UNIVERSIDADE DE SÃO PAULO
FACULDADE DE SAÚDE PÚBLICA
DEPARTAMENTO DE EPIDEMIOLOGIA
Programa de Mestrado Profissional em Entomologia em Saúde Pública
ESP5101– Bioestatística Básica

PERÍODO : 26 de fevereiro a 11 de junho de 2019
HORÁRIO: 8:00h – 12:00h
DOCENTES:

Profª Dra. Denise Pimentel Bergamaschi - denisepb@usp.br
Prof. Colaborador: Fredy Galvis

CONTEÚDO PROGRAMÁTICO

- Coleta de dados, escalas de mensuração, tipos de variáveis

- Tabelas e gráficos

- Medidas de posição e de dispersão

- Fundamentos de correlação linear; estimação da reta de regressão

- Noções de probabilidade, variável aleatória

- Principais modelos de distribuições de probabilidades: Binomial, Normal; distribuição amostral da média, distribuição "t" de Student e qui-quadrado

- Noções de amostragem, conceitos de vício e precisão

- Estimativas por intervalo de confiança

- Tamanho de amostras para estimar média e proporção, por intervalo de confiança

- Teste de hipóteses

- Teste de uma média populacional

- Teste de associação pelo qui-quadrado de Pearson

BIBLIOGRAFIA

1. ARMITAGE, P. & BERRY, G. Statistical methods in medical research. Oxford: Blackwell Scientific Publ; 1994.

2. BERQUÓ, E.S.; SOUZA, J.M.P. & GOTLIEB, S.L.D. Bioestatística. São Paulo, EPU, 1981.

3. COSTA NETO, P.L.L.P. Estatística. São Paulo, Ed. Edgar Blucher, 1977. FISCHER, L.D., van BELLE, G. – Biostatistics: A methodology for the health sciences. New York: John Wiley & Sons, Inc; 1993.

4. DAWSON-SAUNDERS, B. & TRAPP, R.G. Basic and clinical biostatistics 2nd edition. Connecticut: Appleton & Lange; 1994.

5. MORETTIN, P. & BUSSAB, W. Estatística Básica. São Paulo, Atual, 1982.

6. NOETHER, G.E. Introdução à estatística. Rio de Janeiro: Guanabara Dois, 1983.

7. PAGANO, M. & GAUVREAU, K. Principles of biostatistics. 2nd Ed. Pacific Grove, CA; Duxbury; 2000.

8. PEREIRA, J.C. Bioestatística em outras palavras. São Paulo: EDUSP/FAPESP, 2010.

9. Townsend C R, Begon M, Harper JL. Fundamentos em ecologia. Artmed 2ª edição, 2003.

10. Bicudo CEM, Bicudo D. Amostragem em Limnologia. RiMa, 2004.

11. Gotelli NL, Ellison AM. Princípios de Estatística em Ecologia. Artmed, 2011

Cronograma

DATA	AULA	CONTEÚDO PROGRAMÁTICO	DOCENTE/MONITOR
26/02	1	Coleta de dados, níveis de mensuração, variáveis, organização de dados, apresentação tabular	Denise
12/03	2	Apresentação tabular e gráfica	Denise
19/03	3	Apresentação gráfica e medidas de tendência central	Denise
26/03	4	Medidas de tendência central e de posição; medidas de dispersão ou de variabilidade; Box plot	Denise
02/04	5	Noções de correlação e regressão	Denise
09/04	6	Noções de probabilidade; noções de amostragem; distribuição normal, distribuição amostral da média	Denise
16/04	7	Teste de hipóteses de uma média populacional com variância conhecida	Denise
23/04	8	Teste de hipóteses de uma média populacional com variância desconhecida	Fredy
30/04	9	Exercícios - consolidação de conteúdo	Fredy/Fernanda
07/05	10	Exercícios - consolidação de conteúdo	Fernanda
14/05	11	Teste de hipóteses de associação pelo qui-quadrado de Pearson	Fredy
21/05	12	Estimação de uma média populacional com variância conhecida e desconhecida	Fredy
28/05	13	Exercícios - consolidação de conteúdo	Fredy
04/06	14	Exercícios - consolidação de conteúdo	Fredy
11/06	15	Avaliação	Fredy/Denise

CRITÉRIOS DE AVALIAÇÃO:

A nota final será composta pela nota da avaliação acrescida de 1 ponto para os que entregarem a lista de exercícios referentes à primeira parte do conteúdo (aula 1 à aula 6).

Critério:

A (nota final entre 8,5 e 10);
B (nota final entre 7,5 e 8,4);
C (nota final entre 5,0 e 7,4);
D (nota final abaixo de 5,0).

O aluno que obtiver nota final abaixo de 5,0 poderá realizar uma avaliação de recuperação. Neste caso, o aluno terá que ter nota acima de 5,0 e, independente da nota obtida na recuperação, o aluno receberá nível C.

População, amostra, variável, coleta de dados, apuração de dados e apresentação tabular.

Estatística: fornece uma coleção de métodos para planejar experimentos, para obter e organizar dados, resumi-los, analisá-los, interpretá-los e deles extrair conclusões (Triola, 1999).

Bioestatística – Estatística aplicada às ciências da vida.

Níveis de mensuração

Escala nominal

Os indivíduos (ou unidades de análise) são classificados em categorias segundo uma característica.

Exemplos

Sexo (masculino, feminino),

Hábito de fumar (fumante, não fumante),

Sobrepeso (sim, não),

Condição do domicílio (próprio já pago, próprio em pagamento, alugado, cedido, outra condição);

Sexo dos insetos vetores (fêmea, macho);

Tipo de habitat (mata, margem da mata, campo aberto, domicílio);

Local do domicílio (intradomicilio e peridomicilio)

Característica:

Não existe ordem entre as categorias e suas representações, se numéricas, são destituídas de significado numérico.

Ex: Sexo do paciente
Masculino =1, feminino = 2
Os valores 1 e 2 são apenas rótulos

Ex: Sexo do inseto
Fêmea=1, Macho = 2
Os valores 1 e 2 são apenas rótulos e não podem ser tratados como números.

Tipo de habitat de determinado inseto
1= mata, 2 = margem da mata, 3= campo aberto, 4= domicílio
Da mesma forma, os valores 1, 2, 3, e 4 são apenas rótulos.

Escala ordinal

Os indivíduos são classificados em categorias que possuem algum tipo inerente de ordem. Neste caso, uma categoria pode ser "maior" ou "menor" do que outra.

Ex: Nível socioeconômico (A, B, C e D; onde A representa maior poder aquisitivo);

Nível de retinol sérico (alto, aceitável, baixo, deficiente) critérios: *Commitee on Nutrition for National Defense ICNND/USA*, 1963 (in Prado MS et al, 1995).

Alto: maior ou igual a 50,0 μg/dl;
Aceitável: 20,0 a 49,9 μg/dl;
Baixo: 10,0 a 19,9 μg/dl;
Deficiente: menor ou igual a 10,0 μg/dl.

Tamanho da asa de mosquitos culicídeos classificados em categorias (Landry, SV et al., 1988, Journal of the Americam Mosquito Control Association- vol.4 nº 2).

Pequeno (≤ 2,00 mm),
Médio (2,01 – 3,64 mm),
Grande (≥ 3,65 mm).

Embora exista ordem entre as categorias, a diferença entre as categorias adjacentes não tem o mesmo significado em toda a escala. Se as categorias forem representadas por números, estas são destituídas de significado numérico.

Escala numérica intervalar

Este nível de mensuração possui um valor zero arbitrário e, por este motivo não permite calcular a razão entre dois valores, sendo possível, entretanto calcular a soma e subtração.

Como exemplo deste nível de aferição temos a temperatura em graus Celsius e *Fahrenheit*. O exemplo abaixo indica o efeito do zero arbitrário na utilização de operações matemáticas (diferença e divisão) tanto em uma variável aferida pela escala numérica intervalar como por uma em escala de razões contínua.

A variável de aferição é temperatura em graus Celsius.

material	^{0}C	^{0}F	|dif^{0}C|	|dif ^{0}F|	dif^{0}C/dif^{0}F	razão^{0}C	razão^{0}F	Razão^{0}C/razão^{0}F
A	20	68	|A-B|=20	|A-B|=36	0,56	A/B=0,50	A/B=0,65	0,77
B	40	104	|B-C|=20	|B-C|=36	0,56	B/C=0,67	B/C=0,74	0,91
C	60	140	|A-C|=40	|A-C|=72	0,56	A/C=0,33	A/C=0,49	0,67

A variável de aferição é comprimento (cm) em escala de razões contínua

comprimento	cm	polegada	|difcm|	|dif pol|	Difcm/difpol	Razãocm	Razãopol	Razãocm/razãopol
A	20	50,8	|A-B|=15	|A-B|=38,1	0,394	A/B=0,571	A/B=0,571	1
B	35	88,9	|B-C|=5	|B-C|=12,7	0,394	B/C=0,875	B/C=0,875	1
C	40	101,6	|A-C|=20	|A-C|=50,8	0,394	A/C=0,5	A/C=0,5	1

Escala (numérica) de razões discretas

A escala de razões possui zero inerente de acordo com a natureza da característica sendo aferida. No caso de razões discreta, o resultado numérico da aferição é um valor inteiro. Normalmente trata-se de contagem.

Exemplos

Número de refeições em um dia (nenhuma, uma, duas, três, quatro, ...),
Frequência de consumo semanal de determinado alimento (1 vez, 2 vezes, 3 vezes, 4 vezes, 5 vezes, 6 vezes, 7 vezes),
Número de exemplares de *Aedes aegypti* na forma imatura (larvas, pupas)
Número de exemplares capturados (2, 3, 10, 30, 40, 50, 100 ...)
Número de ovos postos (1, 2, 20, 30, ... 50, 100).
Quantidade de repastos sanguíneos realizados por fêmea de inseto (1, 2, 3, 4, 5)

Escala de razões contínua

O resultado numérico é um valor pertencente ao conjunto dos números reais R ={-∞; ...; 0; 0,2; 0,73; 1; 2,48;...; +∞}.

Exemplos

Idade (anos)
Peso (g)
Altura (cm)
Nível de retinol sérico (μg/dl)
Circunferência da cintura (cm)
Precipitação pluviométrica em mm³ (quantidade de chuva por metro quadrado)
Tamanho da asa de um inseto (mm)
Peso seco de fêmeas ou pupas de mosquitos (0,53 mg, 0,43 mg;...)
Volume do repasto sanguíneo em (μl) (4,7; 3,6; 4,0; 4,9 ...)
Comprimento da asa em (mm)
Largura de partes do corpo de Triatomíneos em (mm)

Tipos de variáveis

De acordo com os níveis de mensuração, pode-se classificar a **natureza das variáveis** segundo a escala de mensuração em:

$$\text{VARIÁVEL:} \begin{cases} \text{qualitativa} \begin{cases} \text{nominal} \\ \text{ordinal} \end{cases} \\ \text{quantitativa} \begin{cases} \text{discreta} \\ \text{contínua} \end{cases} \end{cases}$$

O tipo da variável irá indicar a melhor forma para o dado ser apresentado em tabelas e gráficos, em medidas de resumo e, a análise estatística mais adequada.

Tabela 1 - Prevalência (%) de sedentarismo no lazer e global segundo variáveis socioeconômicas e demográficas em homens adultos em áreas do Estado de São Paulo, Brasil.

	HOMEM				
		Inativos no lazer		Inativos IPAQ	
	N	Prevalência (%)	Razão de prevalência (IC 95%)	Prevalência (%)	Razão de prevalência (IC 95%)
Faixa etária					
18 a 29	474	44,7	1	19,8	1
30 a 39	204	59,0	**1,10 (1,03-1,17)**	19,4	1,00 (0,92-1,08)
40 a 49	198	64,9	**1,14 (1,07-1,21)**	29,0	1,08 (0,98-1,19)
50 a 59	144	65,2	**1,14 (1,06-1,23)**	28,5	1,07 (0,99-1,16)
Total	1020	56,2		23,4	
		p=0,000		p=0,150	
Cor*					
Branca	716	54,1	1	26,4	1
Preta/parda	281	61,8	**1,05 (1,00-1,10)**	16,7	**0,92 (0,87-0,98)**
		p=0,050		p=0,016	
Situação conjugal					
Casado	500	58,5	1	28,1	1
Unido	177	62,5	1,03 (0,95-1,10)	21,6	0,95 (0,89-1,01)
Solteiro	267	36,0	**0,86 (0,79-0,93)**	14,2	**0,89 (0,83-0,95)**
Separado	44	58,2	1,00 (0,88-1,14)	7,0	**0,83 (0,75-0,93)**
Viúvo	19	65,4	1,04 (0,88-1,23)	46,6	1,14 (0,91-1,44)
		p=0,002		p=0,005	
Religião					
Evangélica	150	59,1	1	18,3	1
Outras	869	55,6	0,98 (0,90-1,06)	24,1	1,05 (0,96-1,14)
		p=0,602		p=0,291	
Escolaridade (em anos)					
0 a 7	375	70,0	1	20,0	1
8 a 11	478	46,6	**0,86 (0,81-0,91)**	24,1	1,03 (0,97-1,10)
12 ou mais	167	46,1	**0,86 (0,79-0,93)**	29,9	**1,08 (1,00-1,17)**
		p=0,000		p=0,128	
Renda per capita - salário mínimo					
<=2	533	59,3	1	20,2	1
> 2	488	52,8	0,96 (0,90-1,02)	26,5	1,05 (0,98-1,13)
		p=0,205		p=0,165	
Situação de ocupação**					
Ocupações de melhor qualificação	214	47,5	1	33,2	1
Ocupações menos qualificadas	647	62,5	**1,10 (1,02-1,19)**	21,1	**0,91 (0,84-0,98)**
Desempregados	59	37,6	0,93 (0,81-1,08)	18,0	**0,89 (0,79-1,00)**
Estudantes	80	19,3	**0,81 (0,73-0,89)**	11,9	**0,84 (0,75-0,94)**
		p=0,000		p=0,007	
Posse de carro					
Não	384	61,2	1	15,5	1
Sim	635	52,8	**0,95 (0,91-0,99)**	28,3	**1,11 (1,05-1,18)**
		p=0,020		p=0,001	

* Excluídos 15 outros **dois indivíduos declararam ser "do lar" e foram excluídos da amostra.
* *[Excluded 15 others **two individuals declared being "housewives" and were excluded from sample.]*

Fonte:
Zanchetta Luane Margarete, Barros Marilisa Berti de Azevedo, César Chester Luiz Galvão, Carandina Luana, Goldbaum Moisés, Alves Maria Cecília Goi Porto. Inatividade física e fatores associados em adultos, São Paulo, Brasil. Rev. Bras. Epidemiol. 2010;13(3): 387-399.

TABLE 1

Neurological manifestations of dengue in a case series of patients recruited in Central Brazil, 2005-2006.

Patient/ number	Sex	Age (years)	Fever/ days	2009 WHO classification	Neurological/ manifestation	MAC-ELISA	Virus isolation	Molecular test	Outcome
1	F	52	4	Dengue	Paresthesia/hands	Pos	Neg	DENV-3	Cured
2	F	3	5	Dengue	Paresthesia/face	Pos	Neg	---	Cured
3	F	35	6	DWS	Paresthesia/UL+LL	Pos	Neg	---	Cured
4	M	36	3	Severe	Paresthesia/hands	Pos	DENV-3	DENV-3	Cured
5	F	17	5	DWS	Paresthesia/lips	Pos	Neg	---	Cured
6	F	56	8	Severe	Encephalo	Pos	Neg	---	Cured
7	F	24	5	Severe	Paresthesia/UL+LL	Undeter.	DENV-3	DENV-3	Cured
8	F	20	6	DWS	Paresthesia/LL	Pos	Neg	---	Cured
9	F	31	9	DWS	Paresthesia/LL	Pos	Neg	---	Cured
10	F	36	7	DWS	Paresthesia/hands	Pos	Neg	---	Cured
11	F	29	7	Dengue	Paresthesia/hands	Pos	Neg	---	Cured
12	F	47	6	Dengue	Paresthesia/hands	Pos	Neg	---	Cured
13	F	41	10	Severe	Encephalo /Seizures	Pos	Neg	DENV-3	Death
14	F	45	6	Dengue	Paresthesia/LL	Pos	Neg	---	Cured
15	F	25	7	Severe	Paresthesia/LL	Pos	Neg	---	Cured
16	F	17	8	Severe	Meningoenceph	Pos	Neg	---	Cured
17	F	19	1	DWS	Paresthesia/feet	Pos	Neg	---	Cured
18	F	54	7	Dengue	Paresthesia+Paresis	Pos	Neg	---	Cured
19	F	27	6	DWS	Paresthesia	Pos	Neg	---	Cured
20	F	34	7	DWS	Paresthesia/LL	Pos	Neg	---	Cured
21	M	68	10	Severe	Encephalo	Pos	Neg	---	Cured
22	F	39	3	DWS	Paresthesia/UL+LL	Pos	Neg	---	Cured
23	M	14	5	Severe	Encephalo	Pos	Neg	---	Cured
24	M	71	15	Severe	Encephalitis/Paresis	Pos	Neg	---	Cured
25	F	38	7	DWS	Paresthesia/hands	Pos	Neg	---	Cured
26	F	15	16	Severe	Meningoenceph	Pos	Neg	DENV-3	Death
27	M	24	1	Severe	Encephalo/Seizures	Pos	Neg	---	Cured
28	F	18	3	Severe	Seizures	Pos	Neg	---	Cured

WHO: World Health Organization; **MAC-ELISA:** immunoglobulin M (IgM) antibody-capture enzyme-linked immunosorbent assay; **F:** female; **M:** male; **DENV:** dengue virus; **DWS:** dengue warning sign;; **LL:** lower limbs; **UL:** upper limbs;:; **Severe:** severe dengue; **Encephalo:** encephalopathy; **Meningoenceph:** meningoencephalitis.; **Pos**: positive; **Neg**: negative

Fonte: Tassara MP et al. Neurological manifestations of dengue in Central Brazil. Rev Soc Bras Med Trop 50 (3): 379-382, May-June, 2017

TABLE

Characteristics of cases of Zika virus infection, mainland France, 1 January–15 July 2016 (n = 625)

Characteristic	Number (%)
Sex	
Female	357 (57)
Age group in years	
<10	6 (1)
10–19	15 (2)
20–29	83 (13)
30–39	155 (25)
40–49	106 (17)
50–59	122 (20)
60–69	109 (17)
≥70	29 (5)
Regions visited during the incubation period[a]	
French departments and collectivities of the Americas	527 (84)
Caribbean islands	28 (4)
South America	25 (4)
Central America	8 (1)
Asia	1 (0.2)
Pacific	1 (0.2)
Africa	1 (0.2)
Not documented	26 (4)
No travel	8 (1.3)
Complications	
Guillain–Barré syndrome	2 (0.3)
Meningoencephalitis	1 (0.2)
Hospitalisation	29 (5)
Viraemic cases[b]	156 (25)
Month of notification	
January	8 (1)
February	76 (12)
March	74 (12)
April	121 (19)
May	144 (23)
June	158 (25)
July[c]	44 (7)

[a] During the two weeks before symptom onset.

[b] In an area in which the vector *Aedes albopictus* is established and active.

[c] Until 15 July 2016.

Fonte: A Septfons at al. Travel-associated and autochthonous Zika vírus infection in mainland France, 1 January to 15 July 2016. www.eurosurveillance.org

Considerar a pesquisa realizada em 2013, com 50 idosos do município de São Paulo. Entre as características investigadas foram obtidos dados do sexo do participante, peso e altura para construção do índice de massa corporal (imc); perguntou-se sobre doenças crônicas não transmissíveis (diabetes, hipertensão, doenças respiratórias e outras doenças crônicas) registrando-se o número de doenças no momento da pesquisa e nível de triglicérides (mg/dL).

id	idade	sexo	doenças crônicas	imc	triglic	id	idade	sexo	doenças crônicas	imc	triglic
1	94	M	1	26	128	26	82	F	1	24	89
2	74	F	4	31	166	27	82	F	1	34	92
3	74	F	1	24	79	28	85	F	4	25	181
4	64	F	0	22	166	29	87	F	3	20	91
5	61	F	2	27	61	30	74	F	3	27	171
6	89	F	0	27		31	72	F	3	45	176
7	84	F	3	26	211	32	83	F	3	35	165
8	73	M	2	27	157	33	91	F	1	24	38
9	93	F	1	28	124	34	73	F	1	22	46
10	87	F	3	26	111	35	66	F	1	31	
11	83	M	0	24	80	36	82	F	2	27	153
12	78	M	2	27	73	37	82	M	3	23	
13	76	M	1	23	205	38	85	F	2	20	99
14	76	F	1	29	101	39	86	F	2	29	66
15	72	M	3	24		40	92	M	3	29	130
16	65	F	2	35	170	41	71	M	6	27	72
17	68	M	2	29	126	42	75	M	0	30	87
18	66	F	1	37	193	43	74	M	1	34	219
19	91	M	0	19	92	44	61	M	0	25	
20	89	M	1	23	47	45	64	F	2	34	125
21	78	F	3	19	221	46	62	F	4	29	233
22	93	F		28	86	47	80	F	2	27	118
23	71	M	0	28	119	48	80	F	3	23	56
24	88	F	3	26	75	49	91	F	2	29	80
25	80	F	2	28	145	50	86	F	3	27	104

Exercício

Classificar quanto à natureza, as variáveis
Idade:
Sexo:
Doenças crônicas:
IMC:
Triglicérides:

A característica (variável) imc pode ser utilizada em categorias, por exemplo
abaixo de 21 indicando magreza;
de 22 a 27 eutrofia e
28 e mais, excesso de peso

Exercício 1

Classificar quanto à natureza, as seguintes variáveis:

Variável	**Tipo (natureza)**
Condição de saúde (doente, não doente)	
Tipo de parto (normal, cesário)	
Nível de colesterol sérico (mg/100cc)	
Tipo de abrigo (intradomiciliar, peridomiciliar)	
Estado nutricional (desnutrição, eutrofia, sobrepeso, obesidade)	
Altitude da área de coleta (metros)	
Resultado sorológico para presença de vírus (reagente, não reagente)	
Número de larvas e pupas em determinado criadouro	
Peso seco de pupas de exemplares de *Aedes aegypti* (mg)	
Estado de paridade da fêmea de inseto vetor (nulípara e parida)	
Tempo de um procedimento cirúrgico (minutos)	
Número de praias consideradas poluídas	

Coleta de dados

É a observação e registro das categorias ou das medidas das variáveis relacionadas ao objeto de estudo que ocorrem em unidades (indivíduos) de uma amostra ou população.

Definições e notação

População: totalidade de elementos sob estudo. Apresentam uma ou mais características em comum.

Supor o estudo sobre a ocorrência de mosquitos vetores de malária no Parque Estadual da Serra da Cantareira, município de São Paulo.

> População alvo – larvas de anofelinos do subgênero *Kerteszia*
> População de estudo – larvas do gênero *Anopheles* subgênero *Kerteszia* que se criam em bromélias na trilha do Pinheirinho fixadas em até 15 metros de altura e que estejam em condições de identificação.

Supor o estudo sobre a ocorrência de sobrepeso em crianças de 7 a 12 anos no Município de São Paulo.

> População alvo – todas as crianças nesta faixa etária deste município.
> População de estudo – crianças matriculadas em escolas.

Elementos: são unidades de análise por exemplo pessoas, células, domicílios, armadilhas, bromélias ou outro tipo de criadouro.

Amostra: é uma parte da população de estudo.

Amostragem: processo para obtenção de uma amostra. Tem como objetivo estimar parâmetros populacionais.

Parâmetro: Quantidade fixa de uma população.

Ex: Quantidade média de sangue ingerido por uma fêmea de mosquito, em uma picada.

Temperatura média no processo de transformação de larva em pupa.

Estimador: é uma fórmula matemática que permite estimar um parâmetro. Pode ser estimador no ponto e por intervalo.

Ex: Estimador no ponto

Média aritmética: $\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$,

onde $\sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$ e N = número de observações.

Estimador por intervalo

¶

Tabela 1 - Prevalência (%) de sedentarismo no lazer e global segundo variáveis socioeconômicas e demográficas em homens adultos em áreas do Estado de São Paulo, Brasil. ¶

		HOMEM			
		Inativos no lazer		Inativos IPAQ	
	N	Prevalência (%)	Razão de prevalência (IC 95%)	Prevalência (%)	Razão de prevalência (IC 95%)
Faixa etária					
18 a 29	474	44,7	1	19,8	1
30 a 39	204	59,0	**1,10 (1,03-1,17)**	19,4	1,00 (0,92-1,08)
40 a 49	198	64,9	**1,14 (1,07-1,21)**	29,0	1,08 (0,98-1,19)
50 a 59	144	65,2	**1,14 (1,06-1,23)**	28,5	1,07 (0,99-1,16)
Total	1020	56,2		23,4	
		p=0,000		p=0,150	
Cor*					
Branca	716	54,1	1	26,4	1
Preta/parda	281	61,8	**1,05 (1,00-1,10)**	16,7	**0,92 (0,87-0,98)**
		p=0,050		p=0,016	

Estimativa: Valor do estimador calculado em uma amostra. Estima o valor do parâmetro.

Ex: Peso seco médio (mg) de fêmeas de *Culex quinquefasciatus*: $\overline{x}$ = 0,541 mg. Os pesos (mg) para cada indivíduo eram: 0,419; 0,641; 0,592; 0,477; 0,613; 0,501.

Ex: Peso médio ao nascer, calculado em uma amostra de 120000 crianças nascidas no Município de São Paulo no ano de 2000: média amostral = $\overline{x} = 3000\ g$.

Indicações para utilizar uma amostra

População muito grande
Processo destrutivo de investigação
Novas terapias

Vantagens de realizar um estudo com amostragem:

Menor custo
Menor tempo para obtenção dos resultados
Possibilidade de objetivos mais amplos
Dados possivelmente mais fidedignos

Desvantagens

Resultados sujeitos à variabilidade

Se fossem retiradas amostras de uma população e calculada, por exemplo, a média, os valores das médias (estimativas) não seriam necessariamente iguais. Abaixo segue um exemplo de amostras de tamanho 5 retiradas dos dados do exercício 1, para a variável idade.

Amostra 1	Amostra 2	Amostra 3
76	74	94
82	86	76
65	72	65
84	73	72
71	87	83
$\bar{x}$=75,6 anos	$\bar{x}$=78,4 anos	$\bar{x}$=78,0 anos

Considerando-se os 50 idosos como a população, a média de idade **(μ) é 78,7 anos** (soma-se todos os valores e divide-se o resultado por 50)

Tipos de Amostragem

Probabilística: cada unidade amostral tem probabilidade conhecida e diferente de zero de pertencer à amostra. É usada alguma forma de sorteio para a obtenção da amostra.

Não probabilística: não se conhece a probabilidade de cada unidade amostral pertencer à amostra. Algumas unidades terão probabilidade zero de pertencer à amostra.
Ex: amostragem intencional; por voluntários; acesso mais fácil; por quotas.

Tipos de amostragem probabilística:

- aleatória simples (com e sem reposição);
- sistemática;
- com partilha proporcional ao tamanho do estrato;
- por conglomerado.

Amostragem aleatória simples (AAS)

É o processo de amostragem onde qualquer subconjunto de n elementos diferentes de uma população de N elementos tem mesma probabilidade de ser sorteado. Tamanho da população: N; tamanho da amostra: n; fração global de amostragem ou probabilidade de sortear um indivíduo = $\frac{n}{N}$.

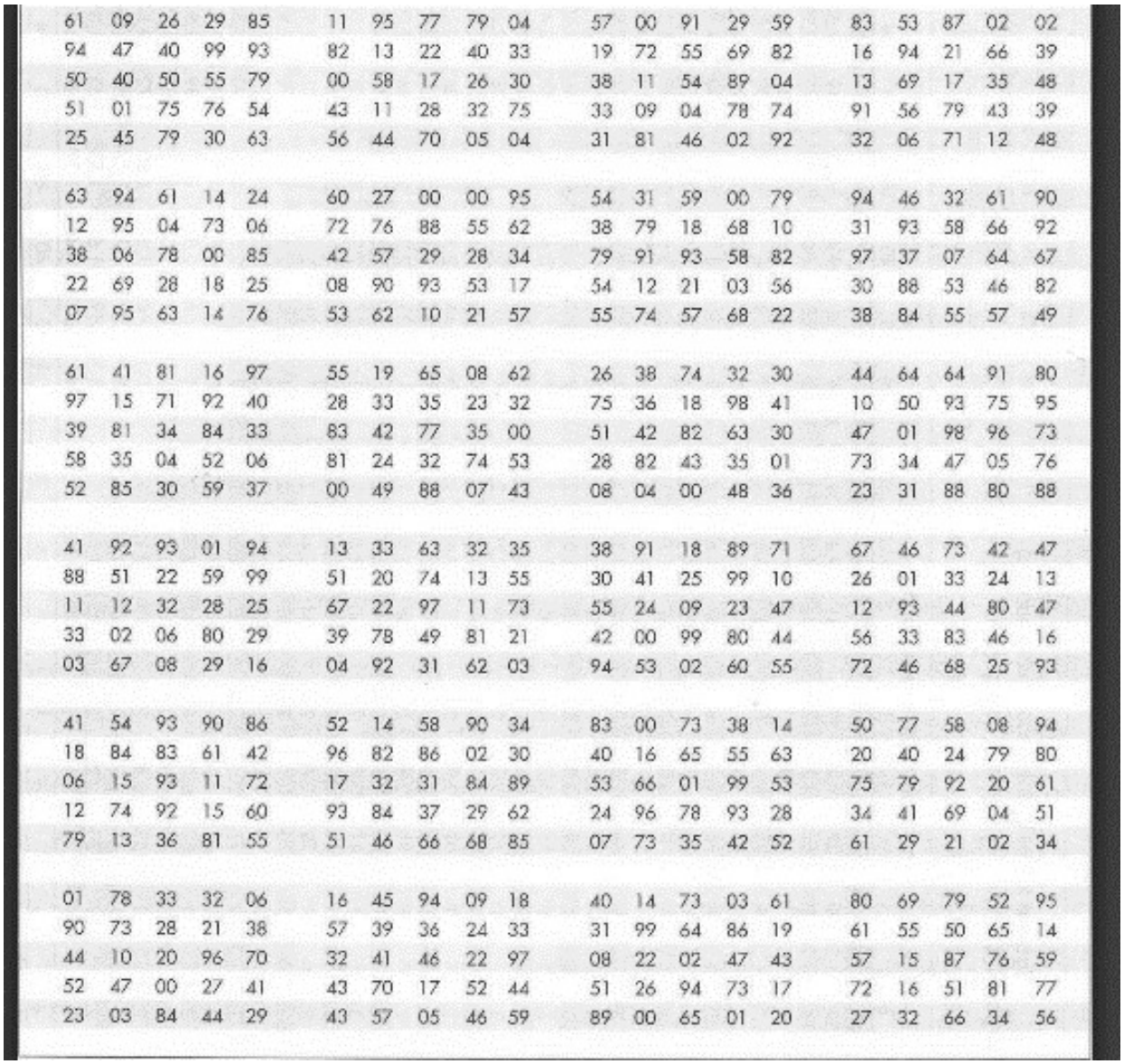
- É necessário ter um sistema de referência que contenha todos os elementos da população da qual será retirada a amostra;
- Utilização da tabela de números aleatórios – mecânica;
- Utilização de programas computacionais.

Considerações a respeito da tabela de números aleatórios

Para a utilização desta tabela é necessário:

- Definir o número de dígitos que serão utilizados
- Sortear um início
- Pré definir um caminho a ser percorrido

Tabela de números equiprováveis (aleatórios)



61	09	26	29	85	11	95	77	79	04	57	00	91	29	59	83	53	87	02	02
94	47	40	99	93	82	13	22	40	33	19	72	55	69	82	16	94	21	66	39
50	40	50	55	79	00	58	17	26	30	38	11	54	89	04	13	69	17	35	48
51	01	75	76	54	43	11	28	32	75	33	09	04	78	74	91	56	79	43	39
25	45	79	30	63	56	44	70	05	04	31	81	46	02	92	32	06	71	12	48
63	94	61	14	24	60	27	00	00	95	54	31	59	00	79	94	46	32	61	90
12	95	04	73	06	72	76	88	55	62	38	79	18	68	10	31	93	58	66	92
38	06	78	00	85	42	57	29	28	34	79	91	93	58	82	97	37	07	64	67
22	69	28	18	25	08	90	93	53	17	54	12	21	03	56	30	88	53	46	82
07	95	63	14	76	53	62	10	21	57	55	74	57	68	22	38	84	55	57	49
61	41	81	16	97	55	19	65	08	62	26	38	74	32	30	44	64	64	91	80
97	15	71	92	40	28	33	35	23	32	75	36	18	98	41	10	50	93	75	95
39	81	34	84	33	83	42	77	35	00	51	42	82	63	30	47	01	98	96	73
58	35	04	52	06	81	24	32	74	53	28	82	43	35	01	73	34	47	05	76
52	85	30	59	37	00	49	88	07	43	08	04	00	48	36	23	31	88	80	88
41	92	93	01	94	13	33	63	32	35	38	91	18	89	71	67	46	73	42	47
88	51	22	59	99	51	20	74	13	55	30	41	25	99	10	26	01	33	24	13
11	12	32	28	25	67	22	97	11	73	55	24	09	23	47	12	93	44	80	47
33	02	06	80	29	39	78	49	81	21	42	00	99	80	44	56	33	83	46	16
03	67	08	29	16	04	92	31	62	03	94	53	02	60	55	72	46	68	25	93
41	54	93	90	86	52	14	58	90	34	83	00	73	38	14	50	77	58	08	94
18	84	83	61	42	96	82	86	02	30	40	16	65	55	63	20	40	24	79	80
06	15	93	11	72	17	32	31	84	89	53	66	01	99	53	75	79	92	20	61
12	74	92	15	60	93	84	37	29	62	24	96	78	93	28	34	41	69	04	51
79	13	36	81	55	51	46	66	68	85	07	73	35	42	52	61	29	21	02	34
01	78	33	32	06	16	45	94	09	18	40	14	73	03	61	80	69	79	52	95
90	73	28	21	38	57	39	36	24	33	31	99	64	86	19	61	55	50	65	14
44	10	20	96	70	32	41	46	22	97	08	22	02	47	43	57	15	87	76	59
52	47	00	27	41	43	70	17	52	44	51	26	94	73	17	72	16	51	81	77
23	03	84	44	29	43	57	05	46	59	89	00	65	01	20	27	32	66	34	56

Amostragem sistemática

Utiliza-se a ordenação natural dos elementos da população (prontuários, casa, ordem de nascimento).

- Intervalo de amostragem $k = \frac{N}{n}$, onde

  N= tamanho da população e n = tamanho da amostra

- Início casual i, sorteado entre 1 e k, inclusive
- Amostra sorteada é composta pelos elementos: i, i+k, i+2k, ..., i+(n-1)k

OBS: É necessário ter cuidado com a periodicidade dos dados, por exemplo se for feito sorteio de dia no mês, pode cair sempre em um domingo onde o padrão de ocorrência do evento pode ser diferente.

Exemplo: N=80; n=10; $k = \frac{N}{n} = \frac{80}{10} = 8$; início casual: $1 \leq i \leq 8$

Começo casual **sorteado**: i=4

Amostra composta dos elementos:

i	4
i+k	12
i+2k	20
i+3k	28
i+4k	36
i+5k	44
i+6k	52
i+7k	60
i+8k	68
i+(n-1)k	76

Se o intervalo de amostragem não for inteiro proceder da seguinte forma:

N= 321 ; n=154; $K = \frac{N}{n} = \frac{321}{154} = 2{,}084$; i deve ser um número sorteado entre 1 e 2,084.

Sortear um número entre 1000 e 2084 e dividir o resultado por 1000

Número sorteado = 1941, portanto i=1,941

Indivíduos:

		Elementos
i	1,941	1
i+k	1,941+2,084 = 4,025	4
i+2k	1,941+4,1680 = 6,109	6
i+3k	1,941+6,252 = 8,193	8
.	.	.
.	.	.
.	.	.
i+(n-1)k	1,941+318,852 = 320,793	320

Amostragem casual simples estratificada com partilha proporcional

A população possui estratos com tamanhos:
N_1; N_2; N_3, onde a soma dos estratos é o tamanho da população, ou seja $\sum N_i = N$

A amostra deve conter os elementos da população nas mesmas proporções dos estratos. Tem-se que os tamanhos dos estratos amostrais são n_1, n_2 e n_3 tal que $\sum n_i = n$

Aplicando-se a proporção:

$$\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = n\frac{N_i}{N}$$

Exemplo:
N=500; N_1=50; N_2=150; N_3=300 e n=40

Estrato i	Tamanho do estrato na população N_i	Tamanho do estrato na amostra n_i	$\frac{n_i}{n} = \frac{N_i}{N}$
1	50	4	0,1
2	150	12	0,3
3	300	24	0,6
Total	500	40	

$$n_1 = 40\frac{50}{500} = 4\,;\; n_2 = 40\frac{150}{500} = 12\,;\; n_3 = 40\frac{300}{500} = 24$$

Amostragem por conglomerado:

O conglomerado é um conjunto de elementos formando uma unidade amostral. Se a unidade amostral for indivíduo e forem sorteados domicílios, então a amostragem é por conglomerado.

Coleta de dados, apuração de dados

Coleta de dados: é a observação e registro da categoria ou medida de variáveis relacionadas ao objeto de estudo que ocorrem em unidades (indivíduos) de uma amostra ou população.

Apuração de dados: é o processo no qual conta-se o número de vezes que a variável assumiu um determinado valor (frequência de ocorrência). Pode ser manual, mecânica ou eletrônica (programas estatísticos: Epi info, Stata, Excel, SPSS, SAS, R, S-Plus).

Distribuição de frequências - correspondência entre categorias (valores) e frequência de ocorrência.

Distribuição de frequências - correspondência entre categorias ou valores da variável e frequência de ocorrência.

