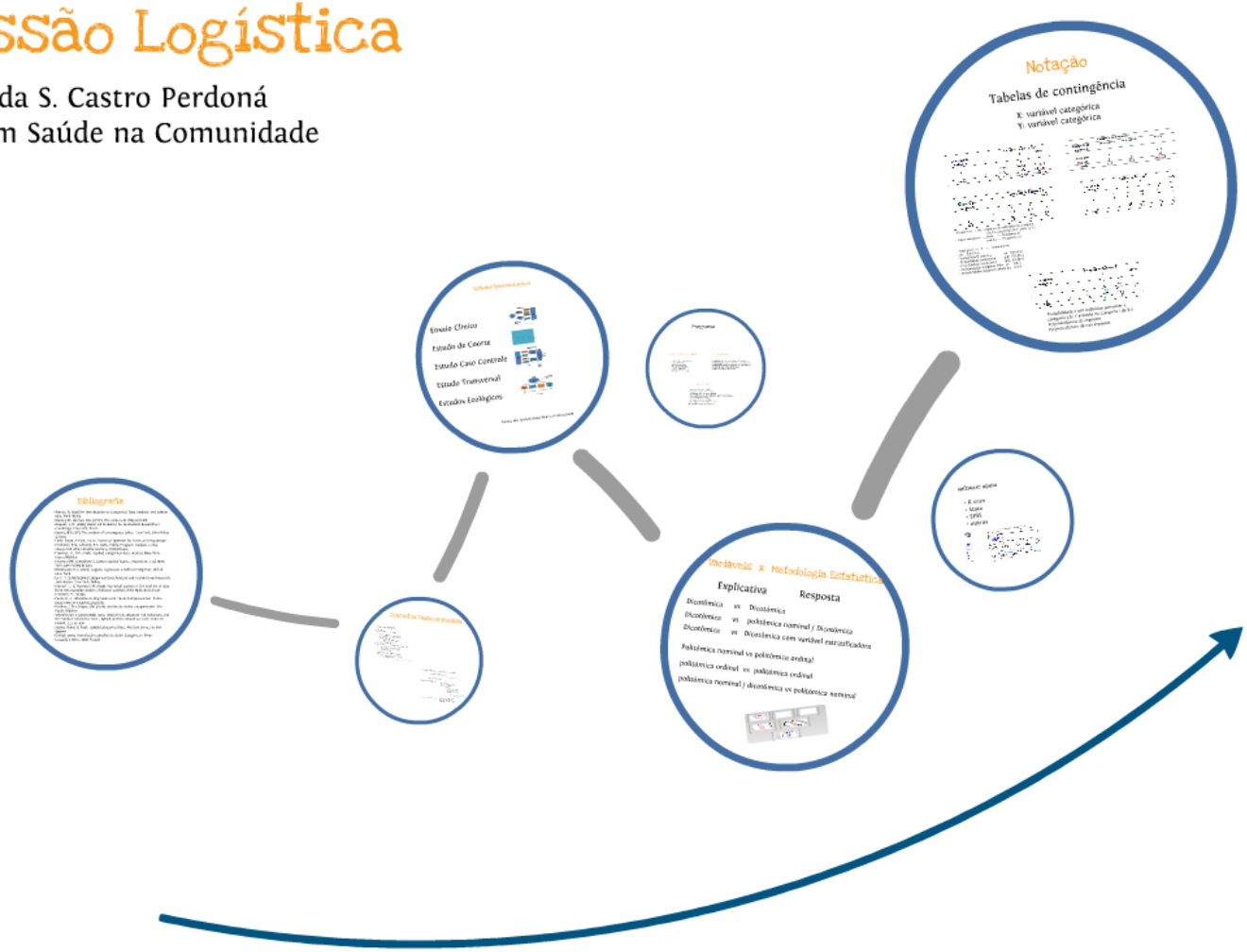
Análise de dados Categóricos e Regressão Logística

Gleici da S. Castro Perdoná
PPG em Saúde na Comunidade



Análise de dados Categóricos e Regressão Logística

Gleici da S. Castro Perdoná
PPG em Saúde na Comunidade



Bibliografia

- Agresti, A. (2007)An Introduction to Categorical Data Analysis. 2nd edition. New York: Wiley.
- Bland, J.M., Altman, D.G. (2000), The odds ratio. BMJ;320:1468.
- Dupont, E. D., (2002).Statistical Modeling for Biomedical Researchers. Cambridge University Press.
- Everitt, B.S.(1977) The analysis of contingency tables. New York, John Wiley & Sons.
- Fleiss, Levin, & Paik, (2003) Statistical Methods for Rates and Proportions.
- Forthofer, R.N.; Lehnem, R.G. (1981). Public Program Analysis: a new categorical data. Lifetime Learning Publications.
- Freeman, Jr., D.H. (1987). Applied categorical data analysis. New York: Marcel Dekker.
- Hosmer DW, Lemeshow S. (2000) Applied logistic regression. 2. ed. New York: John Wiley & Sons.
- Kleinbaum, D.G. (2005). Logistic regression A Self-Learning Text, 2nd ed. New York.
- Le, C. T. (2010)Applied Categorical Data Analysis and Translational Research. 2nd edition. New York: Wiley.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.
- Paula, G. A. , Modelos de Regressão com Apoio Computacional. home-page:www.ime.usp.br/_giapaula.
- Paulino, C.D. e Singer, J.M. (2006). Análise de dados categorizados. São Paulo: Blücher.
- Schechtman E. (2002) Odds ratio, relative risk, absolute risk reduction, and the number needed to treat - Which of these should we use?, Value In Health, 5, 5, 431-436.
- Stokes, Davis, & Koch. (2000) Categorical Data Analysis Using the SAS System.
- Giolo,S. (2012). Introdução a Análise de dados Categóricos. RBras
- Louzada e Diniz, 2007. Fraude.

