Bioestatística

Professor responsável: Denise Pimentel Bergamaschi denisepb@usp.br

Ementa: O objetivo do curso é apresentar conceitos centrais em bioestatística e iniciar os alunos na utilização de técnicas de resumo e análise de dados. A disciplina foi orientada pelo interesse em repassar aos alunos conhecimentos em estatística que facilitassem a compreensão de aspectos metodológicos comumente abordados em artigos científicos da área de epidemiologia. Os seguintes temas serão abordados no curso: estatística descritiva, incluindo apresentação tabular e gráfica de dados e resumo de dados por meio de medidas de tendência central e de dispersão, análise bidimensional incluindo medidas de associação e correlação, inferência estatística, incluindo estimação pontual e intervalar de parâmetros, testes de hipóteses.

Período: 21/09/2018 a 23/11/2018

Carga horária: 30 horas

Número de créditos: 2

Docente: Denise Pimentel Bergamaschi

Estratégias pedagógicas: Aulas expositivas; aulas práticas para realização de exercícios com uso de microcomputadores; apresentação e discussão de artigos científicos.

Avaliação: O aluno será avaliado pela participação em aulas e por trabalhos individuais. Estes serão referentes às atividades propostas nas aulas práticas: exercícios com o uso de computador e análise de artigos científicos focando os aspectos metodológicos (estratégias para coleta de dados e análise estatística).

Programa 2018

Data	Tipo de aula/conteúdo
21/09 manhã	Aula teórica 1: Estatística descritiva Organização de dados em tabelas e gráficos Aula prática: Organização de bancos de dados e construção de tabelas
28/09 manhã	Aula teórica 2: Estatística descritiva Resumo de dados: medidas de tendência central, de posição e de variabilidade Aula prática: Realização de exercícios: cálculo de média, mediana, variância, desvio padrão e quartis e percentis
05/10 manhã	Aula teórica 3: Conceitos de amostragem. Probabilidade, curva normal. Distribuição amostral da média Aula prática: Exercícios
19/10 manhã	Aula teórica 4: Inferência estatística - Estimação - Intervalo de Confiança para uma média populacional - Intervalo de Confiança para uma proporção populacional Aula prática: Exercícios
26/10 manhã	Aula teórica 5: Teste de hipóteses Teste de hipótese de associação pelo qui quadrado de Pearson Aula prática: Exercício Discussão de artigo
09/11 manhã	Aula teórica 6: Fundamentos de correlação linear, estimativa da reta de regressão linear Aula prática: Exercícios
23/11 manhã	Aula prática 7: Exercício Discussão de artigo - seminários

Bibliografia

Berquó ES, Souza JMP, Gotlieb SLD. Bioestatística. São Paulo: EPU, 1981.

Kish L. Survey Sampling. Nova York: John Wiley & Sons, 1995.

Morettin PA, Bussab WO. Estatística Básica. São Paulo: Saraiva, 2003. 5ª edição.

Pereira JCR. Bioestatística em outras palavras. São Paulo: EDUSP. 2010.

Silva NN. Amostragem Probabilística. São Paulo: Editora da Universidade de São Paulo, 1998.

Vieira S. Introdução à Bioestatística. Rio de Janeiro: Campus, 1980. 3ª edição

Aula 1

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

Estatística: é uma coleção de métodos para planejar experimentos, obter e organizar dados, resumi-los, analisá-los, interpretá-los e deles extrair conclusões.

Bioestatística – Estatística aplicada às ciências da vida.

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) (imc=peso/altura^2(m)); 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

O nível de aferição indica como "medir" (aferir) estas características ou fenômenos e eventos.

Exercício 1 –

Como aferir idade?

Como aferir o sexo?

Como aferir o número de doenças crônicas?

O imc é aferido?

Como aferir nível de triglicérides?

A característica (variável) imc pode ser utilizada com valores pontuais ou em categorias, por exemplo
abaixo ou igual a 21 indicando magreza (≤21);
de 22 a 27 eutrofia (22≤IMC≤27) e
28 e mais (≥28), excesso de peso

Para aferir eventos e características é necessário definir o **nível de aferição** de interesse.

Níveis de aferição ou de mensuração

Níveis de aferição

Escala Nominal ⇒ **variável qualitativa nominal**
Ex: sexo (masculino, feminino)
tipo de moradia (própria, alugada, emprestada, outros)

Escala Ordinal ⇒ **variável qualitativa ordinal**
Ex: nível de retinol sérico (μg/dl)
Deficiente (<10,0)
Baixo (10,0 a 19,9)
Aceitável (20,0 a 49,9)
Alto (≥50,0)

Escala de razões discreta ⇒ **variável quantitativa discreta**
Ex: freqüência de consumo semanal de refrigerante
0, 1, 2, 3, ... vezes

Escala de razões contínua ⇒ **variável quantitativa contínua**
Ex: idade; circunferência da cintura (cm)

A forma de apresentação da variável indicará a melhor estratégia de apresentação dos dados em tabelas, em gráficos e a análise estatística mais adequada

Exercício 2 - 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)	
Tempo de um procedimento cirúrgico (minutos)	
Número de praias consideradas poluídas	

Coleta de dados

A coleta de dados é o processo de observação e registro de valores relacionados ao objeto de estudo, mensurados em elementos de uma amostra ou população.

Conceitos básicos de amostragem

População: totalidade de elementos sob estudo. Apresentam uma ou mais características em comum.
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; podem ser pessoas, domicílios, escolas, creches, células ou qualquer outra unidade.

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: peso médio ao nascer de crianças que nascem no município de São Paulo (μ = 3100 g);

Proporção de crianças de 7 a 12 anos classificadas como obesas, no município de São Paulo (π = 12%).

Estimador: é uma fórmula matemática que permite calcular um valor (estimador por ponto) ou um conjunto de valores (estimador por intervalo) para um parâmetro.

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

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

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

Ex 1:
Supor a idade (anos) de 5 pessoas: 3, 5, 8, 12, 12

Estimativa da idade média: Média amostral= $\bar{x} = \frac{3+5+8+12+12}{5} = 8 \ anos$

Ex 2: Peso médio ao nascer, calculado em uma amostra de 120.000 crianças nascidas no Município de São Paulo no ano de 2000: estimativa do peso médio = média amostral = $\bar{x} = 3000\ g$.

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.

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.