Notação:

X : variável

x_i : valor observado para o indivíduo i

Exemplos de distribuição de frequências pontuais

Unidade de observação: mosquito
X: Local de captura de mosquitos (intra domicilio, peridomicilio, campo)

Dados:

Mosquito	i	Local de captura
	1	intra domicilio
	2	intra domicilio
	3	campo
	4	peridomicilio
	5	peridomicilio
	6	campo
	7	peridomicilio
	8	campo
	9	peridomicilio
	10	peridomicilio

Distribuição de frequência dos mosquitos segundo local:

Local de captura	n
Intra domicilio	2
Peridomicilio	5
Campo	3

Exemplo
X: Número de repastos sanguíneos para completar um ciclo gonotrófico
Dados:

Mosquito	i	Valor
	1	1
	2	1
	3	2
	4	4
	5	3
	6	2
	7	3
	8	2
	9	2
	10	4

Distribuição de frequência de mosquitos segundo número de repastos:

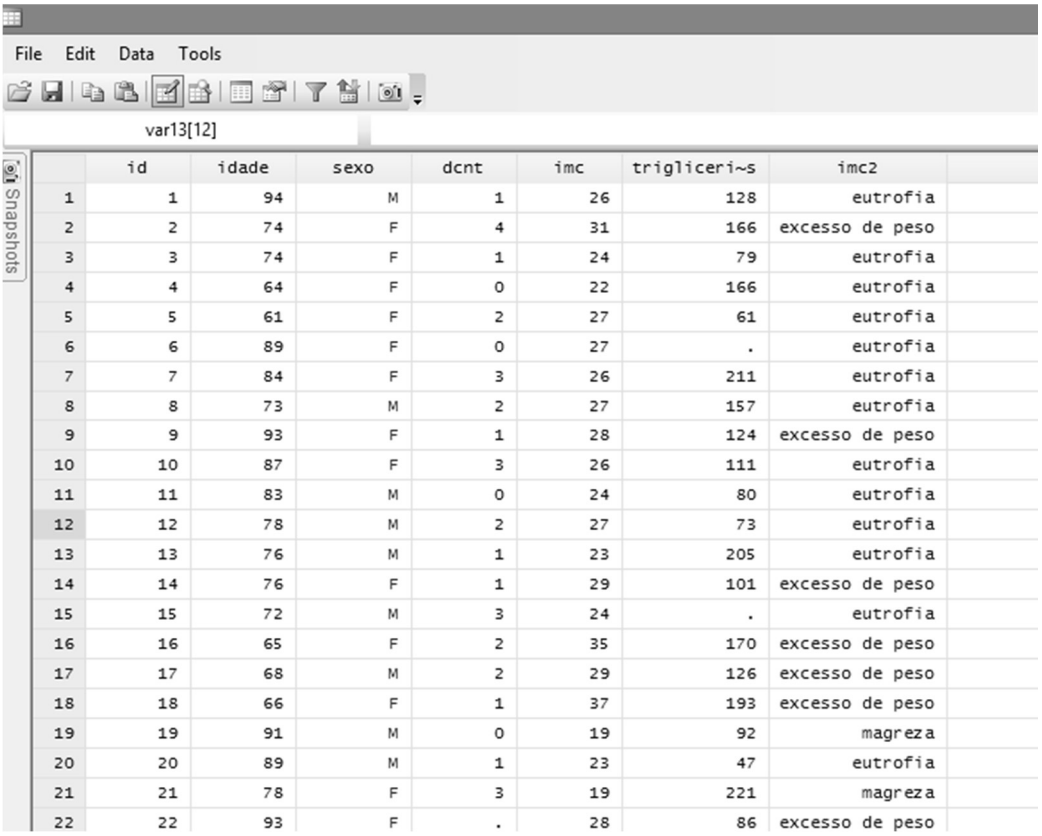
Número de repastos	n
1	2
2	4
3	2
4	2

Banco de dados construído no pacote Stata utilizando o exemplo 1:

Banco de dados construído no pacote Stata utilizando o exemplo 1:

Nome da variável	Detalhamento	Códigos
id	Número de identificação do participante	
idade	Idade (anos)	
sexo	Sexo	1-masculino 2-feminino
imc	Índice de massa corporal	
dcnt	Número de doenças	
triglicerides	Concentração de triglicérides (mg/dL)	

Tela do Stata (banco de dados: idosos, município de São Paulo)



	id	idade	sexo	dcnt	imc	trigliceri~s	imc2
1	1	94	M	1	26	128	eutrofia
2	2	74	F	4	31	166	excesso de peso
3	3	74	F	1	24	79	eutrofia
4	4	64	F	0	22	166	eutrofia
5	5	61	F	2	27	61	eutrofia
6	6	89	F	0	27	.	eutrofia
7	7	84	F	3	26	211	eutrofia
8	8	73	M	2	27	157	eutrofia
9	9	93	F	1	28	124	excesso de peso
10	10	87	F	3	26	111	eutrofia
11	11	83	M	0	24	80	eutrofia
12	12	78	M	2	27	73	eutrofia
13	13	76	M	1	23	205	eutrofia
14	14	76	F	1	29	101	excesso de peso
15	15	72	M	3	24	.	eutrofia
16	16	65	F	2	35	170	excesso de peso
17	17	68	M	2	29	126	excesso de peso
18	18	66	F	1	37	193	excesso de peso
19	19	91	M	0	19	92	magreza
20	20	89	M	1	23	47	eutrofia
21	21	78	F	3	19	221	magreza
22	22	93	F	.	28	86	excesso de peso

Distribuição de frequências com dados pontuais utilizando o comando *tabulate*, do programa Stata

Dados pontuais – variável qualitativa nominal e variável quantitativa discreta.

```
-> tabulation of sexo

       sexo |      Freq.     Percent        Cum.
------------+-----------------------------------
          F |         34       68.00       68.00
          M |         16       32.00      100.00
------------+-----------------------------------
      Total |         50      100.00


    -> tabulation of dcnt

           dcnt |      Freq.     Percent        Cum.
    ------------+-----------------------------------
              0 |          7       14.29       14.29
              1 |         13       26.53       40.82
              2 |         12       24.49       65.31
              3 |         13       26.53       91.84
              4 |          3        6.12       97.96
              6 |          1        2.04      100.00
    ------------+-----------------------------------
          Total |         49      100.00
```

Variável quantitativa contínua utilizando o comando *tabulate* do Stata.

Telas de saída do comando tabulate das variáveis idade e imc

```
-> tabulation of idade

      idade |      Freq.     Percent        Cum.
------------+-----------------------------------
         61 |          2        4.00        4.00
         62 |          1        2.00        6.00
         64 |          2        4.00       10.00
         65 |          1        2.00       12.00
         66 |          2        4.00       16.00
         68 |          1        2.00       18.00
         71 |          2        4.00       22.00
         72 |          2        4.00       26.00
         73 |          2        4.00       30.00
         74 |          4        8.00       38.00
         75 |          1        2.00       40.00
         76 |          2        4.00       44.00
         78 |          2        4.00       48.00
         80 |          3        6.00       54.00
         82 |          4        8.00       62.00
         83 |          2        4.00       66.00
         84 |          1        2.00       68.00
         85 |          2        4.00       72.00
         86 |          2        4.00       76.00
         87 |          2        4.00       80.00
         88 |          1        2.00       82.00
         89 |          2        4.00       86.00
         91 |          3        6.00       92.00
         92 |          1        2.00       94.00
         93 |          2        4.00       98.00
         94 |          1        2.00      100.00
------------+-----------------------------------
      Total |         50      100.00

-> tabulation of imc

        imc |      Freq.     Percent        Cum.
------------+-----------------------------------
         19 |          2        4.00        4.00
         20 |          2        4.00        8.00
         22 |          2        4.00       12.00
         23 |          4        8.00       20.00
         24 |          5       10.00       30.00
         25 |          2        4.00       34.00
         26 |          4        8.00       42.00
         27 |          9       18.00       60.00
         28 |          4        8.00       68.00
         29 |          6       12.00       80.00
         30 |          1        2.00       82.00
         31 |          2        4.00       86.00
         34 |          3        6.00       92.00
         35 |          2        4.00       96.00
         37 |          1        2.00       98.00
         45 |          1        2.00      100.00
------------+-----------------------------------
      Total |         50      100.00
```

```
-> tabulation of triglicerides

triglicerid |
         es |      Freq.     Percent        Cum.
------------+-----------------------------------
         38 |          1        2.22        2.22
         46 |          1        2.22        4.44
         47 |          1        2.22        6.67
         56 |          1        2.22        8.89
         61 |          1        2.22       11.11
         66 |          1        2.22       13.33
         72 |          1        2.22       15.56
         73 |          1        2.22       17.78
         75 |          1        2.22       20.00
         79 |          1        2.22       22.22
         80 |          2        4.44       26.67
         86 |          1        2.22       28.89
         87 |          1        2.22       31.11
         89 |          1        2.22       33.33
         91 |          1        2.22       35.56
         92 |          2        4.44       40.00
         99 |          1        2.22       42.22
        101 |          1        2.22       44.44
        104 |          1        2.22       46.67
        111 |          1        2.22       48.89
        118 |          1        2.22       51.11
        119 |          1        2.22       53.33
        124 |          1        2.22       55.56
        125 |          1        2.22       57.78
        126 |          1        2.22       60.00
        128 |          1        2.22       62.22
        130 |          1        2.22       64.44
        145 |          1        2.22       66.67
        153 |          1        2.22       68.89
        157 |          1        2.22       71.11
        165 |          1        2.22       73.33
        166 |          2        4.44       77.78
        170 |          1        2.22       80.00
        171 |          1        2.22       82.22
        176 |          1        2.22       84.44
        181 |          1        2.22       86.67
        193 |          1        2.22       88.89
        205 |          1        2.22       91.11
        211 |          1        2.22       93.33
        219 |          1        2.22       95.56
        221 |          1        2.22       97.78
        233 |          1        2.22      100.00
------------+-----------------------------------
      Total |         45      100.00
```

Tabelas e gráficos

- Possibilitam conhecer as características da população sob estudo porque resumem e organizam os dados.
- Permitem identificar rapidamente onde a maioria dos indivíduos está e quais são os padrões de ocorrência de valores.
- Fornecem uma idéia prévia de como serão as estimativas dos parâmetros sob investigação.
- Auxiliam na identificação dos testes estatísticos que serão efetuados em fases mais avançadas da análise dos dados.

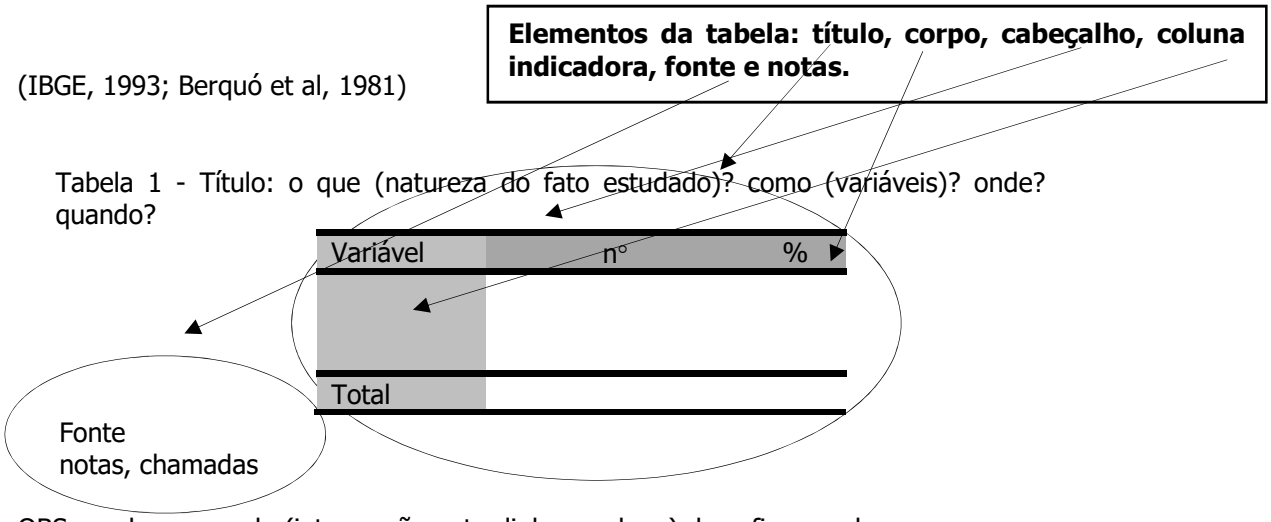
Guia de apresentação tabular do IBGE

http://biblioteca.ibge.gov.br/visualizacao/livros/liv23907.pdf

Apresentação tabular

Elementos da tabela: título, corpo, cabeçalho, coluna indicadora, fonte e notas.

(IBGE, 1993; Berquó et al, 1981)



OBS: nenhuma casela (intersecção entre linha e coluna) deve ficar em branco.

A tabela deve ser uniforme quanto ao número de casas decimais e conter os símbolos **–** ou **0** quando o valor numérico é nulo e **...** quando não se dispõe do dado.

Apresentação de variável qualitativa

Exemplo
Considerando-se a variável imc para classificar indivíduos segundo o estado nutricional

Tabela 1- Distribuição de idosos segundo classificação nutricional. Município de São Paulo,2013.

Estado nutricional[2]	n	%
Magreza	4	8,0
Eutrofia	26	52,0
Excesso de peso	20	40,0
Total	50	100

[2] magreza: $\leq$ 21 kg/m^2; eutrofia: 22-27 kg/m^2; excesso de peso $\geq$ 28 kg/m^2

Exemplo:
Distribuição de crianças[1] segundo níveis séricos de retinol. Cansação – Bahia, 1992

Nível de retinol sérico[2]	n	%
Aceitável	89	55,3
Baixo	65	40,4
Deficiente	7	4,3
Total	161	100

[1] 0 –72 meses
[2] aceitável: 20,0 – 49,9 μg/dl; baixo: 10,0 – 19,9 μg/dl; deficiente: <10,0 μg/dl
Fonte: Prado MS et al., 1995.

Exemplo:
Distribuição de Culicídeos segundo espécie coletados na área de influência indireta da Usina Hidrelétrica de Porto Primavera, SP e MS, Brasil, 1992-1993.

Táxon	n	%
Culex (Culex) quinquefasciatus	244	25,3
Culex (Culex) sp.gr.Coronator	135	14,0
Culex (Culex) spp	131	13,6
Culex (Melanoconion) spp.	111	11,5
Anopheles (Nyssorhynchys) albitarsis	76	7,9
Culex (Culex) sp.pr.inflictus	59	6,1
Outras[*]	210	21,7
Total	966	100

(*) Espécies ou grupos: *Aedeomyia squamipennis, Aedes aegypti, Aedes fluviatilis, Anopheles argyritarsis, Anopheles evansae, Anopheles oswaldoi, Anopheles triannulatus, Culex chidesteri, Culex camposi, Culex dolosus, Culex mollis, Culex saltanensis, Culex surinamensis, Culex bigoti, Culex sp.* gr. Atratrus, *Culex aureonatatus, Culex bastagarius, Culex innovator ou pilosus, Culex oedipus, Culex theobaldi, Culex vaxus, Psorophora albigenu, Psorophora confinnis, Psorophora sp., Psorophora ciliata, Toxorhynchites portoricensis px. Uranotaenia apicalis, Uranotaenia geometrica, Uranotaenia pulcherrima, Uranotaenia lowii, Uranotaenia sp.*
Fonte: Adaptado de Natal D et al., 1995. Revista brasileira de Entomologia. 39(4): 897-899.

Exercício 2

Apresentar e descrever os dados dos idosos relativos as variáveis sexo, número de doenças crônicas em tabelas de distribuição de frequência.

Variável sexo

Variável número de doenças crônicas

Apresentação tabular de variável quantitativa contínua (em intervalos de classe)

Para representar variáveis quantitativas contínuas é necessário construir intervalos de classe definidos como um conjunto de observações contidas entre dois valores limite (limite inferior e limite superior).

Os valores dos limites inferior e superior podem ou não estarem contidos no intervalo. Se um valor estiver contido a representação do intervalo deverá indicar que este é fechado naquele limite.

Por exemplo os intervalos abaixo são fechados no limite inferior. Acrescente um novo intervalo antes e após o intervalo apresentado.

5 | -- 10	Intervalo fechado no limite inferior e aberto no limite superior (contém o valor 5 mas não contém o valor 10)

Os intervalos abaixo são fechados nos limites inferior e superior. Acrescente um novo intervalo antes e após o intervalo apresentado.

5 |--| 10	Intervalo fechado nos limites inferior e superior (contém os valores e 10)

OBS: Representar o intervalo 0 |-- | 11 meses é equivalente a representá-lo como 0 |-- 12 meses.

A **amplitude do intervalo** é o tamanho do intervalo de classe.

Supor a variável idade (anos). O intervalo 5|--10 (anos) tem amplitude 5 que é igual à diferença entre os limites (10-5=5) e inclui as idades 5, 6, 7, 8 e 9 anos. Um indivíduo com 10,3 anos não estaria incluído neste intervalo.

A amplitude do intervalo 5|--|10 é igual a 6 porque o intervalo é fechado no 10 e inclui todos os valores 10, 1; 10,3; ...10,9; ... 10,999999 ...., mas não inclui o 11.

Idade é quantitativa contínua e, portanto, entre dois valores existem infinitos valores. Assim, não é possível saber qual é o valor que antecede o 11 (10,999999.... até o infinito). Neste caso, para o cálculo da amplitude, utiliza-se toda a informação do intervalo e por isso, seu cálculo é feito com o valor 11. A amplitude será 11-5 = 6 (estão incluídos aí os valores 5, 6, 7, 8, 9 e 10) ou então será (10-5)+1 = 6.

A amplitude do intervalo e o número de intervalos dependem basicamente do problema específico e da literatura existente sobre o assunto que será utilizada para se comparar os resultados. O **ponto médio do intervalo** é calculado somando-se o limite inferior e limite superior, dividindo-se o resultado por dois.

Na construção dos intervalos de classe é necessário que eles sejam **mutuamente excludentes** (um indivíduo não pode ser classificado em dois intervalos ao mesmo tempo) e **exaustivos** (nenhum indivíduo pode ficar sem classificação).

Exemplo:
X: Peso seco de mosquitos Culicidae (mg)

Mosquito	i	Valor	i	Valor	i	Valor
	1	0,512	8	0,291	15	0,524
	2	0,670	9	0,334	16	0,389
	3	0,430	10	0,278	17	0,524
	4	0,532	11	0,227	18	0,477
	5	0,789	12	0,432	19	0,625
	6	0,459	13	0,379	20	0,532
	7	0,339	14	0,553		

Distribuição de frequência:

Peso seco (mg)	n
0,200 |-- 0,300	3
0,300 |-- 0,400	4
0,400 |-- 0,500	4
0,500 |-- 0,600	6
0,600 |-- 0,700	2
0,700 |-- 0,800	1

Como idade é variável quantitativa contínua, a melhor forma de apresentá-la em tabelas é utilizando intervalos de valores denominados intervalos de classe.
Ex:
x: 5, 5, 15, 20, 20, 20, 21, 21, 22, 22

idade	frequência	%
5 |-- 10	2	20
10 |-- 15	0	-
15 |-- 20	1	10
20 |-- 25	7	70
Total	10	100

X: Peso ao nascer (g)
x: 2250, 3025, 1600, 2725, 3750, 3950, 2400, 2180, 2520, 2530

Peso (g)	frequência	%
1500|--2000	1	10
2000|--2500	3	30
2500|--3000	3	30
3000|--3500	1	10
3500|--4000	2	20
Total	10	100

X: Altura de adultos (cm)
x: 1,63; 1,60; 1,59; 1,60; 1,45; 1,73; 2,05; 1,85

Altura (cm)	n	%
1,45|--1,55	1	12,5
1,55|--1,65	4	50,0
1,65|--1,75	1	12,5
1,75|--1,85	0	-
1,85|--1,95	1	12,5
1,95|--2,05	0	-
2,05|--2,15	1	12,5
Total	8	100

Ex:
X: Peso seco em (mg) de fêmeas de *Anopheles darlingi* (vetor de malária)

x: 0,10; 0,14, 0,20; 0,24; 0,26; 0,27; 0,30; 0,32; 0,34; 0,37; 0,44

Distribuição de fêmeas de *Anopheles darlingi segundo peso seco. Sítios em Capanema, Pará, Brasil, 1995*

Peso (mg)	Frequência	%
0,10 |-- 0,20	2	20
0,20 |-- 0,30	4	40
0,30 |-- 0,40	3	30
0,40 |-- 0,45	1	10
Total	10	100

Fonte: Adaptado de Lounibos, LP et. al., 1995.

Ex:
X: Tamanho do corpo de *Anopheles darlingi* = comprimento da asa em (mm)
x: 1,7; 2,2; 2,3; 2,4; 2,5; 2,6; 2,7; 2,8; 2,9; 3,0; 3,1; 3,2; 3,3; 3,4

Distribuição de fêmeas de *Anopheles darlingi* segundo tamanho do corpo. Sítios em Capanema, Pará, Brasil, 1995

Comprimento da asa (mm)	Frequência	%
1,5 |-- 2,0	1	7,1
2,0 |-- 2,5	3	21,4
2,5 |-- 3,0	6	42,9
3,0 |-- 3,5	4	28,6
Total	14	100

Fonte: Adaptado de Lounibos, LP et. al., 1995.

Exercício 3

Rev Saúde Pública 2013;47(3):560-70

Artigos Originais DOI: 10.1590/S0034-8910.2013047004379

Maria Helena D'Aquino Benício[I,II]

Ana Paula Bortoletto Martins[II]

Sonia Isoyama Venancio[II,III]

Aluísio Jardim Dornellas de Barros[IV]

Estimativas da prevalência de desnutrição infantil nos municípios brasileiros em 2006

Estimates of the prevalence of child malnutrition in Brazilian municipalities in 2006

Tabela 1. Prevalência de desnutrição infantil nas crianças de seis a 59 meses segundo os fatores de estudo. Brasil, 2006.

Variáveis	n de crianças (n = 3.931)	%	Prevalência de desnutrição (%)	p
Nível Individual				
Idade (meses)				0,014
6 ├ 24	1.290	33,6	10,0	
24 ├ 60	2.641	66,4	6,0	
Sexo				0,012
Masculino	2.020	52,6	8,8	
Feminino	1.911	47,4	5,8	
Nível Domiciliar				
Escore socioeconômico				0,006
Quinto 1	794	21,6	9,5	
Quinto 2	644	18,4	10,0	
Quinto 3	694	20,0	6,7	
Quinto 4	706	20,2	4,3	
Quinto 5	562	19,8	2,8	
Número de pessoas por cômodo				< 0,001
< 2	3.410	92,5	6,6	
2 ou mais	521	7,5	17,4	
Água com canalização interna no domicílio da criança				< 0,001
Sim	3.192	86,5	6,4	
Não	737	13,6	15,5	

Continua

Continuação

Número de crianças menores de cinco anos por domicílio				0,001
1	2.281	67,3	5,6	
2	1.228	27,5	10,7	
3 ou mais	422	5,2	12,7	
Localização do domicílio da criança				0,540
Urbano	1.383	80,9	8,1	
Rural	2.548	19,1	7,2	
Nível Municipal				
Região onde se localiza o município de residência da criança				< 0,001
Norte	881	10,8	16,1	
Nordeste	769	27,3	6,4	
Centro-Sul	788	41,8	6,2	
Percentual de cobertura da ESF (%) – 2006				0,383
0 ├ 15	685	18,6	8,0	
15 ├ 30	722	22,0	4,5	
30 ├ 50	686	18,3	8,2	
50 ├ 70	537	12,9	8,9	
≥ 70	1.301	28,2	7,9	
Porte populacional – 2006				0,002
<15 mil	751	12,3	8,7	
15 ├ 50 mil	1.012	24,3	7,3	
50 ├ 100 mil	502	10,4	14,8	
100 ├ 1 milhão	1.022	33,9	6,7	
mais de 1 milhão	644	19,1	3,7	

ESF: Estratégia Saúde da Família

Quais são as variáveis quantitativas contínuas representadas em intervalos de classe?

Exercício 4

Apresentar e descrever os dados dos idosos relativos a variável triglicérides em uma tabela de distribuição de frequência.

Tabela de dupla entrada

Distribuição de crianças[1] segundo níveis séricos de retinol e idade. Cansação – Bahia, 1992.

Faixa etária (meses)	Aceitável		Inadequado		Total	
	n	%	n	%	n	%
<12	5	45,5	6	54,5	11	100
12|--24	10	43,5	13	56,5	23	100
24|--36	19	54,3	16	45,7	35	100
36|--48	21	65,6	11	34,5	32	100
48|--60	16	43,2	21	56,8	37	100
60|--73	18	78,3	5	21,7	23	100
Total	89	55,3	72	44,7	161	100

[1] 0 –72 meses.
[2] aceitável: 20,0 – 49,9 μg/dl; baixo: 10,0 – 19,9 μg/dl; deficiente: <10,0 μg/dl.
Fonte: Prado MS et al., 1995.

Interpretação:
Faixas etárias menores concentram mais crianças com nível inadequado de retinol sérico

Distribuição de crianças[1] segundo níveis séricos de retinol e idade. Cansação – Bahia, 1992.

Faixa etária (meses)	Aceitável		Inadequado		Total	
	n	%	n	%	n	%
<12	5	5,6	6	8,3	11	6,8
12|--24	10	11,2	13	18,1	23	14,3
24|--36	19	21,3	16	22,2	35	21,7
36|--48	21	23,6	11	15,3	32	19,9
48|--60	16	18,0	21	29,2	37	23,0
60|--73	18	20,2	5	6,9	23	14,3
Total	89	100	72	100	161	100

[1] 0 –72 meses.
[2] aceitável: 20,0 – 49,9 μg/dl; baixo: 10,0 – 19,9 μg/dl; deficiente: <10,0 μg/dl.
Fonte: Prado MS et al., 1995.

Interpretação
Crianças com nível inadequado são mais jovens que crianças com nível adequado.

Exemplo
Distribuição de pneus com coleta de larvas de *Aedes aegypti* segundo número de larvas e predação(*). Dar es Salaan, Tanzânia, 1973.

Número de larvas	Predador ausente		Predador presente	
	n	%	n	%
0	75	47,2	184	83,0
1|--| 10	51	32,1	27	12,1
11|--| 20	16	10,1	8	3,6
21|--| 50	10	6,3	3	1,3
51|--|100	5	3,1	0	-
101|--|300	2	1,2	0	-
Total	159	100	222	100

(*) larva de *Toxorhynchites brevipalpis*
Fonte:Clementes A.N., The biology of Mosquitoes. Vol(2) pag.203, 1999.

Interpretação:

Exercício 5
Os dados a seguir são de um estudo que investiga a relação entre níveis de β-caroteno (mg/L) e hábito de fumar em gestantes.

a) Calcule as frequências relativas. Fixando o 100% no total de fumantes e não fumantes.
b) Calcule as frequências relativas. Fixando o 100% no total do nível de B-caroteno (MG/L).
c) Interprete os resultados. Existe alguma indicação de existência de associação entre as variáveis? Justifique.

a)
Distribuição de gestantes segundo níveis de **β-caroteno (mg/L)** e hábito de fumar, Joinvile, Brasil, 2002.

β-caroteno (mg/L)	Fumante		Não Fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46		74		120	
Normal (0,214 – 1,00)	12		58		70	
Total	58		132		190	

Fonte: Silmara Silva. Tese de Mestrado/FSP/USP.
Interpretação:

b)
Distribuição de gestantes segundo níveis de **β-caroteno (mg/L)** e hábito de fumar, Joinvile, Brasil, 2002.

β-caroteno (mg/L)	Fumante		Não Fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46		74		120	
Normal (0,214 – 1,00)	12		58		70	
Total	58		132		190	

Fonte: Silmara Silva. Tese de Mestrado/FSP/USP.
Interpretação:

Exercício 6

São apresentados na tabela abaixo o local de captura de *Culex quinquefasciatus* – Intradomicílio e peridomicílio e nível socieconômico dos habitantes de setores censitários urbanos do município de Marília, de junho de 2007 a agosto de 2008.

a) Calcule os percentuais;
b) Interprete os dados

Nível	Intradomicilio		Peridomicilio		Total	
Sócio-econômico	n	%	n	%	n	%
Baixo	119		13		132	
Intermediário	71		0		71	
Alto	57		9		66	
Total	247		22		269	

Fonte: Telles-de-Deus, J. Hábito alimentar de *Aedes aegypti* e *Culex quinquefasciatus* e sua implicação na capacidade reprodutiva. São Paulo, 2011 [Tese de Doutorado, Faculdade de Saúde Pública da Universidade de São Paulo].

Apresentação gráfica: diagrama de barras, diagramas de setores circulares, diagrama linear, histograma, polígono de frequência, ogiva de frequências acumuladas.

Diagrama de barras

Utilizado para representar variáveis qualitativa nominal e ordinal, e quantitativa discreta.

Características do diagrama: é construído com figuras geométricas (barras) separadas e bases de mesmo tamanho. A altura das barras é proporcional às frequências.

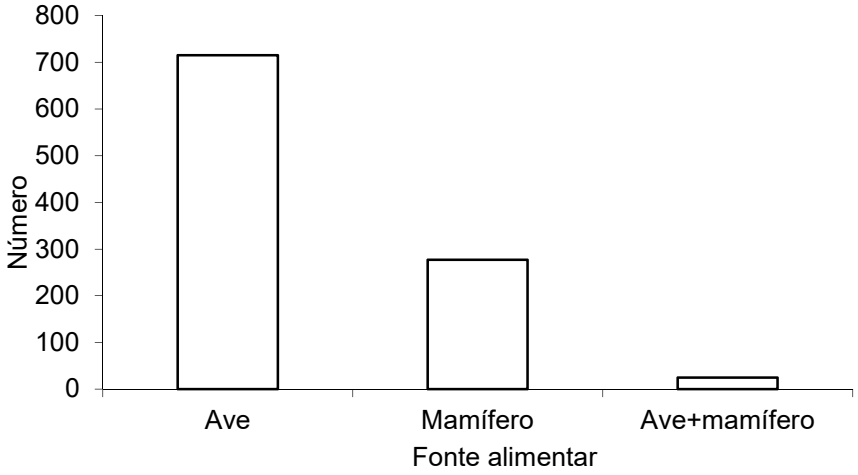
Diagrama de barras representando uma variável qualitativa nominal

Exemplo
Estudo que objetivou detectar o sangue ingerido por fêmeas de mosquitos Culicidae, principalmente das fêmeas da espécie *Culex pipiens*, em área suburbana de Chicago, Illinois de 2005 a 2007. Qual é a preferência alimentar dos mosquitos desta família?

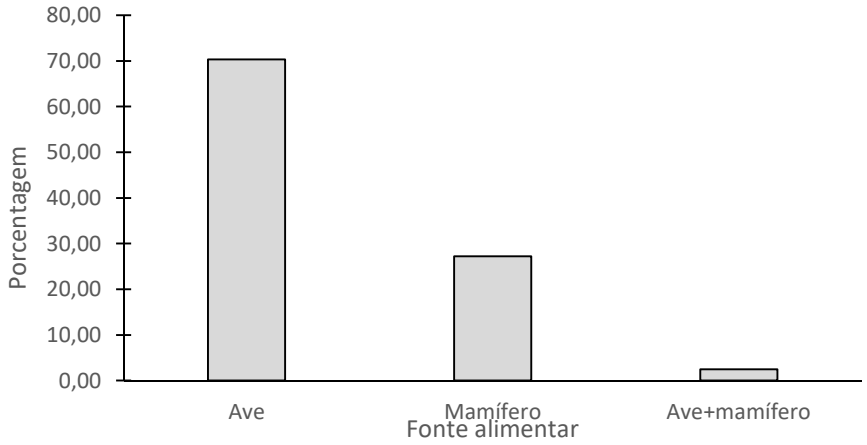
Número e percentual de fêmeas de mosquitos Culicidae, segundo fonte alimentar, coletados em área suburbana do sudoeste de Chicago em Illinois de 2005 a 2007.

Fonte alimentar	n	%
Sangue de ave	715	70,3
Sangue de mamífero	277	27,2
Sangue de ave+mamífero*	25	2,5
Total	995	100

*repasto misto
Fonte: adaptado de Hamer GL et al., 2009. Am. Mosq.Trop.Med.Hyg., 80(2), 2009, PP.268-278.



Ou

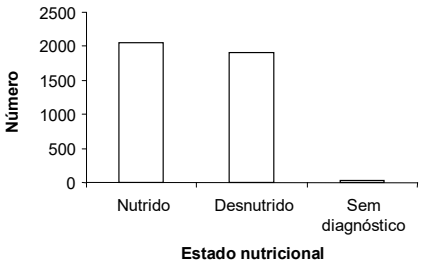


Fonte: adaptado de Hamer GL et al., 2009. Am. Mosq.Trop.Med.Hyg., 80(2), 2009, PP.268-278.
Número e percentual de fêmeas de mosquitos Culicidae, segundo fonte alimentar, coletados em área suburbana do sudoeste de Chicago em Illinois de 2005 a 2007.

Exemplo
O Inquérito Brasileiro de Nutrição (IBRANUTRI) foi um estudo de pacientes maiores de 18 anos, internados em hospitais da rede pública, conveniados, filantrópicos e universitários de 12 estados do Brasil e do Distrito Federal, realizado de maio a novembro de 1996 (in Soares JF, Siqueira AL. Introdução à Estatística Médica, COOPMED,Belo Horizonte, MG 2002). Os dados da tabela são retirados deste estudo.
Distribuição de pacientes segundo estado nutricional. IBRANUTRI, maio a novembro, 1996.

Estado nutricional	n	%
Nutrido	2061	51,5
Desnutrido	1905	47,6
Sem diagnóstico	34	0,9
Total	4000	100,0

Fonte: adaptado de Soares JF, Siqueira AL, 2002.



Fonte: adaptado de Soares JF, Siqueira AL, 2002.
Distribuição de pacientes segundo estado nutricional. IBRANUTRI, maio a novembro, 1996.

Esta representação gráfica está correta?

Atenção: cuidado com a origem!

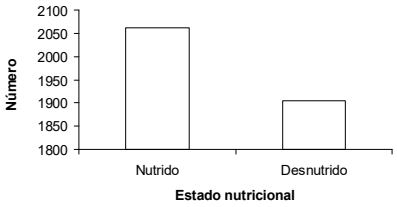


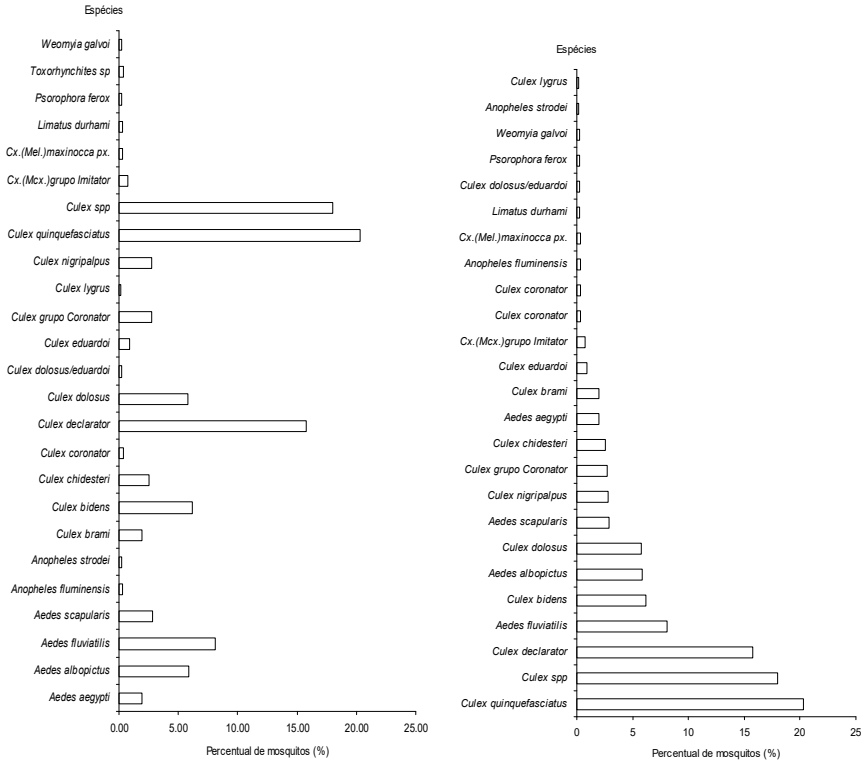
Diagrama de barras da tabela anterior, excluindo-se os registros da categoria sem diagnóstico

Fonte: adaptado de Soares JF, Siqueira AL, 2002.
Distribuição de pacientes segundo estado nutricional. IBRANUTRI, maio a novembro, 1996.

Exemplo
Distribuição do percentual de mosquitos Culicidae segundo espécie, coletados em 35 parques municipais de São Paulo, no período de outubro de 2010 a fevereiro de 2011.

Categoria Taxonômica	Percentual*	Categoria Taxonômica	Percentual*
Culex quinquefasciatus	20,28	*Culex eduardoi*	0,92
Culex spp	17,98	Cx.(Mcx.)grupo Imitator	0,74
Culex declarator	15,75	*Culex coronator*	0,37
Aedes fluviatilis	8,09	*Toxorhynchites spp*	0,37
Culex bidens	6,16	*Anopheles fluminensis*	0,31
Aedes albopictus	5,85	*Cx.(Mel.)maxinocca* px.	0,29
Culex dolosus	5,77	*Limatus durhami*	0,27
Aedes scapularis	2,85	*Psorophora ferox*	0,21
Culex nigripalpus	2,77	*Culex dolosus/eduardoi*	0,21
Culex grupo Coronator	2,73	*Weomyia galvoi*	0,21
Culex chidesteri	2,55	*Anopheles strodei*	0,19
Aedes aegypti	1,95	*Culex lygrus*	0,18
Culex brami	1,95		

* percentual em relação ao total de mosquitos imaturos e adultos coletados.
Fonte: Medeiros-Souza, AR et al., 2011 (adaptado). Biota Neotrop., vol. 13, no. 1, 317-321.