Conceitos Básicos-Revisão

CLASSIFICAÇÃO DOS DADOS

QUALITATIVOS
(característica em estudo é uma qualidade)

CATEGORICAS(BINÁRIAS) – não existe ordem
Estado do paciente, sexo, raça/cor
ORDINAIS – existe ordem
Escolaridade, estadiamento da doença, escalas, desempenho, melhora do paciente

QUANTITATIVOS ou NUMÉRICOS
(característica em estudo é uma medida em uma escala numérica)

DISCRETO – escala composta por nos inteiros
n empregados; n produtos; n células, níveis hematológicos

CONTINUO – escala é contínua(composta por nos reais)
%Fe no tecido, níves de calcio, níveis em geral, tempo, peso, estatura

A classificação é dada pela escala de mensuração

AS VARIÁVEIS AINDA PODEM SER CLASSIFICADAS,SEGUNDO A HIERARQUIA

INDEPENDENTE (EXPLICATIVAS)

ISSO PODE NÃO SER TÃO SIMPLES

DEPENDENTE (RESPOSTA)

ETAPAS DE UMA INVESTIGAÇÃO

Pereira, MG. Epidemiologia Teoria e Prática(2000)

Objetivos da disciplina
Apresentar técnicas estatísticas para análise de dados discretos/categóricos aplicados a área médica.

PORTANTO..

Nossos estudos serão direcionados para problemas que envolvam,
Variável dependente : categórica
Variável independente: categóricas e/ou continuas.

Estudos com estas características são chamados de de Análise de dados categóricos. Análise de dados discretos.

Devido a natureza da variável resposta, associa-se distribuições discretas de probabilidade com : binomial, multinomial, poisson, binomial negativa, geometrica, entre outras.

CLASSIFICAÇÃO DOS DADOS

QUALITATIVOS
(característica em estudo é uma qualidade)

CATEGORICAS(BINÁRIAS) – não existe ordem
Estado do paciente, sexo, raça/cor
ORDINAIS – existe ordem
Escolaridade, estadiamento da doença, escalas, desempenho, melhora do paciente

QUANTITATIVOS ou NUMÉRICOS
(característica em estudo é uma medida em uma escala numérica)

DISCRETO – escala composta por nos inteiros
n empregados; n produtos; n células, níveis hematológicos

CONTINUO – escala é contínua(composta por nos reais)
%Fe no tecido, níves de calcio, níveis em geral, tempo, peso, estatura

A classificação é dada pela escala de mensuração

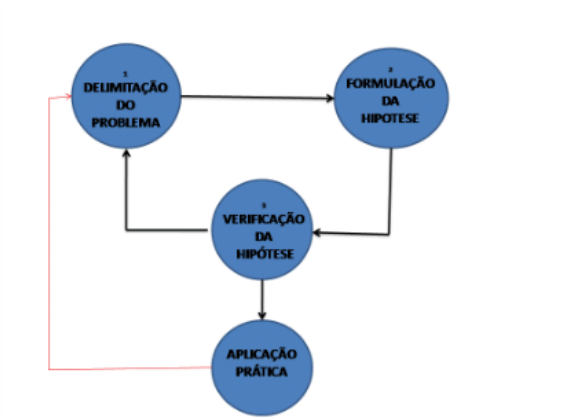
AS VARIÁVEIS AINDA PODEM SER CLASSIFICADAS,SEGUNDO A HIERARQUIA

INDEPENDENTE (EXPLICATIVAS)

ISSO PODE NÃO SER TÃO SIMPLES

DEPENDENTE (RESPOSTA)

ETAPAS DE UMA INVESTIGAÇÃO



Pereira, MG. Epidemiologia Teoria e Prática(2000)