Tabela dos números equiprováveis

TÁBUA VIII NÚMEROS EQÜIPROVÁVEIS – 1

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

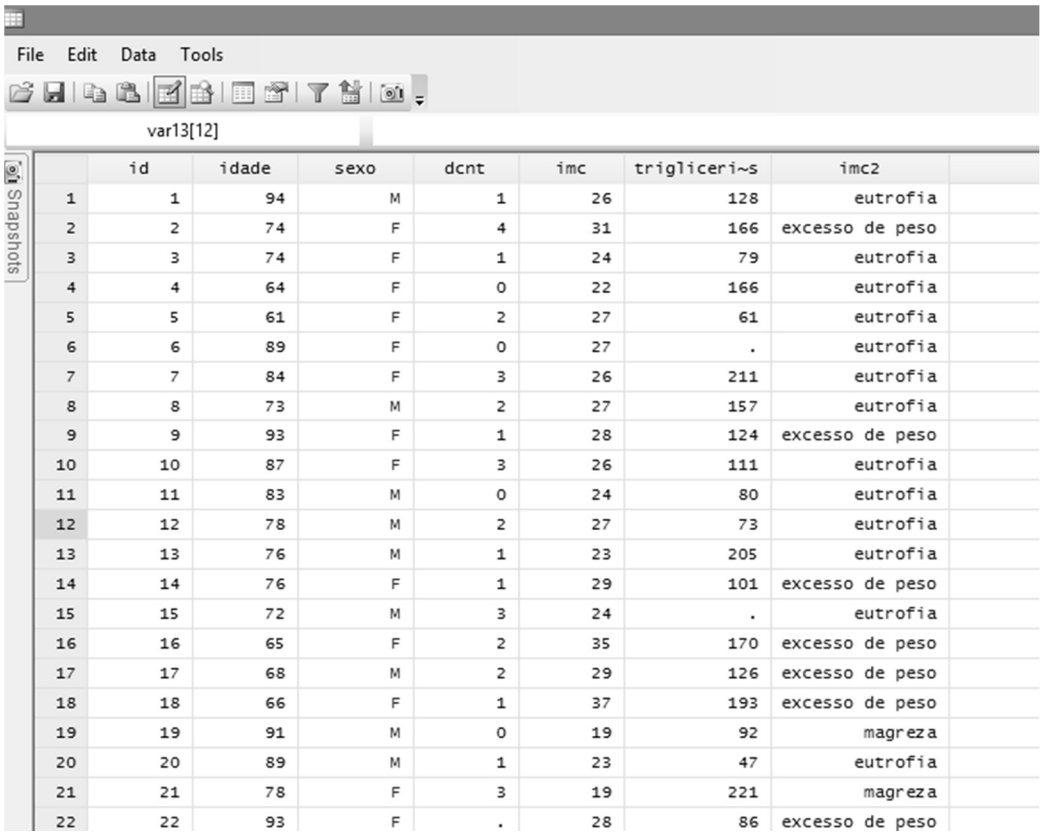
Apuração de dados

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

Distribuição de freqüências - correspondência entre categorias ou valores da variável e freqüência de ocorrência.

Banco de dados construído no pacote Stata utilizando o exemplo do estudo com idosos:

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)	



	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 freqüências 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
```

Valores pontuais – 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.

Obs: regra de aproximação para valores apresentados em casas decimais

```
-> 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	

Valores aproximados para uma casa decimal

```
-> tabulation of dcnt
```

dcnt	Freq.	Percent	Cum.
0	7	14.3	14.29
1	13	26.5	40.82
2	12	24.5	65.31
3	13	26.5	91.84
4	3	6.1	97.96
6	1	2.0	100.00
Total	49	100.0	

Ou

dcnt	Freq.	Percent	Cum.
0	7	14.3	14.29
1	13	26.5	40.82
2	12	24.5	65.31
3	13	26.5	91.84
4	3	6.2	97.96
6	1	2.0	100.00
Total	49	100.0	

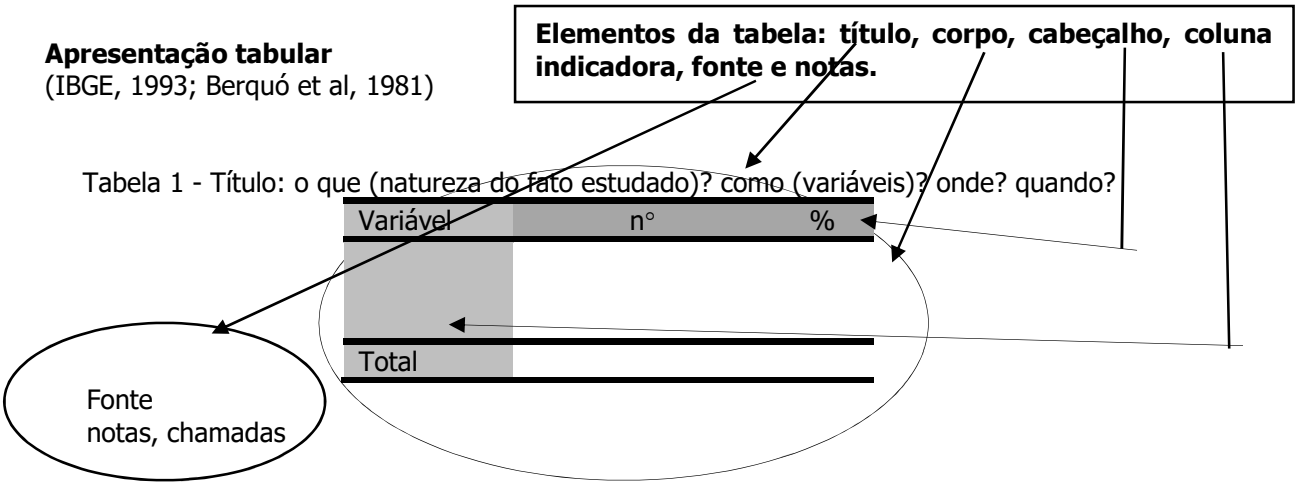
Deseja-se apresentar os valores da porcentagem absoluta, com uma casa decimal.

É necessário olhar para o número que ocupa a segunda casa decimal. Se este for 5, 6, 7, 8 ou 9, o número da esquerda aumenta uma unidade e despreza-se os valores à direita. Se o número da segunda casa decimal for 0, 1, 2, 3 ou 4, o número da esquerda permanece inalterado e despreza-se os valores à direita.

Apresentação de dados em tabelas

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

Apresentação tabular
(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 tabular de uma variável qualitativa

É possível utilizar a imc e construir uma nova variável que permite 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:≤ 21 kg/m^2; eutrofia: 22-27 kg/m^2; excesso de peso ≥28 kg/m^2

Interpretação:

Pode-se observar que a avaliação do estado nutricional indica a presença de excesso de peso em 40% dos idosos.

Ou

Pode-se observar que a avaliação do estado nutricional indica a presença de magreza em 8% dos idosos.

Ou

Pode-se observar que a avaliação do estado nutricional indica eutrofia em 52% dos idosos.

Apresentação tabular de uma variável quantitativa contínua (Berquó ES et al, 1981)

A apresentação deve ser em intervalos de valores - intervalos de classe.

Os intervalos de classe devem ser **mutuamente exclusivos** (um indivíduo não pode ser classificado em dois intervalos ao mesmo tempo) e **exaustivos** (nenhum indivíduo pode ficar sem classificação).

A **amplitude do intervalo** é o tamanho do intervalo de classe. A adoção de determinada amplitude do intervalo e do número de intervalos depende basicamente de cada problema e da literatura existente sobre o assunto.

O **ponto médio do intervalo** é calculado somando-se o limite inferior e limite superior, dividindo-se o resultado por dois.

Tabela 2- Distribuição de idosos segundo triglicérides. Município de São Paulo, 2013.

Triglicerides (mg/dL)	n	%
30 |- 50	3	6,7
50 |- 70	3	6,7
70 |- 90	9	20,0
90 |- 110	6	13,3
110 |- 130	8	17,8
130 |- 150	1	2,2
150 |- 170	6	13,3
170 |- 190	3	6,7
190 |- 210	2	4,4
210 |- 230	3	6,7
230 |- 250	1	2,2
Total	45	100

Interpretação:

Observa-se que os idosos se concentram em níveis de triglicérides que variam de70 a 130mg/dL (51,1%)

Ou
Observa-se que 33,3% dos idosos apresentam níveis de triglicérides 150 mg/dL ou mais.