Distribuição do percentual de mosquitos imaturos e adultos (Diptera:Culicidae) coletados em 35 parques municipais da cidade de São Paulo.
Fonte: Medeiros-Souza, AR et al., 2011 (adaptado). Biota Neotrop., vol. 13, no. 1, 317-321.

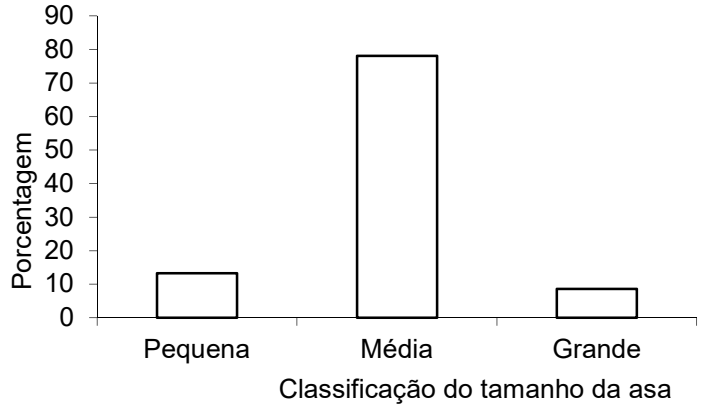
Diagrama de barras representando uma variável qualitativa ordinal

Exemplo

Distribuição do número de fêmeas de *Aedes triseriatus* segundo classificação do comprimento da asa em (mm), Iowa Co., Wisconsin, Madison em 1988.

Classificação do comprimento da asa (mm)	n	%
Pequena	102	13,3
Média	601	78,1
Grande	66	8,6
Total	769	100

Fonte: Adaptado de Landry SV et al., 1988
Journal of the American Mosquito Control Association, 1988, Vol4, nº 2, 121-128.



Fonte: Adaptado de Landry SV et al., 1988. Journal of the American Mosquito Control Association, 1988, Vol4, nº 2, 121-128

Distribuição do número de fêmeas de *Aedes triseriatus* segundo classificação do comprimento da asa em (mm), Iowa Co., Wisconsin, Madison em 1988.

Como você descreveria fêmeas desta espécie segundo o comprimento das asas?

Exemplo

Distribuição de pacientes segundo estado nutricional. IBRANUTRI, maio a novembro, 1996.

Estado nutricional[a]	n	%
Nutrido	2061	52,0
Desnutrido moderado	1407	35,4
Desnutrido grave	498	12,6
Total	3966	100

[a] excluindo-se 34 (0,9%) de pacientes sem diagnóstico.
Fonte: adaptado de Soares JF, Siqueira AL, 2002.

[a] excluindo-se 34 (0,9%) de pacientes sem diagnóstico.
Fonte: adaptado de Soares JF, Siqueira AL, 2002.

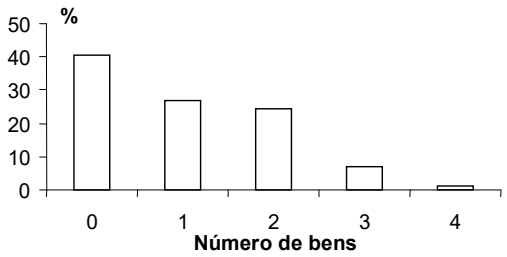
Distribuição de pacientes segundo estado nutricional. IBRANUTRI, maio a novembro, 1996.

Variável quantitativa discreta:
Foi realizada, no período de outubro de 1998 a outubro 1999, a pesquisa "Alimentação no primeiro ano de vida", onde se estudou uma coorte de recém-nascidos da maternidade do Hospital Universitário (HU). Os dados a seguir são parte da caracterização sócio-econômica da amostra estudada.

Distribuição de famílias segundo número de bens* que possuem. Hospital Universitário/USP, São Paulo 1999.

Número de bens	n	%
0	146	40,6
1	97	26,9
2	87	24,2
3	26	7,2
4	4	1,1
Total	360	100

* automóvel, telefone, TV a cabo e computador

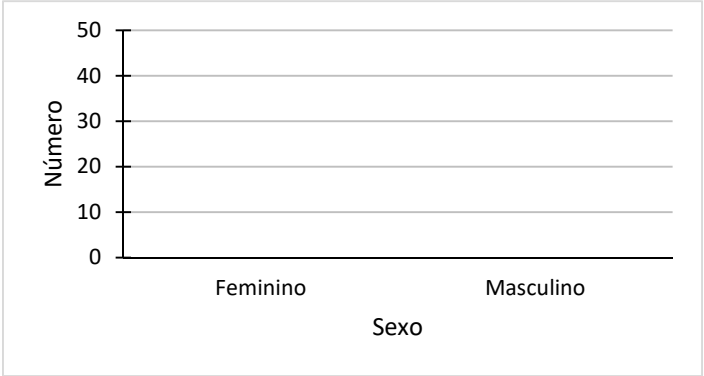


*automóvel, telefone, TV a cabo e computador
Distribuição de famílias segundo número de bens*. Hospital Universitário/USP, São Paulo 1999.

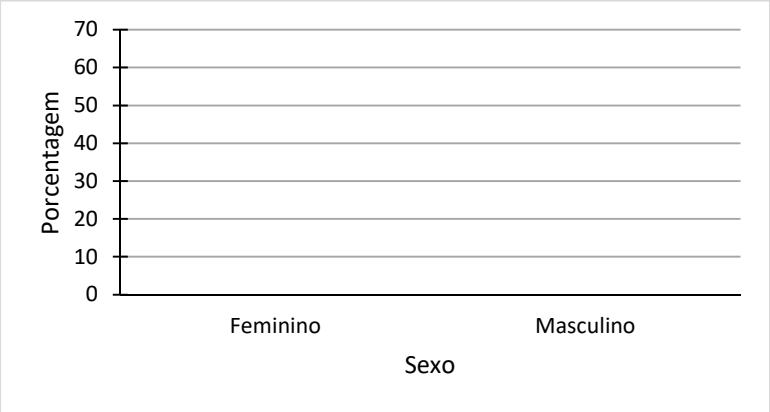
Exercício 7

Apresentar e descrever os dados dos idosos relativos as variáveis sexo e número de doenças crônicas em gráficos.

Variável sexo



Distribuição de segundo . Município de São Paulo, 2013



Variável número de doenças crônicas

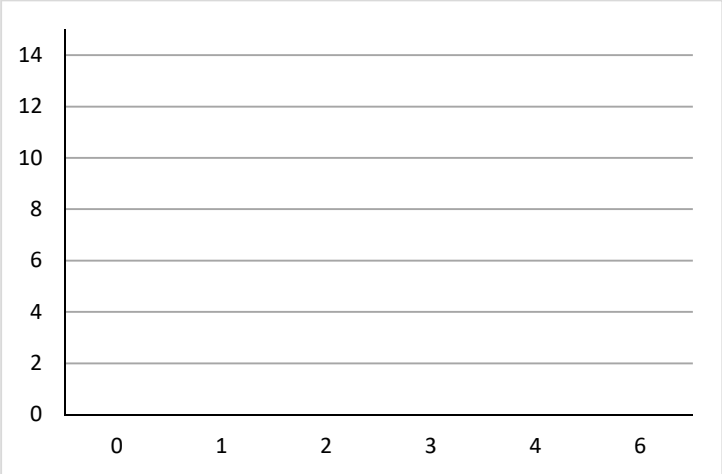


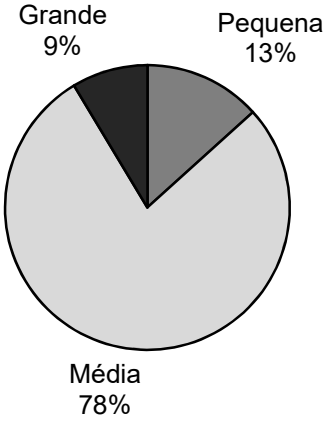
Diagrama de setores circulares

Variáveis: qualitativa nominal e qualitativa ordinal

Exemplo
Distribuição do número de fêmeas de *Aedes triseriatus* segundo classificação do comprimento da asa coletadas in Iowa Co., Wisconsin, Madison em 1988.

Classificação do comprimento da asa (mm)*	n	%
Pequena	102	13,3
Média	601	78,1
Grande	66	8,6
Total	769	100

* equivalente ao tamanho do corpo do mosquito.
Fonte: Adaptado de Landry SV et al., 1988. Journal of the American Mosquito Control Association, 1988, Vol4, nº 2, 121-128.



Fonte: Adaptado de Landry SV et al., 1988. Journal of the American Mosquito Control Association, 1988, Vol4, nº 2, 121-128.

Distribuição do número de fêmeas de *Aedes triseriatus* segundo classificação do comprimento da asa coletadas in Iowa Co., Wisconsin, Madison em 1988.

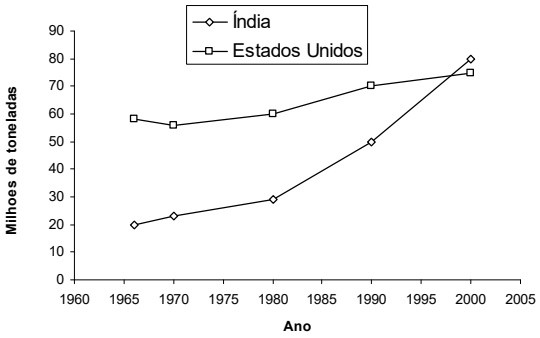
Diagrama linear

Representa variáveis qualitativas ordinais com natureza contínua subjacente às categorias. Por exemplo, a variável dia da semana. As categorias segunda-feira, terça-feira, etc são rótulos (nomes) dados para cada dia da semana caracterizando uma variável qualitativa ordinal e portanto, poderia ser representada por um diagrama de barras. Entretanto, por existir, de modo subjacente uma continuidade entre as categorias (quando termina a segunda-feira, imediatamente começa a terça-feira), esta variável constitui uma exceção na representação das qualitativas podendo-se unir os pontos resultando em uma linha de tendência.

Exemplo
Produção de leite (milhões de toneladas). Índia e Estados Unidos, 1966 – 2000.

Ano	Índia	Estados Unidos
1966	20	58
1970	23	56
1980	29	60
1990	50	70
2000	80	75

Fonte: *State of the World*, 2001.



Produção de leite (milhões de toneladas). Índia e Estados Unidos, 1966 – 2000

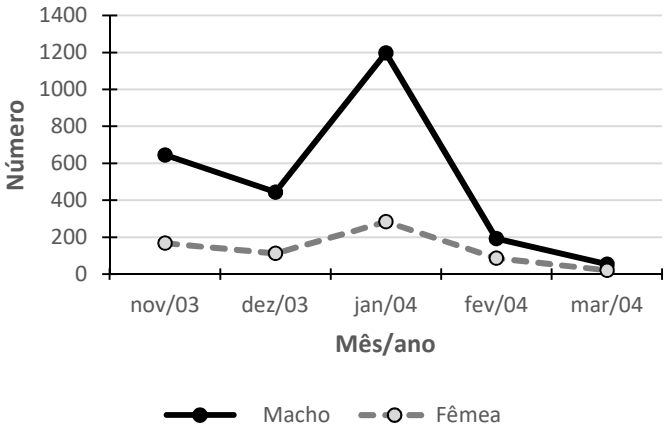
Fonte: *State of the World*, 2001.

Exemplo
Distribuição mensal do número de espécimes de *Culex.quinquefasciatus* segundo sexo. São Paulo, novembro de 2003 a março de 2004.

Mês/Ano	Macho	Fêmea
Nov/2003	644	168
Dez/2003	443	112
Jan/2004	1198	284
Fev/2004	192	87
Mar/2004	53	21

Fonte: Adaptado de Laporta et.al.2006. Revista Brasileira de Entomologia 50(1):125-127, março 2006.

Quais são as variáveis que estão sendo representadas?



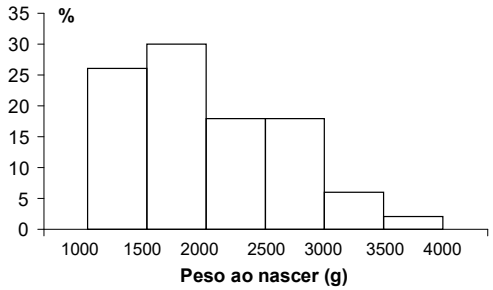
Distribuição mensal do número de espécimes de *Culex.quinquefasciatus* segundo sexo. São Paulo, novembro de 2003 a março de 2004.
Fonte: Adaptado de Laporta et.al. 2006.Revista Brasileira de Entomologia 50(1):125-127, março 2006.

Histograma
Adequado para representar variáveis do tipo quantitativa contínua

Intervalos de classe com mesma amplitude
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g), New South Wales, Australia, 1973*.

Peso(g)	Nº	%
1000 |-- 1500	13	26
1500 |-- 2000	15	30
2000 |-- 2500	9	18
2500 |-- 3000	9	18
3000 |-- 3500	3	6
3500 |-- 4000	1	2
Total	50	100

Fonte: van Vliet PKJ, Gupta JM. (1973).
* ano da publicação do artigo

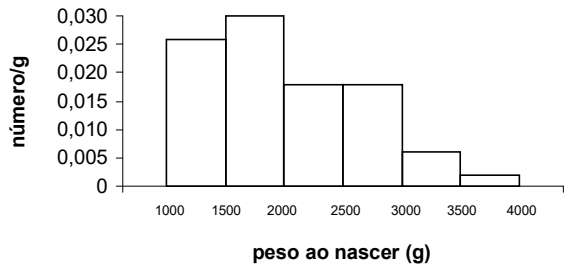


Fonte: van Vliet PKJ, Gupta JM. (1973)

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g), New South Wales, Australia, 1973.

Notar que o gráfico pode ser construído considerando-se pessoas por unidade de medida (densidade)

Peso(g)	Nº	Amplitude	Nº/amplitude	(Nº/amplitude)x10000
1000 |-- 1500	13	500	0,026	26
1500 |-- 2000	15	500	0,030	30
2000 |-- 2500	9	500	0,018	18
2500 |-- 3000	9	500	0,018	18
3000 |-- 3500	3	500	0,006	6
3500 |-- 4000	1	500	0,002	2
Total	50			



Fonte: van Vliet PKJ, Gupta JM. (1973).
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g), New South Wales, Australia, 1973.

OBS: notar que com intervalos iguais, não é necessário fazer ajuste na altura dos retângulos dado que as bases são de mesmo tamanho (mesma amplitude) e, portanto, com proporcionalidade assegurada.

Intervalos de classe com amplitudes diferentes

Distribuição de mulheres idosas segundo a altura, Bangladesh, 1992.

Altura (cm)	Nº	%
140|--150	12	3,4
150|--155	52	14,8
155|--160	109	31,1
160|--170	156	44,4
170|--180	22	6,3
Total	351	100

Fonte: Hand DJ et al., 1994.

Ajuste

Altura (cm)	Nº	Amplitude	Nº/amplitude
140|--150	12	10	1,2
150|--155	52	5	10,4
155|--160	109	5	21,8
160|--170	156	10	15,6
170|--180	22	10	2,2
Total	351		

Fonte: Hand DJ et al., 1994.

Distribuição de mulheres idosas segundo a altura, Bangladesh, 1992.

Cuidado: Sem fazer o ajuste, o gráfico fica errado e pode levar a conclusões incorretas.

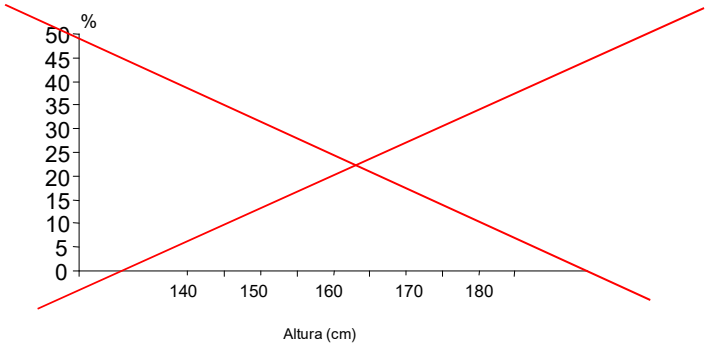
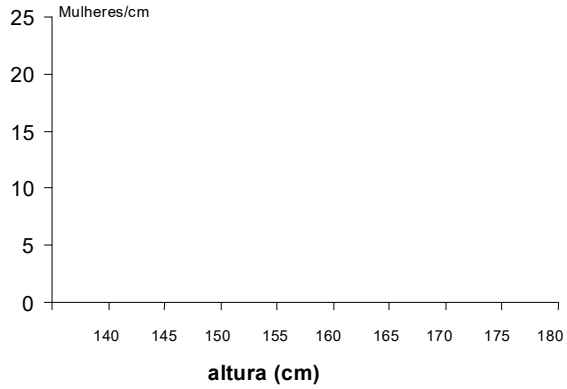


Gráfico correto, com o ajuste para intervalos de classe com amplitudes diferentes.



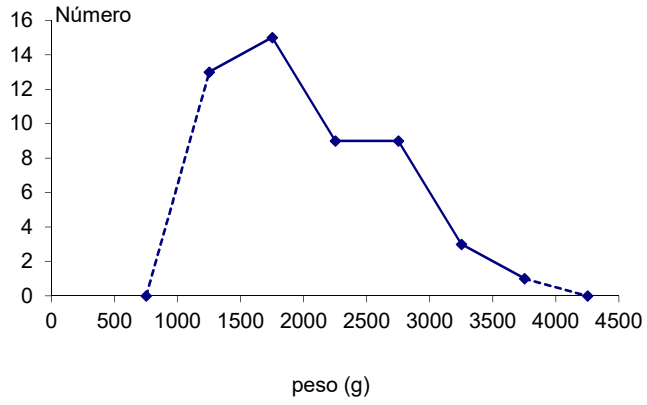
Exercício 8
Representar em um histograma com intervalos de amplitudes iguais a variável triglicérides utilizando os dados de idosos

Polígono de frequência simples

Intervalos de classe com mesma amplitude
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g), New South Wales, Australia, 1973.

Peso(g)	Nº	%
1000 |-- 1500	13	26
1500 |-- 2000	15	30
2000 |-- 2500	9	18
2500 |-- 3000	9	18
3000 |-- 3500	3	6
3500 |-- 4000	1	2
Total	50	100

Fonte: Hand DJ et al., 1994.



Fonte: Hand DJ et al., 1994.
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g), New South Wales, Australia, 1973.

Exercício 9

Representar em um polígono de frequências simples, com intervalos de amplitudes iguais, a variável triglicérides utilizando os dados de idosos

Intervalos de classe com amplitudes diferentes

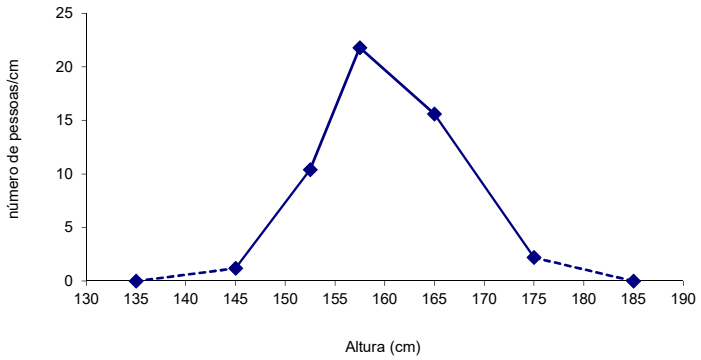
Distribuição de mulheres idosas segundo a altura, Bangladesh, 1992.

Altura (cm)	Nº	%
140|--150	12	3,4
150|--155	52	14,8
155|--160	109	31,1
160|--170	156	44,4
170|--180	22	6,3
Total	351	100

Fonte: Hand DJ et al., 1994.

Ajuste

Altura (cm)	Nº	Amplitude	Nº/amplitude
140|--150	12	10	1,2
150|--155	52	5	10,4
155|--160	109	5	21,8
160|--170	156	10	15,6
170|--180	22	10	2,2
Total	351		



Fonte: Hand DJ et al., 1994.

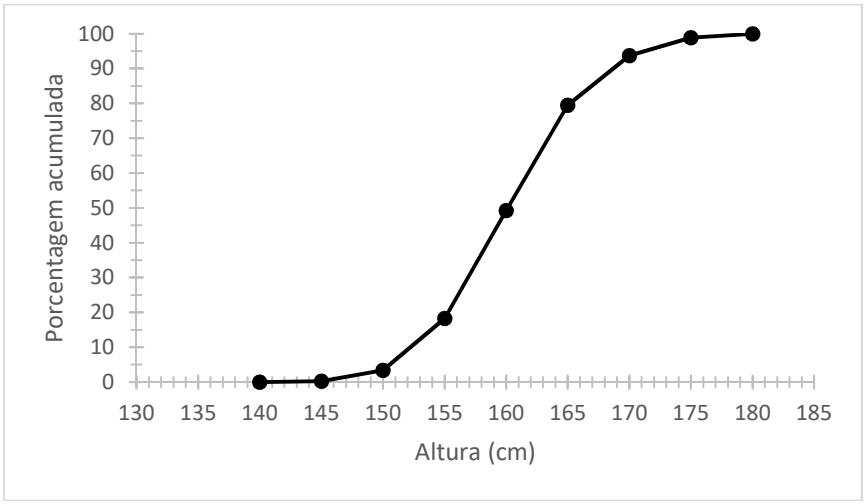
Distribuição de mulheres idosas segundo a altura (cm), Bangladesh, 1992.

Polígono (ogiva) de frequências acumuladas

Distribuição de mulheres idosas segundo a altura, Bangladesh, 1992.

Altura (cm)	Nº	%	% acumulado
140|-145	1	0,29	0,29
145|-150	11	3,13	3,42
150|-155	52	14,81	18,23
155|-160	109	31,05	49,28
160|-165	106	30,20	79,48
165|-170	50	14,25	93,73
170|-175	18	5,13	98,86
175|-180	4	1,14	100
Total	351	100	

Fonte: Hand DJ et al., 1994.



Fonte: Hand DJ et al., 1994.

Distribuição acumulada de mulheres idosas segundo a altura.

Percentil	Valor da variável	Medidas estatísticas
25%	156 cm	Q1 – primeiro quartil
50%	160 cm	Q2 - segundo quartil ou mediana
75%	164 cm	Q3 – terceiro quartil

Exercício 10
Representar em um gráfico de frequências acumuladas a variável triglicérides utilizando os dados de idosos

Representação gráfica de duas variáveis qualitativas

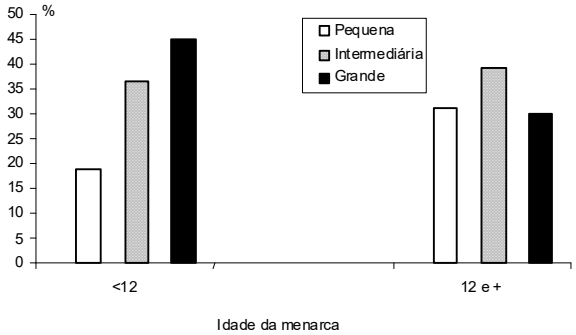
Os dados são de um estudo de obesidade em mulheres da zona urbana de Trinidade e Tobago, realizado em 1985, que estuda a relação entre idade da menarca e a medida do tríceps.

Distribuição de mulheres segundo idade da menarca e medida do tríceps. Trinidade e Tobago,1985.

Idade da menarca	Medida do tríceps		
	Pequena	Intermediária	Grande
< 12 anos	15	29	36
12 anos e mais	156	197	150

Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.

Idade (anos)	Medida do tríceps							
	Pequena		Intermediária		Grande		Total	
	N	%	n	%	n	%	n	%
<12	15	18,8	29	36,2	36	45,0	80	100
12 e +	156	31,0	197	39,2	150	29,8	503	100
Total	171	29,3	226	38,8	186	31,9	583	100



Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.

Distribuição de mulheres segundo idade da menarca e medida do tríceps. Trinidade e Tobago,1985.

Exemplo

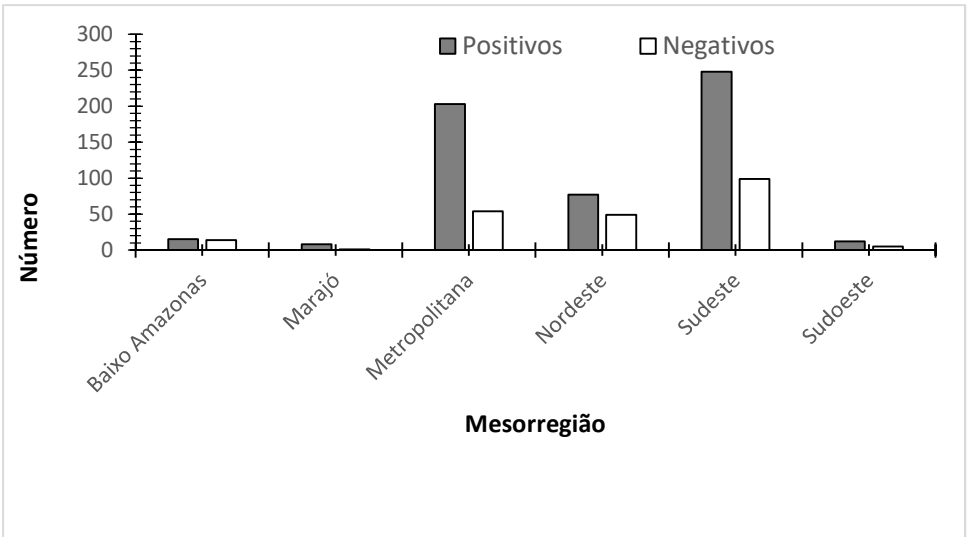
A tabela abaixo foi extraída de estudo que objetivou estudar pacientes com suspeita de dengue ou febre amarela de seis messorregiões do Estado do Pará.

Tabela 1 - Distribuição da positividade do teste de inibição da hemaglutinação para Flavivirus, por mesorregião, Pará, jun/dez 1999.

	Positivos		Negativos		Total
Mesorregião	nº	%	nº	%	nº
Baixo Amazonas	15	51,7	14	48,3	29
Marajó	8	88,9	1	11,1	9
Metropolitana	203	79,0	54	21,1	257
Nordeste	77	61,1	49	38,9	126
Sudeste	248	71,5	99	28,5	347
Sudoeste	12	70,6	5	29,4	17
Total	563	71,7	222	28,3	785

Fonte: Fichas epidemiológicas – Seção de Arbovírus/IEC, junho a dezembro de 1999

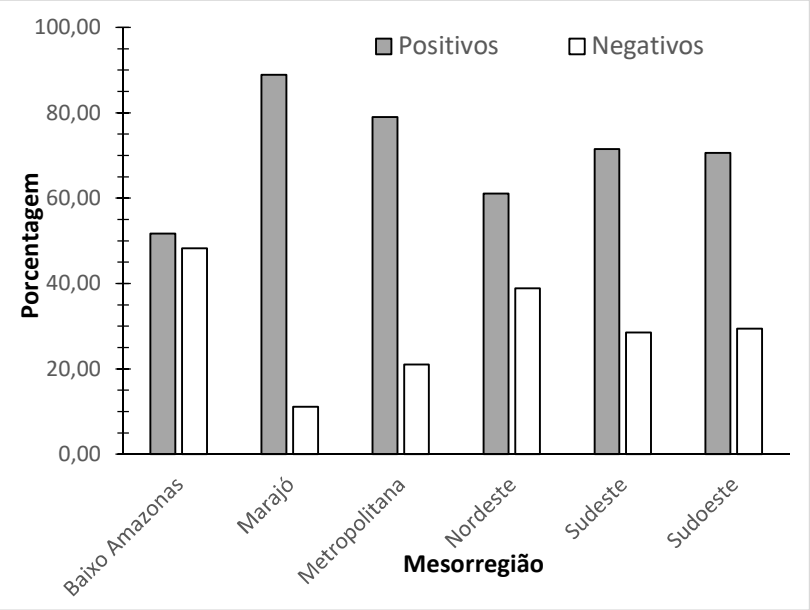
Fonte: Araújo TP et al., 2002. Revista Brasileira de Medicina Tropical 35(6):579-584, 2002.



Fonte: Araújo TP et al., 2002. Revista Brasileira de Medicina Tropical 35(6):579-584, 2002.

Distribuição da positividade de teste de inibição da hemaglutinação para Flavivirus, por mesorregião, Pará, jun/dez 1999

Calculando-se as porcentagens, tomando-se as categorias da variável mesorregião como 100%, tem-se:



Fonte: Araújo TP et al., 2002. Revista Brasileira de Medicina Tropical 35(6):579-584, 2002.

Distribuição da positividade de teste de inibição da hemaglutinação para Flavivirus, por mesorregião, Pará, jun/dez 1999

Representação gráfica de duas variáveis quantitativas

Exemplo

Os dados a seguir são relativos ao peso seco (mg) de fêmeas de *Culex quinquefasciatus* cujas larvas foram tratadas com mistura de ração de peixe, leite ninho e ração de cão, submetidas a temperaturas médias de 20ºC (*) e acima de 20ºC.

0,62*	0,77*	0,84*		0,48	0,61	0,64	0,72
0,64*	0,77*	0,94*		0,50	0,61	0,64	0,72
0,65*	0,79*			0,50	0,62	0,64	0,72
0,70*	0,79*			0,59	0,62	0,66	0,73
0,72*	0,80*			0,59	0,62	0,66	0,73
0,73*	0,80*			0,59	0,62	0,70	0,73
0,73*	0,81*			0,60	0,62	0,70	0,74
0,74*	0,83*			0,61	0,64	0,70	0,75

Fonte: Marchi MJ. Padronização de técnica para produção em massa de *Culex quinquefasciatus* (Diptera:Culicidae). São Paulo, 2014 [Dissertação de Mestrado, Faculdade de Saúde Pública da Universidade de São Paulo]. (Adaptado).

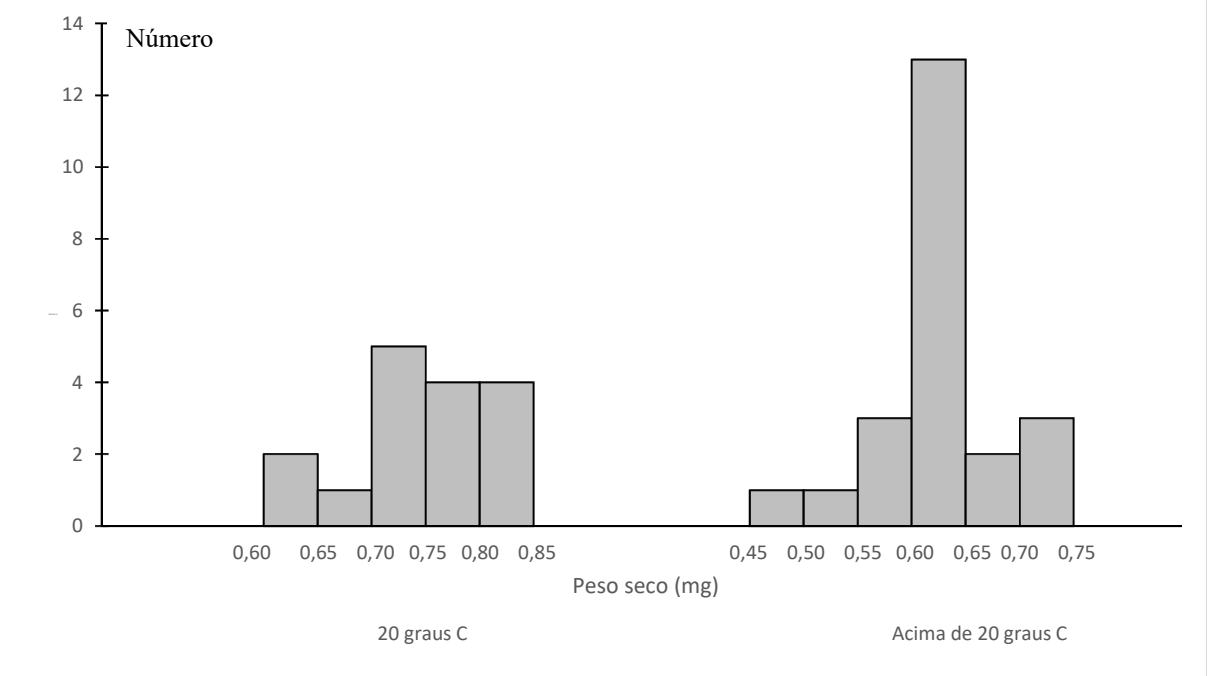
(*) = larvas submetidas a temperatura de 20ºC.

Distribuição de fêmeas de *Culex quinquefasciatus* segundo peso seco (mg) e temperatura

Peso seco (mg)	20 graus C		>de 20 graus C		Total	
	n	%	n	%	n	%
0,45 |-- 0,50	0	-	1	4,35	1	2,56
0,50 |-- 0,55	0	-	1	4,35	1	2,56
0,55 |-- 0,60	0	-	3	13,04	3	7,69
0,60 |-- 0,65	2	12,50	13	56,52	15	38,46
0,65 |-- 0,70	1	6,25	2	8,70	3	7,69
0,70 |-- 0,75	5	31,25	3	13,04	8	20,51
0,75 |-- 0,80	4	25,00	0	-	4	10,26
0,80 |-- 0,85	4	25,00	0	-	4	10,26
Total	16	100	23	100	39	100

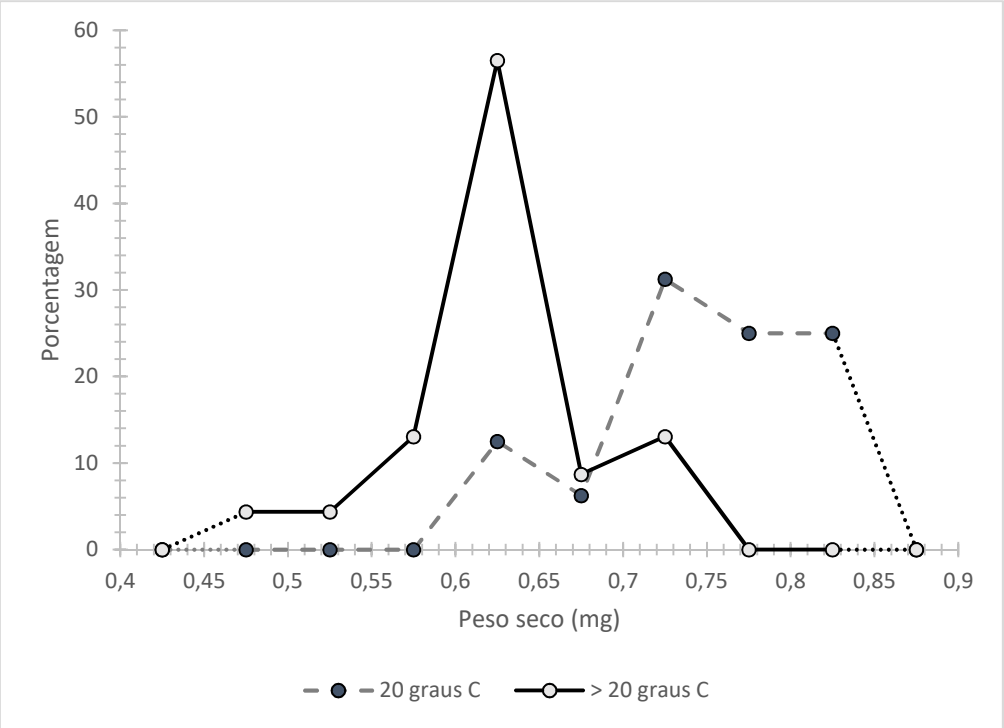
Fonte: Marchi MJ. Padronização de técnica para produção em massa de *Culex quinquefasciatus* (Diptera:Culicidae). São Paulo, 2014 [Dissertação de Mestrado, Faculdade de Saúde Pública da Universidade de São Paulo]. (Adaptado).

Distribuição de fêmeas de *Culex quinquefasciatus* segundo peso seco (mg) e temperatura



Fonte: Marchi MJ. Padronização de técnica para produção em massa de *Culex quinquefasciatus* (Diptera:Culicidae). São Paulo, 2014 [Dissertação de Mestrado, Faculdade de Saúde Pública da Universidade de São Paulo]. (Adaptado).

Distribuição de fêmeas de *Culex quinquefasciatus* segundo peso seco (mg) e temperatura



Fonte: Marchi MJ. Padronização de técnica para produção em massa de *Culex quinquefasciatus* (Diptera:Culicidae). São Paulo, 2014 [Dissertação de Mestrado, Faculdade de Saúde Pública da Universidade de São Paulo]. (Adaptado).

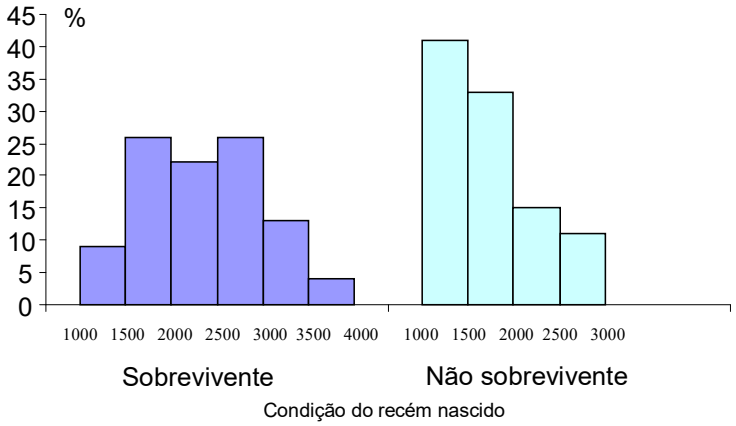
Distribuição de fêmeas de *Culex quinquefasciatus* segundo peso seco (mg) e temperatura

Exemplo
Fixando-se os percentuais na condição do recém-nascido:

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascido, New South Wales, Australia, 1973.