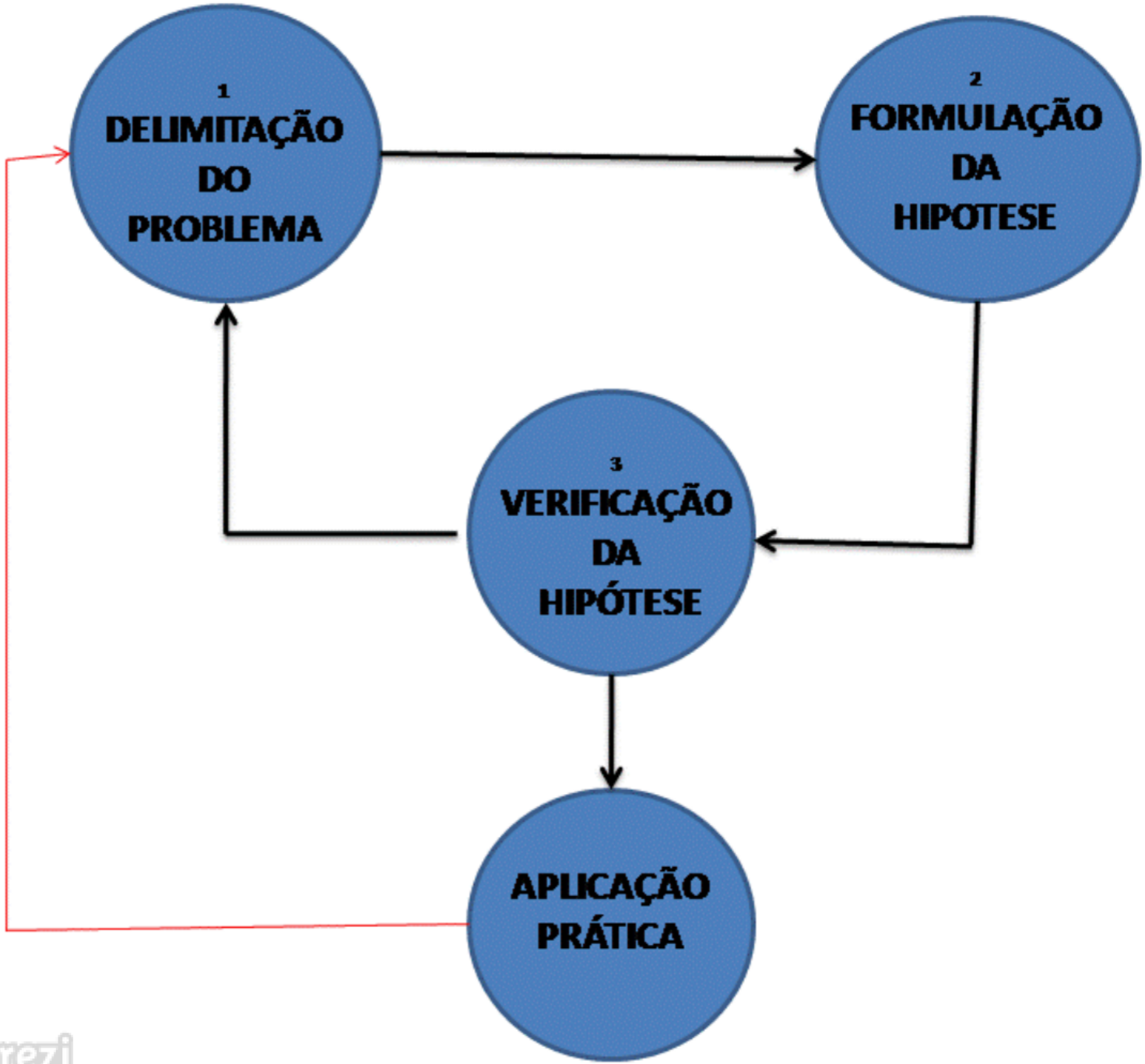
Objetivos da disciplina
Apresentar técnicas estatísticas para análise de dados discretos/categóricos aplicados a área médica.

PORTANTO...

Nossos estudos serão direcionados para problemas que envolvam,
Variável dependente : categórica
Variável independente: categóricas e/ou continuas.

Estudos com estas características são chamados de de Análise de dados categóricos, Análise de dados discretos.

Devido a natureza da variável resposta, associa-se distribuições discretas de probabilidade com : binomial, multinomial, poisson, binomial negativa, geometrica, entre outras.


1
DELIMITAÇÃO DO PROBLEMA
2
FORMULAÇÃO DA HIPOTESE
3
VERIFICAÇÃO DA HIPÓTESE
APLICAÇÃO PRÁTICA

Objetivos da disciplina
Apresentar técnicas estatísticas para análise de dados discretos/categóricos aplicados a área médica.

PORTANTO...

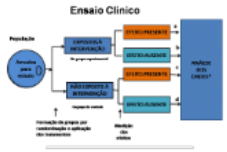
Nossos estudos serão direcionados para problemas que envolvam,
Variável dependente : categórica
Variável independente: categóricas e/ou continuas.

Estudos com estas características são chamados de de Análise de dados categóricos, Análise de dados discretos.

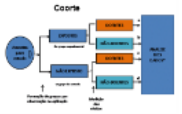
Devido a natureza da variável resposta, associa-se distribuições discretas de probabilidade com : binomial, multinomial, poisson, binomial negativa, geometrica, entre outras.