Exercício 3

Apresentar e descrever os dados dos idosos em tabelas.

Variável sexo

Sexo	n	%
Feminino		
Masculino		
Total		

Interpretação:

Variável número de doenças crônicas

Número de doenças crônicas	n	%
0		
1		
2		
3		
4		
6		
Total		

Interpretação:

```
       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 idade

Idade (anos)	n	%
60 |-- 65		
65 |-- 70		
70 |-- 75		
75 |-- 80		
80 |-- 85		
85 |-- 90		
90 |-- 95		
Total		

Interpretação:

Ou

Idade (anos)	n	%
60 |-- 70		
70 |-- 80		
80 |-- 90		
90 |--100		
Total		

Interpretação:

Tabela de dupla entrada | **São apresentadas duas variáveis ➡ com que objetivo?**

Investigar a existência de associação entre as variáveis

Pergunta: Independente do sexo do idosos, observa-se que 8% apresenta como diagnóstico nutricional, magreza; 52% eutrofia e 40% excesso de peso. Será que esta distribuição se alteraria segundo sexo? Se a distribuição marginal da variável "estado nutricional" for igual em pessoas do sexo feminino e masculino então não existe associação entre as variáveis. Se a distribuição marginal da variável "estado nutricional" for diferente em pessoas do sexo feminino e masculino então deve existir associação entre as variáveis.

Tabela 3 - Distribuição de idosos segundo classificação nutricional e sexo. Município de São Paulo, 2013.

Classificação nutricional	Feminino		Masculino		Total	
	n	%	n	%	n	%
Magreza	3	8,8	1	6,3	4	8,0
Eutrofia	16	47,1	10	62,5	26	52,0
Excesso de peso	15	44,1	5	31,2	20	40,0
Total	34	100	16	100	50	100

Cálculo das porcentagens (%)

$\frac{3}{34} = (0{,}0882) * 100 = 8{,}8$

$\frac{16}{34} = (0{,}4706) * 100 = 47{,}1$

$\frac{15}{34} = (0{,}4412) * 100 = 44{,}1$

$\frac{1}{16} = (0{,}0625) * 100 = 6{,}3$

$\frac{10}{16} = (0{,}625) * 100 = 62{,}5$

$\frac{5}{16} = (0{,}3125) * 100 = 31{,}3$

Interpretação:

Observa-se que independente do sexo, os idosos apresentam 8% de magreza e 40% de excesso de peso. É possível que exista associação entre estado nutricional e sexo. Entre os idosos do sexo feminino a situação nutricional parece pior uma vez que 8,8% apresentam magreza e 44,1% excesso de peso contra 6,3% e 31,2% respectivamente entre os homens.

Outra possibilidade de apresentar os percentuais

Tabela 3 - - Distribuição de idosos segundo classificação nutricional e sexo. Município de São Paulo, 2013.

Classificação nutricional	Feminino		Masculino		Total	
	n	%	n	%	n	%
Magreza	3	75,0	1	25,0	4	100
Eutrofia	16	61,5	10	38,5	26	100
Excesso de peso	15	75,0	5	25,0	20	100
Total	34	68,0	16	32,0	50	100

Cálculo dos percentuais (%)

$$\frac{3}{4} = (0{,}750) * 100 = 75{,}0$$

$$\frac{1}{4} = (0{,}250) * 100 = 25{,}0$$

$$\frac{16}{26} = (0{,}6154) * 100 = 61{,}5$$

$$\frac{10}{26} = (0{,}3846) * 100 = 38{,}5$$

$$\frac{15}{20} = (0{,}750) * 100 = 75{,}0$$

$$\frac{5}{20} = (0{,}250) * 100 = 25{,}0$$

Interpretação:

Observa-se que independente do estado nutricional, 68% dos idosos são do sexo feminino e 32% são do sexo masculino. É possível que exista associação entre estado nutricional e sexo. Entre os idosos classificados como magreza, 75% são do sexo feminino o mesmo sendo observado entre os idosos classificados com excesso de peso contra 25% do sexo masculino entre os classificados como magreza e como excesso de peso.

Exercício 4

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 freqüências relativas. Fixando o 100% no total de fumantes e não fumantes.
b) Calcule as freqüências 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.