Peso(g)	Sobrevivente		Não sobrevivente		Total	
	nº	%	nº	%	nº	%
1000 |-- 1500	2	9	11	41	13	26
1500 |-- 2000	6	26	9	33	15	30
2000 |-- 2500	5	22	4	15	9	18
2500 |-- 3000	6	26	3	11	9	18
3000 |-- 3500	3	13	0	-	3	6
3500 |-- 4000	1	4	0	-	1	2
Total	23	100	27	100	50	100

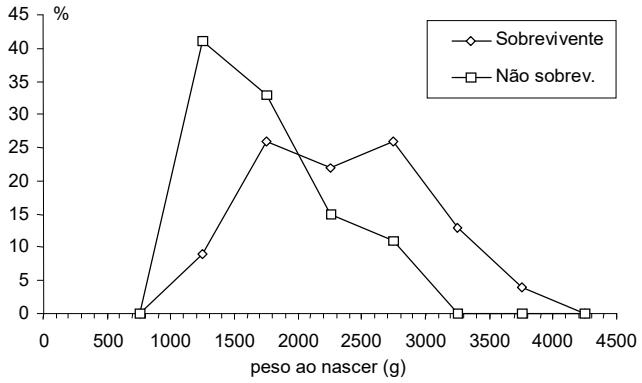
Fonte: Hand DJ et al., 1994.



Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascidos, New South Wales, Australia, 1973.

Polígono de frequências



Fonte: Hand DJ et al., 1994.

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascido, New South Wales, Australia, 1973.

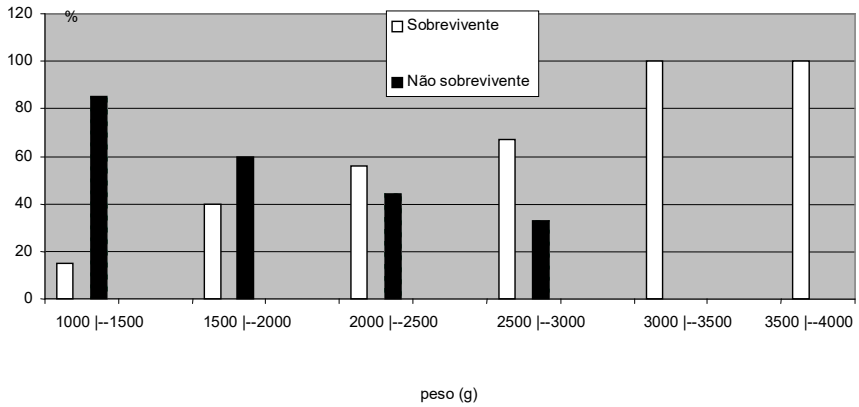
Fixando-se os percentuais no peso ao nascer:

Diagrama de barras

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascido, New South Wales, Australia, 1973.

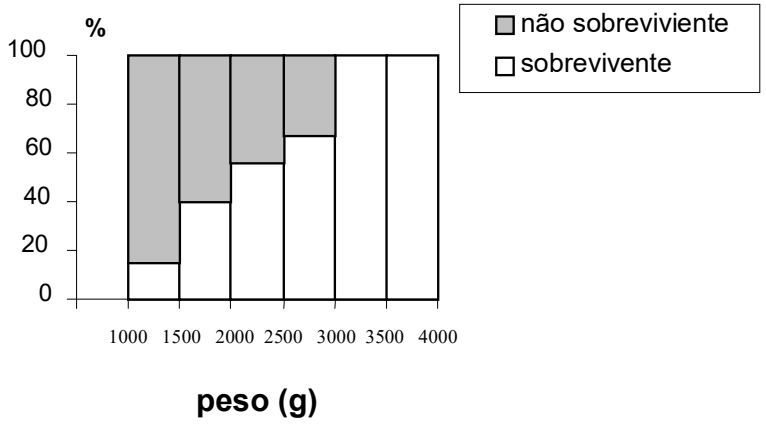
Peso(g)	Sobrevivente		Não sobrevivente		Total	
	nº	%	nº	%	nº	%
1000 |-- 1500	2	15	11	85	13	100
1500 |-- 2000	6	40	9	60	15	100
2000 |-- 2500	5	56	4	44	9	100
2500 |-- 3000	6	67	3	33	9	100
3000 |-- 3500	3	100	0	-	3	100
3500 |-- 4000	1	100	0	-	1	100
Total	23	46	27	54	50	100

Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.



Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascido, New South Wales, Australia, 1973.



Fonte: Hand DJ et al. *A handbook of small data sets*. Chapman&Hall, 1994.
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g) e condição do recém-nascido, New South Wales, Australia, 1973.

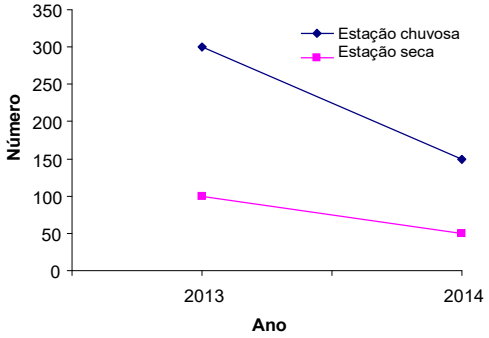
Escala aritmética e escala logarítmica

Número de mosquitos (Culicidae) segundo estação seca e chuvosa. , Habitat X. 2013 e 2014.

Ano	Estação chuvosa	Estação seca
2013	300	100
2014	150	50

Fonte: dados hipotéticos.

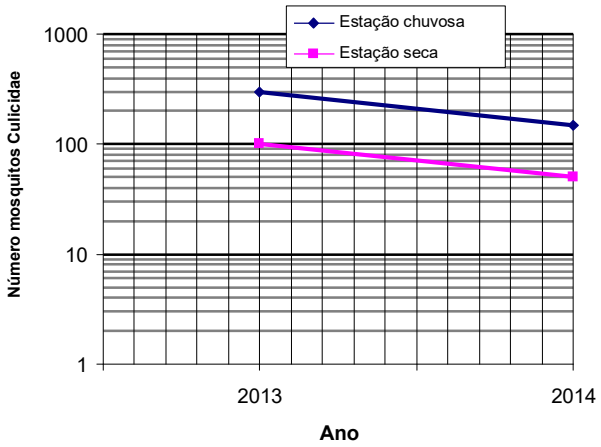
Gráfico em escala aritmética



Fonte: dados hipotéticos

Número de mosquitos Culicidae segundo estação do ano. Habitat X, 2013 e 2014.

Gráfico em escala logarítmica



Fonte: dados hipotéticos.

Número de mosquitos Culicidae segundo estação do ano. Habitat X, 2013 e 2014.

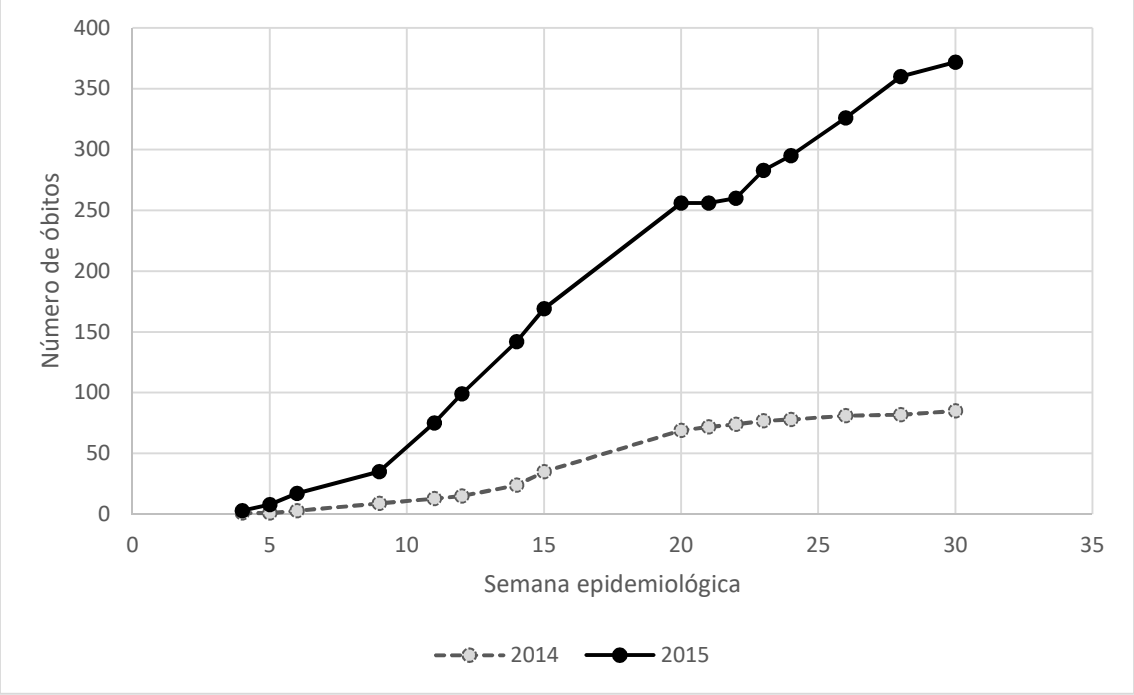
Exemplo

Mortalidade* por dengue segundo semana epidemiológica. Estado de São Paulo**, 2014 e 2015

Semana epidemiológica	Óbitos		Coeficiente*	
	2014	2015	2014	2015
4	1	3	0,002	0,007
5	1	8	0,002	0,018
6	3	17	0,007	0,038
9	9	35	0,020	0,079
11	13	75	0,029	0,169
12	15	99	0,034	0,223
14	24	142	0,054	0,320
15	35	169	0,079	0,381
20	69	256	0,155	0,577
21	72	256	0,162	0,577
22	74	260	0,167	0,586
23	77	283	0,173	0,637
24	78	295	0,176	0,664
26	81	326	0,182	0,734
28	82	360	0,185	0,811
30	85	372	0,191	0,838

* Por 100000 habitantes

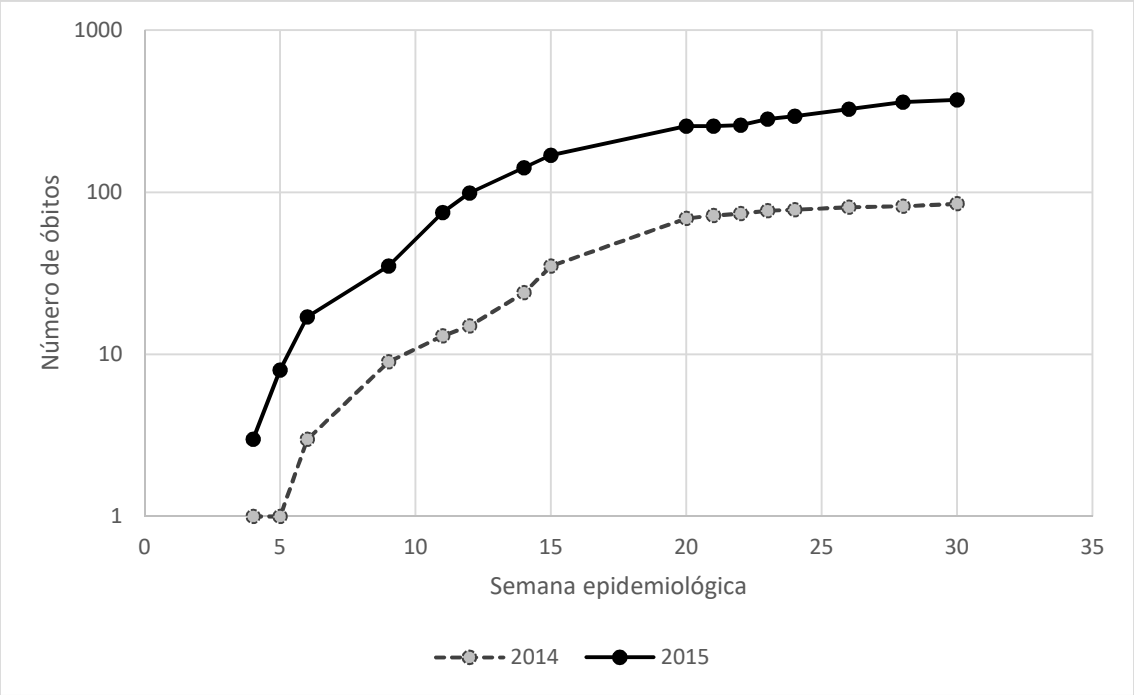
**População do estado de São Paulo estimada (IBGE) para 2015 = 44396484 hab.



Fonte: Boletim Epidemiológico. Secretaria de Vigilância em saúde. MS, 2015

Mortalidade* por dengue segundo semana epidemiológica. Estado de São Paulo, 2014 e 2015

Gráfico em escala logarítmica



Fonte: Boletim Epidemiológico. Secretaria de Vigilância em saúde. MS, 2015

Mortalidade* por dengue segundo semana epidemiológica. Estado de São Paulo, 2014 e 2015

Exemplo

Apresente os dados abaixo graficamente.

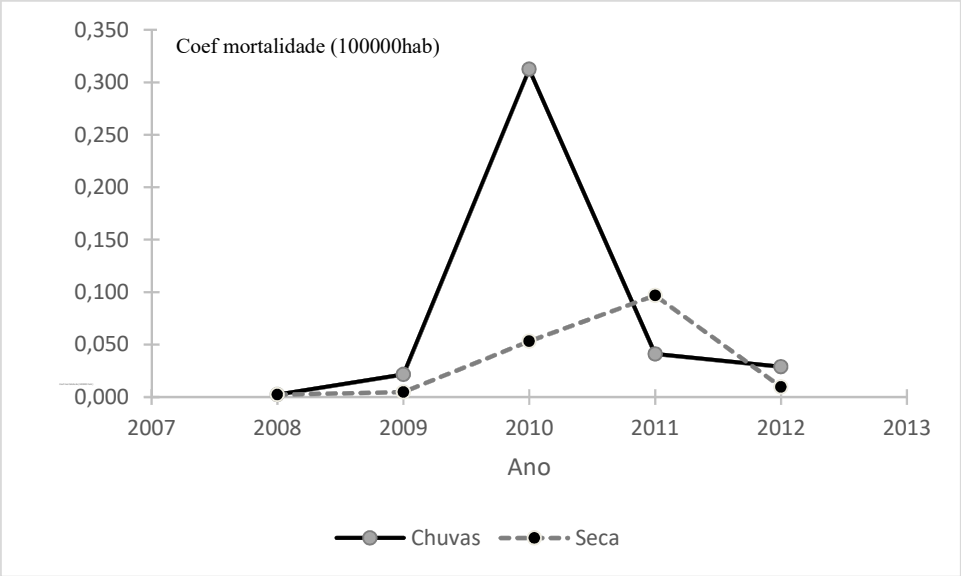
Mortalidade* por dengue segundo período de chuvas e de seca(1). Estado de São Paulo, 2014 e 2015

Ano	Óbitos		Coeficiente	
	Chuvas	Seca	Chuvas	Seca
2008	1	1	0,002	0,002
2009	9	2	0,022	0,005
2010	129	22	0,313	0,053
2011	17	40	0,041	0,097
2012	12	4	0,029	0,010

* Por 100000 habitantes

(1) chuvas: outubro a março; seca: abril a setembro

**População do estado de São Paulo estimada (IBGE) para 2010 = 41262199 hab

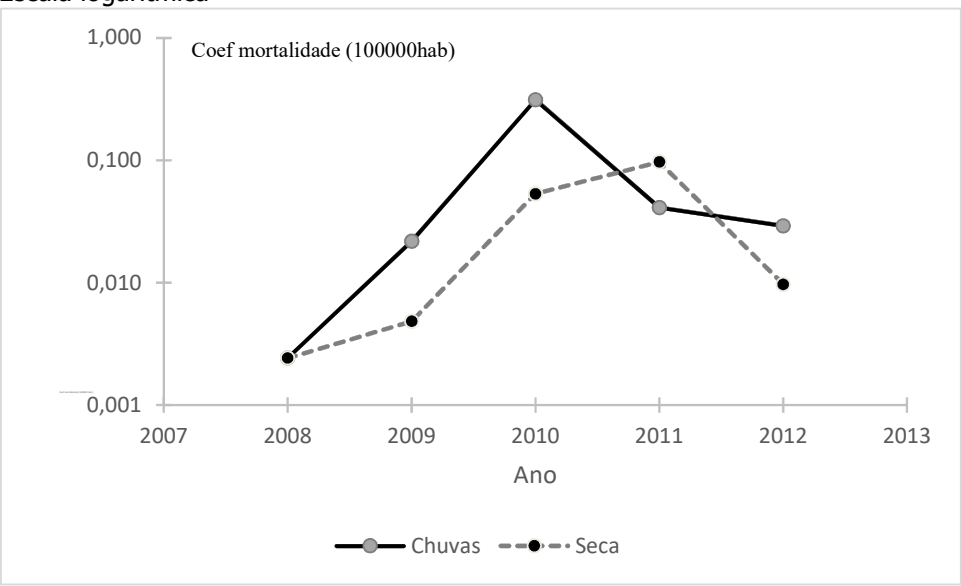


Fonte: Boletim Epidemiológico. Secretaria de Vigilância em saúde. MS, 2015

* Por 100000 habitantes; (1) chuvas: outubro a março; seca: abril a setembro

Mortalidade* por dengue segundo período de chuvas e de seca. Estado de São Paulo, 2014 e 2015

Escala logarítmica



Medidas de tendência central (média e mediana)

Medidas de tendência central

Média aritmética

Notação:

X → variável

N → tamanho da população

n → tamanho da amostra

μ →Média populacional (parâmetro, geralmente desconhecido)

$\overline{X}$ → Estatística (fórmula)

$\overline{x}$ → Média amostral (estimativa, valor calculado na amostra)

Média aritmética

Considerar

X: Número de ovos de Aedes aegypti

3	2	5	6	4

Para calcular a média soma-se os valores de uma variável e divide-se a soma pelo número de valores.

$$\text{Média aritmética} = \frac{3+2+5+6+4}{5} = 4 \text{ ovos}$$

2	3	4	5	6
		média		

2-4=	-2
3-4=	-1
4-4=	0
5-4=	1
6-4=	2
Soma=	0

Média aritmética é o valor que ocupa o centro de equilíbrio de uma distribuição de frequências de uma variável quantitativa. Portanto, a soma das diferenças entre cada valor e a média é igual a zero.

Apresentação em fórmula
Em uma amostra aleatória simples de tamanho n, composta pelas observações x_1, x_2, ..., x_n, a média aritmética ($\overline{x}$) é igual a:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

No exemplo, x_1=3; x_2=2, x_3=5,x_4=6,x_5=4; n=5. Portanto, $\bar{x} = \frac{3+2+5+6+4}{5} = \frac{20}{5} = 4$ ovos

OBS: a média aritmética

- só existe para variáveis quantitativas e seu valor é único;
- é da mesma natureza da variável considerada;
- sofre influência dos valores aberrantes (outlier).

Ex: x_1=3; x_2=2, x_3=5,x_4=6,x_5=20; n=5. Portanto, $\bar{x} = \frac{3+2+5+6+24}{5} = \frac{40}{5} = 8$ ovos

Exemplo:

Considerar os valores de número de doenças crônicas para idosos do sexo masculino e feminino

Masculino	3	0	1	3	2	1	3	0	2	1	0	6	0	0	1	2	
(n=16)																	
Feminino	1	4	4	0	2	1	2	3	2	1	3	1	2	3	3	2	3
(n=33)	1	3	3	1	3	2	3	1	3	1	0	2	2	1	2	4	

Calcular o número médio ($\bar{x}$) de doenças crônicas para

Homens n=16

	Masculino (X)
	3
	0
	1
	3
	2
	1
	3
	0
	2
	1
	0
	6
	0
	0
	1
	2
Total	

$\bar{x} = \frac{}{16} =$ doenças

Mulheres

	Feminino (X)
	1
	1
	4
	3
	4
	3
	0
	1
	2
	3
	1
	2
	2
	3
	3
	1
	2
	3
	1
	1
	3
	0
	1
	2
	2
	2
	3
	1
	3
	2
	2
	4
	3
Total	69

Exercício

Os dados a seguir são provenientes de um estudo que avaliou o tempo médio de vida em dias de 22 machos e 31 fêmeas de *Triatoma sordida*, nos estágios de ninfa e adulto, em condições de laboratório. Utilizou-se neste exemplo apenas os dados de tempo de vida em estágio de ninfa. [Fonte: Souza JMP de, 1978. *Triatoma sórdida* – Considerações sobre o tempo de vida das formas adultas e sobre a oviposição das fêmeas. Revista de Saúde Pública. São Paulo, 12:291-6.]

Calcule o número médio de dias no estágio de ninfa para machos e fêmeas:

Machos

136	157	154	129	247	164	133	126	247	139
139	148	221	248	131	139	135	143	249	173
241	241								

$\overline{x}_{Machos} =$

Fêmeas

126	126	127	130	129	128	131	126	132	136
146	128	150	136	158	134	126	128	128	139
203	208	242	241	250	244	259	241	253	234
250									

$\overline{x}_{Fêmeas} =$

Média geométrica

É a raiz n-ésima do produto de n observações $\overline{X}_G = \sqrt[n]{X_1 X_2 X_3 .... X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$

A média geométrica também pode ser calculada como o anti logaritmo da média aritmética dos logaritmos dos valores, onde o logaritmo pode estar em qualquer base.

$$\overline{X}_G = anti\log(\frac{\log X_1 + \log X_2 + ... + \log X_n}{n}) = \frac{\sum_{i=1}^{n} \log X_i}{n}$$

É apropriada somente para valores positivos. Se os valores forem todos iguais, a média aritmética e a geométrica serão idênticas, caso contrário, $\overline{X}_G < \overline{X}$. É útil para razões onde se deseja dar pesos iguais a razões e para o cálculo de mudanças relativas percentuais.

Lembrando sobre logaritmo

$\log_a b = x$ então $a^x = b$; em que $a > 0; a \neq 1; b > 0$

$\log_e a = \ln a$ é o logaritmo natural ou neperiano, em que *e* é um número irracional e vale aproximadamente 2,71.

mudança de base : $\log_e a = \frac{\log_{10} a}{\log_{10} e}$

mas $\log_{10} e \cong 0{,}43$, então $\log_e a = \frac{1}{0{,}43} \log_{10} a$; assim, $\log_e a = 2{,}3 \log_{10} a$

Exemplo

Valor (a)	$\log_{10} a$	$\ln_e a$
3	0,477	1,099
2	0,301	0,693
5	0,699	1,609
6	0,778	1,792
4	0,602	1,386
Soma	2,857	6,579

Antilogaritmo
Se $\log_a b = x$ então $anti\log_a x = b$
Se $anti\log_a x = b$, então $a^x = b$

Exemplo:
$anti\log_3 2 = b$
$3^2 = b$ portanto $b = 9$

$\log_3 9 = x$; $3^x = 9$; portanto $x = 2$

Exemplo:

X: Número de ovos de Aedes aegypti

3	2	5	6	4

$$\bar{x}_G = \sqrt[5]{3x2x5x6x4} = \sqrt[5]{\prod_{i=1}^{5} X_i} = \sqrt[5]{720} = 3,7 \text{ ovos}$$

ou

$$\bar{x}_G = anti\log(\frac{\log 3 + \log 2 + ... + \log 6}{5}) = \frac{\sum_{i=1}^{5} \log X_i}{5}$$

$$\bar{x}_G = anti\log(\frac{2,857}{5}) = anti\log(0,5714) = 3,73 \text{ ovos}$$

$$\bar{x}_G = anti\ln(\frac{6,579}{5}) = anti\ln(1,3158) = 3,73 \text{ ovos}$$

$\bar{x} = 4,0$ ovos

Exercício

Considere as observações; calcule e compare as medidas de resumo

3	2	5	6	47

Média aritmética
$\bar{x} = 12,6$ ovos

Média geométrica
$\bar{x}_G = 6,1$ ovos

Mediana

Considerar os valores de número de doenças crônicas para idosos do sexo masculino e feminino

Masculino	3	0	1	3	2	1	3	0	2	1	0	6	0	0	1	2

n=16

Ordenando-se os valores

	0	0	0	0	0	1	1	**1**	**1**	2	2	2	3	3	3	6
Posto	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Postos centrais

Mediana= $\frac{(1+1)}{2}=1$

Feminino	1 1	4 3	4 3	0 1	2 3	1 2	2 3	3 1	2 3	1 1	3 0	1 2	2 2	3 1	3 2	2 4	3

n=33

Ordenando-se os valores

Valor	Posto
0	1
0	2
1	3
1	4
1	5
1	6
1	7
1	8
1	9
1	10
1	11
2	12
2	13
2	14
2	15
2	16
2	17
2	18
2	19
2	20
3	21
3	22
3	23
3	24
3	25
3	26
3	27
3	28
3	29
3	30
4	31
4	32
4	33

Posto central $\frac{n+1}{2} = \frac{33+1}{2} = 17$

Portanto, mediana=2 doenças crônicas não infecciosas

<u>Mediana</u>

É o valor que ocupa a posição central de uma série de n observações, quando estas estão ordenadas de forma crescente ou decrescente.

Quando número de observações (n) for **ímpar**:

a mediana é o valor da variável que ocupa o posto $\frac{n+1}{2}$

Quando o número de observações (n) for **par**:

a mediana é a média aritmética dos valores da variável que ocupam os postos $\frac{n}{2}$ e $\frac{n+2}{2}$

OBS:

- existe para variável quantitativa e qualitativa ordinal;
- é da mesma natureza da variável considerada;
- torna-se inadequada quando há muitos valores repetidos;
- não sofre influência de valores aberrantes.

Exemplo:

Os dados a seguir são relativos à quantidade mensal de larvas de *Aedes albopictus* coletadas em dois ambientes do Parque Ecológico do Tietê, Guarulhos, SP, no período de abril de 2001 a março de 2003. Os dados foram extraídos de Urbinatti PR. "Observações ecológicas de *Aedes albopictus* (Diptera:Culicidae) em áreas de proteção ambiental e urbana da periferia na Grande São Paulo. São Paulo, 2004". [Tese de Doutorado, Faculdade de Saúde Pública da Universidade de São Paulo]. (Adaptado).

Ambiente A (n=20)

111	117	170	113	163	212	173	155	114	167
109	158	220	118	129	112	130	128	135	119

Ordenando-se os valores:

109	112	114	118	128	130	155	163	170	212
111	113	117	119	129	135	158	167	173	220

Número mediano de larvas de *Ae.albopictus* =

Valor mediano: (129+130)/2= 129,5 larvas

Ambiente B (n=20)

118	105	159	113	149	72	76	83	92	104
137	84	87	158	138	137	130	112	122	142

Ordenando-se os valores:

72	76	83	84	87	92	104	105	112	113
118	122	130	137	137	138	142	149	158	159

n=20

Número mediano de larvas de Ae.albopictus =

Exercício

Os dados a seguir são provenientes de um estudo que avaliou o tempo médio de vida em dias de 22 machos e 31 fêmeas de *Triatoma sordida*, nos estágios de ninfa e adulto, em condições de laboratório (Souza JMP de, 1978. *Triatoma sórdida* – Considerações sobre o tempo de vida das formas adultas e sobre a oviposição das fêmeas. Revista de Saúde Pública. São Paulo, 12:291-6).

Utilizou-se neste exemplo apenas os dados de tempo de vida em estágio de ninfa.

Calcule o número mediano de dias no estágio de ninfa para machos:

Machos

136	157	154	129	247	164	133	126	247	139
139	148	221	248	131	139	135	143	249	173
241	241								

Ordenando-se os valores

126	133	139	143	157	21	247	249
129	135	139	148	164	241	247	
131	136	139	154	173	241	248	

Valor mediano=

Número mediano de dias no estágio de ninfa para fêmeas:

Fêmeas

126	126	127	130	129	128	131	126	132	136
146	128	150	136	158	134	126	128	128	139
203	208	242	241	250	244	259	241	253	234
250									

Valor mediano=

Ordenando-se os valores

126	127	128	132	139	203	241	250
126	128	129	134	146	208	242	253
126	128	130	136	150	234	244	259
126	128	131	136	158	241	250	

n=31

Valor mediano=

Medidas de dispersão (variância, desvio-padrão, coeficiente de variação e percentis)

Constituem medidas de dispersão

- Valores mínimo e máximo
- Amplitude de variação
- Variância
- Desvio padrão
- Coeficiente de variação de Pearson

Valores mínimo e máximo: valores extremos da distribuição.

Ambiente A

109	112	114	118	128	130	155	163	170	212
111	113	117	119	129	135	158	167	173	220

Valor mínimo = 109 larvas; valor máximo = 220 larvas

Ambiente B

118	105	159	113	149	72	76	83	92	104
137	84	87	158	138	137	130	112	122	142

Ordenando-se os valores:

72	76	83	84	87	92	104	105	112	113
118	122	130	137	137	138	142	149	158	159

Valor mínimo = 72 larvas; valor máximo = 159 larvas

Amplitude de variação: é a diferença entre os 2 valores extremos da distribuição.

Ambiente A
Valor máximo - valor mínimo = 220 - 109= 111 larvas

Ambiente B
Valor máximo - mínimo = 159 - 72 = 87 larvas

Variância

É uma medida de dispersão que fornece a distância média ao quadrado das observações em relação á media. As distâncias de cada observação em relação à media são denominados desvios em relação à média. Se forem elevados ao quadrado, são denominados desvios quadráticos. Então a variância também pode ser entendida como a média dos dos desvios quadraticos de cada observação em relação à média aritmética.
Considerar os valores

3	2	5	6	4

$\bar{x} = 4$ ovos

Valor	(valor-média)	(valor–média)	(valor-média)^2
3	3-4=	-1 ovos	1 ovos^2
2	2-4=	-2 ovos	4 ovos^2
5	5-4=	1 ovos	1 ovos^2
6	6-4=	2 ovos	4 ovos^2
4	4-4=	0 ovos	0 ovos^2
Soma =		0 ovos	10 ovos^2

2	3	4	5	6

Variância = $\frac{10}{5} = 2$ ovos^2

Desvio padrão

É uma medida de dispersão calculada a partir da variância sedo a raiz quadrada desta. Indica o quanto em média "erramos em média" ao representarmos um conjunto de dados pela média. É portanto, o desvio médio dos valores em relação à media

Desvio padrão= $\sqrt{2} = 1{,}4$ ovos
O erro médio que se comete ao resumir os dados pela média é de 1,4 ovos.

Apresentando as fórmulas:

Na população a variância é representada pelo parâmeto σ^2 que pode ser estimado por dois estimadores:

Se os dados forem referentes á toda a população, o estimador é $S^2_{(N)} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$

É a soma dos desvios quadráticos dos valores em relação à média dividida por N, onde N é o número de observações

Se os dados forem referentes a uma amostra, o estimador é $S^2_{(N-1)} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$

É a soma dos desvios quadráticos dos valores em relação à média dividida por N-1, onde N é o número de observações

Desvio padrão

Na população, o desvio padrão é um parâmetro com notação σ sendo igual à a raiz quadrada da variância, ou seja $\sigma = \sqrt{\sigma^2}$.

O estimador do desvio padrão é representado por $S = \sqrt{S^2}$

Notação, resumo:

Estatística	População Parâmetro	Estimador	Estimativa (com dados da amostra)
Média	μ	$\overline{X} = \frac{\sum X_i}{N}$	$\overline{x} = \frac{\sum x_i}{n}$
Variância	σ^2	$S^2_{(N)} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$	$s^2_{(N)} = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N}$
		$S^2_{(N-1)} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}$	$s^2_{(n-1)} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$
Desvio padrão	σ	$S = \sqrt{S^2}$	$s = \sqrt{s^2}$

Coeficiente de Variação de Pearson (CV):

É uma medida de dispersão que relaciona a média e o desvio padrão. É representado em porcentagem. Será próximo de zero quando a dispersão for pequena, próxima a zero. Pode ser maior do que 100%. Isto ocorrerá quando a dispersão for maior que a média.

$\mathrm{CV} = \frac{\mathrm{S}}{\overline{\mathrm{X}}} x100$, onde S é o desvio padrão e $\overline{X}$, a média.

Exercício

Calcule as medidas de dispersão da variável "número de doenças crônicas" para cada um dos sexos

Masculino	3	0	1	3	2	1	3	0	2	1	0	6	0	0	1	2	

Masculino (X)	$(x-\bar{x})$	$(x-\bar{x})^2$
3		
0		
1		
3		
2		
1		
3		
0		
2		
1		
0		
6		
0		
0		
1		
2		

Valor mínimo

Valor máximo

Variância (n)

Variância (n-1)

Desvio padrão (n)

Desvio padrão (n-1)

Coeficiente de variação de Pearson

Sexo

Feminino	1	4	4	0	2	1	2	3	2	1	3	1	2	3	3	2	3
	1	3	3	1	3	2	3	1	3	1	0	2	2	1	2	4	

Feminino (X)	$(x-\bar{x})$	$(x-\bar{x})^2$
1		
4		
4		
0		
2		
1		
2		
3		
2		
1		
3		
1		
2		
3		
3		
2		
3		
1		
3		
3		
1		
3		
2		
3		
1		
3		
1		
0		
2		
2		
1		
2		
4		

Valor mínimo

Valor máximo

Variância (n)

Variância (n-1)

Desvio padrão (n)

Desvio padrão (n-1)

Coeficiente de variação de Pearson

Apresentação das medidas-resumo

A tabela abaixo foi extraída do artigo: Diagnóstico de sobrepeso em adolescentes: estudo do desempenho de diferentes critérios para o Índice de Massa Corporal de MONTEIRO POA *et al*. (*Rev. Saúde Pública* , 2000;.34(5):506 - 13).
Discuta os resultados obtidos ignorando a coluna do valor de p.

Tabela 1 – Estatística descritiva da população em estudo, por sexo (n=493). Pelotas, RS, Brasil. 1998.

Variável	Meninos (n=242) Média	DP	Meninas (n=251) Média	DP	p - valor
Idade (anos)	16,1	0,2	16,1	0,2	0,6
Peso (kg)	65,2	12,3	57,5	10,5	<0,001
Altura (cm)	170,6	6,6	159,8	6,2	<0,001
IMC (kg/m)	22,1	3,7	22,1	3,5	0,8
Dobra subescapular (mm)*	19,9	7,5	23,7	6,3	<0,001
Dobra tricipital (mm)*	19,6	6,3	26,3	5,4	<0,001

*As dobras cutâneas foram medidas apenas nos 92 meninos e 96 meninas cujo IMC foi igual ou superior ao percentil 85 para idade e sexo conforme Nhanes I (OMS).[8]

Tabela 3 – Medidas de tendência central, de dispersão e intervalos de confiança do consumo alimentar dos escolares estimados pelos DA. Escola de Aplicação da USP, São Paulo, 2009.

Estatística	Energia (Kcal)	Carboidrato (g)	Proteína (g)	Lipídios (g)
Média	1730,7	238,6	64,1	59,1
Mediana	1702,0	236,8	61,1	56,4
Desvio padrão	493,2	71,0	21,1	20,8
Valor mínimo	480,0	97,8	12,7	4,6
Valor máximo	3711,3	465,8	157,2	139,4
Q1; Q3	1408,6; 1947,3	179,0; 271,9	51,9; 72,5	46,7; 67,4
IC 95%	(1624,3 -1837,1)	(223,3 - 253,9)	(59,6 – 68,7)	(54,6 – 63,6)

(n=85)

HINNIG PF. Construção de um Questionário de Frequência Alimentar Quantitativo para crianças de 7 a 10 anos [dissertação de mestrado]. São Paulo: Faculdade de Saúde Pública da USP; 2010.
4

Quartil

Valores da variável que dividem a distribuição em quatro partes iguais.

	¼	½	¾	
25%	25%	25%	25%	

Q1: deixa abaixo 25% das observações

25%	75%

Q2: deixa abaixo 50% das observações

50%	50%

Q3: deixa abaixo 75% das observações

75%	25%

$$Q_1 = x_{(\frac{1}{4}(n+1))} \quad \text{e} \quad Q_3 = x_{(\frac{3}{4}(n+1))}$$

onde x é o valor da variável e $(\frac{1}{4}(n+1))$ e $(\frac{3}{4}(n+1))$ são índices que representam as posições ocupadas por x.

Os dados abaixo são referentes ao peso ao nascer de 50 recém-nascidos que tiveram síndrome de desconforto respiratório idiopático grave. 23 crianças sobreviveram e 27 foram a óbito (*).

1.050*	2.500*	1.890*	1.760	2.830
1.175*	1.030*	1.940*	1.930	1.410
1.230*	1.100*	2.200*	2.015	1.715
1.310*	1.185*	2.270*	2.090	1.720
1.500*	1.225*	2.440*	2.600	2.040
1.600*	1.262*	2.560*	2.700	2.200
1.720*	1.295*	2.730*	2.950	2.400
1.750*	1.300*	1.130	2.550	3.160
1.770*	1.550*	1.575	2.570	3.400
2.275*	1.820*	1.680	3.005	3.640

Ordenando-se os dados, em cada grupo, obtém-se:

1.030*	1.310*	2.200*	1.680	2.550
1.050*	1.500*	2.270*	1.715	2.570
1.100*	1.550*	2.275*	1.720	2.600
1.175*	1.600*	2.440*	1.760	2.700
1.185*	1.720*	2.500*	1.930	2.830
1.225*	1.750*	2.560*	2.015	2.950
1.230*	1.770*	2.730*	2.040	3.005
1.262*	1.820*	1.130	2.090	3.160
1.295*	1.890*	1.410	2.200	3.400
1.300*	1.940*	1.575	2.400	3.640

Fonte: van Vliet PK; Gupta JM. Sodium bicarbonate in idiopatic respiratory distress syndrome. *Arch. Diseases in Childhood*,1973:48, 249-255.