Estudos Epidemiológicos

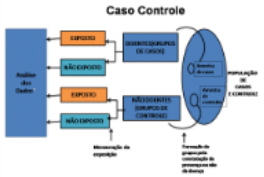
Ensaio Clinico



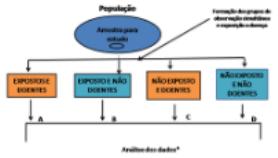
Estudo de Coorte



Estudo Caso Controle



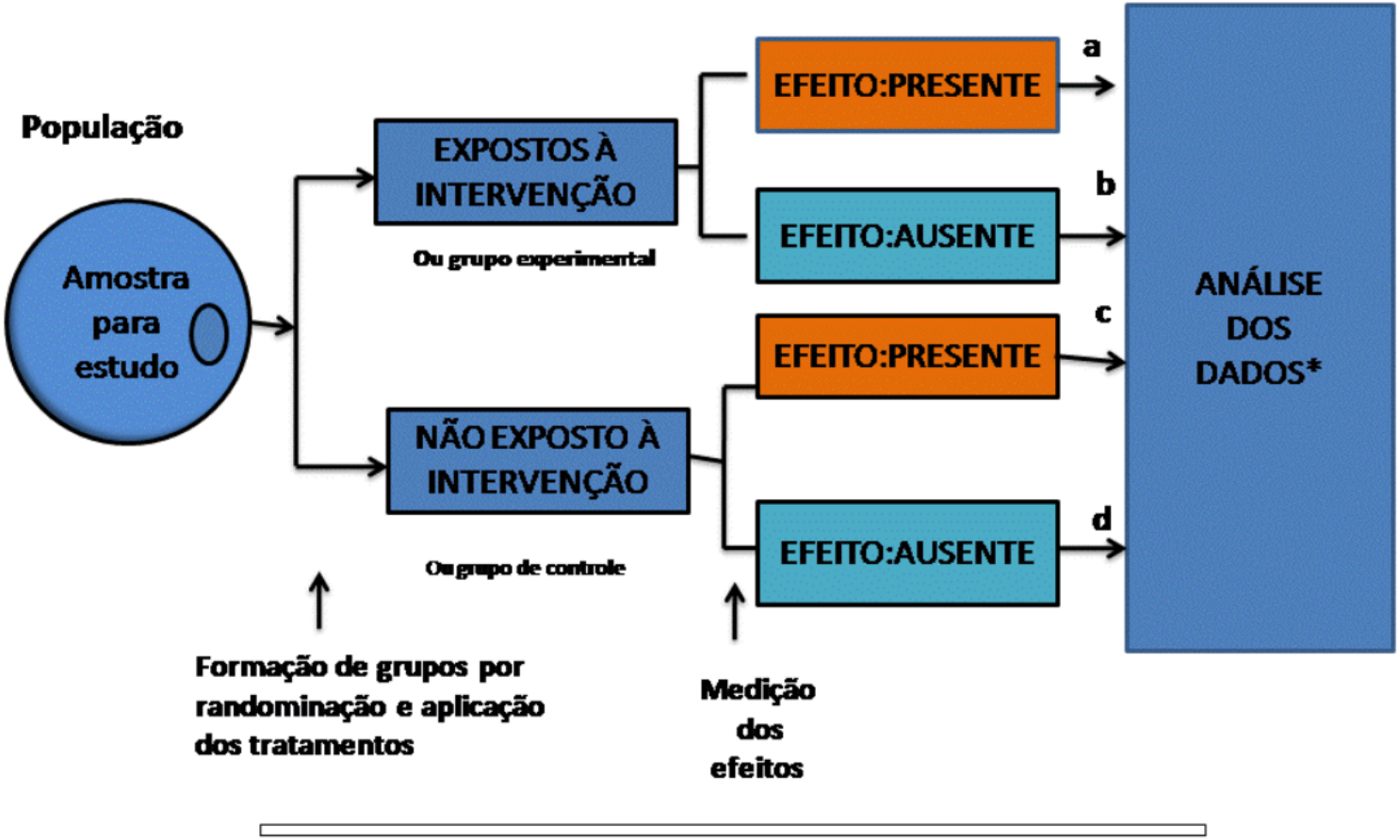
Estudo Transversal



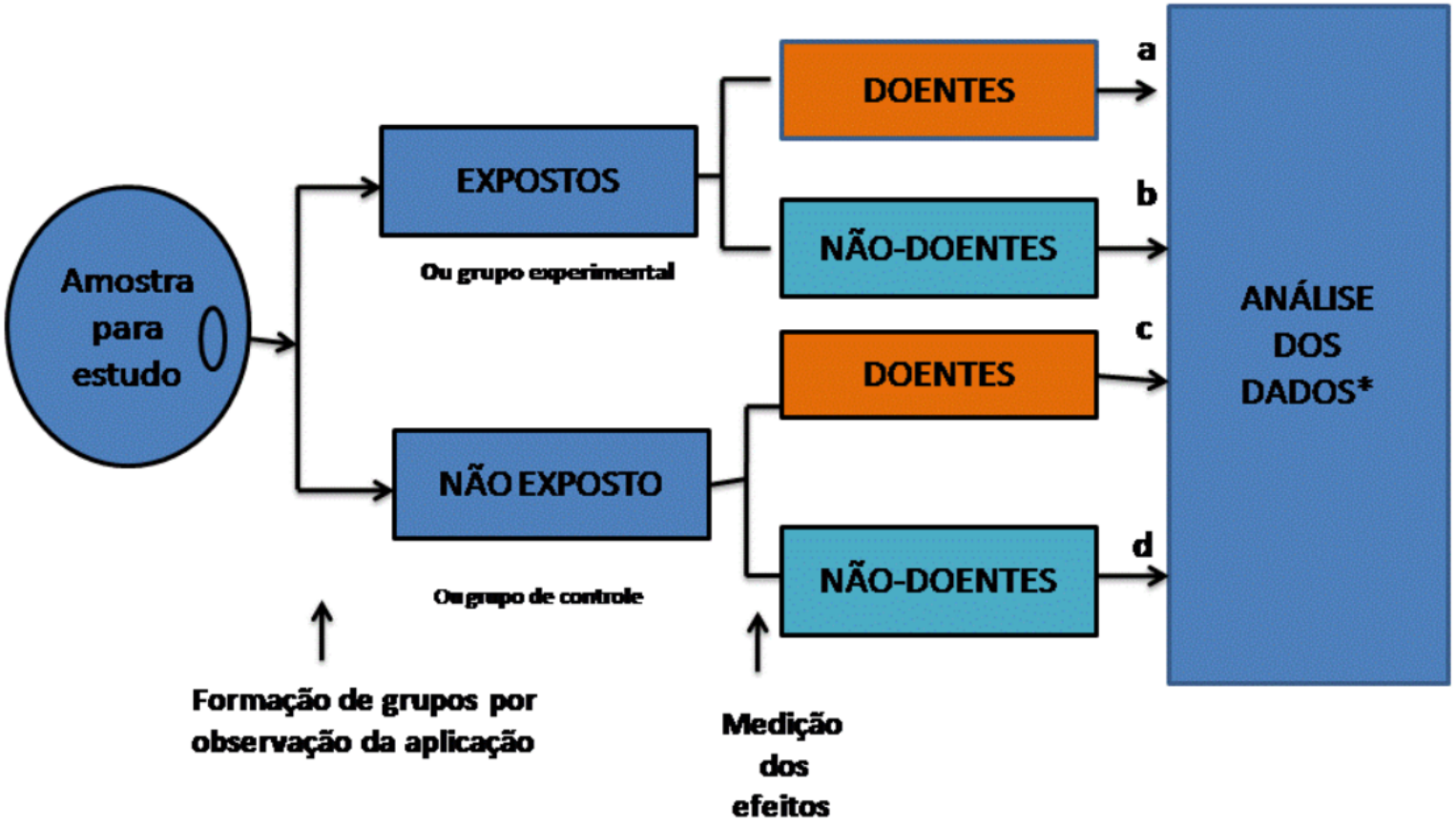
Estudos Ecológicos

Pereira, MG. Epidemiologia Teoria e Prática(2000)