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.

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:

Aula 2:

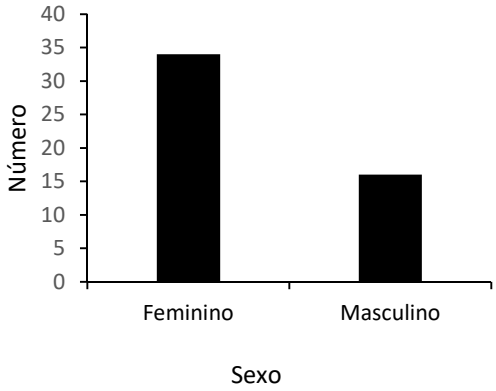
Apresentação gráfica , medidas de tendência central e de dispersão

Apresentação gráfica (Berquó et al., 1981; Chambers et al, 1983)

- ✓ Diagrama de barras
- ✓ Diagrama linear
- ✓ Histograma
- ✓ Outros tipos

Diagrama de barras

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



Distribuição de idosos segundo sexo. Município de São Paulo, 2013.

Interpretação:

Observa-se por meio do gráfico que o número de idosos do sexo feminino é maior que o número de idosos do sexo masculino.

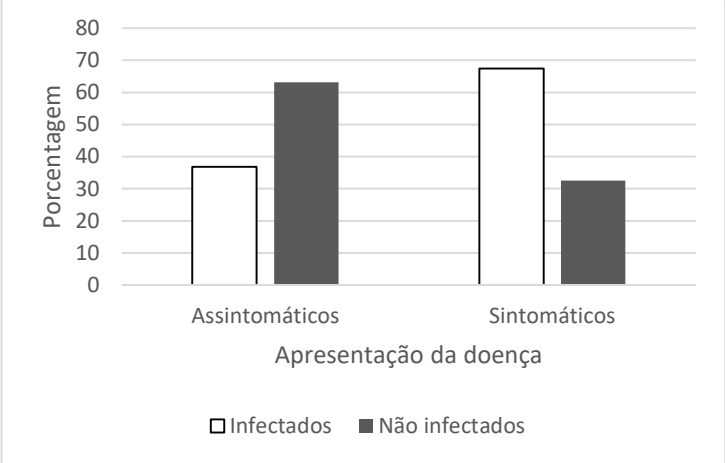
Características do diagrama de barras: as frequências de ocorrência são representadas por figuras geométricas (barras) separadas e bases de mesmo tamanho. A altura das barras é proporcional ao número de ocorrências ou à porcentagem.

Diagrama de barras com duas variáveis

Laranjeira DF et al. Serological and infection status of dogs from a visceral leishmaniasis-endemic area. Rev Saúde Pública 2014;48(4):563-570.

Table 1. Absolute number and percentage of infected and uninfected dogs enrolled in the study, in relation to clinical status. Araçatuba, SP, Southeastern Brazil, 2006. (N = 134)

Clinical status	Infected		Uninfected		Total	
	n	%	n	%	n	%
Asymptomatic dogs	21	36.8	36	63.2	57	42.5
Symptomatic dogs	52	67.5	25	32.5	77	57.5
Total	73	54.5	61	45.5	134	100



Distribuição de cães segundo status clínico de infecção e resultado do teste. Araçatuba, São Paulo, 2006

Interpretação:

Observa-se que entre os cães assintomáticos a proporção de animais não infectados é maior que a proporção entre os sintomáticos.

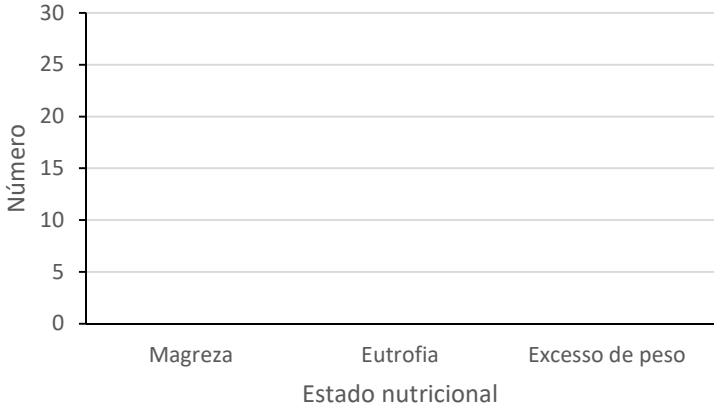
Variável qualitativa ordinal

Exercício 5 – Apresente o diagrama de barras para a variável imc em três categorias

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:≤ 21 kg/m^2; eutrofia: 22-27 kg/m^2; excesso de peso ≥28 kg/m^2



Interpretação:

Diagrama linear

Variável qualitativa ordinal (ano de ocorrência) que permite a união dos pontos pois subjacente às categorias existe continuidade – exceção das variáveis qualitativas.

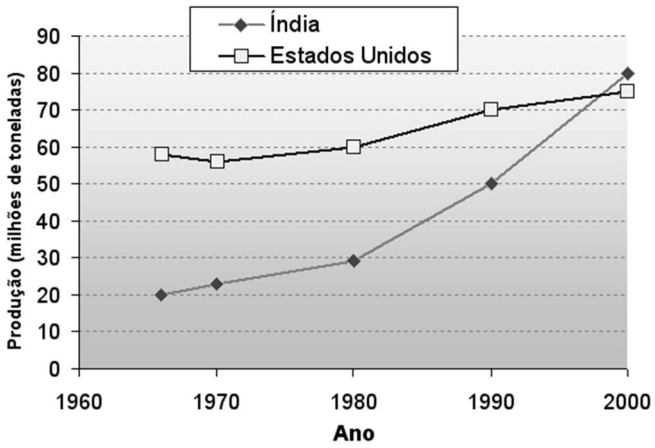
Ex1 -

Tabela 5 - 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: Brown LR et al., 2001.

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



Fonte: Brown LR et al, 2001

Ex2

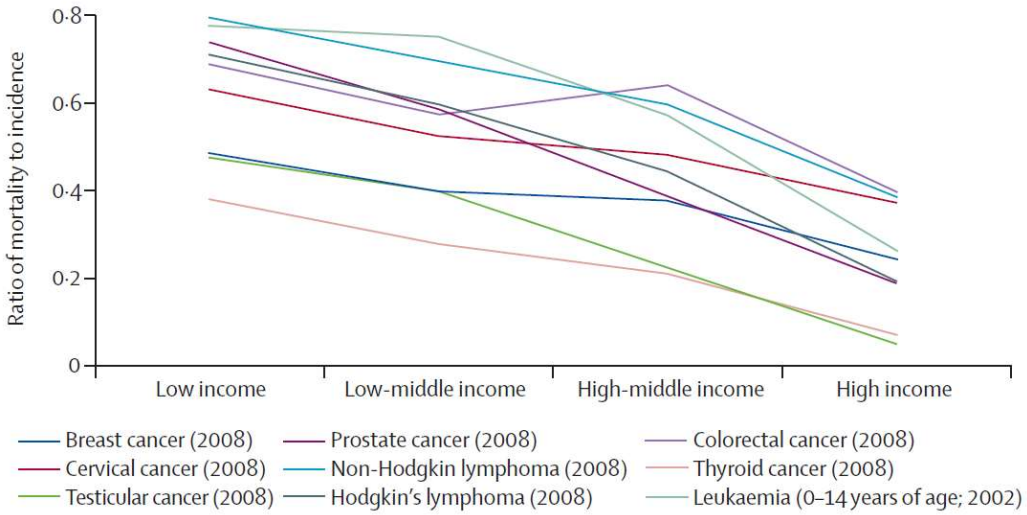


Figure: **Ratio of mortality to incidence in a specific year by cancer type and country income**
Case fatality (calculated by approximation from the ratio of mortality to incidence in a specific year) is much lower in high-income countries than in low-income countries for cancers that are treatable, such as childhood leukaemia (0·26 *vs* 0·78) and testicular cancer (0·05 *vs* 0·47), treatable if detected early, such as breast cancer (0·24 *vs* 0·48), or preventable, such as cervical cancer (0·37 *vs* 0·63). Estimates are based on International Agency for Research on Cancer GLOBOCAN data for 2002 and 2008 (http://globocan.iarc.fr).[3,6]

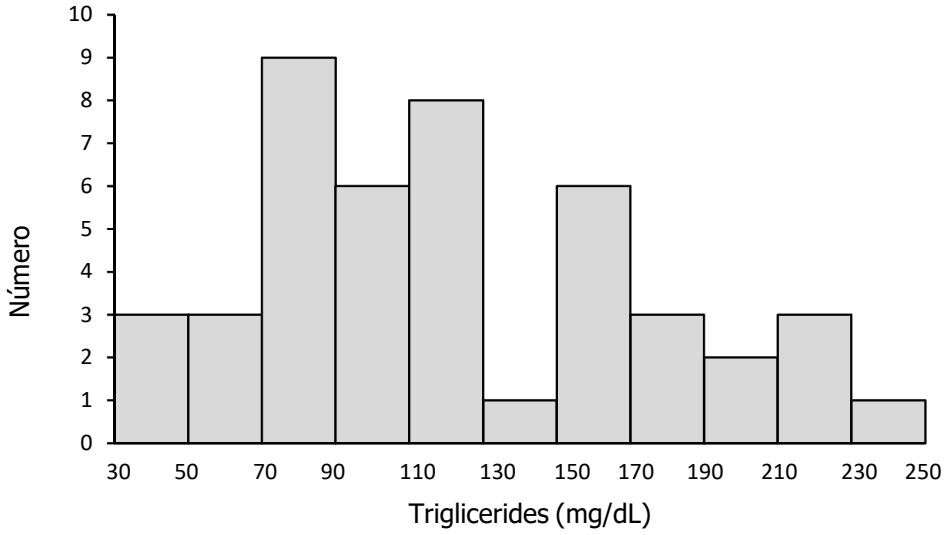
Tipo
(azul água) Linfoma não Hodgkin – linha 1 (low income)
(verde claro) Testículo – linha 2
(vinho) Próstata - linha 3
(verde escuro) Linfoma Hodgkin – linha 4
(lilás) Colorretal – linha 5
(vermelho) Cérvico – uterino - linha 6

(azul escuro) Mama – linha 7