Entre os recém-nascidos que sobreviveram:

$$Q_1 = x_{(\frac{1}{4}(23+1))} = x_6 = 1720g \; ; \quad Q_3 = x_{(\frac{3}{4}(23+1))} = x_{18} = 2830g$$

$$Q_2 = x_{(\frac{1}{2}(23+1))} = x_{12} = 2200g$$

Entre os recém-nascidos que foram a óbito

$$Q_1 = x_{(\frac{1}{4}(27+1))} = x_7 = 1230g \; ; \quad Q_3 = x_{(\frac{3}{4}(27+1))} = x_{21} = 2200g$$

$$Q_2 = x_{(\frac{1}{2}(27+1))} = x_{14} = 1600g$$

Se o resultado for um valor fracionário:

Por exemplo, para n=22

$$Q_1 = x_{(\frac{1}{4}(22+1))} = x_{(\frac{23}{4})} = x_{(5\frac{3}{4})}$$

que é ¾ do caminho entre x_5=1715 e x_6=1720

$$Q_1 = 1715 + \frac{3}{4}(1720 - 1715) = 1718{,}8g$$

$$Q_3 = x_{(\frac{3}{4}(22+1))} = x_{(17\frac{1}{4})}$$

que é ¼ do caminho entre x_{17}=2700 e x_{18}=2830

$$Q_3 = 2700 + \frac{1}{4}(2830 - 2700) = 2732{,}5g$$

Decil
Valores da variável que dividem a distribuição em dez partes iguais.

Percentil
Valores da variável que dividem a distribuição em cem partes iguais.

<u>Entre os recém-nascidos que sobreviveram</u>

Percentil 5:

$$P_5 = x_{(\frac{5}{100}(23+1))} = x_{(\frac{120}{100})} = x_{(1\frac{1}{5})}$$

que é 1/5 do caminho entre x_1=1130 e x_2=1410

$$P_5 = 1130 + \frac{1}{5}(1410 - 1130) = 1186g$$

Percentil 10:

$$P_{10} = x_{(\frac{10}{100}(23+1))} = x_{(\frac{240}{100})} = x_{(2\frac{2}{5})};\ P_{10} = 1410 + \frac{2}{5}(1575 - 1410) = 1476g$$

Percentil 50:

$$P_{50} = x_{(\frac{50}{100}(23+1))} = x_{(\frac{1200}{100})} = x_{(12)};\ P_{50} = 2200g$$

Percentil 75:

$$P_{75} = x_{(\frac{75}{100}(23+1))} = x_{(\frac{1800}{100})} = x_{(18)};\ P_{75} = 2830g$$

Percentil 90:

$$P_{90} = x_{(\frac{90}{100}(23+1))} = x_{(\frac{2160}{100})} = x_{(21\frac{3}{5})};\ P_{90} = 3160 + \frac{3}{5}(3400 - 3160) = 3304g$$

Percentil 95:

$$P_{95} = x_{(\frac{95}{100}(23+1))} = x_{(\frac{2280}{100})} = x_{(22\frac{4}{5})};\ P_{95} = 3400 + \frac{4}{5}(3640 - 3400) = 3592g$$

Box plot e identificação de valores aberrantes (*outliers*)

O Box plot representa graficamente dados de forma resumida em um retângulo onde as linhas da base e do topo são o primeiro e o terceiro quartis, respectivamente. A linha entre estas é a mediana. Linhas verticais que iniciam no meio da base e do topo do retângulo, terminam em valores denominados adjacentes inferior e superior (Chambers *et al.*, 1983, pag 60).

O valor adjacente superior é o maior valor das observações que é menor ou igual a Q3+1,5(Q3-Q1).

O valor adjacente inferior é definido como o menor valor que é maior ou igual a Q1-1,5(Q3-Q1), sendo a diferença Q3-Q1 denominada intervalo inter-quartil (IIQ).

Valores *outliers* (discrepantes ou aberrantes) são valores que "fogem" da distribuição dos dados. O box plot além de apresentar a dispersão dos dados torna-se útil também para identificar a ocorrência destes valores como sendo os que caem fora dos limites estabelecidos pelos valores adjacentes superior e inferior.

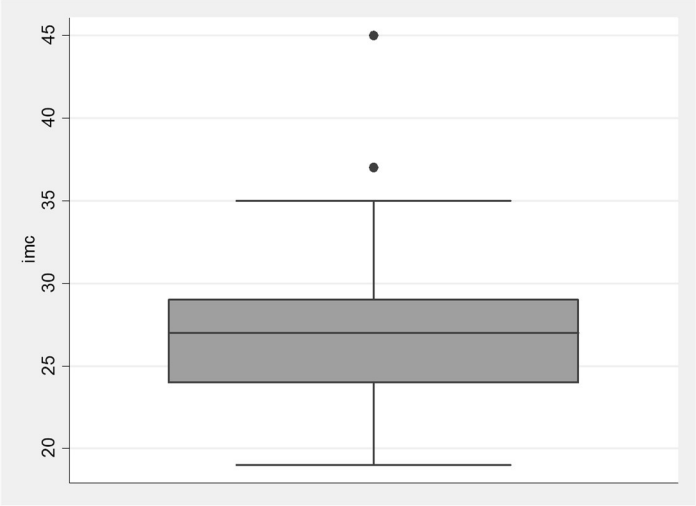
O box plot permite também investigar a dispersão e simetria dos dados.

Comentários sobre o gráfico:

Utilizando-se os dados de imc tem-se quartil 1 = 24; quartil 2 = 27 e quartil 3 = 29
Intervalo Inter quartil = 29-24= 5

VAI: Menor valor dos dados que é maior ou igual a 16,5 (24-(1,5x5)). **VAI =16,5**
VAS: Maior valor dos dados que é menor ou igual a 36,5 (24-(1,5x5)). **VAS = 35**

Não existem valores abaixo do VAI, mas existem valores acima do VAS indicando existência de dois outliers.



id	imc	id	imc
1	26	26	24
2	31	27	34
3	24	28	25
4	22	29	20
5	27	30	27
6	27	31	45
7	26	32	35
8	27	33	24
9	28	34	22
10	26	35	31
11	24	36	27
12	27	37	23
13	23	38	20
14	29	39	29
15	24	40	29
16	35	41	27
17	29	42	30
18	37	43	34
19	19	44	25
20	23	45	34
21	19	46	29
22	28	47	27
23	28	48	23
24	26	49	29
25	28	50	27

Ordenando-se os dados

id	imc	
19	19	
21	19	
29	20	
38	20	
4	22	
34	22	
13	23	
20	23	
37	23	
48	23	
3	24	
11	24	
15	24	
26	24	
33	24	
28	25	
44	25	
1	26	
7	26	
10	26	
24	26	
5	27	
6	27	

8	27	
12	27	
30	27	
36	27	
41	27	
47	27	
50	27	
9	28	
22	28	
23	28	
25	28	
14	29	
17	29	
39	29	
40	29	
46	29	
49	29	
42	30	
2	31	
35	31	
27	34	
43	34	
45	34	
16	35	
32	35	
18	37	
31	45	

Calcular Q1, Q2, IIQ e VAI e VAS e construir o box plot.

Exercício

Fazer o gráfico box plot para triglicérides. Existem valores outlier?

id	triglic	id	triglic
1	128	26	89
2	166	27	92
3	79	28	181
4	166	29	91
5	61	30	171
6		31	176
7	211	32	165
8	157	33	38
9	124	34	46
10	111	35	
11	80	36	153
12	73	37	
13	205	38	99
14	101	39	66
15		40	130
16	170	41	72
17	126	42	87
18	193	43	219
19	92	44	
20	47	45	125
21	221	46	233
22	86	47	118
23	119	48	56
24	75	49	80
25	145	50	104

Valores ordenados

id	triglic	
33	38	
34	46	
20	47	
48	56	
5	61	
39	66	
41	72	
12	73	
24	75	
3	79	
11	80	
49	80	
22	86	
42	87	
26	89	
29	91	
19	92	
27	92	
38	99	
14	101	

50	104	
10	111	
47	118	
23	119	
9	124	
45	125	
17	126	
1	128	
40	130	
25	145	
36	153	
8	157	
32	165	
2	166	
4	166	
16	170	
30	171	
31	176	
28	181	
18	193	
13	205	
7	211	
43	219	
21	221	
46	233	
6		
15		
35		
37		
44		

Calcular Q1, Q2, IIQ e VAI e VAS e construir o box plot.

Exercício

Os dados a seguir são adaptados de artigo publicado por Honório NA & Lourenço-de-Oliveira R. 2001, cujo estudo avaliou a frequência mensal de larvas e pupas de *Aedes aegytpi* e *Aedes albopictus* coletadas em pneus, no período de novembro de 1997 a outubro de 1998, em Nova Iguaçu, Rio de Janeiro.

Número de larvas – *Ae.albopictus*

123	243	215	142	153	118	164	194	160	120
128	122	151	155	137	129	216	157	145	182

Número de larvas – Ae.aegypti

90	104	140	72	67	78	60	61	101	81
117	89	111	83	98	101	74	88	132	66

a) Calcule a o número médio de larvas em cada grupo utilizando a média aritmética
b) Calcule a o número médio de larvas em cada grupo utilizando a média geométrica
c) Calcule o número mediano de larvas em cada grupo.
d) Desenhe o *box plot* do número de larvas representando os dois grupos em um só gráfico.
e) Comente o gráfico *box plot* quanto a dispersão dos dados, existência de valores aberrantes e simetria dos dados.

Atenção: é necessário ordenar os valores para fazer os itens c e d.

Exercício 11

Os dados a seguir são adaptados de estudo, publicado por Devicari et al. 2013, que avaliou o tamanho das asas em (mm) de *Aedes scapularis* para machos e fêmeas da espécie, capturados no Município de São Paulo:

Machos

1,78	1,91	2,02	2,11	2,13	2,21	2,30	2,41	1,87	2,01
2,03	2,11	2,15	2,21	2,31	3,50	1,90	2,01	2,10	2,11
2,15	2,21	2,32							

Fêmeas

1,01	1,62	2,30	2,40	2,56	2,61	2,71	2,80	1,52	1,89
2,31	2,45	2,60	2,65	2,75	2,89	1,58	1,97	2,34	2,45
2,60	2,70	2,78							

a) Calcule o tamanho médio da asa (mm) para cada sexo. Utilize a média aritmética;
b) Calcule o tamanho mediano da asa (mm) para cada sexo;
c) Calcule a variância, o desvio-padrão e o coeficiente de variação de Pearson do tamanho da asa em (mm) para cada sexo.
d) Machos e fêmeas são parecidos quanto ao tamanho da asa (mm)?
e) E quanto à variabilidade?
f) Apresente o box plot do tamanho da asa (mm) para os sexos e interprete o gráfico.

Exercício

A tabela abaixo foi extraída do artigo: Influência da Altitude, Latitude e Estação de Coleta (Regra de Bergmann) na dimensão de *Lutzomyia intermedia* (Lutz & Neiva, 1912) (Diptera:Psychodidae, Phlebotominae). Marcondes CB et al. (Memórias do Instituto Oswaldo Cruz, 1999;. vol94(5):693-700).

Discuta os resultados obtidos.

Dimensions (in μm) of females and some of their respective ratios for *Lutzomyia intermedia* from Viana, a low altitude and lower latitude locality in the State of Espírito Santo (ES), and from low altitude and higher latitude localities in the states of Rio de Janeiro and São Paulo

	Viana (ES)				Rio de Janeiro and São Paulo			
Structures and ratios	Mean	s	N	C. V.	Mean	s	N	C. V.
Width of head[b]	342.6	12.2	17	3.6	374.1	21.3	61	5.7
Length of eye[b]	200.4	11	17	5.5	222.1	14.2	56	6.4
Width of eye[b]	110.6	8.8	17	7.9	126.7	8.3	55	6.6
Length of palpomere 3[a]	163.5	6.6	15	4	171.6	7.1	60	7.1
Length of palpomere 5[b]	135.3	11.5	14	8.5	145.9	12.3	58	8.5
Total length of palpus[b]	552	21.5	14	3.9	580	29.3	58	5
Maximum width of wing[b]	576	29.9	13	5.2	620	36.4	64	5.9
Length wing/maximum width of wing[a]	3.53	0.201	13	5.7	3.35	0.225	61	6.7
Length of R_2[a]	547	30.4	13	5.5	585	58.9	65	9.7
δ[a]	278	34.2	12	12.3	317	58.6	65	18.5
Length of R_3 [a]	683	30.1	13	4.41	723	60.2	65	8.33
Length of anterior femur[a]	692	33.8	11	4.9	722.5	45.5	50	6.3
Maximum width of spermathecal head[a]	10.49	1.8	18	17.5	11.87	2.13	59	17.9

s: standard deviation; N: number of observations; C.V.: coefficient of variation; *a*: significant at 5%; *b*: significant at 1%; δ: distance between the distal extremity of R_1 and the fork of R_{2+3}.

Exercício

A tabela abaixo foi extraída do artigo: Influência da Altitude, Latitude e Estação de Coleta (Regra de Bergmann) na dimensão de *Lutzomyia intermedia* (Lutz & Neiva, 1912) (Diptera:Psychodidae, Phlebotominae). Marcondes CB et al. (Memórias do Instituto Oswaldo Cruz, 1999;. vol94(5):693-700).

Discuta os resultados obtidos.

Tabela - Distribuição das médias, desvios padrão e variâncias das absorbâncias de amostras de* Lutzomyia longipalpis *coletadas no campo, alimentadas em laboratório e de* Lutzomyia almerioi *procedentes de campo no período de 2002 a 2004.

		Lutzomyia longipalpis			*Lutzomyia almerioi*		
Amostras/Ano	Local	X	S	V	X	S	V
2002	campo	0,518	0,247	0,061	0,781	0,167	0,028
	laboratório	0,815	0,030	0,001	-	-	-
2003	campo	0,652	0,148	0,022	0,764	0,031	0,001
2004	campo	0,668	0,197	0,039	-	-	-

X = média; S = desvio-padrão; V = variância

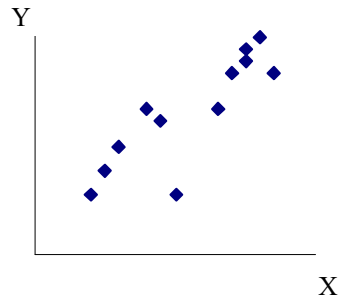
Correlação e regressão linear simples

Análise simultânea entre duas variáveis quantitativas

Gráfico de dispersão: deve ser feito antes da análise numérica dos dados.

É construído com conjuntos de pontos formados por pares de valores (x,y). Pode indicar correlação linear positiva, negativa ou inexistência de correlação. Também é útil para identificar existência de valores aberrantes.

Ex: X: coeficiente de mortalidade por câncer gástrico
Y: consumo médio de sal

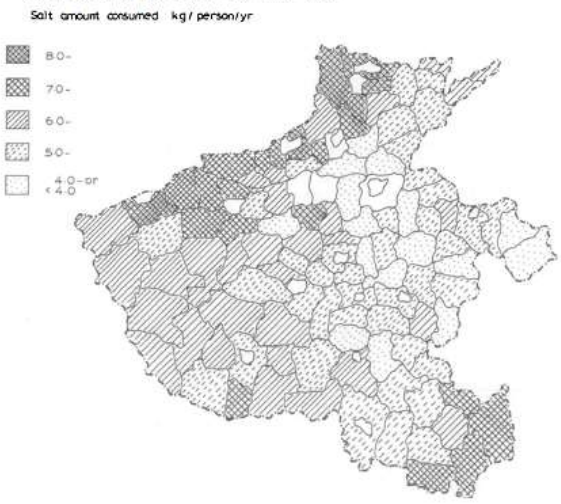


Observar a direção da nuvem de pontos
correlação positiva

International Journal of Epidemiology, 1987. Vol. 16, No. 2

Correlation between High Salt Intake and Mortality Rates for Oesophageal and Gastric Cancers in Henan Province, China

JIAN-BANG LU AND YU-MIN QIN



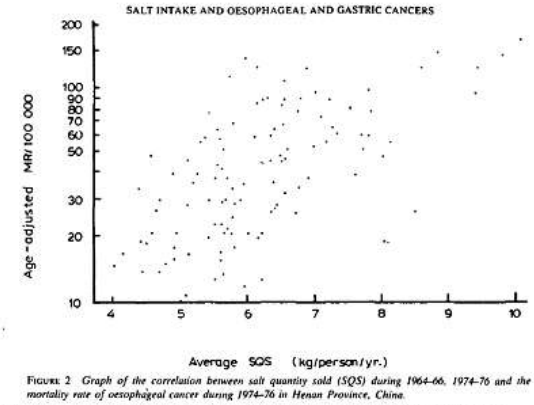
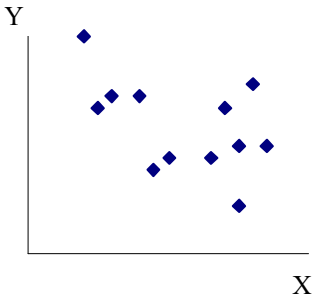


FIGURE 2 *Graph of the correlation between salt quantity sold (SQS) during 1964–66, 1974–76 and the mortality rate of oesophageal cancer during 1974–76 in Henan Province, China.*

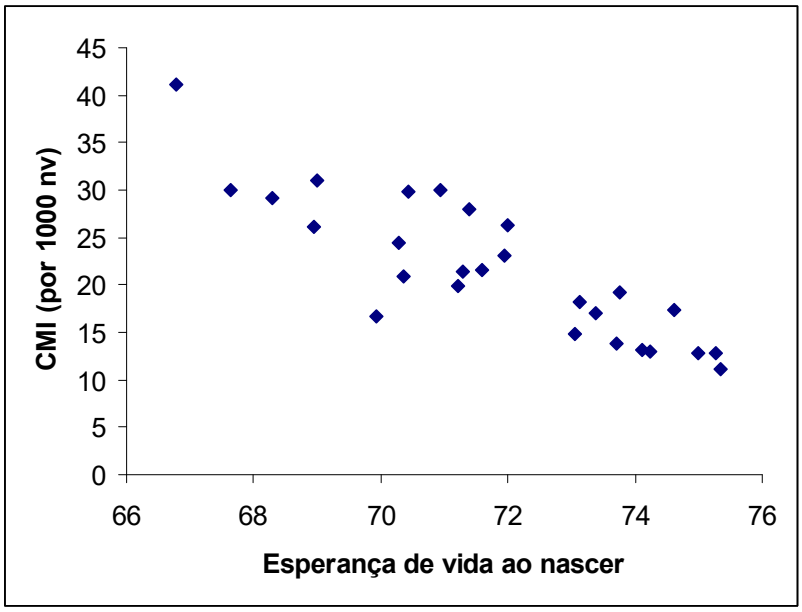
TABLE 1 *The rank correlation coefficient between the SQS in 1964–66, 1974–76 and mortality rate from malignant neoplasms selected in 1974–76 in Henan, China.*

Cancer site	Sex	Σd_i^2	r_s	p value
Oesophagus	M	81945.5	0.6097	<0.01
	F	11820.5	0.4674	<0.01
Stomach	M	77933.5	0.6288	<0.01
	F	96028.3	0.5426	<0.01
Liver	M	185273.5	0.1175	>0.05
Cervix	F	183543.3	0.1257	>0.05
Lung	M	185329.5	0.1172	>0.05
Leukaemia	M	216721.3	0.0323	>0.05

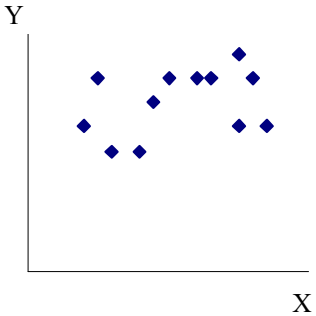
Ex: X: Esperança de vida ao nascer
Y: Coeficiente de mortalidade infantil (por 1000 nascidos vivos)



Observar a direção da nuvem de pontos
correlação negativa



X: coeficiente de mortalidade por câncer de colo de útero
Y: consumo de sal



correlação inexistente

Distinção entre associação e causação: duas variáveis podem estar associadas mas uma não será necessariamente a causa da outra.

Na correlação é comum investigar se mudanças na magnitude de uma variável são acompanhadas de mudanças na magnitude da outra sem significar que uma variável causa a outra.

Coeficiente de correlação de Pearson (ρ), lê-se *rhô*

Mede o grau de associação entre 2 variáveis X e Y.

Definição: $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, onde

σ_{XY} é a covariância de X e Y (dispersão conjunta)

σ_X é o desvio padrão de X (dispersão de X)

σ_Y é o desvio padrão de Y (dispersão de X)

Covariância: É o valor médio do produto dos desvios de X e Y, em relação às suas respectivas médias.

$$\sigma_{XY} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{N}$$

Substituindo-se as fórmulas:

Parâmetro

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}}{\sqrt{\frac{\sum(X-\bar{X})^2}{N}} x \sqrt{\frac{\sum(Y-\bar{Y})^2}{N}}} = \frac{\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}}{\sqrt{\frac{\sum(X-\bar{X})^2}{N}\frac{\sum(Y-\bar{Y})^2}{N}}} = \frac{\frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}}{\frac{1}{N}\sqrt{\sum(X-\bar{X})^2 x \sum(Y-\bar{Y})^2}}$$

estimador (r) $$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\left[\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2\right]}}$$

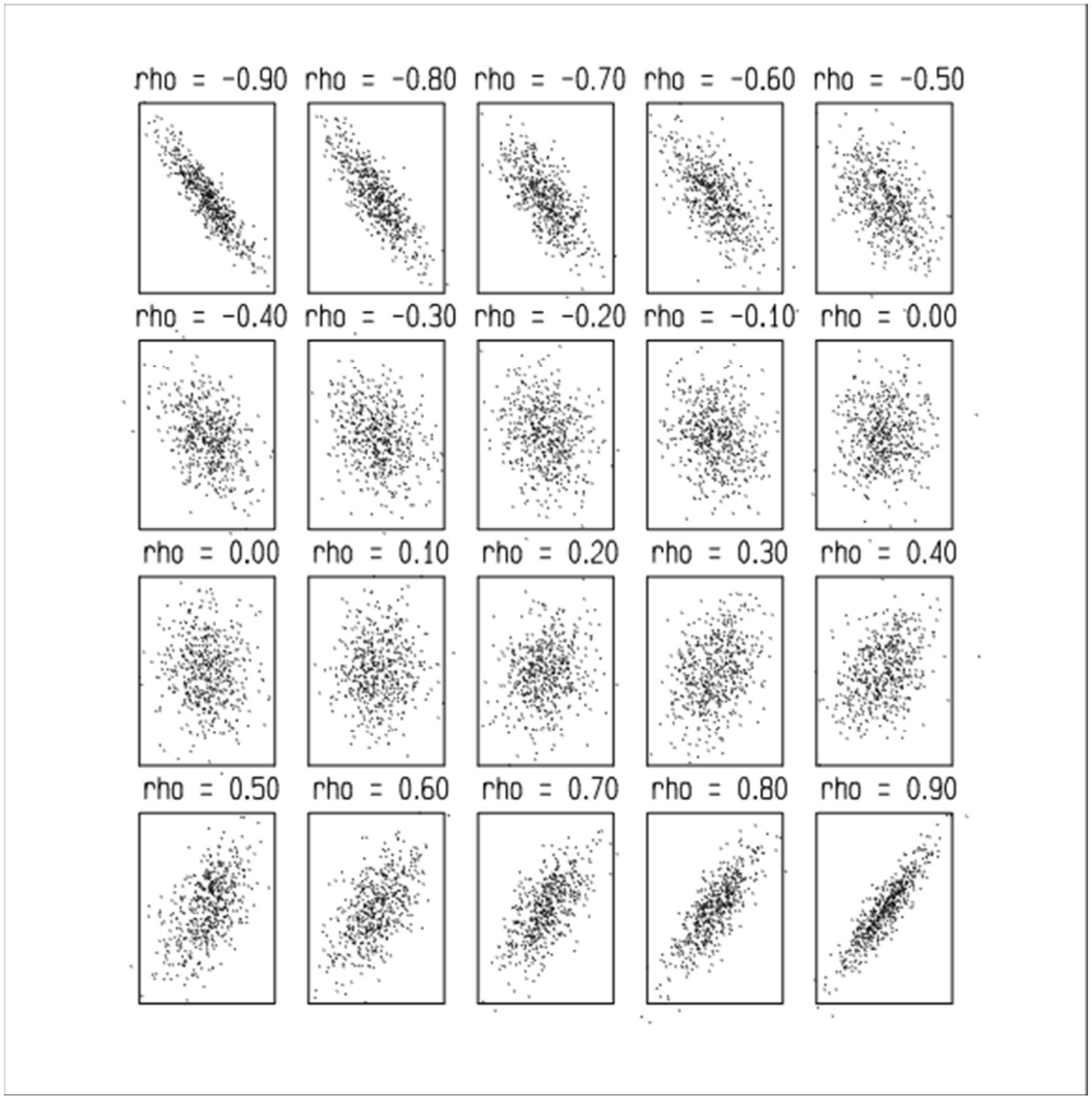
Propriedades

a) $-1 \leq \rho \leq +1$;

b) ρ não possui dimensão, isto é, não depende da unidade de medida das variáveis X e Y;

c) $\rho_{XY} = \rho_{YX}$.

Gráficos de dispersão para diferentes valores do coeficiente de correlação ρ (rho).



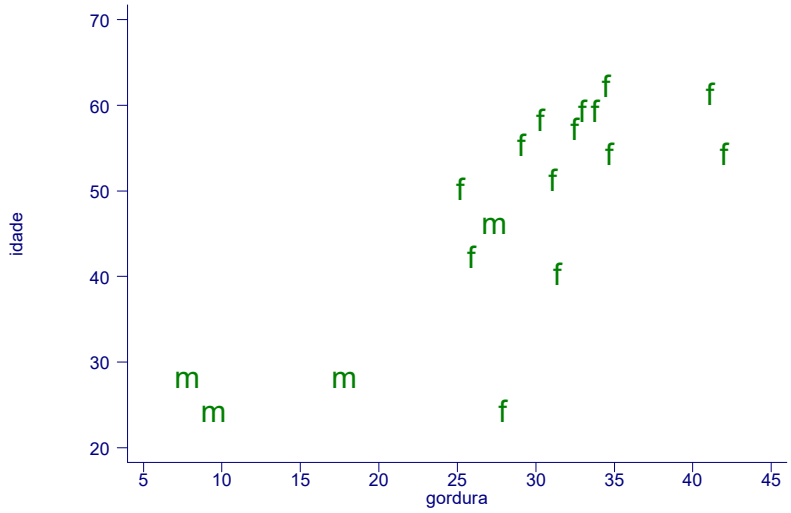
Exemplo:
Os dados a seguir são provenientes de um estudo que investiga a composição corporal e fornece o percentual de gordura corporal (%), idade e sexo para 18 adultos com idades entre 23 e 61 anos.

a) Qual a relação entre a idade e o % de gordura? Existe alguma evidência de que a relação é diferente entre pessoas do sexo masculino e feminino? Explore os dados graficamente.
b) Calcule o coeficiente de correlação de Pearson entre a idade e o % de gordura para homens e mulheres. Interprete os resultados.

Idade	% Gordura	Sexo	Idade	% Gordura	Sexo
23	9,5	M	53	34,7	F
23	27,9	F	53	42,0	F
27	7,8	M	54	29,1	F
27	17,8	M	56	32,5	F
39	31,4	F	57	30,3	F
41	25,9	F	58	33,0	F
45	27,4	M	58	33,8	F
49	25,2	F	60	41,1	F
50	31,1	F	61	34,5	F

M=masculino ; F= feminino
Fonte: Hand DJ et al. A handbook of small data sets. Chapman&Hall, 1994.

Dispersão entre % de gordura e idade



Fonte: Hand DJ et al. A handbook of small data sets. Chapman&Hall, 1994.

Cálculo do coeficiente de correlação de Pearson
Sexo: masculino

Idade (X)	% gordura (Y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
23	9,5	-7,5	-6,13	45,94	56,25	37,52
27	7,8	-3,5	-7,83	27,39	12,25	61,23
27	17,8	-3,5	2,18	-7,61	12,25	4,73
45	27,4	14,5	11,78	170,74	210,25	138,65
30,5	15,625			236,45	291,00	242,13

Coeficiente de correlação (idade,%gordura) masculino: $r = \dfrac{236,45}{\sqrt{291 x 242,13}} = 0,89$

Sexo: feminino

Idade (X)	% gordura (Y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
23	27,9	-27,86	-4,42	123,17	776,02	19,55
39	31,4	-11,86	-0,92	10,93	140,59	0,85
41	25,9	-9,86	-6,42	63,30	97,16	41,23
49	25,2	-1,86	-7,12	13,23	3,45	50,71
50	31,1	-0,86	-1,22	1,05	0,73	1,49
53	34,7	2,14	2,38	5,10	4,59	5,66
53	42	2,14	9,68	20,74	4,59	93,67
54	29,1	3,14	-3,22	-10,12	9,88	10,38
56	32,5	5,14	0,18	0,92	26,45	0,03
57	30,3	6,14	-2,02	-12,42	37,73	4,09
58	33	7,14	0,68	4,85	51,02	0,46
58	33,8	7,14	1,48	10,56	51,02	2,19
60	41,1	9,14	8,78	80,26	83,59	77,06
61	34,5	10,14	2,18	22,10	102,88	4,75
50,86	32,32			333,64	1389,71	312,12

Coeficiente de correlação (idade,%gordura) feminino: $r = \frac{333,64}{\sqrt{1389,71 x 312,12}} = 0,51$;

Coeficiente de correlação considerando o grupo todo (homens e mulheres)

Idade (X)	% gordura (Y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
23	9,5	-23,33	-19,11	445,93	544,44	365,23
27	7,8	-19,33	-20,81	402,35	373,78	433,10
27	17,8	-19,33	-10,81	209,01	373,78	116,88
45	27,4	-1,33	-1,21	1,61	1,78	1,47
23	27,9	-23,33	-0,71	16,59	544,44	0,51
39	31,4	-7,33	2,79	-20,45	53,78	7,78
41	25,9	-5,33	-2,71	14,46	28,44	7,35
49	25,2	2,67	-3,41	-9,10	7,11	11,64
50	31,1	3,67	2,49	9,13	13,44	6,19
53	34,7	6,67	6,09	40,59	44,44	37,07
53	42	6,67	13,39	89,26	44,44	179,26
54	29,1	7,67	0,49	3,75	58,78	0,24
56	32,5	9,67	3,89	37,59	93,44	15,12
57	30,3	10,67	1,69	18,01	113,78	2,85
58	33	11,67	4,39	51,20	136,11	19,26
58	33,8	11,67	5,19	60,54	136,11	26,92
60	41,1	13,67	12,49	170,68	186,78	155,97
61	34,5	14,67	5,89	86,37	215,11	34,68
			Soma	1627,53	2970,00	1421,54

$\bar{x} = 46,33$; $\bar{y} = 28,61$; $S_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n-1}} = \sqrt{\frac{1421,54}{17}} = 9,14\%$; $S_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{2970,0}{17}} = 13,22 anos$

$$r = \frac{1627,53}{\sqrt{2970,0x1421,54}} = 0,79$$

Estudando a relação entre duas variáveis quantitativas por meio da análise de regressão (retirado de Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods, PWS-Kent Publishing Company. Boston, Second Edition)

Considerando-se os dados de pressão arterial sistólica (PAS) e idade para 30 indivíduos
Y: Pressão Arterial sistólica (mmHg) – variável dependente
X: idade (anos) – variável independente

Indivíduo	PAS (Y)	Idade (X)	Indivíduo	PAS (Y)	Idade (X)
1	144	39	16	130	48
2	220	47	17	135	45
3	138	45	18	114	17
4	145	47	19	116	20
5	162	65	20	124	19
6	142	46	21	136	36
7	170	67	22	142	50
8	124	42	23	120	39
9	158	67	24	120	21
10	154	56	25	160	44
11	162	64	26	158	53
12	150	56	27	144	63
13	140	59	28	130	29
14	110	34	29	125	25
15	128	42	30	175	69

Os pares de valores são apresentados por meio do diagrama de dispersão num plano cartesiano como pares (x,y) onde X:Idade e Y: PAS.

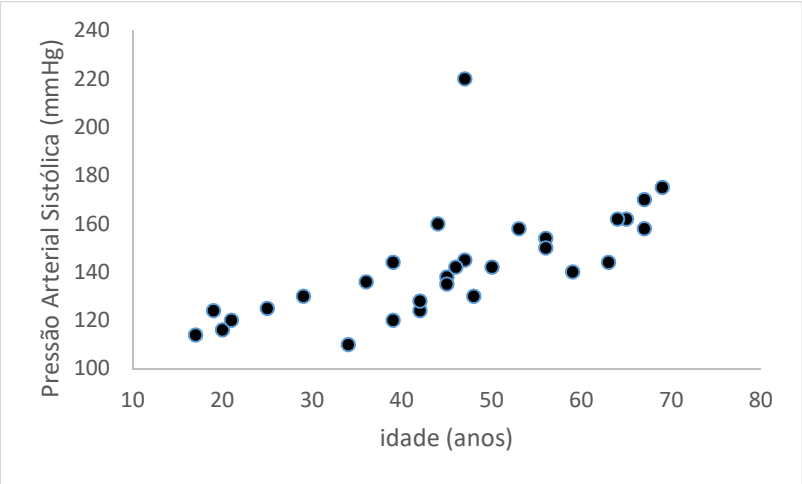


Diagrama de dispersão da Idade (anos) e pressão arterial sistólica (mmHg)
Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

Pode-se investigar qual é o melhor modelo que se ajusta aos dados, por exemplo, uma reta e, neste caso, qual é a equação da reta estimada pelos dados?

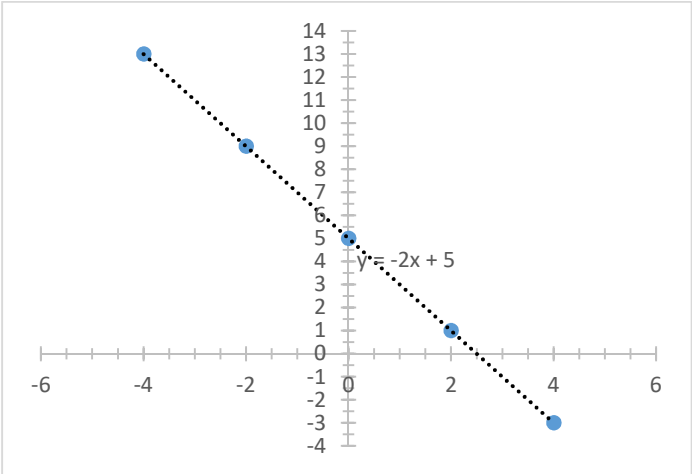
Segundo Kleinbaum DG, Kupper LL , Muller KE uma estratégia geral para ajustar um modelo aos dados seria:

1) Inicia-se assumindo que a reta é o modelo apropriado;
2) Encontra-se a equação da reta que melhor se ajusta aos dados;
3) Determina-se se a reta encontrada ajuda de modo significante a descrever a variável dependente Y;
4) Examina-se se assumir que o modelo é uma reta está correto. Pode-se fazer aqui o teste para ajuste do modelo;
5) Se a reta não for um modelo válido, ajustar outro modelo (ex: parábola) e determinar se Y está bem descrita (passo 3);
6) Repete-se até que se encontre o modelo apropriado.

Propriedades da reta

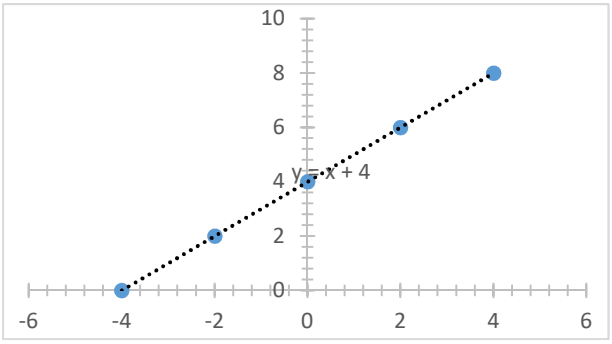
Matematicamente, a equação da reta, é $y = \beta_0 + \beta_1 x$ em que y é o valor de uma variável estimado como função de uma variável X. Os parâmetros da reta, β_0 e β_1 precisam ser estimados. β_0 é denominado intercepto e β_1 é o coeficiente angular (*slope*) da reta e indica quanto y muda para a mudança de uma unidade em X.

Se a equação da reta for y=5-2x; então o intercepto é igual a 5 e o coeficiente angular é igual a 2 e indica que se X aumenta uma unidade, então Y diminui 2 unidades. Isto pode ser visto no gráfico. Para x=2, y=1; para x=3, y=-1 (aumenta-se X em uma unidade e Y diminui 2 unidades). Notar que a reta é decrescente porque o coeficiente angular é negativo.



Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

Os coeficientes da reta y=4+1x indicam que para aumento de uma unidade em X, Y aumenta uma unidade pois o coeficiente angular é igual a 1. Notar que a reta é crescente pois o coeficiente angular é positivo.