Ensaio Clinico

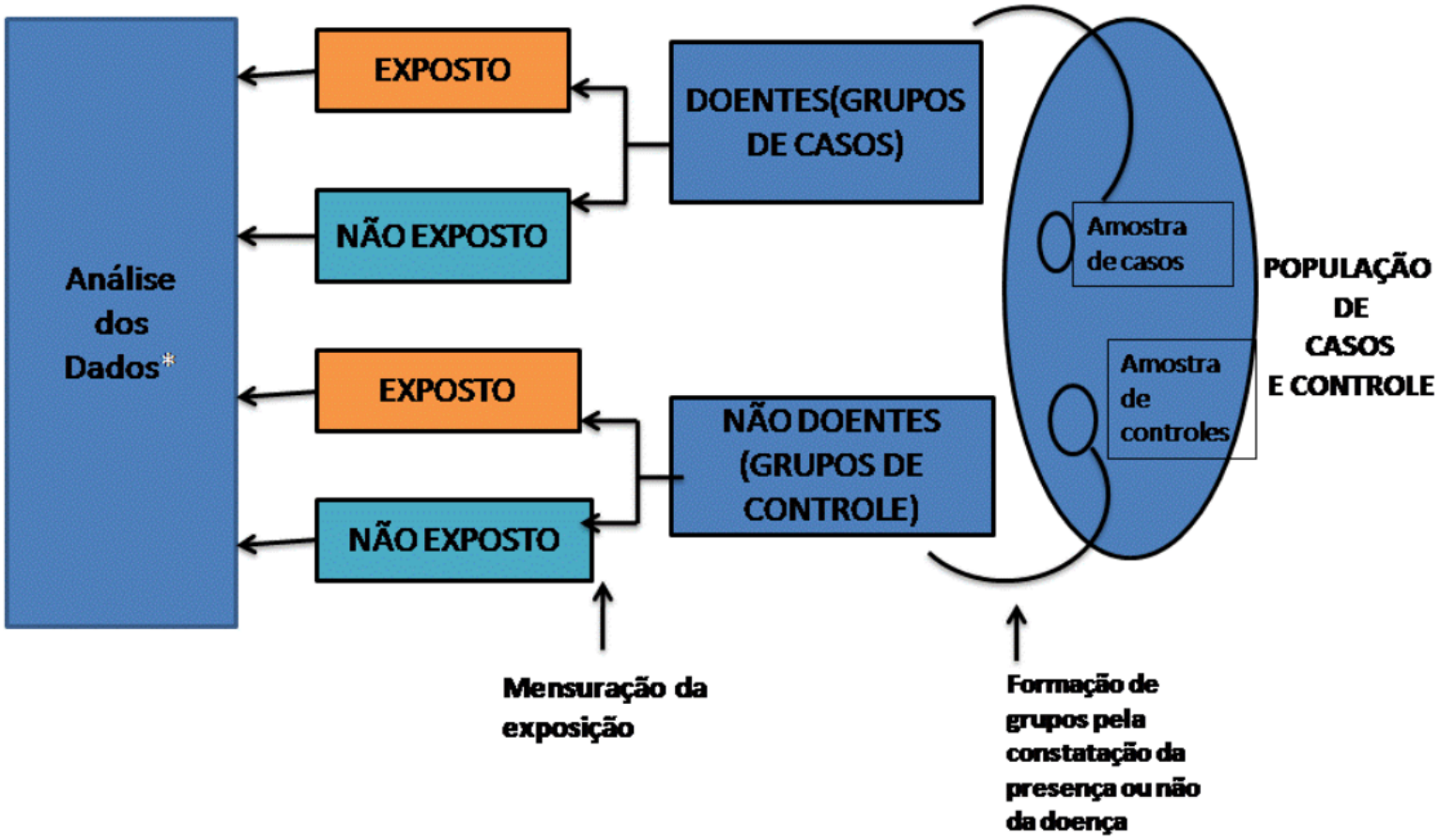


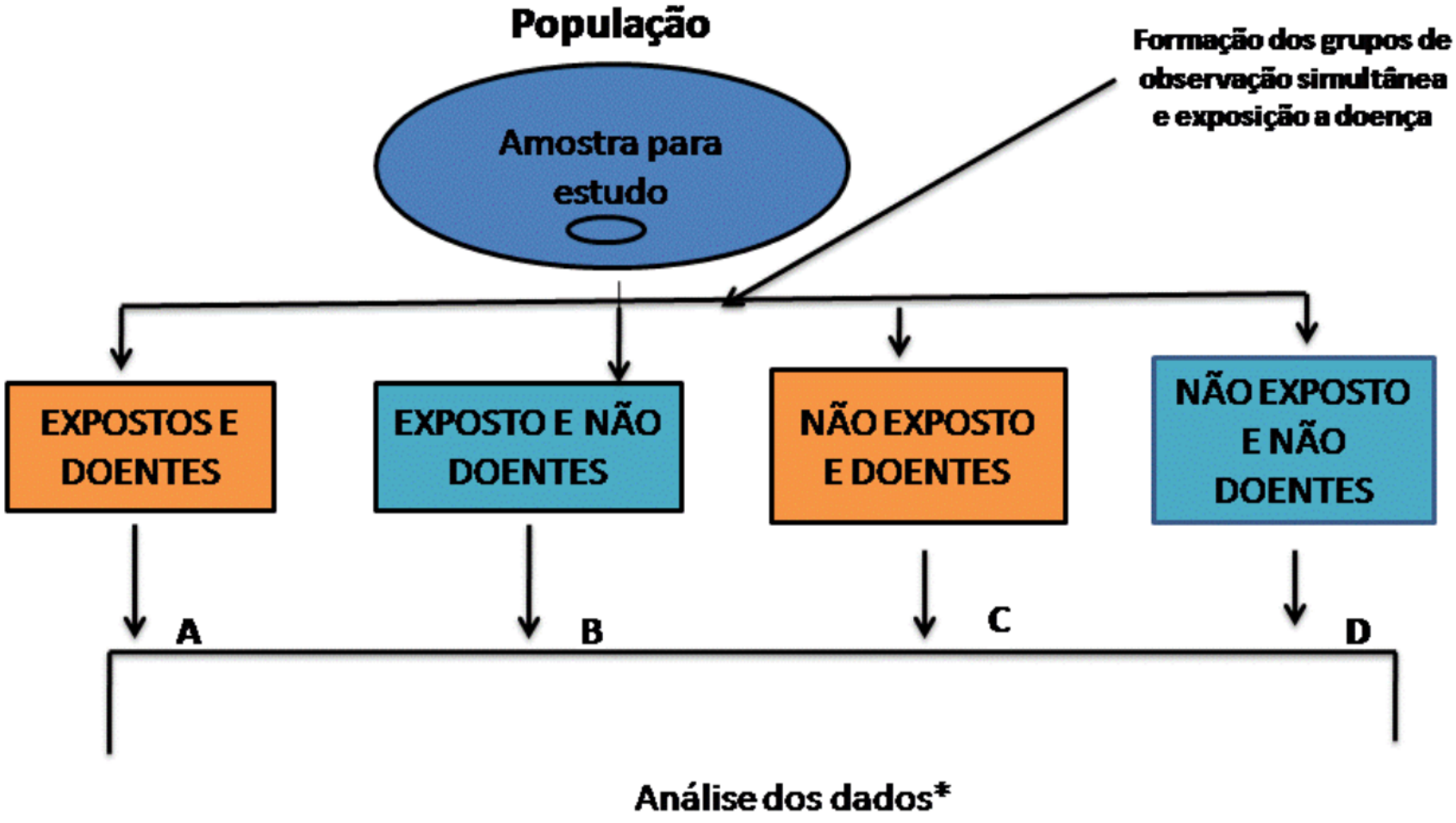
Coorte


Amostra para estudo
EXPOSTOS
Ou grupo experimental
NÃO EXPOSTO
Ou grupo de controle
DOENTES
NÃO-DOENTES
DOENTES
NÃO-DOENTES
a
b
c
d
ANÁLISE DOS DADOS*
Formação de grupos por observação da aplicação
Medição dos efeitos

efeitos

Caso Controle




População
Amostra para estudo
Formação dos grupos de observação simultânea e exposição a doença
EXPOSTOS E DOENTES
EXPOSTO E NÃO DOENTES
NÃO EXPOSTO E DOENTES
NÃO EXPOSTO E NÃO DOENTES
A
B
C
D
Análise dos dados*

Programa

Análise e tabelas de contingência.

- Tabelas de contingências r x c
- Teste Qui quadrado
- Análise de resíduos
- Partição de tabelas rxc
- Tabelas ordenadas
- Medidas de Associação

Modelos loglineares

- Modelos loglineares para tabelas de contingência
- Modelo loglinear de independência para tabela 2x2
- Modelos loglineares para tabelas de tripla entrada
- Inferência para modelos loglineares
- Método de seleção de Modelos

Regressão logística

- Regressão logística simples
- Razão de chances (Odds Ratio)
- Inferência para regressão logística
- Regressão logística com preditores categóricos (teste de CochranMantelHaenszel)
- Regressão logística múltipla
- Estratégias para seleção do modelo

(AIC Akaike Information Criterion)

Análise e tabelas de contingência.