(verde oliva) Testículo – linha 8
(rosa) Tireoide – linha 9

Farmer P et al. Expansion of câncer care and control in countries of low and middle income: a call to action. The Lancet. Vol 376. Outubro 2, 2010.

Histograma
Adequado para representar variáveis quantitativas contínuas. As alturas das barras são proporcionais à freqüência de ocorrência. OBS: é necessário fazer o ajuste se as amplitudes dos intervalos forem diferentes.

Intervalos de classe com mesma amplitude



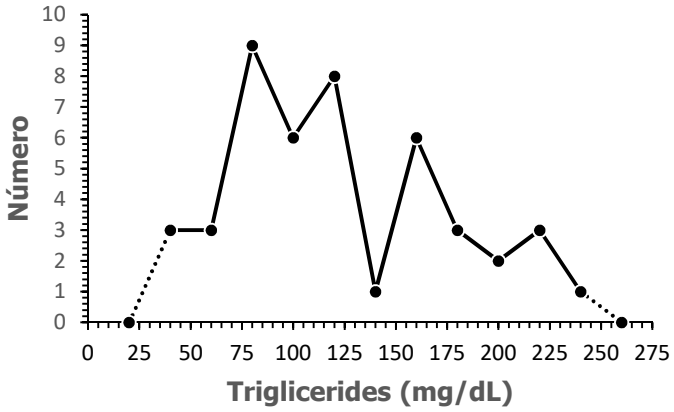
Distribuição de idosos segundo triglicérides. Município de São Paulo, 2013

Interpretação:

Observa-se maior número de idosos em níveis de triglicérides entre 70 e 130 mg/dL. Chama a atenção o número de idosos com níveis de triglicérides acima de 150 mg/dL.

Polígono de freqüência simples – adequado para representar uma variável quantitativa contínua

Intervalos de classe com mesma amplitude



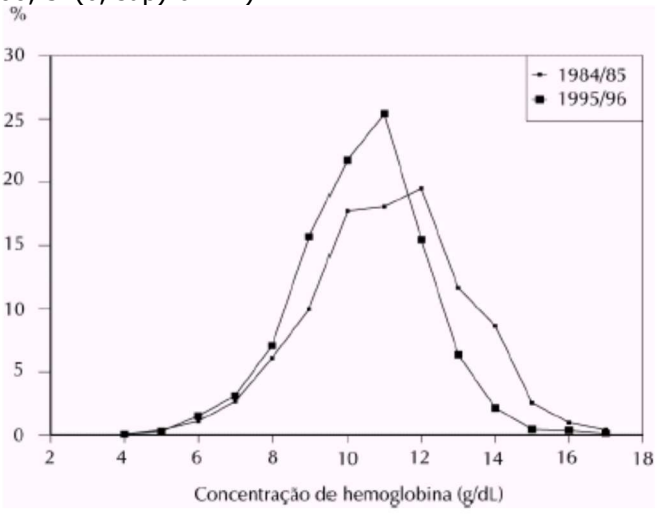
Distribuição de idosos segundo triglicérides. Município de São Paulo, 2013

Características:
O gráfico é construído a partir da união dos pontos médios dos intervalos de classe. O primeiro e último intervalos são construídos unindo-se os pontos médios ao eixo X nos pontos médios de classes hipotéticas construídas com a mesma amplitude do primeiro e último intervalos de classe.

Interpretação:
Observa-se concentração de idosos entre valores de triglicérides de 60 a 120 mg/dL. O gráfico sugere uma concentração importante de idosos acima de 150mg/dL.

Exercício 6

Artigo: Tendência secular da anemia na cidade de São Paulo (1984-1996) de MONTEIRO CA *et al*. (*Rev. Saúde Pública*, 2000; 34(6, sup):62-72).



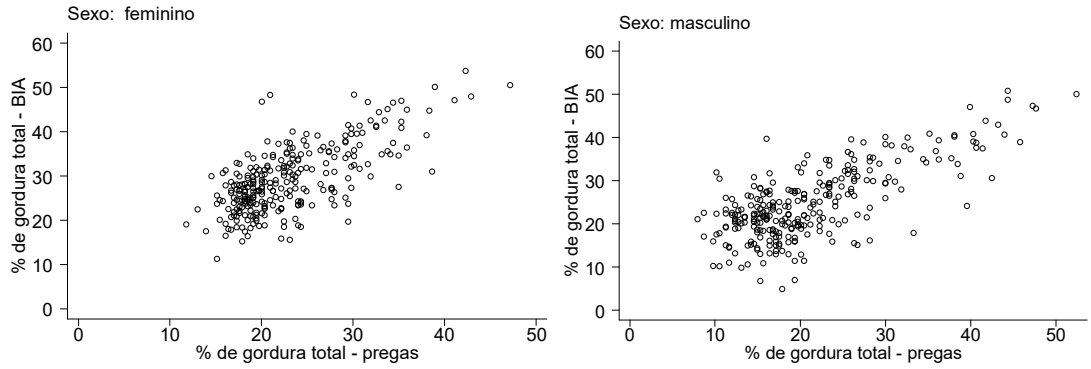
Distribuição de pessoas segundo concentração de hemoglobina. Cidade de São Paulo, 1984/85 e 1995/96.

Interpretação:

Observa-se que em 1984/85 existia concentração de pessoas com taxa de hemoglobina entre 9 e 13g/dL e em 1995/96, os indivíduos se concentram em níveis ao redor de 11 g/dL indicando possível piora da anemia.

Outros tipos de gráficos

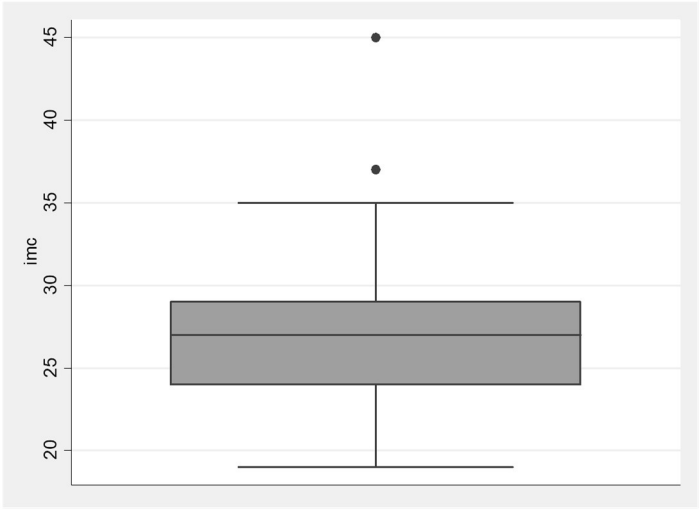
Diagrama de dispersão
Permite investigar a relação entre duas variáveis quantitativas.



(a) (b)

Gráficos 17a e 17b – Distribuição de escolares segundo percentual de gordura total estimado pela BIA e pelas pregas cutâneas tricipital e da panturrilha, segundo sexo. Duas escolas públicas de São Paulo, 2004

Box plot



Box plot da variável imc. Idosos do município de São Paulo, 2013

<u>Box plot e identificação de valores aberrantes (*outliers*)</u>

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

imc
26
31
24
22
27
27
26
27
28
26
24
27
23
29
24
35
29
37
19
23
19
28
28
26

	posto
19	1
19	2
20	3
20	4
22	5
22	6
23	7
23	8
23	9
23	10
24	11
24	12
24	13
24	14
24	15
25	16
25	17
26	18
26	19
26	20
26	21
27	22
27	23
27	24

28	27	25
24	27	26
34	27	27
25	27	28
20	27	29
27	27	30
45	28	31
35	28	32
24	28	33
22	28	34
31	29	35
27	29	36
23	29	37
20	29	38
29	29	39
29	29	40
27	30	41
30	31	42
34	31	43
25	34	44
34	34	45
29	34	46
27	35	47
23	35	48
29	37	49
27	45	50

quartil 1 = 24;

n= número de observações =50
Quartil 1= valor que está na posição 1/4 de (n+1)
Q1=(1/4)x51=12,75 Valor que está na posição 12,75

Q1=24+(0,75x(24-24))=24

quartil 2 = 27
Quartil 2= valor que está na posição 1/2 de (n+1)
Q2=(1/2)x51=25,5; Valor que está na posição 25,5
Q2=27+(0,5x(27-27))=27

e quartil 3 = 29
Quartil 3= valor que está na posição 3/4 de (n+1)