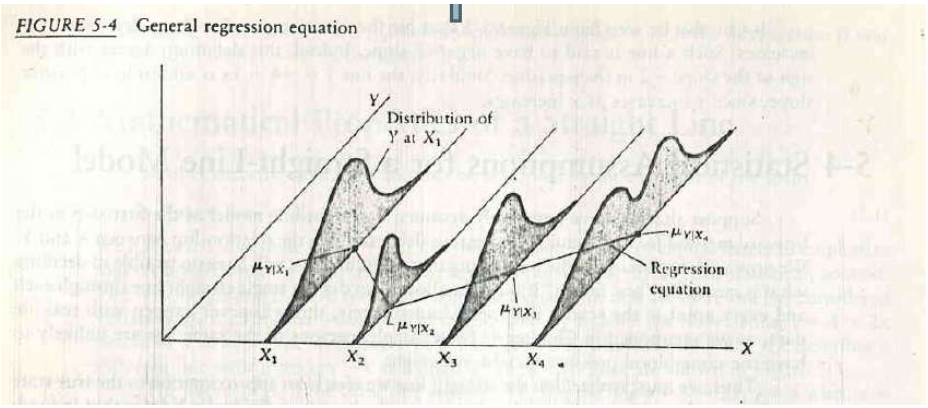


Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

Ao se utilizar o exemplo onde Y representa a pressão arterial e X, a idade, pode-se observar que a relação entre as variáveis é linear e então pode-se perguntar qual é a melhor reta que passa pelos pontos? Na vida real, nem todos os pontos vão cair sobre a reta; indivíduos com a mesma idade podem ter pressão arterial sistólica diferentes pois terão peso, estatura e condições de vida diferentes. Então a reta que se procura não irá predizer a pressão arterial sistólica precisamente de todos os indivíduos mesmo que se considere toda a população. **O que se faz é predizer o valor esperado (valor médio).**

Para ajustar a reta utilizando técnicas estatísticas é necessário fazer pressuposições (estatísticas)

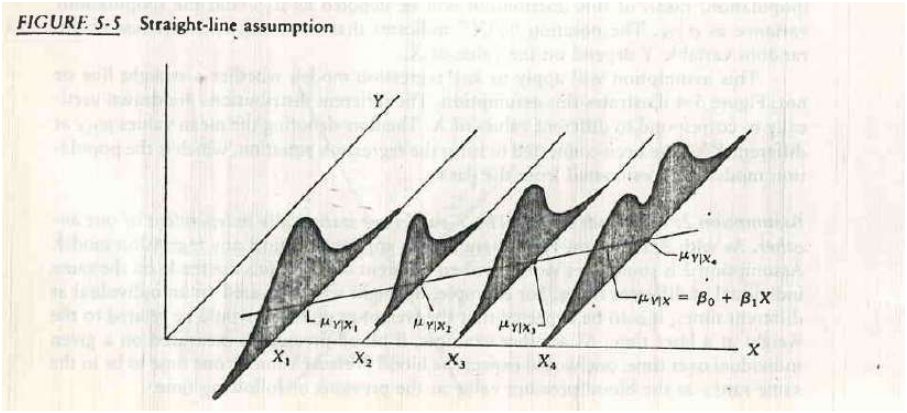
1) Para qualquer valor fixado de X, Y será uma variável aleatória com determinada distribuição de probabilidade e terá média e variância finitas. Na população existirá uma média denotada $\mu_{Y|X}$ e uma variância $\sigma^2_{Y|X}$ indicando que tanto a média como a variância de Y dependem de X. Lê-se Y|X como Y dado X.



Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

2) Os valores de Y são independentes uns dos outros. Outros métodos são utilizados quando existe dependência entre os valores de Y.

3) O valor médio de Y ($\mu_{Y|X}$) é uma reta, função de X. É formada quando são conectados os valores diferentes de $\mu_{Y|X}$. Pode-se escrever $\mu_{Y|X} = \beta_0 + \beta_1 X$, em que β_0 e β_1 são o intercepto e o coeficiente angular da linha reta populacional.



Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

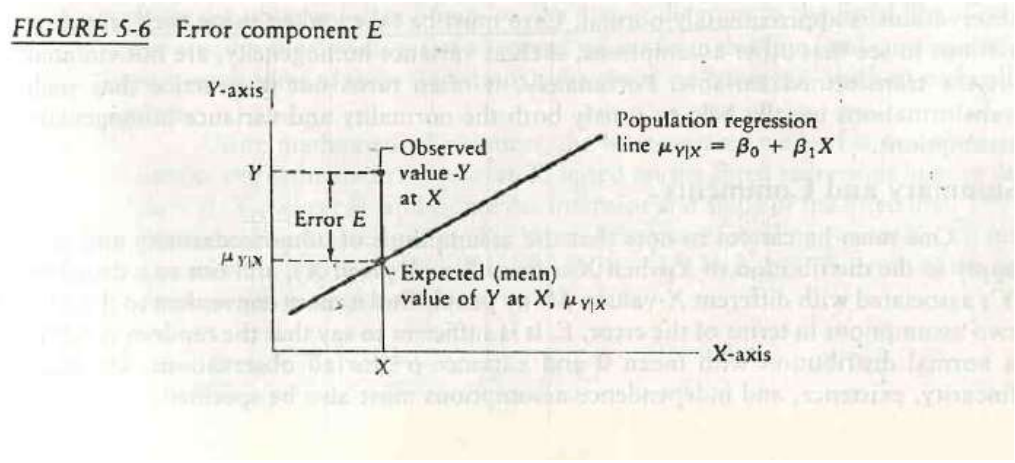
A expressão $\mu_{Y|X} = \beta_0 + \beta_1 X$ também pode ser escrita como $Y = \beta_0 + \beta_1 X + E$ em que *E* representa uma variável aleatória que tem média zero para qualquer valor fixado de X $(\mu_{E|X} = 0)$. Quando se fixa X é como se a característica X fosse aferida sem erro existindo um erro aleatório somente para o termo *E*.

X não é uma variável aleatória porque é fixada, então $Y = (\beta_0 + \beta_1 X) + E$ é a soma de um termo constante $(\beta_0 + \beta_1 X)$ e um erro aleatório (*E*). A distribuição de Y e de E só diferem no termo constante. Como E tem média zero, Y tem média $(\beta_0 + \beta_1 X)$.

Assim, tanto a equação $\mu_{Y|X} = \beta_0 + \beta_1 X$ como a equação $Y = (\beta_0 + \beta_1 X) + E$, descrevem o modelo estatístico pois consideram Y como variável aleatória.

A variável *E* descreve quão longe a resposta de um indivíduo está da linha de regressão populacional. Constitui um erro médio esperado para o que se observa de Y, dado um X. E é denominado componente de erro do modelo. Matematicamente, *E* pode ser escrito como $E = Y - (\beta_0 + \beta_1 X)$ ou por $E = Y - \mu_{Y|X}$

Componente do erro E.



Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

O componente do erro é importante para definir uma reta bem ajustada. Esta deverá ter erros pequenos entre os valores observados e os preditos pelo modelo ajustado.

4- Igualdade de variância de Y para qualquer valor de X (homocedasticidade). A ausência de igualdade de variância pode ser vista na figura 5-5. A distribuição de Y para X_1 é mais espalhada que a distribuição de Y para X_2. Se existir homocedasticidade então $\sigma^2_{Y|X} = \sigma$ para todo X.

5- Para qualquer valor de X, Y tem que ter distribuição normal. A figura 5-5 apresenta uma violação desta pressuposição. A violação desta pressuposição pode ser vista na figura 5-5. Esta violação não é tão problemática como a violação da homocedasticidade e as conclusões do modelo podem ainda ser consideradas válidas. Também é possível transformar os dados e utilizar uma distribuição com valores transformados que seguem uma distribuição normal. Neste caso a pressuposição de homogeneidade de variâncias precisa ser reverificada.

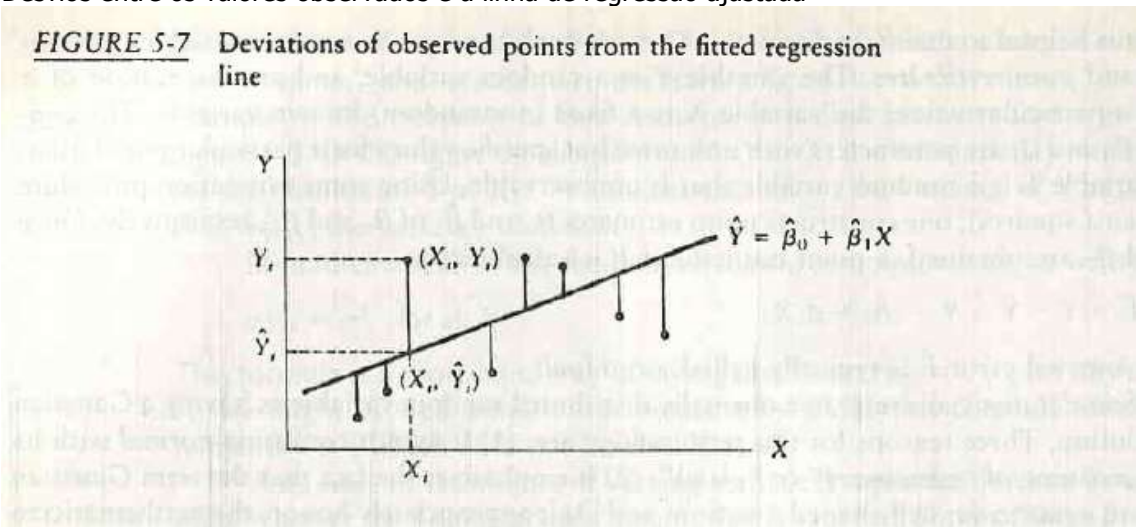
Resumindo:
Y: pressão arterial é uma variável aleatória e a observação desta para um indivíduo leva a um valor particular observado;
X: é uma variável fixada
As constantes β_0 e β_1 são parâmetros desconhecidos da população de onde os dados foram retirados;
E é uma variável aleatória não observável e representa o termo de erro de aferição de Y.

$\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores (no ponto) de β_0 e β_1. Uma vez que se obtém os valores estimados para $\hat{\beta}_0$ e $\hat{\beta}_1$, então pode-se estimar $\hat{E}$ que é denominado resíduo. $\hat{E} = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$.

Determinando a melhor reta que passa pelos dados

O método dos mínimos quadrados determina a melhor reta ajustada que minimiza a soma de quadrados das distâncias dos segmentos de linha verticais desenhados a partir dos pontos dos valores observados e apresentados no diagrama de dispersão, até a reta ajustada.

Desvios entre os valores observados e a linha de regressão ajustada



Fonte: Kleinbaum DG, Kupper LL , Muller KE. Applied Regression Analysis and Other Multivariable Methods,.PWS-Kent Publishing Company. Boston, Second Edition

O método de mínimos quadrados

Considerando-se um valor de X denotado genericamente por X_i

Tem-se o valor de $\hat{Y}_i$ estimado para X_i: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

A distância vertical entre o ponto observado (X_i, Y_i) e o ponto estimado pela reta $(X_i, \hat{Y}_i)$ é, em valor absoluto escrito como: $\left|Y_i - \hat{Y}_i\right|$ ou $\left|Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right|$.

A soma de quadrados de cada distância é dada por $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

A solução do método é dada ao se encontrar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a soma de quadrados descrita acima. Para tanto, é necessário derivar a equação em relação a cada um dos parâmetros e igualar a zero

$$\frac{\partial \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i$$

Igualando a zero cada uma das equações para encontrar os valores estimados.

$$-2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$-2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0$$

Duas equações e dois parâmetros. Resolvendo o sistema:

Divide-se ambos os lados das equações por 2

A primeira equação resulta em $-\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}\hat{\beta}_0 + \sum_{i=1}^{n}\hat{\beta}_1 X_i = 0$ ou

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

A segunda equação resulta em $-\sum_{i=1}^{n} Y_i X_i + \sum_{i=1}^{n}\hat{\beta}_0 X_i + \sum_{i=1}^{n}\hat{\beta}_1 X_i^{\,2} = 0$ ou

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2}$$

Da primeira equação obtém-se $\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i}{n} + \hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n}$ ou $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Da segunda equação obtém-se $-\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} = \hat{\beta}_0 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i X_i$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} = \sum_{i=1}^{n} Y_i X_i - (\overline{Y} + \hat{\beta}_1 \overline{X}) \sum_{i=1}^{n} X_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} = \sum_{i=1}^{n} Y_i X_i - \overline{Y} \sum_{i=1}^{n} X_i + \hat{\beta}_1 \overline{X} \sum_{i=1}^{n} X_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} = \sum_{i=1}^{n} Y_i X_i - \frac{\sum_{i=1}^{n} Y_i}{n} n \frac{\sum_{i=1}^{n} X_i}{n} + \hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n} n \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} = \sum_{i=1}^{n} Y_i X_i - n\overline{YX} + \hat{\beta}_1 n\overline{XX}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_i^{\,2} - \hat{\beta}_1 n\overline{X}^2 = \sum_{i=1}^{n} Y_i X_i - n\overline{YX}$$

$$\hat{\beta}_1 (\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2) = \sum_{i=1}^{n} Y_i X_i - n\overline{YX}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{YX}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2}$$

Também pode-se escrever $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{YX}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$ uma vez que

$$\frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}) Y_i}{\sum_{i=1}^{n} (X_i - \overline{X}) X_i} = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{YX}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^2}$$

Assim,

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$ e

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ pode também ser escrito como $\hat{Y} = \overline{Y} + \beta_1 (X - \overline{X})$

Utilizando-se os dados de pressão arterial temos

Indivíduo	PAS (Y)	Idade (X)	$(y_i - \bar{y})$	$(x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	144	39	1,47	-6,13	-9,00	2,15	37,62
2	220	47	77,47	1,87	144,60	6001,08	3,48
3	138	45	-4,53	-0,13	0,60	20,55	0,02
4	145	47	2,47	1,87	4,60	6,08	3,48
5	162	65	19,47	19,87	386,74	378,95	394,68
6	142	46	-0,53	0,87	-0,46	0,28	0,75
7	170	67	27,47	21,87	600,60	754,42	478,15
8	124	42	-18,53	-3,13	58,07	343,48	9,82
9	158	67	15,47	21,87	338,20	239,22	478,15
10	154	56	11,47	10,87	124,60	131,48	118,08
11	162	64	19,47	18,87	367,27	378,95	355,95
12	150	56	7,47	10,87	81,14	55,75	118,08
13	140	59	-2,53	13,87	-35,13	6,42	192,28
14	110	34	-32,53	-11,13	362,20	1058,42	123,95
15	128	42	-14,53	-3,13	45,54	211,22	9,82
16	130	48	-12,53	2,87	-35,93	157,08	8,22
17	135	45	-7,53	-0,13	1,00	56,75	0,02
18	114	17	-28,53	-28,13	802,74	814,15	791,48
19	116	20	-26,53	-25,13	666,87	704,02	631,68
20	124	19	-18,53	-26,13	484,34	343,48	682,95
21	136	36	-6,53	-9,13	59,67	42,68	83,42
22	142	50	-0,53	4,87	-2,60	0,28	23,68
23	120	39	-22,53	-6,13	138,20	507,75	37,62
24	120	21	-22,53	-24,13	543,80	507,75	582,42
25	160	44	17,47	-1,13	-19,80	305,08	1,28
26	158	53	15,47	7,87	121,67	239,22	61,88
27	144	63	1,47	17,87	26,20	2,15	319,22
28	130	29	-12,53	-16,13	202,20	157,08	260,28
29	125	25	-17,53	-20,13	353,00	307,42	405,35
30	175	69	32,47	23,87	774,87	1054,08	569,62
Média=	142,53	45,13		Soma=	6585,87	14787,47	6783,47

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{6585,87}{6783,47} = 0,97$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_0 = 142,53 - (0,97x45,13) = 98,71$$

A reta estimada é escrita como:

$\hat{Y} = 98,71 + 0,97X$ ou

$$\hat{Y} = 142,53 + (0,97)(X - 45,13)$$

Esta reta pode ser desenhada definindo-se dois pontos no diagrama de dispersão. Escolhe-se um valor para X e calcula-se o valor de Y estimado.

Por exemplo para x=25, y=98,71+(0,97x25)=122,96
x=55, y=98,71+(0,97x55)=152,06

Desenha-se no diagrama estes pontos (x=25; y=122,95) e (x=55; y= 152,06) e desenha-se a reta estimada.

Interpretação: A reta indica tendência de crescimento da pressão arterial com o aumento da idade. Pela reta estimada pode-se dizer que com o aumento de 1 ano na idade, a pressão arterial sistólica aumentaria em média 0,97 mmHg.

Exercício

Utilizando-se o Y: gordura abdominal e X: idade, estimar

$\hat{\beta}_1 =$

$\hat{\beta}_0 =$

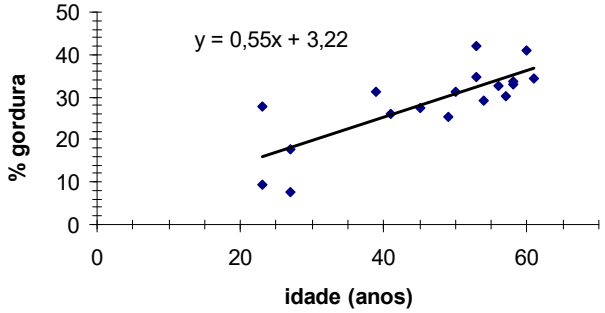
Confira se a reta estimada é dada por: $\hat{Y} =$ 3,22 + 0,55 X

Interpretação:

Para aumento de 1 ano, o percentual de gordura aumenta 0,55%.

Desenhando-se a reta

Para x = 30; y = 19,7; para x = 50, y = 30,7



OBS: o coeficiente angular depende das unidades de medida de X e Y. Isto deve ser considerado na decisão da importância do coeficiente angular.

O coeficiente angular da equação de Y=f(X) é diferente do coeficiente angular de X=f(Y), a menos que os desvios padrão de X e Y sejam iguais.

Usos da reta de regressão:

- Predição - utilizar X para predizer Y; quando a correlação for forte, melhor é a predição;
- Correlação – mede o grau de relacionamento linear entre X e Y;
- Resumir os dados – cada valor de X tem um valor médio de Y.

Exercício

Considere os dados de 11 alunos sendo Y: nota obtida na prova e X: número de horas de estudo. Investigue a relação entre as variáveis:

a) Construa o diagrama de dispersão
b) Calcule o coeficiente de correlação linear de Pearson
c) Estime a reta de regressão
d) Desenhe a reta no gráfico e
e) Interprete os coeficientes da reta.

Aluno	nota	horas de estudo
1	10	30
2	9	13
3	8	11
4	4	4
5	6	5
6	7	8
7	8	15
8	6	17
9	5	10
10	7	17
11	7	13

Análise de duas variáveis qualitativas

Medidas de Associação

- Razão de prevalências (estudos de prevalência ou transversais))
- Razão de incidências ou risco relativo (estudos de coorte ou de seguimento)
- Razão de odds ou odds ratio (estudos caso-controle)
- Qui quadrado de Pearson e coeficiente de associação de Yule (estudos de prevalência, incidência e caso-controle)

Razão de prevalências

Estudo de prevalência

São apresentados dados sobre o estado nutricional de 1226 crianças brasileiras de 2 anos de idade, segundo sexo. Local X, Ano Y.

Estado nutricional	Masculino	Feminino	Total
Desnutridas	29	20	49
Normais	574	603	1177
Total	603	623	1226

Fonte: dados hipotéticos.

Prevalência de desnutrição: $\frac{49}{1226} = 0{,}040$ ou 4%.

Prevalência de desnutrição segundo sexo:

Masculino: $\frac{29}{603} = 0{,}05$ ou 5,0%; Feminino: $\frac{20}{623} = 0{,}032$ ou 3,2%.

Razão de prevalências: $\frac{\frac{29}{603}}{\frac{20}{623}} = 1{,}5$

Se a razão de prevalências for igual a 1 ou a diferenças de prevalências for igual a 0 então diz-se que as variáveis não estão associadas. Na inferência estatística é possível testar se o valor observado vem de uma população com parâmetro igual a 1. Até agora estamos construindo a estatística que permite verificar associação.

Interpretação como medida de efeito:
A prevalência de desnutrição entre meninos é 1,5 vezes (uma vez e meia) a prevalência de desnutrição entre meninas.

A prevalência de desnutrição parece ser maior entre as crianças do sexo masculino. Os meninos apresentam uma prevalência 50% maior do que as meninas, calculado como (1,5-1)x100.

Se para o cálculo da razão de prevalências for utilizada a prevalência entre meninas no numerador, a razão de prevalências se altera para:

Razão de prevalências: $\frac{\frac{20}{623}}{\frac{29}{603}} = \frac{603x20}{623x29} = 0{,}68$

De modo semelhante,

Se a razão de prevalências for igual a 1 então diz-se que as variáveis não estão associadas. Na inferência estatística é possível testar se o valor observado vem de uma população com parâmetro igual a 1.

Interpretação como medida de efeito:

A prevalência de desnutrição entre meninas é 0,68 vezes a prevalência de desnutrição entre meninos.

A prevalência de desnutrição parece ser menor entre as crianças do sexo feminino. As meninas apresentam uma prevalência 32% menor do que os meninos, calculado como (1-0,68)x100.

De forma geral

Y: variável resposta (Ex: desnutrição)

X: variável explicativa ou de confusão (Ex: sexo)

	Variável Y		
Variável X	Y_1	Y_0	Total (%)
X_1	a	b	n_1 (100
X_0	c	d	n_0 (100
Total	m_1	m_2	n (100)

p= prevalência de Y_1= m_1/n

p_1= prevalência de $Y_1|x_1$= a/n_1

p_0= prevalência de $Y_1|x_0$= c/n_0

rp= razão de prevalências= p_1/p_0; dp=diferença de prevalências= p_1-p_0

Exercício

Considerando-se os dados apresentados na tabela 4, retirado do artigo: ERICA: padrões de consumo de bebidas alcoólicas em adolescentes brasileiros. Evandro SF Coutinho et al, Suplemento ERICA; Artigo Original Rev Saúde Pública 2016;50(supl 1):8s

Construa a razão de prevalências de consumo de cerveja entre homens e mulheres. Os dados sugerem que existe associação entre as variáveis?

Tabela 4. Percentual dos tipos de bebida consumida pelos adolescentes na maioria das vezes nos últimos 30 dias, por sexo e idade. ERICA, Brasil, 2013-2014.

Caracte-rísticas	Cerveja		Vinho		*Ice*		Cachaça		Vodca, tequila e rum		Outros	
	%	IC95%	%	IC95%	%	IC95%	%	IC95%	%	IC95%	%	IC95%
Masculino	25,1	22,4-27,9	11,1	9,5-13,0	12,4	10,9-14,2	3,6	2,7-4,7	35,3	32,3-38,5	8,4	7,1-10,0
12-14	23,1	19,3-27,4	13,6	11,0-16,7	13,3	11,4-15,5	3,4	2,1-5,5	30,1	26,3-34,3	10,9	8,5-13,8
15-17	27,3	24,8-29,9	8,3	7,0-10,0	11,5	9,2-14,2	3,7	2,5-5,4	41,1	37,6-44,8	5,7	4,6-6,9
Feminino	19,5	17,7-21,5	12,9	11,0-15,0	19,9	16,9-23,4	2,7	2,0-3,6	33,3	30,4-36,4	10,0	8,5-11,7
12-14	15,5	13,5-17,8	14,4	11,4-18,1	22,8	18,4-27,8	2,2	1,3-3,7	29,5	25,7-33,7	13,4	10,9-16,3
15-17	24,0	21,2-27,0	11,2	9,6-12,9	16,8	13,8-20,3	3,2	2,3-4,5	37,5	33,6-41,5	6,3	5,0-7,8

Razão de riscos ou risco relativo ou razão de incidências

Estudo de incidência

Distribuição de pessoas segundo hábito de fumar e morte em 5 anos por DIC. Local X. Ano Y

	Morte em 5 anos por DIC		
Fumar	Sim	Não	Total
Sim	208	850	1058
Não	264	1467	1731
Total	472	2317	2789

Fonte: dados hipotéticos.

r= 472/2789= 0,17 = 17%

r_1=208/1058= 0,20= 20%

r_0=264/1731= 0,15=15%

rr=0,20/0,15= 1,33

ra= 0,20- 0,15= 0,05= 5%

Se a razão de incidências ou RR for igual a 1 ou a diferenças de incidências for igual a 0 então diz-se que as variáveis não estão associadas. Na inferência estatística é possível testar se o valor observado da razão de incidências vem de uma população com parâmetro igual a 1.

Interpretação como medida de efeito:

A incidência de óbitos entre fumantes é 1,33 vezes a incidência de óbitos entre não fumantes.

A incidência de mortes parece ser maior entre as pessoas que fumam. Os fumantes apresentam uma incidência 33% maior do que os não fumantes.

Pela diferença diz-se que 5% dos óbitos excedentes são devidos ao fumo. É possível dizer que 5% dos óbitos poderiam ser evitados na ausência do fumo.

De forma geral

Y: variável resposta
X: variável explicativa ou de confusão

	Variável Y		
Variável X	Y_1	Y_0	Total (%)
X_1	a	b	n_1 (100
X_0	c	d	n_0 (100
Total	m_1	m_2	n (100)

r= incidência de Y1= m_1/n
r_1= incidência de Y1 entre os expostos (x_1)= a/n_1
r_0= incidência de Y1 entre os não expostos (x_0)= c/n_0
r_i= razão de incidências= r_1/r_0
d_i= diferença de incidências= r_1-r_0

incidência		risco	
r_1	r_0	r_1/r_0	r_1-r_0

r_i=rr=**razão de riscos=risco relativo**=r_1/r_0

d_i= ra= risco atribuível= r_1-r_0

Exercício

Investigação de toxinfecção alimentar

Tomou sorvete de baunilha	Toxinfecção					
	Sim		Não		Total	
	n	%	n	%	n	%
Sim	43	79,6	11	20,4	54	100
Não	3	14,3	18	85,7	21	100
Total	46	61,3	29	38,7	75	100

Fonte: Epi Info, 2000.

Calcule

a) a incidência global ou taxa de ataque global
b) incidência entre quem tomou sorvete
c) incidência entre quem não tomou sorvete
d) o risco relativo ou a razão de incidências
e) Você diria que tomar sorvete de baunilha consistiu em fator de risco para toxinfecção alimentar?

Razão de odds ou odds ratio

Estudo do tipo caso-controle

Odds ratio

Odds e probabilidade – duas formas diferentes de quantificar incertezas

Supor que durante um jogo de basquete um jogador acerta a cesta 2 vezes em 5 tentativas. Chamando $\hat{p}$ (p chapéu) de probabilidade de acerto tem-se que $\hat{p} = \frac{2}{5}$ = 0,4 ou 40% e a probabilidade de erro, $\hat{q} = \frac{3}{5}$ = 0,6 ou 60%.

Considerando-se que a probabilidade de acerto ou de erro = p+q= 1; então $\hat{q} = 1 - \hat{p}$.

Odds ratio

Define-se *odds* como a razão entre a probabilidade de acerto e a probabilidade de erro, ou seja, $\frac{p}{1-p}$.

No exemplo acima, o *odds* a favor de acerto é $\frac{p}{1-p} = \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2x5}{3x5} = \frac{2}{3} = 0{,}67$ ou 0,67:1 (0,67 acertos para 1 erro).

Exemplo

Estudo do tipo caso-controle

Os dados a seguir são de um estudo sobre câncer de esôfago e consumo de álcool. Local X. Ano Y.

Condição	Consumo médio de álcool (g/dia)		Total
	80 e +	0-79	
Casos	96	104	200
Controles	109	666	775
Total	205	770	975

Fonte: Tuyns et al.,1977.

(entre expostos) odds a favor de casos entre consumidores de 80 e + g/dia: $\frac{96}{205} : \frac{109}{205} = \frac{96}{109} = 0{,}88$

(entre não expostos) odds a favor de casos entre consumidores de 0-79g/dia: $\frac{104}{770}:\frac{666}{770}=\frac{104}{666}=0{,}16$

odds ratio: $\frac{96}{109}:\frac{104}{666}=\frac{96x666}{109x104}=5{,}6$

Se a razão de odds for igual a 1 então diz-se que as variáveis não estão associadas. Na inferência estatística é possível testar se o valor observado vem de uma população com parâmetro igual a 1. Até agora estamos construindo a estatística que permite verificar associação.

Interpretação:
A força de morbidade de câncer de esôfago entre consumidores de 80 e + g/dias de bebida alcoólica é 5,6 a força de morbidade entre os que consomem de 0 a 79g/dia.

Em casos especiais, o *odds ratio* pode ser um bom estimador do risco (quando a doença de estudo é rara).

De forma geral
Y: variável resposta
X: variável explicativa ou de confusão

	Variável Y		
Variável X	Y_1	Y_0	Total (%)
X_1	a	b	n_1 (100
X_0	c	d	n_0 (100
Total	m_1	m_2	n (100)

odds a favor de Y_1:

na categoria X_1= $(a/n_1)\div(b/n_1)$

na categoria X_0= $(c/n_0)\div(d/n_0)$

odds ratio: $[(a/n_1)\div(b/n_1)]\div[(c/n_0)\div(d/n_0)]=\frac{a/b}{c/d}=\frac{ad}{bc}$

Exercício

Os dados a seguir são de um estudo sobre câncer de mama realizado na região sul do Brasil.

Idade	Grupo	
	Caso	Controle
18 – 35	2	13
36 – 40	4	4
41 – 49	26	8
50 – 59	34	21
60 e +	36	56
Total	102	102

Fonte: Dagmar S Lauter et al. *Revista Ciência & Saúde,* Porto Alegre, v. 7, n. 1, p. 19-26, jan./abr. 2014

a) Calcule o odds de caso para cada grupo de idade
b) Calcule o odds ratio a favor de caso considerando a menor idade como referência
c) Você diria que as variáveis estão associadas?

QUI-QUADRADO DE PEARSON

É uma estatística que permite verificar se existe ou não associação entre duas variáveus qualitativas.

Os exemplos são retirados de BUSSAB, Wilson de O; MORETTIN, Pedro A. *Estatística básica.* 5ª Ed. São Paulo: Saraiva, 2004.

X - curso universitário e
Y – sexo do aluno

Questão: sexo do indivíduo influi na escolha do curso?

Situação 1

Curso	Masculino (A)	Feminino	Total
	n	n	n
Economia (B)	24	36	60
Administração	16	24	40
Total	40	60	100

Definição de independência:

A – Ser do sexo masculino;
B – Estar cursando economia.

A e B são independentes se P(A e B) = P(A) x P(B).

P(A e B) = Probabilidade (ser homem e estar cursando Economia)

P(A e B) = $\frac{24}{100} = 0{,}24$

P(A) = $\frac{40}{100} = 0{,}4$ P(B) = $\frac{60}{100} = 0{,}6$

Como $\frac{24}{100} = \frac{40}{100} x \frac{60}{100}$, então A e B são independentes e, portanto, não existe associação.

Notar que no caso de independência, as proporções de escolha dos cursos não diferem segundo sexo do estudante

Curso	Masculino (A)		Feminino		Total	
	n	proporção	n	proporção	n	proporção
Economia (B)	24	0,6	36	0,6	60	0,6
Administração	16	0,4	24	0,4	40	0,4
Total	40	1	60	1	100	1

Situação 2

Curso	Masculino (B) n	Feminino n	Total n
Física (A)	100 (a)	20 (b)	120
Ciências Sociais	40 (c)	40 (d)	80
Total	140	60	200

Lembrando a definição de independência

A e B são independentes se P(A e B) = P(A) x P(B).

P(A e B) = Probabilidade (ser homem e estar cursando Economia)

P(A e B) = $\frac{100}{200} = 0{,}5$

$$P(A) = \frac{100}{140} = 0{,}71$$

$$P(B) = \frac{140}{200} = 0{,}7$$

Como, A e B não são independentes e, portanto, é possível que exista associação.

Notar que as proporções de escolha dos cursos diferem segundo sexo do estudante e a distribuição de alunos em cada curso, segundo sexo não é a mesma, sexo e curso podem estar associados

Curso	Masculino		Feminino		Total	
	n	proporção	n	proporção	n	proporção
Física	100	0,7	20	0,3	120	0,6
Ciências Sociais	40	0,3	40	0,7	80	0,4
Total	140	1	60	1	200	1

Até agora foi apresentada a definição de independência utilizando-se o conceito probabilístico.

Pearson propôs uma estatística que compara os valores observados de uma distribuição conjunta observada com valores esperados calculados a partir da imposição que as variáveis sejam independentes.

Se a variável sexo não fosse associada à escolha do curso, quantos indivíduos poder-se-ia esperar em Física, entre os homens?

Aplicar a proporção marginal utilizando o raciocínio da regra de três:
x (valor desconhecido) estará para 140 assim como 120 está para 200

$$\frac{x}{140} = \frac{120}{200} \text{ e } x = \frac{140x120}{200} = 84$$

O mesmo raciocínio é feito para cada casela da distribuição conjunta.

Para os demais valores esperados observar os cálculos abaixo.

Curso	Sexo	Valor Esperado sob a condição de independência
Física	Masculino (a)	$\frac{120}{200} x140 = 84$
Física	Feminino (b)	$\frac{120}{200} x60 = 36$
Ciências Sociais	Masculino (c)	$\frac{80}{200} x140 = 56$
Ciências Sociais	Feminino (d)	$\frac{80}{200} x60 = 24$

Tabela de frequências esperadas, sob a condição de independência.

Curso	Masculino n	Feminino n	Total n
Física	84	36	120
Ciências Sociais	56	24	80
Total	140	60	200

Estatística qui quadrado de Pearson

Valores observados O	Valores esperados E	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
100	84	16	256	3,048
40	56	-16	256	4,571
20	36	-16	256	7,11
40	24	16	256	10,667
			Qui-quadrado=	25,397

O Qui-quadrado é obtido somando-se a diferença ao quadrado entre as frequências observadas e as esperadas, dividido pelas frequências esperadas.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Quando a estatística qui quadrado será igual a zero?

Se o qui quadrado for igual a 0 então diz-se que as variáveis não estão associadas ou que existe independência entre elas. Na inferência estatística é possível testar se o valor observado vem de uma população com parâmetro igual a zero. Até agora estamos construindo a estatística que permite verificar associação.

Como o qui quadrado indica somente se existe ou não associação, é necessário calcular um coeficiente de associação que ajuda a verificar a força de associação.

Coeficiente de associação de Yule

Coeficiente de associação de Yule – permite investigar a força (magnitude) da associação

$$Y = \frac{a.d - b.c}{a.d + b.c} \quad \text{, onde:} \qquad -1 \le Y \le +1$$

$$Y = \frac{100x40 - 20x40}{100x40 + 20x40} = +0{,}67$$

Exemplo:
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo condição de sobrevivência e peso ao nascer (g).

Peso ao nascer	Óbito		Sobrevivente		Total	
	n	%	n	%	n	%
Baixo peso (<2500)	24	64,9	13	35,1	37	100
Não baixo peso (2500 e mais)	3	23,1	10	76,9	13	100
Total	27	54,0	23	46,0	50	100

Fonte: Hand DJ et al. A handbook of small data sets. Chapman&Hall, 1994.

Cálculo do qui-quadrado de Pearson

Valores observados O	Valores esperados E	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
24	19,98	4,02	16,16	0,809
3	7,02	-4,02	16,16	2,302
13	17,02	-4,02	16,16	0,949
10	5,98	4,02	16,16	2,702

Qui-quadrado=6,762

Coeficiente de associação de Yule

$$Y = \frac{24x10 - 3x13}{24x10 + 3x13} = \frac{240 - 39}{240 + 39} = \frac{201}{279} = +0{,}72$$

Portanto pode-se dizer que a situação de sobrevivência dos recém-nascidos pode estar associada ao peso ao nascer porque o qui-quadrado é diferente de zero. Pelo valor do coeficiente de Yule

pode-se dizer que a associação é forte. Pode-se notar pela tabela a proporção de recém-nascidos que vão a óbito entre os recém-nascidos de baixo peso é maior do que a proporção de óbitos entre os que não nasceram com baixo peso. Isto indica que o peso ao nascer modifica as proporções de ocorrência de óbito indicando existência de associação entre os eventos.

Exercício

Investigação de toxinfecção alimentar

Tomou sorvete de baunilha	Toxinfecção					
	Sim		Não		Total	
	n	%	n	%	n	%
Sim	43	79,6	11	20,4	54	100
Não	3	14,3	18	85,7	21	100
Total	46	61,3	29	38,7	75	100

Fonte: Epi Info, 2000.

Calcular

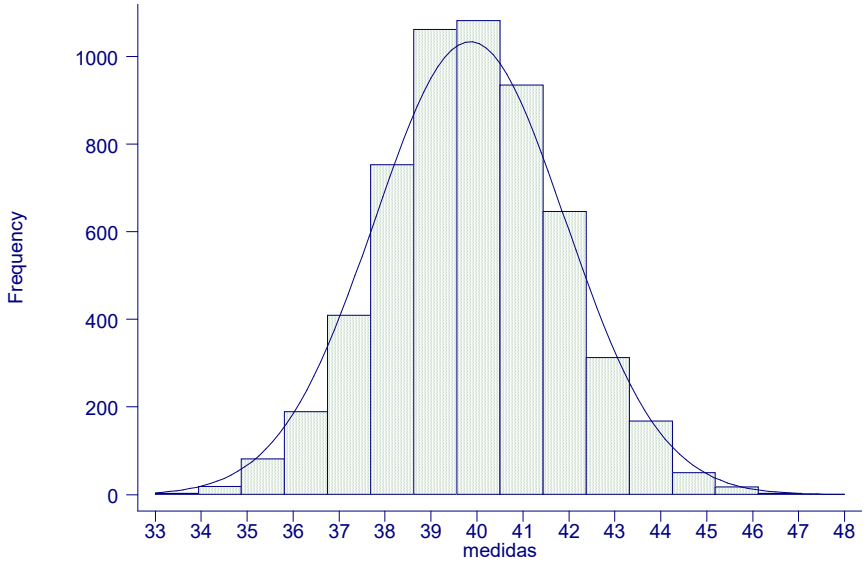
a) O Qui quadrado de Pearson
b) O coeficiente de associação de Yule
c) Você diria que as variáveis estão associadas?

Distribuição normal ou de Gauss; distribuição amostral da média

Os dados abaixo são medidas do tórax (polegadas) de 5732 soldados escoceses, tomadas pelo matemático belga, Adolphe Quetelet (1796-1874).

medidas	Freq,	Percent	Cum,
33	3	0,05	0,05
34	19	0,33	0,38
35	81	1,41	1,80
36	189	3,30	5,09
37	409	7,14	12,23
38	753	13,14	25,37
39	1062	18,53	43,89
40	1082	18,88	62,77
41	935	16,31	79,08
42	646	11,27	90,35
43	313	5,46	95,81
44	168	2,93	98,74
45	50	0,87	99,62
46	18	0,31	99,93
47	3	0,05	99,98
48	1	0,02	100,00
Total	5732	100,00	

Distribuição de medidas do tórax (polegadas) de soldados escoceses.