- Tabelas de contingências r x c
- Teste Qui quadrado
- Análise de resíduos
- Partição de tabelas rxc
- Tabelas ordenadas
- Medidas de Associação

Modelos loglineares

- Modelos loglineares para tabelas de contingência
- Modelo loglinear de independência para tabela 2x2
- Modelos loglineares para tabelas de tripla entrada
- Inferência para modelos loglineares
- Método de seleção de Modelos

Regressão logística

- Regressão logística simples
- Razão de chances (Odds Ratio)
- Inferência para regressão logística
- Regressão logística com preditores categóricos (teste de CochranMantelHaenszel)
- Regressão logística múltipla
- Estratégias para seleção do modelo

(AIC Akaike Information Criterion)

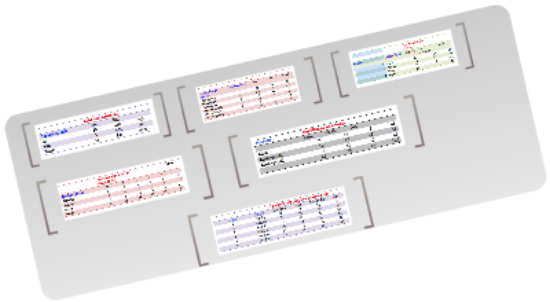
Variáveis x Metodologia Estatística

Explicativa		Resposta
Dicotômica	vs	Dicotômica
Dicotômica	vs	politômica nominal / Dicotômica
Dicotômica	vs	Dicotômica com variável estratificadora

Politômica nominal vs politômica ordinal

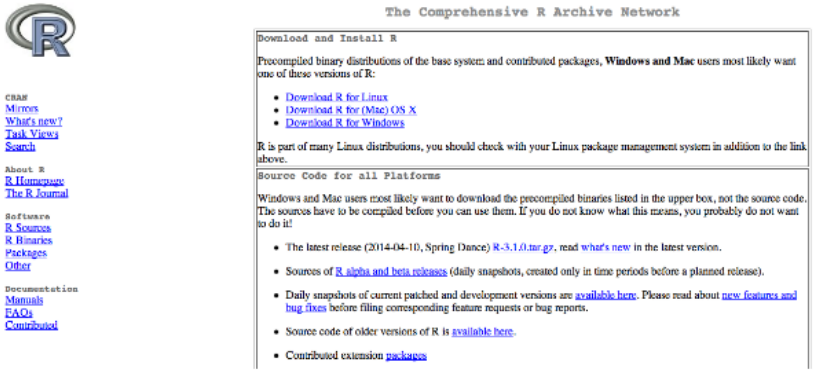
politômica ordinal vs politômica ordinal

politômica nominal / dicotômica vs politômica nominal



software apoio

- R cran
- Stata
- SPSS
- outros



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2014-04-10, Spring Dance) R-3.1.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Pontomic

Bebê com baixo peso			
Tabagismo mãe	Sim	Não	
sim	50	450	500
Não	5	495	500
Totais	55	945	1000

		Cura		
Diagnostico de infecção	Medicamentos	Sim	Não	Totais
Complicada	A	78	28	106
Complicada	B	101	11	112
Complicada	C	68	46	114
Não complicada	A	40	5	45
Não complicada	B	54	5	59
Não complicada	C	34	6	40

		Tipo do resultado (Efeito)		
Centros	Medicaments	Favorável	Não Favorável	Totais
1	Novo	29	16	45
1	Padrão	14	31	45
2	Novo	37	8	45
2	Padrão	24	21	45

	Horas de promoção de alivio da dor					Totais
Medicamentos	0	1	2	3	4	
Placebo	6	9	6	3	1	25
Padrão	1	4	6	6	8	25
Novo	2	5	6	8	6	27
Totais	9	18	18	17	15	77

Produto	Classificação da limpeza			
	Baixa	Média	Alta	
Agua	27	14	5	46
Agua +prod1	10	17	26	53
Agua +prod2	5	12	50	67
	42	43	81	166

		Preferencia pelo programa de aprendizado			
Escola	Periodo	Individual	Grupo	Sala de Aula	Totais
1	Padrão	10	17	26	53
1	Integral	5	12	50	67
2	Padrão	21	17	26	64
2	Integral	16	12	36	64
3	Padrão	15	15	16	46
3	Integral	12	12	20	44

	Bebê com baixo peso		
Tabagismo mãe	Sim	Não	
sim	50	450	500
Não	5	495	500
Totais	55	945	1000

	Horas de promoção de alivio da dor	**Totais**

		Cura		
Diagnostico de infecção	Medicamentos	Sim	Não	Totais
Complicada	A	78	28	106
Complicada	B	101	11	112
Complicada	C	68	46	114
Não complicada	A	40	5	45
Não complicada	B	54	5	59
Não complicada	C	34	6	40

Produto	Classificação d

		Tipo do resultado (Efeito)		
Centros	Medicaments	Favoravel	Não Favorável	Totais
1	Novo	29	16	45
1	Padrão	14	31	45
2	Novo	37	8	45
2	Padrão	24	21	45

Não	5	495	500
Totais	55	945	1000

	Horas de promoção de alivio da dor					Totais
Medicamentos	0	1	2	3	4	
Placebo	6	9	6	3	1	25
Padrão	1	4	6	6	8	25
Novo	2	5	6	8	6	27
Totais	9	18	18	17	15	77