Q3=(3/4)x51=38,25; Valor que está na posição 38,25

Q3=29+(0,25x(29-29))=29

Intervalo Inter quartil = 29-24= 5

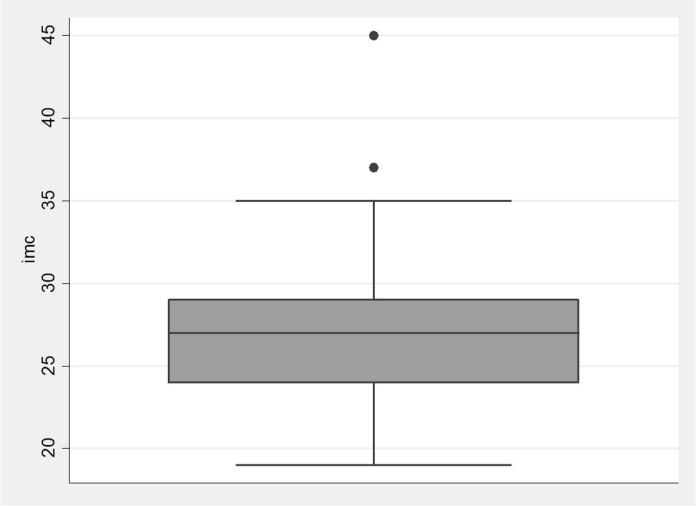
VAI:

Menor valor dos dados que é maior ou igual a Q1-1,5(IIQ)
(24-(1,5x5)) =16,5
VAI =19

VAS: Maior valor dos dados que é menor ou igual a Q3+1,5(IIQ)

(29+(1,5x5)) = 36,5

VAS = 35



Box plot da variável imc. Idosos do município de São Paulo, 2013

Interpretação:

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

Medidas de tendência central e de dispersão

Medidas de tendência central

Média aritmética

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.

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

Ordenando-se os valores,

2	3	(4)	5	6
		média		

Calculando-se os desvios em torno da média

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

Média aritmética é o valor que indica 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 ($\bar{x}$) é igual a:

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

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=24; n=5. Portanto, $\bar{x} = \frac{3+2+5+6+24}{5} = \frac{40}{5} = 8$ ovos

Notação:

X → variável (número de ovos)

N → tamanho da população (desconhecido)

n → tamanho da amostra (n=5)

μ → Média populacional (parâmetro, geralmente desconhecido)

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

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

Exercício 7

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	
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	

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	25

$\overline{x} = \frac{25}{16} = 1{,}56$ 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

$\overline{x} = \frac{69}{33} = 2{,}09$ doenças

Mediana

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

Quando o 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;

Exercício 8

Utilizando-se os valores da variável número de doenças crônicas, calcular o valor mediano para pessoas do sexo masculino e feminino.

Inicie ordenando os valores

Homens

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

Número de observações (n=16) é par

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

mediana= 1 doença

Mulheres

X	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

Número de observações (n=33) é ímpar

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

a mediana é o valor da variável que ocupa o posto 34/2=17

mediana= 2 doenças

Medidas de dispersão

Valores mínimo e máximo: valores extremos da distribuição.

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

Variância: indica o quanto, em média, os quadrados dos desvios de cada observação em relação à média aritmética estão afastados desta média.

Variância

É uma **medida de dispersão** que fornece a distância média ao quadrado das observações em relação à média. As distâncias de cada observação em relação à média são denominadas 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 desvios quadráticos 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

Variância = $\frac{10}{5} = 2$ ovos^2

Desvio padrão

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

Desvio padrão= $\sqrt{2} = 1{,}4$ ovos

Indica o erro médio que se comete ao resumir os dados pela média.

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 à toda a população, o estimador é $S^2_{(N)} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$