Fonte: Daly F et al. Elements of Statistics, 1999.

Função densidade de probabilidade da distribuição normal: Se a variável aleatória X é normalmente distribuída com média μ e desvio padrão σ (variância σ^2), a função densidade de probabilidade de X é dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{[-\frac{(x-\mu)^2}{2\sigma^2}]},$$

$-\infty < x < +\infty$;

onde

π : constante $\cong$ 3,1416; e: constante $\cong$ 2,718

μ : constante (média aritmética da população)

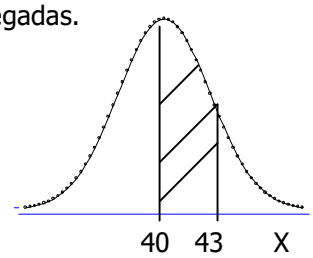
σ : constante (desvio padrão populacional)

Propriedades:

- campo de variação : $-\infty < X < +\infty$;
- é simétrica em torno da média m (ou μ);
- a média e a mediana são coincidentes;
- a área total sob a curva é igual a 1 ou 100%;
- a área sob a curva pode ser entendida como medida de probabilidade.

$$\begin{cases} \mu \pm 1\sigma \ \ inclui \ \ 68{,}2\% \ das \ observações \\ \mu \pm 1{,}96\sigma \ inclui \ \ 95{,}0\% \ das \ observações \\ \mu \pm 2{,}58\sigma \ \ inclui \ \ 99{,}0\% \ das \ observações \end{cases}$$

Exemplo:
Depois de tomarmos várias amostras, decidiu-se adotar um modelo para as medidas de perímetro do tórax de uma população de homens adultos com os parâmetros: média (μ) = 40 polegadas e desvio padrão (σ) = 2 polegadas.



Qual a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax entre 40 e 43 polegadas?

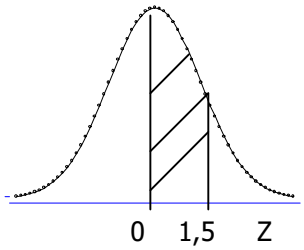
$$P(40 \le X \le 43) = \int_{40}^{43} \frac{1}{2\sqrt{2\pi}} e^{[-\frac{(x-40)^2}{2x4}]} dx$$

Quantos desvio padrão 43 está em torno da média?

Normal reduzida:

$$Z \sim N(0;1) \quad \text{onde } Z = \frac{x - \mu}{\sigma}$$

$$P(40 \le X \le 43) = P(\frac{40-40}{2} \le \frac{X-\mu}{\sigma} \le \frac{43-40}{2}) = P(0 \le Z \le 1{,}5)$$



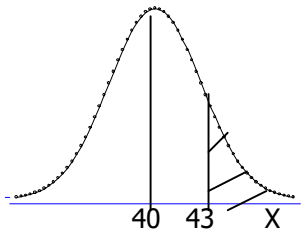
Utilizando a tabela da curva normal reduzida,

$P(0 \le Z \le 1{,}5)$ = 0,43319 =43,3%

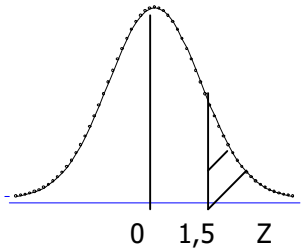
Exemplo

Com base na distribuição de X~N(μ =40, σ =2), calcular:

a) a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax maior ou igual a 43 polegadas.



$$P(X \ge 43) = P(\frac{X-\mu}{\sigma} \ge \frac{43-40}{2}) = P(Z \ge 1{,}5)$$



Utilizando a tabela da curva normal reduzida,

$P(Z \ge 1{,}5)$ = 0,5-0,43319=0,06681= 6,7%.

b) a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax entre 35 e 40 polegadas.

c) a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax menor que 35.

d) Qual o valor do perímetro do tórax, que seria ultrapassado por 25% da população?

Exercício

Considere que fêmeas de *Anopheles darlingi*, criadas em laboratório apresentam peso seco (mg) com média 0,20 mg e desvio padrão 0,06 mg. Sorteia-se um exemplar; qual a probabilidade de que ele tenha

a) Peso seco entre 0,18 e 0,21 mg?
b) Peso seco maior que 0,27 mg
c) Peso seco maior que 0,20 mg
d) Peso seco menor que 0,15 mg
e) Calcule o valor do peso seco que deixaria 25% da população de fêmeas abaixo dele.
f) Calcule o valor do peso seco que deixaria 25% da população de fêmeas acima dele.

Exercício 14

Considere que fêmeas de *Aedes triseriatus*, apresentam comprimento da asa (mm) média 3,25 mm e desvio padrão 0,30 mm. Sorteia-se um exemplar; qual a probabilidade de que ele tenha

a) Comprimento da asa entre 3,0 e 3,50 mm?
b) Comprimento da asa maior que 3,75 mg
c) Comprimento da asa maior que 3,00 mm
d) Comprimento da asa menor que 3,55 mm
e) Calcule o valor do comprimento da asa que deixaria 5% da população de fêmeas abaixo dele.
f) Calcule o valor do comprimento da asa que deixaria 95% da população de fêmeas abaixo dele.

Distribuição amostral da média

Supor a situação onde uma população é composta por 6 elementos, para os quais observou-se a característica X, cujos valores estão apresentados abaixo.

elementos	Xi
A	11
B	16
C	12
D	15
E	16
F	14

Fonte: Dixon WJ e Massey FJ. Introduction to Statistical Analysis. 2nd edit. The Maple Press Company, York, 1957.

Média populacional (μ) = 14;

Variância populacional (σ^2) = 3,667;

Desvio padrão populacional (σ) = 1,9149.

Parâmetros População	valor	Estimador amostra	Valor (estimativa) Par(A,D)=(11,15)
Média (μ)	14	$\bar{x}$	13
Variância (σ^2)	3,67	S^2	8
Desvio padrão (σ)	1,91	S	2,828

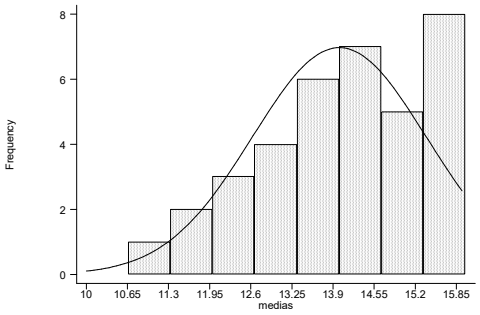
Todas as possíveis amostras de tamanho 2, determinadas pelo processo de amostragem aleatório, com reposição (N=6, n=2):

Amostra	Elementos que compõem a amostra	valores	Média($\bar{x}_i$)
1	A,A	(11,11)	11
2	A,B	(11,16)	13,5
3	A,C	(11,12)	11,5
4	A,D	(11,15)	13
5	A,E	(11,16)	13,5
6	A,F	(11,14)	12,5
7	B,A	(16,11)	13,5
8	B,B	(16,16)	16
9	B,C	(16,12)	14
10	B,D	(16,15)	15,5
11	B,E	(16,16)	16
12	B,F	(16,14)	15
13	C,A	(12,11)	11,5
14	CB	(12,16)	14
15	CC	(12,12)	12
16	C,D	(12,15)	13,5
17	C,E	(12,16)	14
18	C,F	(12,14)	13
19	D,A	(15,11)	13
20	D,B	(15,16)	15,5
21	D,C	(15,12)	13,5
22	D,D	(15,15)	15
23	D,E	(15,16)	15,5
24	D,F	(15,14)	14,5
25	E,A	(16,11)	13,5
26	E,B	(16,16)	16
27	E,C	(16,12)	14
28	E,D	(16,15)	15,5
29	E,E	(16,16)	16
30	E,F	(16,14)	15
31	F,A	(14,11)	12,5
32	F,B	(14,16)	15
33	F,C	(14,12)	13
34	F,D	(14,15)	14,5
35	F,E	(14,16)	15
36	F,F	(14,14)	14

Distribuição de frequência de todas as possíveis médias:

Distribuição amostral da média

i	$\bar{x}_i$	frequência
1	11	1
2	11,5	2
3	12	1
4	12,5	2
5	13	4
6	13,5	6
7	14	5
8	14,5	2
9	15	5
10	15,5	4
11	16	4
Total		36



Média das médias $(\bar{\bar{x}}) = \dfrac{\sum_{i=1}^{11} \bar{x}_i f_i}{n} = 14$

Variância das médias $\sigma^2_{\bar{x}} = \dfrac{\sum_{i=1}^{11} (\bar{x}_i - \bar{\bar{x}})^2 f_i}{n} = 1{,}833$;

Desvio padrão das médias = erro padrão da média = $\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}}$;

Erro padrão da média = $\sqrt{1{,}833} = 1{,}354$.

Teorema central do limite: X é variável aleatória com média μ e variância σ^2, então

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

No exemplo, $X \sim N(\mu = 14, \sigma = 1{,}915)$, portanto $\bar{X} \sim N(\mu_{\bar{x}} = 14, \sigma_{\bar{x}} = \dfrac{1{,}915}{\sqrt{2}} = 1{,}354)$.

Exemplo

Os valores de ácido úrico em homens adultos sadios seguem distribuição aproximadamente Normal com média 5,7mg% e desvio padrão 1mg%. Encontre a probabilidade de que uma amostra aleatória de tamanho 9, sorteada desta população, tenha média

a) maior do que 6 mg%.
b) menor do que 5,2 mg%.

X~N(μ =5,7; $\sigma = 1$)

a) $P(\bar{X} \geq 6) = P(Z_{\bar{X}} \geq \dfrac{6 - 5{,}7}{\frac{1}{\sqrt{9}}}) = P(Z_{\bar{X}} \geq 0{,}91) = 0{,}5 - 0{,}31859 = 0{,}18141.$

b) $P(\overline{X} \leq 5{,}2) = P(Z_{\overline{X}} \leq \dfrac{5{,}2 - 5{,}7}{\dfrac{1}{\sqrt{9}}}) = P(Z_{\overline{X}} \leq -1{,}52) = 0{,}5 - 0{,}43574 = 0{,}064$.

Exercício

Considere uma amostra de 25 fêmeas de *Anopheles darlingi*, capturadas na floresta Amazônica. Sabe-se que a população desta espécie apresenta peso seco (mg) com média 0,20 mg e desvio padrão 0,06 mg. Calcule a probabilidade de que a amostra apresente

a) Peso seco médio maior que 0,22 mg
b) Peso seco médio maior que 0,19 mg
c) Peso seco médio menor que 0,195 mg
d) Peso seco médio entre 0,18 e 0,21 mg?

Exercício 15

Considere uma amostra de 9 fêmeas de *Aedes triseriatus*, capturadas em ambiente silvestre nos Estados Unidos. A população desta espécie é descrita na literatura como apresentando comprimento da asa (mm) com média 3,25 mm e desvio padrão 0,30 mm. Calcule a probabilidade de que a amostra apresente comprimento médio

a) Entre 3,0 e 3,3 mm
b) Maior que 3,35
c) Maior que 3,28
d) Menor que 3,20
e) Entre 3,0 e 3,3

Estatística inferencial

Estimação de parâmetros populacionais

Estimação por ponto

X é uma característica que na população possui distribuição normal com média μ e variância σ^2 (desvio padrão σ).

Seja X_1, X_2, X_3, ...X_n uma amostra aleatória de tamanho **n** extraída desta população.

Os parâmetros μ e σ^2 podem ser estimados com base na amostra.

Se o estimador for um único valor, a estimação é chamada de estimação por ponto.
Se o estimador for um conjunto de valores, a estimação é chamada de estimação por intervalo.

Média aritmética

Populacional Parâmetro μ estimador : $\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$

Variância

Populacional Parâmetro σ^2 estimador : $S^2_{(N)} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N}$ *ou*

$$S^2_{(N-1)} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N-1}$$

Atenção: Antes dos dados serem coletados, **os estimadores são variáveis aleatórias**.

Estimação por intervalo

Intervalo de confiança: É um conjunto de valores calculados com base na amostra. Pressupõe-se que cubra o parâmetro de interesse com um certo grau (nível) de confiança.

O grau de confiança tem origem na probabilidade associada ao processo de construção do intervalo antes de se obter o resultado amostral.

O grau de confiança mais comumente utilizado é o de 95%.

Seria impossível construir um intervalo de 100% de confiança a menos que se medisse toda a população.

Na maioria das aplicações não sabemos se um intervalo de confiança específico cobre o verdadeiro valor. Só podemos aplicar o conceito frequentista de probabilidade e dizer que se realizarmos a amostragem infinitas vezes e construirmos intervalos de confiança de 95%, em 95% das vezes os intervalos de confiança estarão corretos (cobrirão o parâmetro) e 5% das vezes estarão errados.
Exemplos de intervalo de confiança:

IMC médio, desvio padrão (dp) e IC de 95% segundo sexo e idade (anos). Duas escolas públicas de São Paulo, 2004.

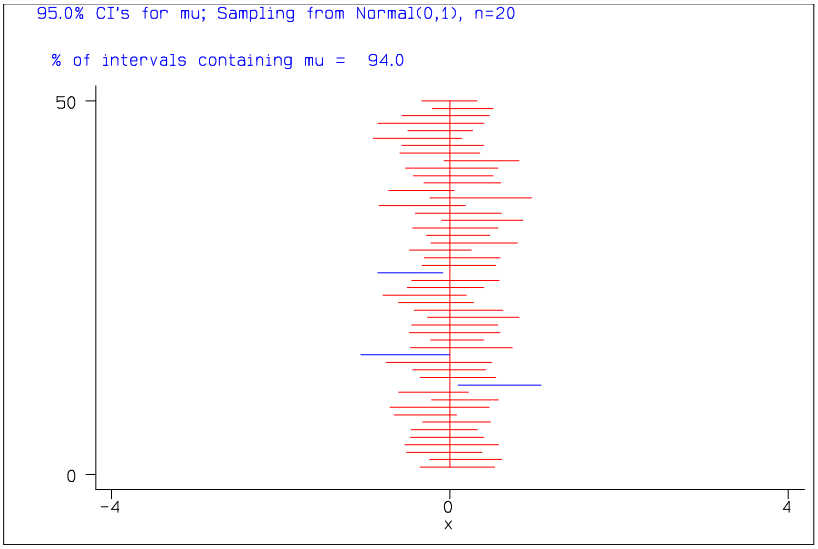
Sexo[1]	Idade (anos)[2]			
	7	8	9	10
	IMC (kg/m^2) médio e desvio padrão (dp) (IC 95%)			
Masculino	16,8 (2,5) (16,2 – 17,4)	17,9 (4,0) (17,0 – 18,9)	17,3 (3,1) (16,5 – 18,1)	18,9 (4,0) (17,9 – 19,8)
Feminino	16,4 (2,30) (15,9 – 17,0)	16,9 (2,9) (16,2 – 17,6)	17,4 (3,3) (16,6 – 18,2)	18,7 (3,1) (17,9 – 19,5)
Total	16,6 (2,4) (16,2 – 17,0)	17,4 (3,5) (16,8 – 18,0)	18,7 (3,2) (17,9 – 19,5)	18,8 (3,7) (18,2 – 19,4)

(1) Masculino (n=281), Feminino (n=275);
(2) 7 anos (n=151); 8 anos (n=138); 9 anos (n=126); 10 anos (n=141)
Fonte: Claudia Regina Koga. Dissertação de Mestrado (dados preliminares)

IC para a proporção populacional

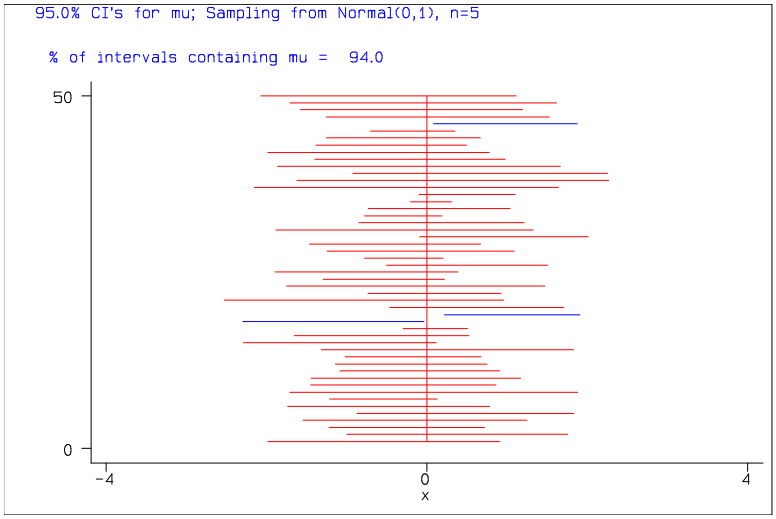
"Os dados de composição corporal obtidos pela utilização da BIA, classificados em duas categorias: sem risco de doença cardiovascular e com risco de DCV, resultaram em prevalência de risco de DCV igual a 42,3% (IC95%: 38,1 - 46,5%)."

Representação gráfica



A linha vertical representa o parâmetro populacional. O gráfico foi gerado via programa de computador. São apresentados 50 intervalos de confiança para amostras de tamanho n=20. As linhas horizontais representam os intervalos de confiança. Se o intervalo de confiança não contiver o parâmetro, a linha horizontal não cruzará a linha vertical. A linha vertical é o parâmetro. No exemplo, 3 intervalos não cobrem ("capturam") o parâmetro.

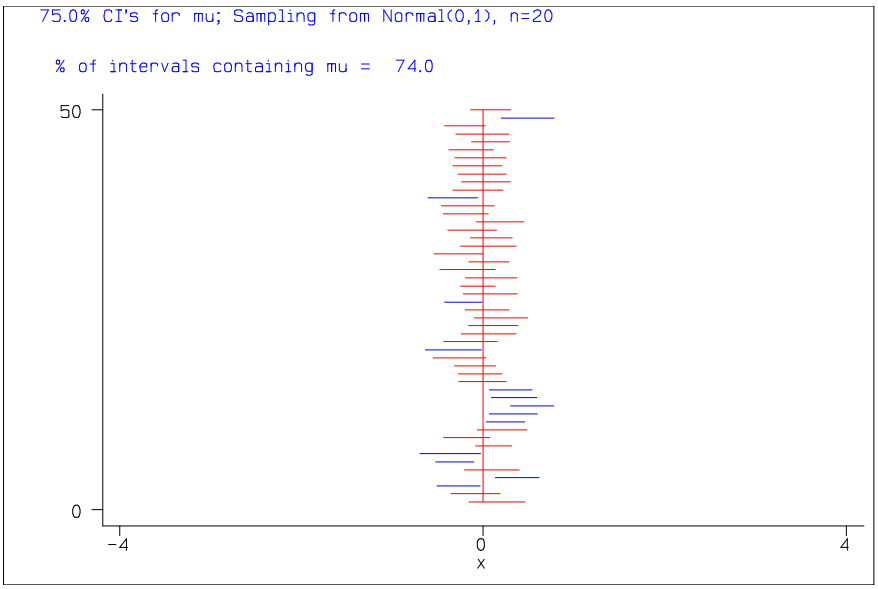
Apresentação gráfica do efeito do tamanho da amostra:



Para amostras menores (n=5), as larguras dos intervalos são maiores a proporção de intervalos que "capturam" o parâmetro é parecida com a anterior (para n=20). Portanto, o tamanho da amostra não interfere na proporção de "captura" do parâmetro mas sim na precisão do estimador.

Efeito do grau de confiança

Para n=20 e α =0,25, obtém-se intervalos com os apresentados a seguir:



Os intervalos são mais estreitos do que para n=20 e α =0,05. Uma porcentagem bem maior não contém o parâmetro. Isto é o que 75% de confiança significa. Do total de todas as possíveis amostras, 75% delas resultará em intervalos de confiança que contêm o verdadeiro valor do parâmetro.

Interpretando Intervalos de Confiança (IC)

Um intervalo de confiança para um parâmetro é um intervalo de valores no qual pode-se depositar uma confiança que o intervalo cobre (contém) o valor do parâmetro. Por exemplo, se com base em uma amostra encontrarmos que o intervalo (3200 ; 3550) é um intervalo de 95% de confiança para a média (μ) da população de valores do peso médio ao nascer de recém-nascidos no Município de São Paulo, então podemos estar 95% confiantes que o conjunto de valores 3220 – 3500 gramas cobre (contém) o verdadeiro peso médio ao nascer da população.

Pode-se também pensar no IC a partir da seleção de milhares de amostras de uma população. Para cada amostra calcula-se um intervalo de confiança com grau de confiança 100(1- α)%, para um parâmetro da população. A porcentagem de intervalos que contém o verdadeiro valor do parâmetro é 100(1- α). Para α =0,05, o grau de confiança será igual a 100(1-0,05)% = 100(0,95)% = 95%.

Na prática, tomamos somente uma amostra e obtemos somente um intervalo. Mas sabemos que 100(1- α)% de todas as amostras tem um intervalo de confiança contendo o verdadeiro valor do parâmetro, portanto depositamos uma confiança 100(1- α)% que o particular intervalo contém o verdadeiro valor do parâmetro.

Amplitude do intervalo

Para um grau de confiança especificado (por exemplo, 95%), desejamos o intervalo tão pequeno quanto possível.

Ex: o intervalo de confiança de 95% para o peso médio ao nascer (gramas) de recém-nascidos no Município de São Paulo de (2500, 4000) traz pouca informação prática porque sabe-se, da experiência, que a média populacional está neste intervalo. Deseja-se um intervalo com amplitude de poucas gramas. É o tamanho da amostra que determina a amplitude do intervalo. Quanto maior a amostra, menor será o intervalo.

Fórmulas para construção dos intervalos de confiança:

As fórmulas dos intervalos de confiança são derivadas da distribuição amostral da estatística;

Construção do intervalo de confiança para a média populacional μ;

Pressuposição: A amostra deve ser obtida de forma aleatória;

É necessário utilizar as propriedades do teorema central do limite :

$$X \sim N(\mu, \sigma)\,;\ \overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Padronizando-se a média $\overline{X}$, obtém-se $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, que permite calcular

$$P(-z \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z) = 1 - \alpha \cdot$$

Para $\alpha = 5\%$, $P(-1{,}96 \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le +1{,}96) = 0{,}95$

$$P(-1{,}96\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le +1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

$$P(-\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Multiplicando tudo por -1

$$P(\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Reescrevendo a equação tem-se

$$P(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Obtém-se um intervalo aleatório **centrado na média amostral** o qual possui 95% de probabilidade de conter a verdadeira média populacional.

O parâmetro será estimado por um conjunto de valores provenientes de uma amostra. Quando isto é feito, a média é estimada por um determinado valor ($\hat{\overline{X}} = \bar{x}$), e o intervalo $\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}$ deixa de ser uma variável aleatória.

Este intervalo cobre (contém) ou não cobre (não contém) a verdadeira média (parâmetro). Diz-se então que a confiança que se deposita neste intervalo é de 95% porque antes de coletar a amostra de tamanho n, existia, associada a ele, uma probabilidade de 95% de que contivesse a média populacional. Por isso chama-se intervalo de confiança para a média populacional.

$$\text{IC(95\%)} : (\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}};\ \ \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}})$$

Intervalo de confiança para a média populacional com variância populacional conhecida

Pressuposição: A amostra deve ser obtida de forma aleatória.

Estatística: média populacional - μ.

$$\text{IC}(\mu) = \bar{x} - z_{\alpha/2}.\frac{\sigma_x}{\sqrt{n}}; \bar{x} + z_{\alpha/2}.\frac{\sigma_x}{\sqrt{n}}$$

Exemplo:
Construa um intervalo de 95% de confiança para estimar a pressão diastólica média populacional (μ), sabendo que em uma amostra de 36 adultos a pressão média amostral ($\bar{x}$) foi igual a 85mmHg e o desvio padrão populacional (σ) foi 9 mm de Hg. Interprete o significado desse intervalo
Solução:

$85 - 1{,}96\frac{9}{\sqrt{36}}; 85 + 1{,}96\frac{9}{\sqrt{36}}$, ou seja, (82,06; 87,94mmHg)

Exercício
Em uma amostra de 16 gestantes com diagnóstico clínico de pré-eclâmpsia, a taxa média de ácido úrico no plasma foi de 5,3 mg sabendo que a variabilidade na população é igual a 0,6 mg. Estime, com 95% de confiança, a taxa média de ácido úrico no plasma da população de gestantes com diagnóstico de pré-eclâmpsia.

Intervalo de confiança para a média populacional com variância populacional desconhecida

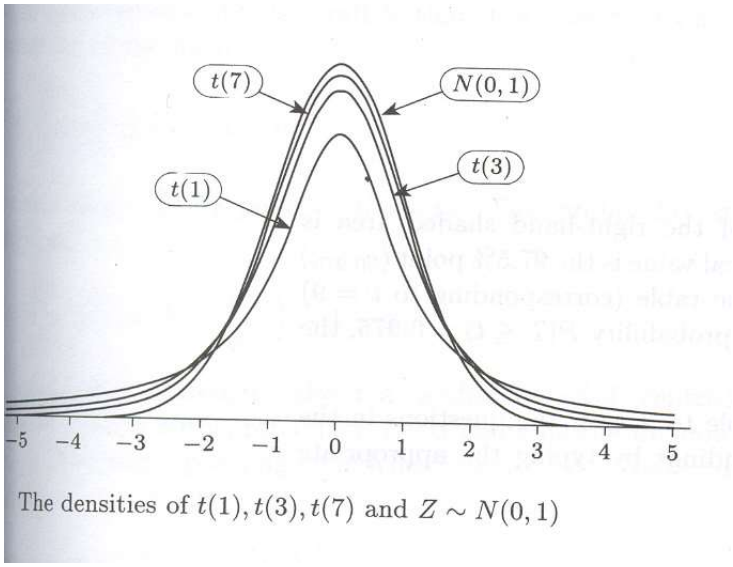
$$\mathrm{IC}(\mu): \bar{x} - t_{n-1,\alpha/2} \cdot \frac{S_x}{\sqrt{n}} ; \bar{x} + t_{n-1,\alpha/2} \cdot \frac{S_x}{\sqrt{n}}$$

Distribuição t de Student

A família de distribuições t de Student

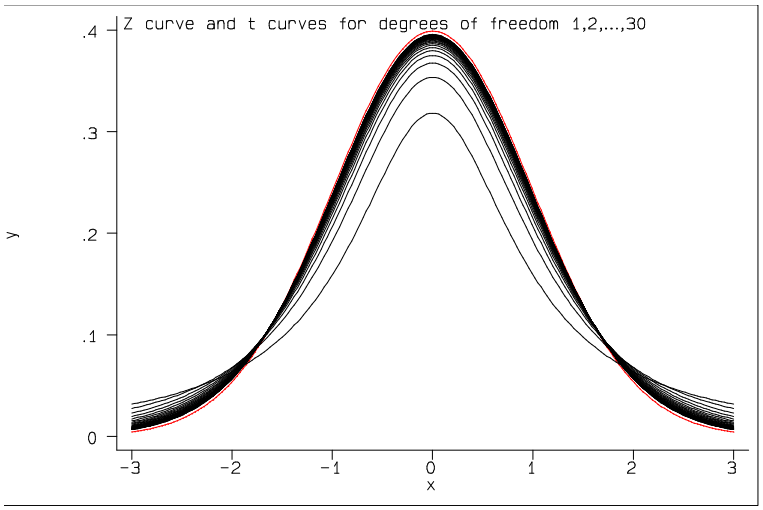
Student é o pseudônimo de W. S. Gosset que, em 1908, propôs a distribuição t. Esta distribuição é muito parecida com a distribuição normal. A família de distribuições t é centrada no zero e possui formato em sino. A curva não é tão alta quanto a curva da distribuição normal e as caudas da distribuição t são mais altas que as da distribuição normal. O parâmetro que determina a altura e largura da distribuição t depende do tamanho da amostra (n) e é denominado graus de liberdade (gl), denotado pela letra grega (ν) (lê-se ni). A notação da distribuição t é t_ν .

Curvas t para graus de liberdade (tamanhos de amostra) diferentes.



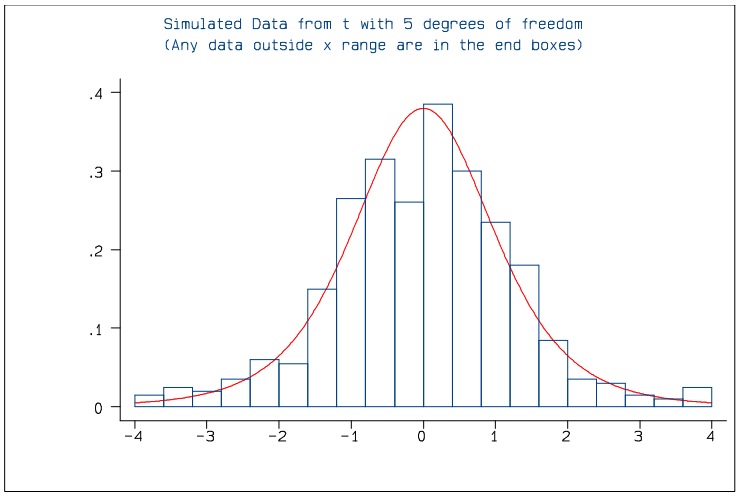
The densities of $t(1), t(3), t(7)$ and $Z \sim N(0,1)$

Quando o número de graus de liberdade da distribuição t aumenta, a distribuição se aproxima de uma distribuição normal.



Esta família t não descreve o que acontece na natureza mas sim o que aconteceria se selecionássemos milhares de amostras aleatórias de uma população normal com média μ e fosse calculado $t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ para cada amostra.

Calculando o valor de t para 500 amostras de tamanho 6 de uma população com distribuição normal, obtém-se o gráfico a seguir:



Obs: utilização da tabela t de Student

A tabela da distribuição de *Student* apresenta um valor de probabilidade dividido em duas partes iguais. Para n=50, o número de graus de liberdade (gl) é 49; como não existe este valor na tabela, deve-se trabalhar com o número de gl mais próximo e dependendo se o teste é mono ou bicaudal, utiliza-se respectivamente o valor de p/2 ou p, apresentados na primeira linha da tabela.

Exemplo de utilização da tabela t de Student:

n=10; teste **bicaudal**, α=0,05; $t_{crítico}$=-2,262 e $t_{crítico}$= 2,262 (p da tabela =0,05)
n=10; teste **monocaudal** a esquerda, α=0,05; $t_{crítico}$=-1,833 (p da tabela = 0,10)
n=10; teste **monocaudal** a direita, α=0,05; $t_{crítico}$= 1,833 (p da tabela = 0,10)

Exemplo:
Construa um intervalo de 95% de confiança para estimar a pressão diastólica média populacional (μ), sabendo que em uma amostra de 36 adultos a pressão média amostral ($\bar{x}$) foi igual a 85mmHg e o desvio padrão amostral (s) foi 12 mm Hg. Interprete o significado desse intervalo.

$$85 - 2{,}03\frac{12}{\sqrt{36}};85 + 2{,}03\frac{12}{\sqrt{36}}$$, ou seja, (80,94; 89,06 mmHg)

Exercício
Uma amostra de 25 adolescentes meninos apresenta peso médio de 56 kg e desvio padrão 8

kg.

- **a)** encontre o intervalo de confiança de 95% para o peso médio da população da qual esta amostra foi sorteada.
- **b)** interprete o intervalo de confiança encontrado.

Exercício 16
São apresentadas medidas de pressão arterial sistólica de uma amostra de 20 pacientes.

a) Construa o intervalo de confiança de 90% para a pressão sistólica média populacional.
b) Interprete o intervalo de confiança encontrado.

98	136	128	130	114	123	134	128	107	123
160	125	129	132	154	115	126	132	136	130

Valores de média e desvio padrão das observações:

Média ($\bar{x}$)	128
Desvio padrão (S_{n-1})	13,91

Exercício 17

O nível médio de protrombina em populações normais é 20 mg/100ml de sangue. Uma amostra de 40 pacientes que tinham deficiência de vitamina K tiveram nível médio observado de protrombina de 18,5mg/100ml e desvio padrão 4mg/100ml. Seria razoável concluir que a verdadeira média de pacientes com deficiência de vitamina K é a mesma que a da população normal?

Resumo: Intervalo de Confiança

Média populacional: μ

Com variância conhecida σ^2 : $\bar{x} - Z_{\alpha/2}\frac{\sigma^2}{\sqrt{n}}; \bar{x} + Z_{\alpha/2}\frac{\sigma^2}{\sqrt{n}}$

Com variância σ^2 desconhecida: $\bar{x} - t_{\alpha/2,\nu}\frac{s}{\sqrt{n}}$, $\bar{x} + t_{\alpha/2,\nu}\frac{s}{\sqrt{n}}; \nu = n-1$

Teste de hipóteses

Estatística descritiva

Descreve eventos por meio de:

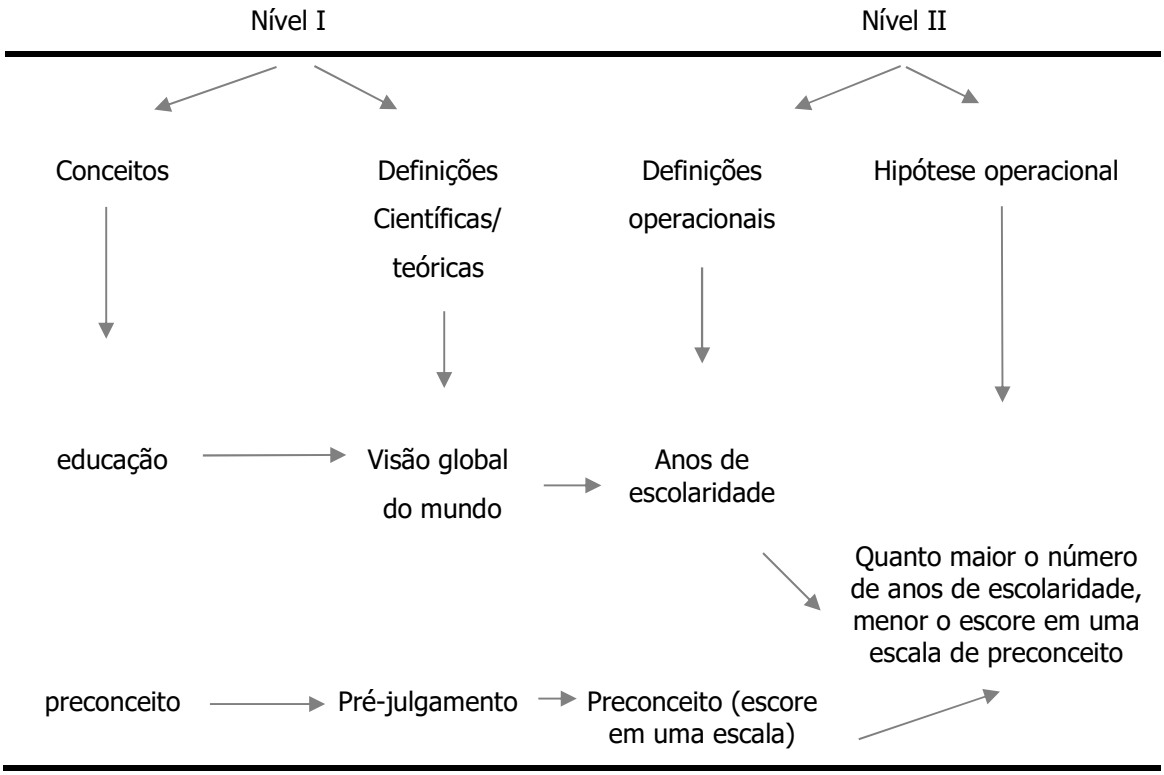
tabelas
gráficos
razões e índices
parâmetros típicos (medidas de posição e dispersão)

Estatística analítica

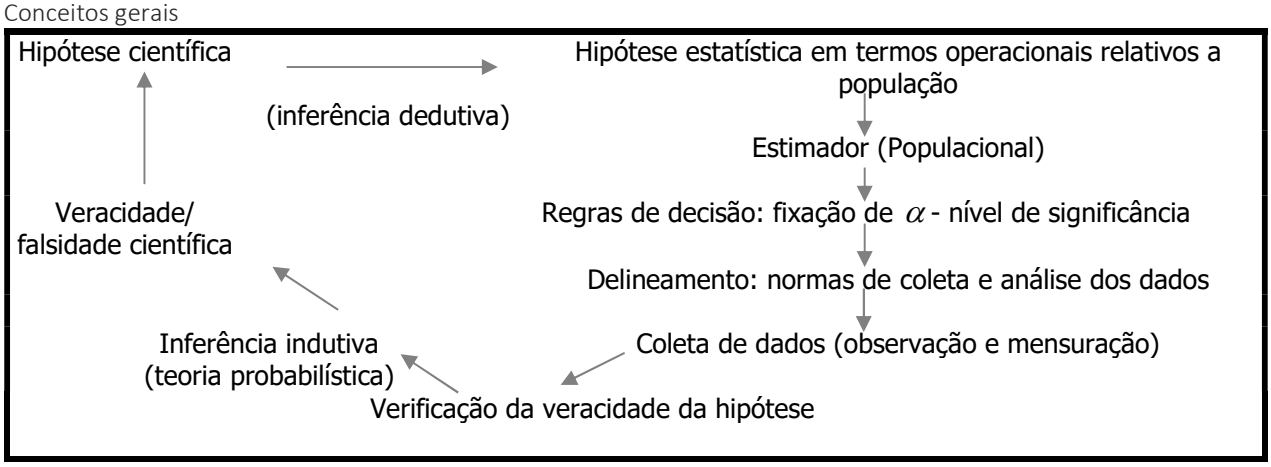
Nível I - Teórico (conceitos, hipóteses científicas)
Nível II - operacional (hipótese estatística)

Situação

Quanto mais bem educada uma pessoa, menor o seu preconceito em aceitar certa campanha sanitária



Conceitos gerais



Inferência estatística: É qualquer procedimento que se utiliza para se generalizar afirmações sobre determinada população, baseadas em dados retirados de uma amostra.

Parâmetro: É a medida usada para se descrever uma característica de uma população.