	A	78	28	106
	B	101	11	112
	C	68	46	114
la	A	40	5	45
la	B	54	5	59
la	C	34	6	40

Centros	Medicaments
1	Novo
1	Padrão
2	Novo
2	Padrão

Produto	Classificação da limpeza			
	Baixa	Média	Alta	
Agua	27	14	5	46
Agua +prod1	10	17	26	53
Agua +prod2	5	12	50	67
	42	43	81	166

Preferencia pelo programa de aprendizado				
Periodo	Individual	Grupo	Sala de Aula	Totais
Padrão	10	17	26	53
Meeprol	5	12	50	67
Padrão	21	17	26	64

Agua +prod1	10	17
Agua +prod2	5	12
	42	43

		Preferencia pelo programa de aprendizado			
Escola	Periodo	Individual	Grupo	Sala de Aula	Totais
1	Padrão	10	17	26	53
1	Integral	5	12	50	67
2	Padrão	21	17	26	64
2	Integral	16	12	36	64
3	Padrão	15	15	16	46
3	Integral	12	12	20	44

Notação

Tabelas de contingência

X: variável categórica
Y: variável categórica

Categoria da Variável X		Categoria da Resposta Y: j = 1	j = 2	Totais
i	1	a=n11	b=n12	n1.=a+b
i	2	c=n21	d=n22	n2.=c+d
Totais		n.1 = a+c	n.2=b+d	n=a+b+c+d

Categoria da Variável X	Categoria da Resposta Y: Doente sim	Doente não	Totais
expo sim	a	b	a+b
exp nao	c	d	c+d
Totais	a+c	b+d	a+b+c+d

Categoria da Variável X		Categoria da Resposta Y: j = 1	j = 2	Totais
i	1	n11	n12	n1.
i	2	n21	n22	n2.
Totais		n.1	n.2	n

Categoria da Variável X		Categoria da Resposta Y: j = 1	j = 2	..	j = c	Totais
i	1	n11	n12	..	n1c	n1.
i	2	n21	n22		n2c	n2.
i	r	nr1	nr2		nrc	nr.
Totais		n.1	n.2		n.c	n

- Frequencia --› nij --›Numero de individuos na categoria i de X e categoria j de Y onde i,j =1,2
- Totais marginais --›Linha => frequência ni.
  coluna => frequência n.j

- Total geral => n => soma dos nij
- pij P(X=i,Y=j)
- Probabilidade conjunta pij P(X=i,Y=j)
- Probabilidade condicional pi(j) P(X=i|Y=j)
- Probabilidade condicional p(i)j P(Y=j|X=i)
- Probabilidades marginais linha pi. P(X=i)
- Probabilidades marginais coluna p.j P(X=j)

Categoria da Variável X			Categoria da Resposta Y: j = 1 (D)	j = 2 (ND)	Totais
i	1	E	P(1)1	P(1)2	1
i	2	NE	P(2)1	P(2)2	1
Totais			p.1	p.2	1

Probabilidade e um indíviduo pertencer a categoria j de Y estando na categoria i de X é
P(1)1=incidencia de expostos
P(2)1=incidencia de nao expostos

Y: variáv

Categoria da Variável X		Categoria da Resposta Y j	j	Totais
		1	2	
i	1	a=n11	b=n12	n1.=a+b
i	2	c=n21	d=n22	n2.=c+d
Totais		n.1 = a+c	n.2=b+d	n=a+b+c+d

Categoria da Variável X	Categoria da Resposta Y j	j	Totais

categórica

	Categoria da Resposta Y		
Categoria da Variável X	Doente sim	Doente não	Totais
expo sim	a	b	a+b
exp nao	c	d	c+d
Totais	a+c	b+d	a+b+c+d

		Categoria da Resposta Y				
Categoria da Variável X		j	j		j	Totais
		1	2	..	c	
i	1	n11	n12	..	n1c	n1.

exp sim	a	b	a+b
exp nao	c	d	c+d
Totais	a+c	b+d	a+b+c+d

Categoria da Variável X		Categoria da Resposta Y				
		j	j		j	Totais
		1	2	..	c	
i	1	n11	n12	..	n1c	n1.
i	2	n21	n22		n2c	n2.
i	r	nr1	nr2		nrc	nr.
Totais		n.1	n.2		n.c	n

		Categoria da Resposta Y		
Categoria da Variável X		j	j	Totais
		1	2	
i	1	n11	n12	n1.
i	2	n21	n22	n2.
Totais		n.1	n.2	n

- Frequencia --› nij -->Numero de individuos na categoria i de X e categoria j de Y onde i,j =1,2
- Totais marginais -->Linha => frequência ni.
  coluna => frequência n.j

- Total geral => n => soma dos nij
- pij P(X=i,Y=j)
- Probabilidade conjunta pij P(X=i,Y=j)
- Probabilidade condicional pi(j) P(X=i|Y=j)
- Probabilidade condicional p(i)j P(Y=j|X=i)
- Probabilidades marginais linha pi. P(X=i)
- Probabilidades marginais coluna p.j P(X=j)

Categoria da Variável X			Categoria da Resposta Y j	j	Totais
			1	2	
			D	ND	
i	1	E	P(1)1	P(1)2	1
i	2	NE	P(2)1	P(2)2	1
Totais			p.1	p.2	1

Probabilidade e um indíviduo pertencer a categoria j de Y estando na categoria i de X é
P(1)1=incidencia de expostos
P(2)1=incidencia de nao expostos