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

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

$$CV = \frac{Desvio\ padrão}{Média} x100$$

Exercício 9

Calcule as medidas de dispersão da variável "número de doenças crônicas" para cada sexo.

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	1,4375	2,066406
0	-1,5625	2,441406
1	-0,5625	0,316406
3	1,4375	2,066406
2	0,4375	0,191406
1	-0,5625	0,316406
3	1,4375	2,066406
0	-1,5625	2,441406
2	0,4375	0,191406
1	-0,5625	0,316406
0	-1,5625	2,441406
6	4,4375	19,69141
0	-1,5625	2,441406
0	-1,5625	2,441406
1	-0,5625	0,316406
2	0,4375	0,191406
	soma	39,9375

$\bar{x}$ = 1,5625

Variância (n)= $s^2(n)$ = (39,9375/16)= 2,5 doenças^2

Variância (n-1)= $s^2(n-1)$ = (39,9375/15)= 2,7 doenças^2

Desvio padrão (n)= s(n) = $\sqrt{2,5}$= 1,58 doenças

Desvio padrão (n-1)= s(n-1) = $\sqrt{2,7}$= 1,63 doenças

Valor mínimo

Valor máximo

Variância (n)

Variância (n-1)

Desvio padrão (n)

Desvio padrão (n-1)

Coeficiente de variação de Pearson

$\overline{x}$ = 1,5625 doenças

Valor mínimo = 0 doenças

Valor máximo = 6 doenças

Variância (n) = 2,5 doenças^2

Variância (n-1) = 2,7 doenças^2

Desvio padrão (n) = 1,58 doenças

Desvio padrão (n-1) = 1,63 doenças

CV= $\frac{1,63}{1,5625} x100$ = 104,3%

Feminino

Feminino (X)	$(x-\bar{x})$	$(x-\bar{x})^2$
1	-1,09091	1,190083
1	-1,09091	1,190083
4	1,909091	3,644628
3	0,909091	0,826446
4	1,909091	3,644628
3	0,909091	0,826446
0	-2,09091	4,371901
1	-1,09091	1,190083
2	-0,09091	0,008264
3	0,909091	0,826446
1	-1,09091	1,190083
2	-0,09091	0,008264
2	-0,09091	0,008264
3	0,909091	0,826446
3	0,909091	0,826446
1	-1,09091	1,190083
2	-0,09091	0,008264
3	0,909091	0,826446
1	-1,09091	1,190083
1	-1,09091	1,190083
3	0,909091	0,826446
0	-2,09091	4,371901
1	-1,09091	1,190083
2	-0,09091	0,008264
2	-0,09091	0,008264
2	-0,09091	0,008264
3	0,909091	0,826446
1	-1,09091	1,190083
3	0,909091	0,826446
2	-0,09091	0,008264
2	-0,09091	0,008264
4	1,909091	3,644628
3	0,909091	0,826446
	Soma	38,72727

$\bar{x}$ =2,09 doenças

Valor mínimo = 0 doenças

Valor máximo = 4 doenças

Variância (n) = 1,17 doenças^2

Variância (n-1) = 1,21 doenças^2

Desvio padrão (n) = 1,08 doenças

Desvio padrão (n-1) = 1,1 doenças

CV= $\frac{1,1}{2,09} x100$ = 52,6%

Apresentação das medidas-resumo

Tabela 1 *- Valores mínimo e máximo e médias dos parâmetros dietéticos obtidos através de dois recordatórios de 24-horas.*

Dietetic Variables	Crude values			
	Mean	Standard Deviation	Minimum	Maximum
Energy (kcal)	2,326.18	883.50	1,045.20	5,938.42
Fat (g)	89.03	38.30	35.42	253.08
Carbohydrate (g)	305.31	121.36	117.68	744.61
Protein (g)	82.15	32.84	28.89	202.29

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.

Aula 3

Correlação, regressão linear simples e Medidas de associação

Correlação

Análise simultânea entre duas variáveis quantitativas (associação entre duas variáveis quantitativas).

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

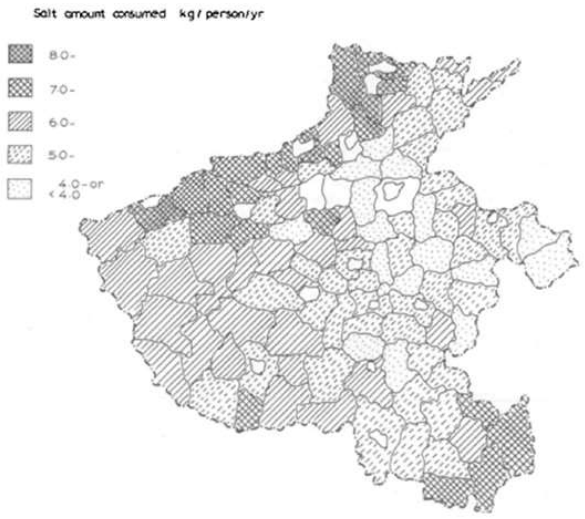
É construído com conjuntos de pontos formados por pares de valores (x,y). Pode indicar correlação linear positiva, negativa ou inexistência de correlação. Também é útil para identificar existência de valores aberrantes.

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

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



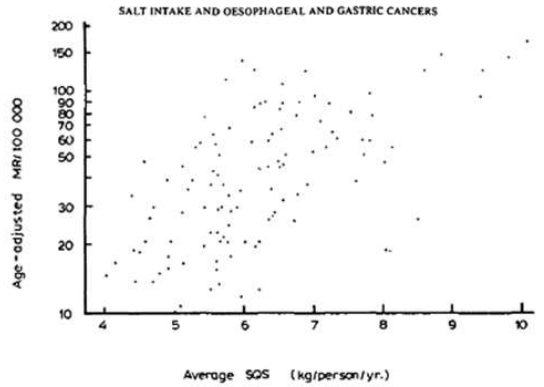
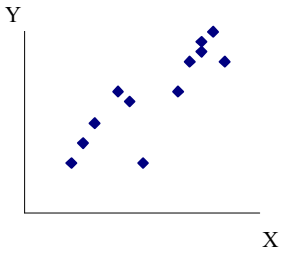