Estatística: É uma função dos valores amostrais.

Estimação: É o processo através do qual estima-se o valor de um parâmetro de uma população com base no valor obtido em uma amostra.

Hipótese: É uma forma de especulação relativa a um fenômeno estudado (qualquer que seja). É qualquer afirmação sobre a distribuição de probabilidade de uma variável aleatória (afirmação sobre um parâmetro).

Hipótese estatística: É uma especulação feita em relação a uma proposição, porém relativa a uma população definida.

Teste de hipóteses de uma média populacional (μ) com variância conhecida

Proposta clássica de Neyman e Pearson

Situação de interesse
Tomando-se como exemplo os dados de recém-nascidos com Síndrome de Desconforto Idiopático Grave (SDIG), é possível elaborar a hipótese de que crianças que nascem com esta síndrome possuem peso médio ao nascer menor do que o peso médio ao nascer de crianças sadias.

A variável de estudo X é peso ao nascer (quantitativa contínua).

Com base em conhecimento prévio (da literatura) sabe-se que a distribuição do peso ao nascer em crianças sadias segue uma distribuição normal com média 3000 gramas e desvio padrão 500 gramas, ou seja, $X \sim N(\mu_X = 3000; \sigma_X = 500)$.
Recordando-se, para a realização do teste de hipóteses segundo Neyman e Pearson é necessário:

- Formular as hipóteses estatísticas;
- Fixar a probabilidade do erro tipo I;
- Calcular o tamanho da amostra necessária para detectar uma diferença que se suspeita existente o que é equivalente a fixar a probabilidade do erro tipo II;
- Apresentar a distribuição de probabilidade da estatística do teste;
- Estabelecer a(s) região(ões) de rejeição e aceitação (regiões críticas) do teste;
- Realizar o estudo, ou seja, coletar os dados e calcular a estatística do teste;
- Confrontar a estatística do teste observada com a região crítica;
- Tomar a decisão;
- Elaborar a conclusão.

Formulação das hipóteses

$H_0 : \mu_{SDIG} = \mu_{Sadia}$ ou $H_0 : \mu_{SDIG} = 3000$

$H_a : \mu_{SDIG} < \mu_{Sadia}$ ou $H_a : \mu_{SDIG} < 3000$

Possíveis erros na tomada da decisão:

Decisão	Verdade	
	H_0	H_a
H_0	não cometeu erro	*erro* tipo II
há	erro tipo I	não cometeu erro

$\alpha = \Pr obabilidade(erro\ tipoI)$ = Probabilidade (Rejeitar H_0 e H_0 é verdade)

$\beta = \Pr obabilidade(erro\ tipoII)$ = Probabilidade (Aceitar H_0 e H_0 é falsa)

$(1-\beta)$ = poder do teste = Probabilidade (Rejeitar H_0 e H_0 é falsa)

Poder de revelar a falsidade de H_0 quando a verdade é H_a

Conduta: Antes do experimento, fixa-se α e trabalha-se com o menor β possível.

Na situação de estudo, fixando-se o nível de significância $\alpha = 0{,}05$

Supor um tamanho de amostra n=50 recém-nascidos com SDIG

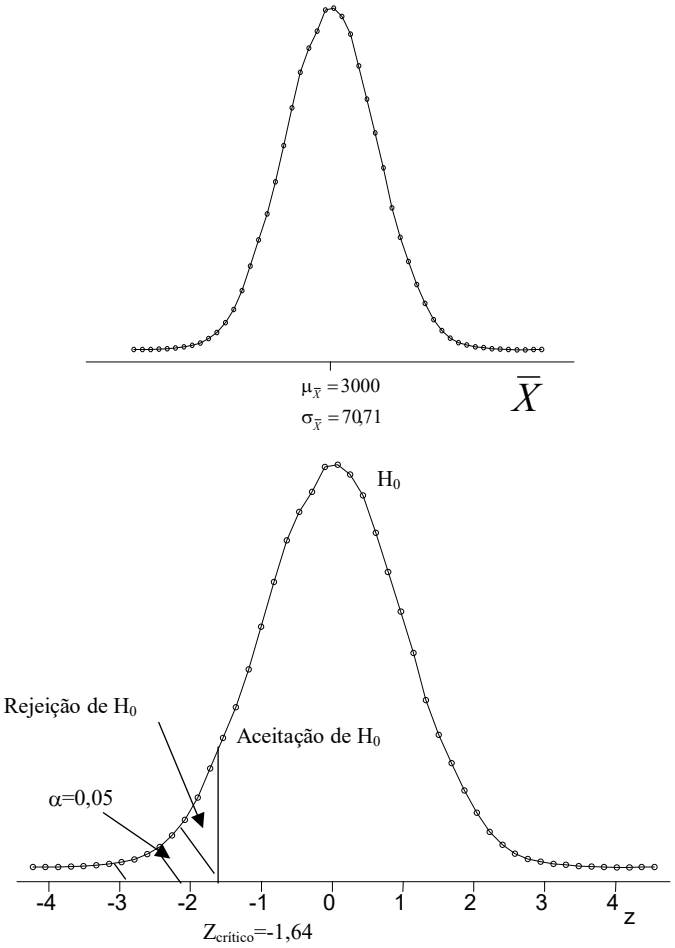
Distribuição de probabilidade

Como as hipóteses envolvem a média populacional, é necessário utilizar a distribuição de probabilidade da média.

Pelo Teorema Central do Limite tem-se que $\overline{X} \sim N(\mu_{\overline{X}} = \mu_X ; \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}})$, portanto, se H_0 for verdade, e admitindo-se que as crianças com SDIG possuem distribuição do peso ao nascer com mesma dispersão que as crianças sadias, pode-se afirmar (em H_0) que a distribuição do peso médio de crianças com a síndrome é $\overline{X} \sim N(\mu_{\overline{X}} = 3000; \sigma_{\overline{X}} = \frac{500}{\sqrt{50}})$.

Pode-se utilizar $Z_{\overline{X}}$ ou $\overline{x}_{obs}$ para a tomada de decisão.

Região de rejeição e aceitação da hipótese H_0.



Cálculo do peso médio na amostra de crianças com SDIG.

Supor que na amostra de 50 crianças, foi observado peso médio ao nascer igual a 2800 gramas $(\bar{x}_{obs} = 2800)$.

Cálculo do peso médio observado em número de desvios padrão:

$$Z_{\bar{X}obs} = \frac{\bar{x}_{obs} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{2800 - 3000}{70{,}71} = -2{,}83$$

Confrontar o valor da estatística do teste com a região de rejeição e aceitação de H_0.

Como Z_{obs} está à esquerda de $Z_{crítico}$ (região de rejeição), decide-se por rejeitar H_0.

Decisão
Rejeita-se H_0.

Conclusão
Foi encontrada diferença estatisticamente significante entre os pesos ao nascer de crianças sadias e com SDIG para nível de significância $\alpha = 0{,}05$. Crianças com SDIG nascem com peso menor do que crianças sadias.

Regra geral:

Rejeita-se H_0 se

$Z_{obs}>Z_{crítico}$ para $H_a : \mu_{SDIG} > \mu_{Sadias}$

$Z_{obs}<-Z_{crítico}$ para $H_a : \mu_{SDIG} < \mu_{Sadias}$

$Z_{obs}>Z_{crítico}$ ou $Z_{obs}<-Z_{crítico}$ para $H_a : \mu_{SDIG} \neq \mu_{Sadias}$

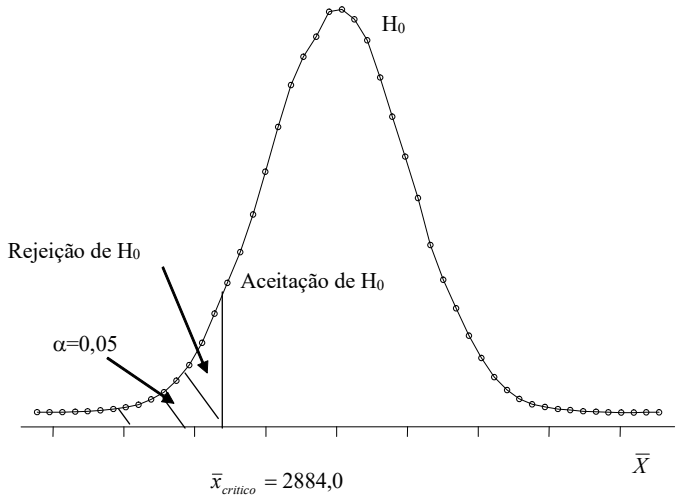
Utilizando-se a estatística $\overline{X}$ para a tomada de decisão.

É possível realizar o teste comparando a média observada na amostra $(\overline{x}_{obs} = 2800)$ e o valor de peso médio ao nascer que deixa, no caso deste exemplo, uma área α=0,05 à sua esquerda. O valor de peso médio que limita esta área é denominado $\overline{x}_{critico}$.

Pode-se calcular o valor de $\overline{x}_{critico}$ que limita a área de 5% utilizando a estatística

$$Z = \frac{\overline{x} - \mu_{\overline{x}}}{\frac{\sigma_X}{\sqrt{n}}}$$

Substituindo-se os valores, $-1{,}64 = \frac{\overline{x} - 3000}{70{,}71}$, então $\overline{x} = -1{,}64x70{,}71 + 3000$ =2884,0



Tomando-se a decisão

Comparando-se $\bar{x}_{obs} = 2800$ e $\bar{x}_{critico} = 2884{,}0$ pode-se ver que $\bar{x}_{obs} < \bar{x}_{critico}$ indicando que a hipótese H_0 deve ser rejeitada.

Regra geral para a tomada de decisão utilizando a estatística $\bar{x}_{observada}$:

Rejeita-se H_0 se

$\bar{X}_{obs} > \bar{X}_{critico}$ para $H_a : \mu_{SDIG} > \mu_{Sadias}$

$\bar{X}_{obs} < -\bar{X}_{critico}$ para $H_a : \mu_{SDIG} < \mu_{Sadias}$

$\bar{X}_{obs} < -\bar{X}_{critico}$ *ou* $\bar{X}_{obs} > \bar{X}_{critico}$ para $H_a : \mu_{SDIG} \neq \mu_{Sadias}$

Cálculo do tamanho mínimo da amostra

Para uma hipótese monocaudal, onde $\begin{cases} H_0 : \mu_{SDIG} = 3000 \\ H_a : \mu_{SDIG} < 3000 \end{cases}$

$n = \dfrac{(Z_\alpha + Z_\beta)^2}{d^2}$, em que

Z_α é o valor de Z que deixa α à direita

Z_β é o valor de Z que deixa β à direita

$$d = \frac{|\mu_{SDIG} - 3000|}{500}$$

Supondo que a média populacional para recém-nascidos com a síndrome seja igual a 2900,

$$d = \frac{|2900 - 3000|}{500} = 0{,}2$$

Pela tabela da N(0,1) tem-se que para $\alpha = 0{,}05$, Z_α =1,64

Pela tabela da N(0,1) tem-se que para $\beta = 0{,}20$, Z_β =0,845

Substituindo-se os valores, tem-se

$$n = \frac{(Z_\alpha + Z_\beta)^2}{d^2} = \frac{(1{,}64 + 0{,}845)^2}{0{,}2^2} = 154{,}4$$

Portanto, seria necessário obter uma amostra mínima de 155 recém-nascidos com SDIG para localizar uma diferença de 0,2 desvios padrão do valor médio da população sem esta síndrome.

Exercício

a) Altere o valor de $\beta = 0{,}10$ e recalcule o tamanho da amostra.

b) Altere os valores de $\beta = 0{,}10$ e d=0,3 e recalcule o tamanho da amostra

Teste de hipóteses de uma média populacional (μ) (com variância conhecida): abordagem de Fisher

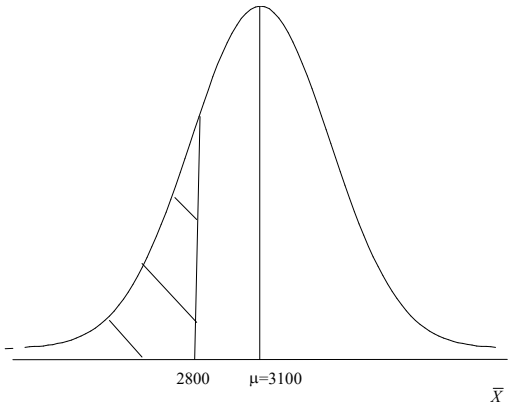
Situação:
Estudos mostram que crianças sadias possuem peso médio (m) ao nascer igual a 3100 gramas e desvio padrão $\sigma = 610 gramas$. Suspeita-se que crianças que nascem com síndrome de desconforto idiopático grave possuem peso ao nascer abaixo do peso ao nascer da população de crianças sadias.

Proposição (equivalente à H_0): Crianças com síndrome vêm de uma população com peso médio =3100 gramas.

Realiza-se um estudo em uma amostra de n=50 crianças que nasceram com esta síndrome, onde observou-se peso médio ($\overline{x}$) igual a 2800 gramas.

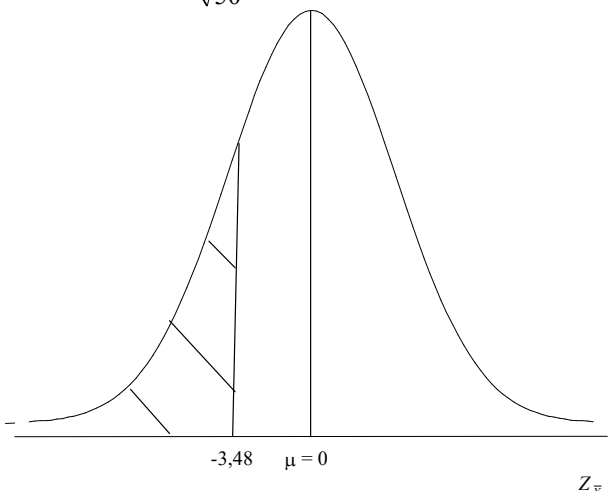
Supondo-se que as crianças da amostra (com síndrome) vêm de uma população com mesma dispersão do peso ao nascer de crianças sadias, teste a hipótese de que crianças com síndrome de desconforto idiopático grave possuem peso médio ao nascer igual ao peso médio ao nascer de crianças sadias.

Distribuição de probabilidade:
Distribuição do peso médio: segue uma distribuição normal com média m=3100 gramas e desvio padrão $\frac{\sigma}{\sqrt{n}} = \frac{610}{\sqrt{50}} = 86{,}27$ gramas



Cálculo da probabilidade de observar um peso médio ao nascer igual ou menor que 2800 se H_0 for verdade.

$$P(\overline{X} \leq 2800) = P(\frac{\overline{X} - m}{\sigma_{\overline{X}}} \leq \frac{2800 - 3100}{\frac{610}{\sqrt{50}}}) = P(Z_{\overline{X}} \leq \frac{-300}{86{,}27}) = P(Z_{\overline{X}} \leq -3{,}48)$$



Pela distribuição Normal reduzida tem-se que $P(Z \leq 3{,}48) = 0{,}5 - 0{,}49975 = 0{,}00025$ ou 0,025%

Os resultados não são compatíveis com uma distribuição que tem peso médio igual a 3100. Possivelmente a amostra vem de uma população com média menor que 3100. Pode-se dizer que crianças com síndrome de desconforto idiopático grave possivelmente possuem peso ao nascer menor do que o peso médio de crianças sadias (p<0,001).

Exercício

Sabe-se que o consumo mensal per capita de um determinado produto tem distribuição normal com desvio padrão $\sigma = 2kg$. A diretoria da indústria que fabrica este produto está desconfiada que a procura pelo produto caiu muito e resolveu tirar este item de produção caso o consumo mensal per capita fosse menor que 8kg (consumo médio). Assim, realizou uma pesquisa com 25 indivíduos e observou um consumo médio mensal igual a 7,2kg. Faça um teste de hipóteses com nível de significância de 5% para auxiliar a diretoria em sua decisão. Tome a decisão também utilizando a estratégia de Fisher calculando o valor de p.

Exercício

O nível médio de protrombina em populações normais é 20 mg/100ml de sangue com desvio padrão $\sigma = 4mg/100ml$. Em uma amostra de 40 pacientes que tinham deficiência de vitamina K foi observado nível médio de protrombina de 18,5mg/100ml. Seria razoável concluir que a verdadeira média de pacientes com deficiência de vitamina K é a mesma que a da população normal? Faça um teste de hipóteses com nível de significância de 5% para responder a pergunta. Utilize também a estratégia de Fisher calculando o valor de p.

Teste de hipóteses para uma média populacional com variância desconhecida

Supor a situação anterior, só que a variância (desvio padrão) populacional do peso ao nascer de crianças sadias é desconhecida sendo conhecido somente o peso médio populacional de crianças sadias (μ_{Sadias} =3000 gramas).

Formulação das hipóteses

$$H_0 : \mu_{SDIG} = 3000$$
$$H_a : \mu_{SDIG} < 3000$$

Fixando-se o nível de significância $\alpha = 0{,}05$

Cálculo do tamanho da amostra: supor um tamanho de amostra n=50 recém-nascidos com SDIG

Distribuição de probabilidade

Como as hipóteses envolvem a média populacional, é necessário utilizar a distribuição de probabilidade da média.

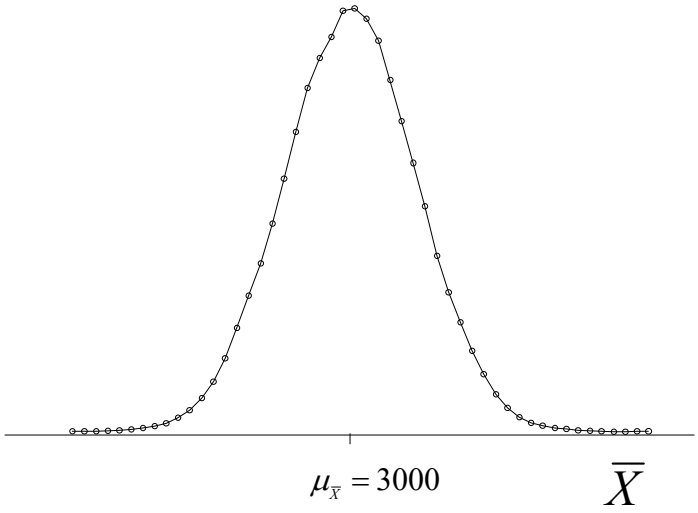
Pelo Teorema Central do Limite tem-se que $\overline{X} \sim N(\mu_{\overline{X}} = \mu_X ; \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}})$.

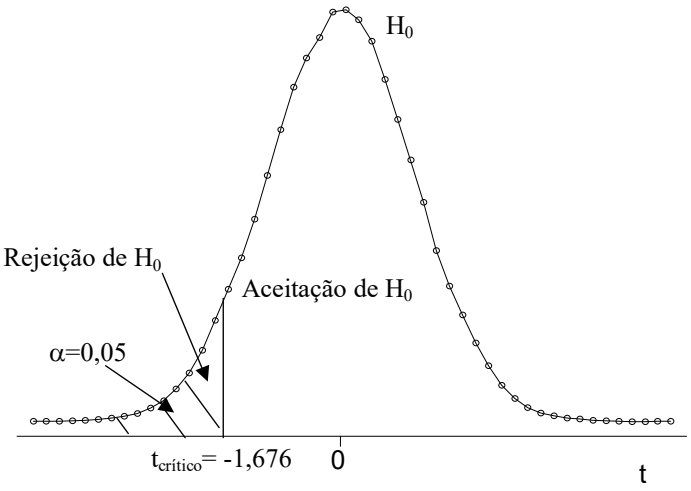
Admitindo-se que H_0 é verdade, resta um problema que é o fato de não se conhecer o valor da dispersão do peso ao nascer das crianças sadias. Neste caso não é possível utilizar a estatística Z.

Utiliza-se, então, a estatística T onde $T = \frac{\overline{X} - \mu_{\overline{X}}}{S_{\overline{X}}} = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{S_X}{\sqrt{n}}}$ sendo S_X o desvio padrão da população de estudo, estimado com os dados da amostra de crianças com SDIG.

T segue uma distribuição *t* de *Student*, com (n-1) graus de liberdade. Quando o tamanho da amostra é grande, a estatística T tende para uma distribuição normal com média 0 e desvio padrão 1 ($n \to \infty \Rightarrow T \sim N(0;1)$).

Região de rejeição e aceitação da hipótese H_0.





Cálculo do peso médio na amostra de crianças com SDIG

Supor que na amostra de 50 crianças, foi observado peso médio ao nascer igual a 2800 gramas e desvio padrão igual a 610g $(\bar{x}_{obs} = 2800; s_{\bar{X}} = 610)$.

Cálculo do peso médio observado em número de desvios:

$$t_{obs} = \frac{\bar{x}_{obs} - \mu_{\bar{X}}}{S_{\bar{X}}} = \frac{2800 - 3000}{\frac{610}{\sqrt{50}}} = -2{,}318$$

Comparação entre o valor da estatística do teste e a região de rejeição e aceitação de H_0

Como t_{obs} está à esquerda de $t_{crítico}$ (região de rejeição), decide-se por rejeitar H_0.

Decisão
Rejeita-se H_0.

Conclusão
Foi encontrada diferença estatisticamente significante entre os pesos ao nascer de crianças sadias e com SDIG para nível de significância $\alpha = 0{,}05$. Crianças com SDIG nascem com peso menor do que crianças sadias.

Exercício
Uma companhia de produtos alimentícios utiliza uma máquina para embalar salgadinhos cujas embalagens especificam 454gramas. Com o propósito de verificar se a máquina está trabalhando corretamente, selecionou-se 50 pacotes de salgadinhos, obtendo-se os seguintes valores de peso:

464	450	450	456	452	433	446	446	450	447
442	438	452	447	460	450	453	456	446	433
448	450	439	452	459	454	456	454	452	449
463	449	447	466	446	447	450	449	457	464
468	447	433	464	469	457	454	451	453	443

média da amostra, $\bar{x}$ =451,22 gramas e s=8,40 gramas

Testar a hipótese de que a máquina está trabalhando corretamente, para α = 0,05.

Exercício

Deseja-se saber se o consumo calórico médio de determinada população adulta de zona rural é menor que 2000. Uma amostra de 500 pessoas apresentou consumo médio igual a 1985 e desvio padrão igual a 210. Faça um teste de hipóteses para tomar a decisão; considere o nível de significância igual a 5%.

Teste de hipóteses para uma média populacional com variância desconhecida - Abordagem de Fisher

Supor a mesma situação anterior, só que neste caso somente a média populacional é conhecida. O peso médio de crianças sadias (μ) é igual a 3100 gramas.

H_0: Crianças com síndrome de desconforto idiopático grave vêm de uma população com peso médio = 3100 gramas

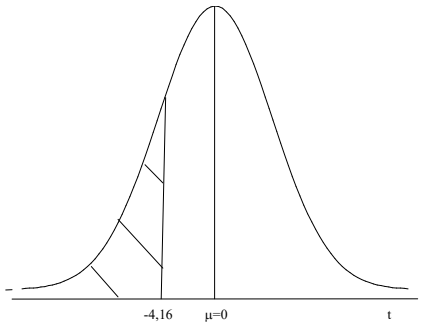
Seleciona-se uma amostra de 50 crianças com a síndrome e calcula-se o peso médio e o desvio padrão do peso, obtendo-se n=50; $\bar{x} = 2800$ e s=510

Distribuição de probabilidade:

Distribuição do peso médio ao nascer de crianças sadias: como não se sabe o desvio padrão populacional, este é estimado utilizando-se os dados da amostra.

Neste caso a variável a variável segue uma distribuição t de Student com n-1=50-1=49 graus de liberdade.

$$P(\overline{X} \leq 2800) = P(\frac{\overline{X} - m}{S_{\overline{X}}} \leq \frac{2800 - 3100}{\frac{510}{\sqrt{50}}}) = P(t_{\overline{X}} \leq \frac{-300}{72{,}12}) = P(t_{\overline{X}} \leq -4{,}159)$$



Pela distribuição *t de Student* com 49 graus de liberdade, tem-se $P(t_{\overline{X}} \leq -4{,}159) < 0{,}05\%$

Os resultados não são compatíveis com uma distribuição que tem peso médio igual a 3100. Pode-se dizer que crianças com desconforto idiopático grave provavelmente vêm de uma população com peso médio ao nascer menor do que o peso médio ao nascer de crianças sadias.

Suspeita-se que crianças que nascem com síndrome de desconforto idiopático grave possuem peso ao nascer abaixo do peso ao nascer da população de crianças sadias.

Proposição: Crianças com síndrome vêm de uma população com peso médio =3100 gramas

Exercício

O conteúdo de iodo em pacotes de sal é recomendado que seja igual a 590 μg . Determinada indústria, tendo recebido reclamações de que estava vendendo seu produto com teor de iodo abaixo do recomendado, realizou um estudo com dosagem de iodo em 15 amostras de sal. Os resultados das quantidades de iodo são apresentados a seguir. Realize um teste de hipóteses pela abordagem de Neyman e Pearson para verificar se a reclamação procedia. Utilize nível de significância de 5%. Tome a decisão utilizando também a abordagem de Fisher com cálculo do valor de p

555	590	500	550	620
570	610	530	530	600
610	600	580	533	575

Exercício

Vacas da raça Jersey (J) produzem porcentagem média de gordura para manteiga igual a 3,5%. Suspeita-se que vacas Holstein-Fresian (HF), se não forem criadas de um modo especial, produzem quantidades menores deste tipo de gordura. É fornecida a porcentagem média de gordura de manteiga de uma amostra de 10 vacas da raça Holstein-Fresian. Os dados sugerem que as que as vacas Holstein-Fresian produzem a mesma quantidade de gordura do que as vacas Jersey? Conduza um teste de hipóteses pela abordagem de Neyman e Pearson. Utilize nível de significância de 5%. Tome a decisão utilizando também a abordagem de Fisher com cálculo do valor de p

Percentuais de gordura de uma amostra de 10 vacas Holstein-Fresian:

3,0	3,6	3,8	4,0	4,4	4,7	4,3	4,0	3,9	3,5

Teste de hipóteses de associação pelo Qui-quadrado de Pearson (χ^2)

O qui-quadrado é obtido somando-se razões dadas pelos quadrados das diferenças entre freqüências observadas e as esperadas, divididos pelas freqüências esperadas.

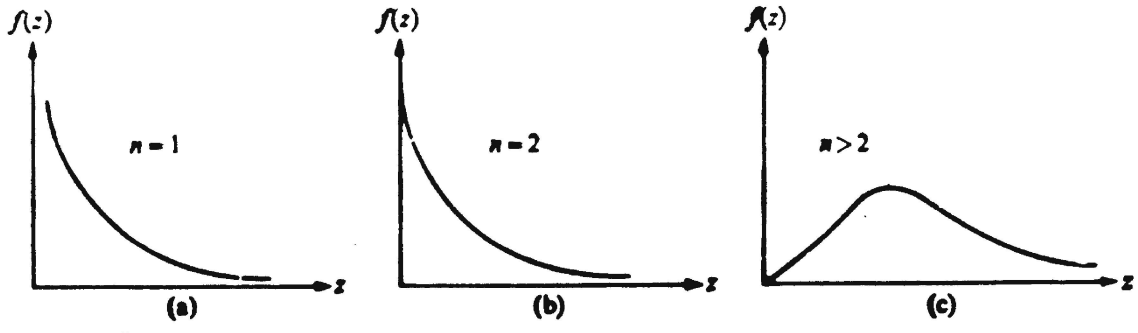
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Quando as variáveis são independentes, é equivalente a dizer que não existe associação, e neste caso, o valor do qui-quadrado será zero. O qui-quadrado não mede força de associação e não é suficiente para estabelecer relação de causa e efeito.

Distribuição qui-quadrado ($\chi^2_{(n-1)}$) com (n-1) graus de liberdade

Seja uma população com distribuição normal $N(\mu, \sigma)$. Se desta população se obtiver um número infinito de amostras de tamanho n, calculando-se as quantidades $\overline{x}$ e S^2 em cada amostra, a variável aleatória $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$, onde $\chi^2_{(n-1)}$ se lê "qui-quadrado com n-1 graus de liberdade" Berquó (1981).

A distribuição qui-quadrado é assimétrica e se torna menos assimétrica a medida que os graus de liberdade aumentam. Os valores da distribuição são sempre positivos (maior ou igual a zero). Existe uma família de distribuições qui-quadrado, dependendo do número de graus de liberdade. Para grandes amostras, a distribuição qui-quadrado tende para uma distribuição normal.



Abordagem de Neyman e Pearson

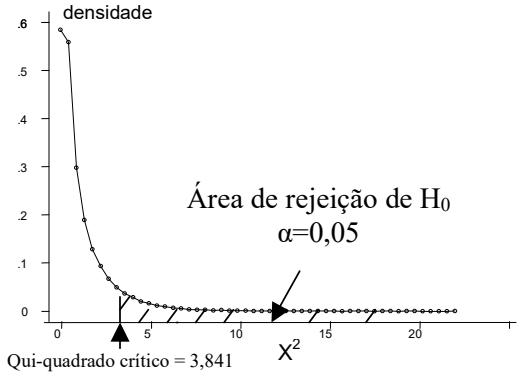
Estabelecimento das hipóteses:

H_0: Não existe associação
H_a: Existe associação

Fixando-se a probabilidade de erro tipo I:
Nível de significância (α) = 0,05

Para a tomada de decisão, utiliza-se a regra: rejeita-se H_0 se o valor calculado do qui-quadrado for maior do que o valor crítico para um nível de significância pré definido.

Área de rejeição do teste:



Estatística do teste**:**

$$Qui-quadrado = \sum \frac{(O-E)^2}{E} \sim \chi^2_{(r-1)(c-1)}$$

onde r e c representam o número de linhas e de colunas, respectivamente.

Correção de continuidade**:**

$$Qui-quadrado_{\text{correcao de Yates}} = \sum \frac{(|O-E|-0{,}5)^2}{E} \sim \chi^2_{(r-1)(c-1)}$$

Limitações:
Para n<20, utilizar o teste exato de Fisher
Para $20 \leq n \leq 40$, utilizar o qui-quadrado somente se os valores esperados forem maiores ou iguais a 5

Exemplo

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo condição de sobrevivência e peso ao nascer (g). Local? Ano?

Peso ao nascer (g)	Óbito	Sobrevida	Total
Baixo peso (<2500g)	24	13	37
Não baixo peso (2500g e mais)	3	10	13
Total	27	23	50

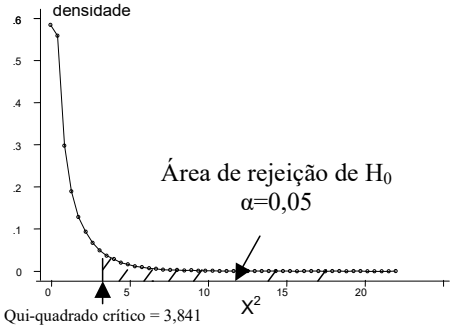
Fonte: Hand DJ et al. A handbook of small data sets. Chapman & Hall, 1994.

Hipóteses:

H_0: Não existe associação
H_a: Existe associação

Fixando-se a probabilidade de erro tipo I:
Nível de significância (α) = 0,05

Área de rejeição do teste:



Cálculo do qui-quadrado de Pearson

Valores observados O	Valores esperados E	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$	$\frac{(\lvert O-E\rvert -0{,}5)^2}{E}$
24	19,98	4,02	16,16	0,809	0,62
3	7,02	-4,02	16,16	2,302	1,77
13	17,02	-4,02	16,16	0,949	0,73
10	5,98	4,02	16,16	2,702	2,07
				$\chi^2 = 6{,}762$	$\chi^2_{corrigido} = 5{,}19$

Decisão: rejeita-se H_0
Conclusão: As variáveis estão associadas para nível de significância de 5%

Coeficiente de associação de Yule

Coeficiente de associação de Yule – permite investigar a força (magnitude) da associação

$$Y = \frac{a.d - b.c}{a.d + b.c} \quad \text{, onde:} \qquad -1 \le Y \le +1$$

$$Y = \frac{24x10 - 3x13}{24x10 + 3x13} = +0{,}72$$

Calculando-se os percentuais:

Peso ao nascer (g)	Óbito		Sobrevida		Total	
	n	%	n	%	n	%
Baixo peso (<2500g)	24	64,9	13	35,1	37	100
Não baixo peso (2500g e mais)	3	23,1	10	76,9	13	100
Total	27	54,0	23	46,0	50	100

A associação entre peso ao nascer e condição de sobrevivência é forte. A proporção de óbitos é maior entre recém-nascidos de baixo peso se comparados aos de não baixo peso.

Exemplo utilizando a estatística qui quadrado corrigido

Com o objetivo de investigar a associação entre história de bronquite na infância e presença de tosse diurna ou noturna em idades mais velhas, foram estudados 1319 adolescentes com 14 anos. Destes, 273 apresentaram história de bronquite até os 5 anos de idade sendo que 26 apresentaram tosse diurna ou noturna aos 14 anos.

Número de adolescentes segundo história de bronquite aos 5 anos e tosse diurna ou noturna aos 14 anos de idade. Local X, ano Y.

Tosse	Bronquite		Total
	Sim	Não	
Sim	26	44	70
Não	247	1002	1249
Total	273	1046	1319

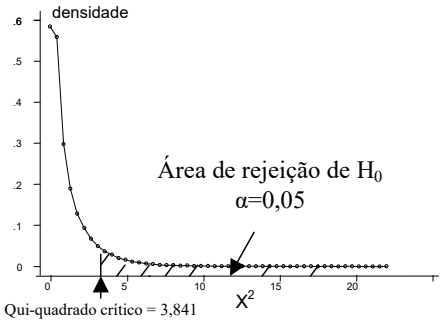
Fonte: Holland, WW et al.. Long-term consequences of respiratory disease in infancy*. Journal of Epidemiology and Community Health* 1978; 32: 256-9.

Hipóteses:

H_0: Não existe associação
H_a: Existe associação

Fixando-se a probabilidade de erro tipo I:
Nível de significância (α) = 0,05

Área de rejeição do teste:



Cálculo do qui quadrado

Valores observados (O)	Valores esperados (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$	$\frac{(\lvert O-E\rvert -0{,}5)^2}{E}$
26	14,488	11,512	132,526	9,147	8,37
247	258,512	-11,512	132,526	0,513	0,469
44	55,512	-11,512	132,526	2,387	2,184
1002	990,488	11,512	132,526	0,134	0,122
				$\chi^2 = 12{,}181$	$\chi^2_{corrigido} = 11{,}145$

Decisão:
O valor do qui-quadrado calculado é maior do que o valor do qui-quadrado crítico para 1 grau de liberdade e nível de significância de 5%, portanto, rejeita-se H_0.

Conclusão: Pode-se dizer que na população existe associação entre bronquite na infância e tosse na adolescência.

Coeficiente de associação de Yule

Coeficiente de associação de Yule

$$Y = \frac{a.d - b.c}{a.d + b.c} \quad \text{, onde:} \qquad -1 \le Y \le +1$$

$$Y = \frac{26x1002 - 44x247}{26x1002 + 44x247} = +0{,}41$$

Calculando-se os percentuais:

Tosse	Bronquite (Sim)		Bronquite (Não)		Total	
	n	%	n	%	n	%
Sim	26	9,5	44	4,2	70	5,3
Não	247	90,5	1002	95,8	1249	94,7
Total	273	100	1046	100	1319	100

A associação entre tosse na adolescência e bronquite, mas não é forte, mas também não pode ser desprezada. Adolescentes apresentam mais tosse na adolescência se tiveram bronquite na infância, comparados aos que não tiveram bronquite na infância.

Teste de associação pelo qui quadrado de Pearson - Abordagem de Fisher

Pela tabela da distribuição qui-quadrado, com 1 gl, p<0,001 (na tabela, menor que 0,1%)

Calculando-se o valor de p pelo Excel, para 1 gl, o valor de $p_{\text{não corrigido}}$ = 0,0004829
No Excel utilizar a função DIST.QUI tendo como argumentos o valor calculado do qui-quadrado e o número de graus de liberdade: = DIST.QUI(12,181;1))

Conclusão: Existe forte evidência contrária à independência. Portanto a associação observada ocorre não devido ao acaso. Pode-se dizer que os dados são compatíveis com existência de associação entre bronquite na infância e tosse na adolescência, na população.

Exercício

Considere os dados apresentados a seguir. Investigue a existência de associação entre níveis de β-caroteno (mg/L) e hábito de fumar, em puérperas. Utilize as abordagens de Neyman e Pearson (nível de significância de 5%) e de Fisher.

Distribuição de mulheres no período pós parto, segundo hábito de fumar e nível de β-caroteno sérico.

β-caroteno (mg/L)	Fumante	Não Fumante	Total
Baixo (0 – 0,213)	56	84	140
Normal (0,214 – 1,00)	22	68	90
Total	78	152	230

Fonte: Silmara Salete de Barros Silva, tese de Doutorado [2003]

Exercício

A tabela abaixo apresenta o número de crianças classificados segundo nível de retinol sérico e sexo. Utilize as abordagens de Neyman e Pearson (nível de significância de 5%) e de Fisher para investigar a existência de associação entre as variáveis.

Sexo	Nível de retinol aceitável		Nível de retino Inadequado		Total	
	n	%	n	%	n	%
Masculino	50		40		90	
Feminino	39		32		71	
Total	89		72		161	

Fonte: Prado MS et al. ,Revista de Saúde Pública, 29(4)295 – 300, 1995

Exercício 17

Os dados a seguir são de pesquisa que estuda a associação entre amamentação ao seio e Diabetes Mellitus tipo I . Local X. Ano Y. Utilize as abordagens de Neyman e Pearson (nível de significância de 5%) e de Fisher para investigar a existência de associação entre as variáveis.

Amamentação ao seio	Casos	Controles	Total
Não	35	17	52
Sim	311	329	640
Total	346	346	692

Fonte: Gimeno SGA. Consumo de leite e o Diabetes Mellitus insulino-dependente: um estudo caso-controle. Tese de doutorado, 1996.