FIGURE 2 *Graph of the correlation between salt quantity sold (SQS) during 1964–66, 1974–76 and the mortality rate of oesophageal cancer during 1974–76 in Henan Province, China.*

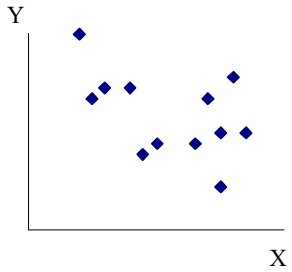
TABLE 1 *The rank correlation coefficient between the SQS in 1964–66, 1974–76 and mortality rate from malignant neoplasms selected in 1974–76 in Henan, China.*

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



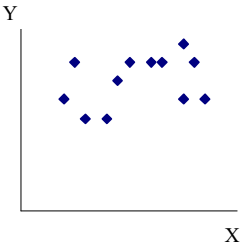
correlação positiva

Ex: X: Peso ao nascer (gramas)
Y: Aumento de peso entre 70 e 100 dias, como percentual de X



Correlação negativa

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



correlação inexistente

A existência de associação não é condição suficiente para se afirmar sobre a existência de relação causal.

Correlação permite responder se mudanças na magnitude de uma variável são acompanhadas de mudanças na magnitude da outra. Atenção: caso exista correlação, não se pode dizer que uma variável causa a outra.

Coeficiente de correlação de Pearson (ρ) - Mede o grau de associação entre 2 variáveis quantitativas 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 Y).

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)

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

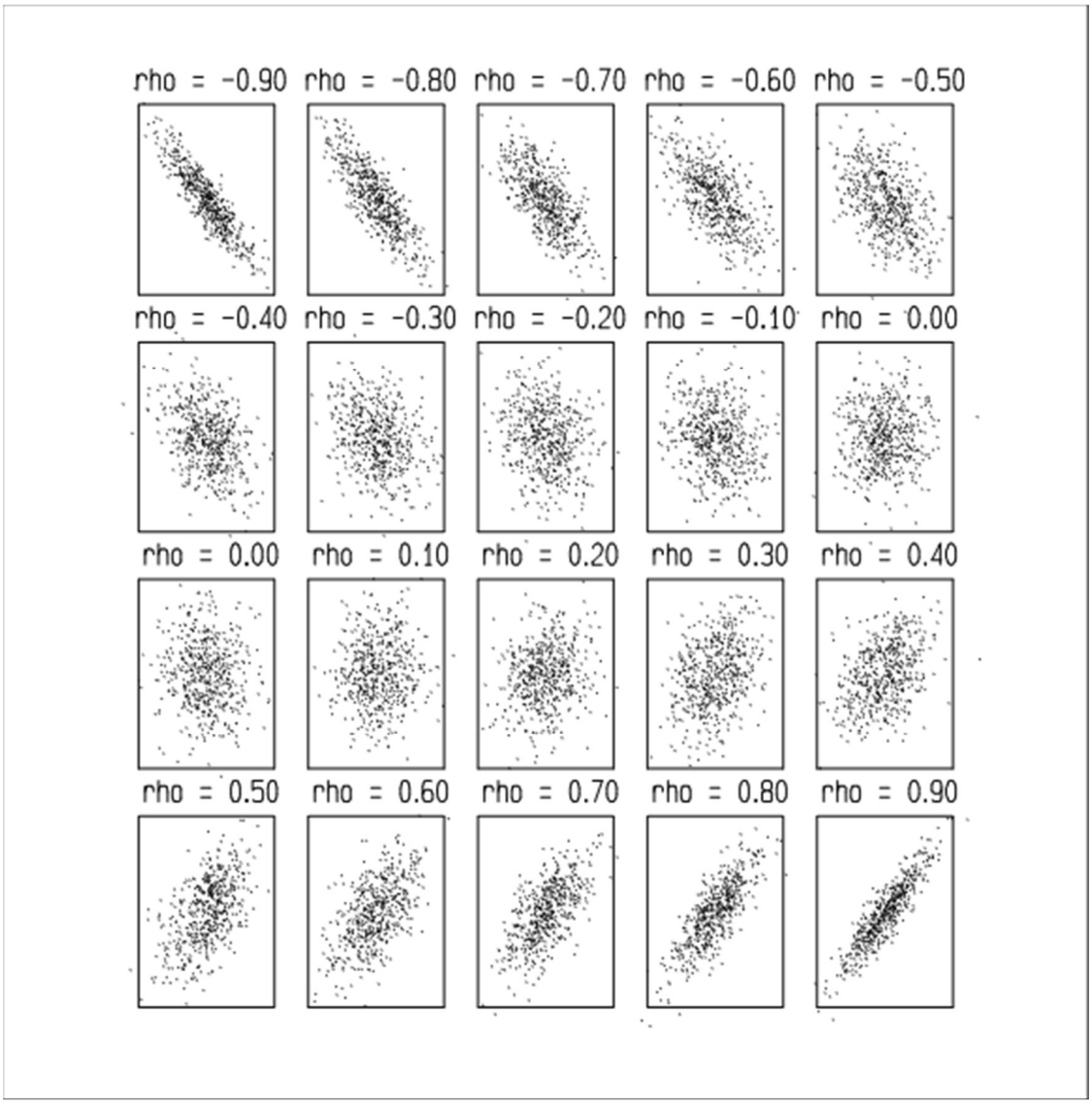
Propriedades

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

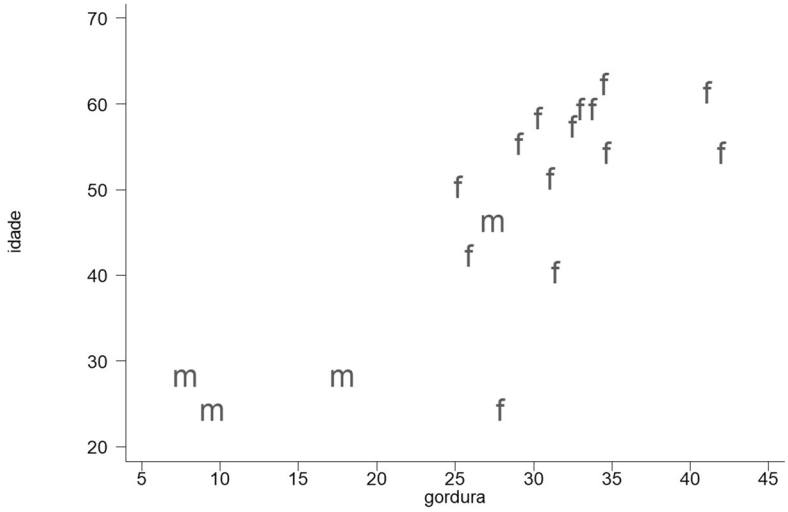
Os dados a seguir são provenientes de um estudo que investiga a composição corporal e fornece o percentual de gordura corporal (%), idade e sexo de 18 adultos com idades entre 23 e 61 anos.

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

Dispersão entre gordura corporal (%) e idade (anos)



Fonte:

Cálculo do coeficiente de correlação de Pearson

Sexo: masculino

Idade	% gordura	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
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	% gordura	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
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: $r = \dfrac{333,64}{\sqrt{1389,71x312,12}} = 0,51$

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

$\bar{x} = 46,33$

$\bar{y} = 28,61$

$S_y = \sqrt{\dfrac{\sum (y-\bar{y})^2}{n-1}} = \sqrt{\dfrac{1421,54}{17}} = 9,14\%$

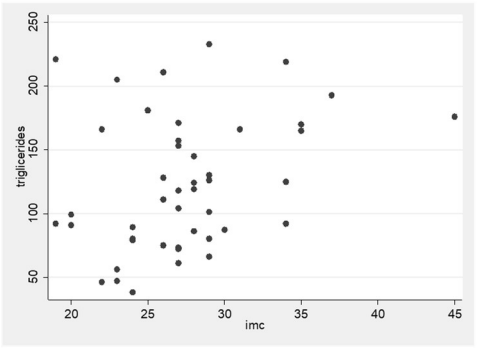
$$S_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{2970,0}{17}} = 13,22 anos$$

Coeficiente de correlação considerando-se homens e mulheres

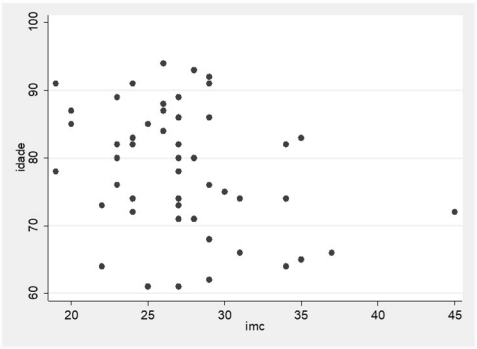
$$r = \frac{1627,53}{\sqrt{2970,0 x 1421,54}} = 0,79$$

Exemplos de investigação de correlação

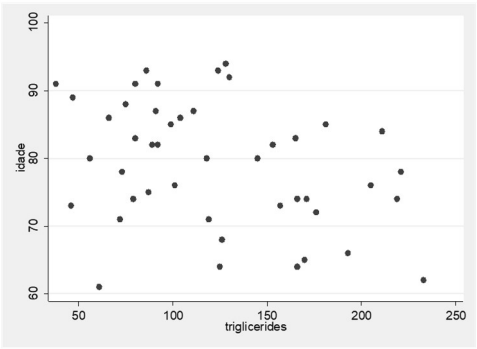
Diagramas de dispersão entre idade e imc, idade e triglicérides e imc e triglicérides. Idosos do município de São Paulo, 2013



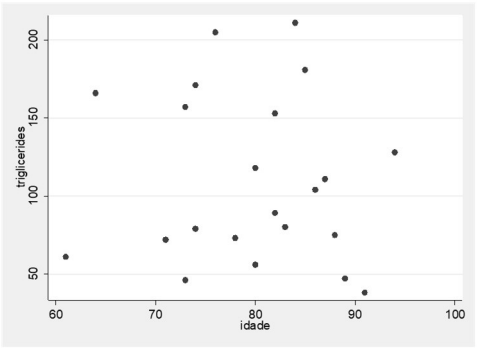
Coeficiente de correlação de Pearson (r) = 0,312 (p=0,037)



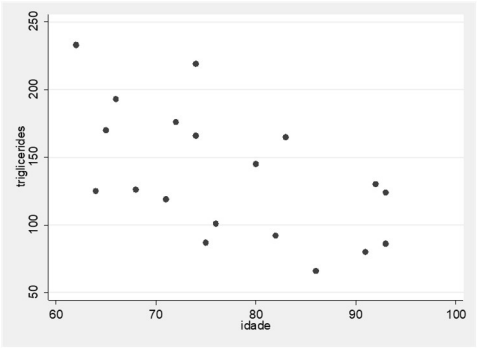
Coeficiente de correlação de Pearson (r) = -0,300 (p=0,036)



Coeficiente de correlação de Pearson (r) = -0,312 (p=0,037) para todos os indivíduos



Coeficiente de correlação de Pearson (r) = -0,073 (p=0,747) para indivíduos eutróficos segundo imc



Coeficiente de correlação de Pearson (r) = -0,575 (p=0,010) para indivíduos com excesso de peso segundo imc

Exemplo: influência de valores outlier

Os gráficos abaixo foram extraídos do artigo: Excesso de peso e gordura abdominal para a síndrome metabólica em nipo-brasileiros de LERARIO DG *et al*. (*Rev. Saúde Pública*, 2002;36(1):4-11). Interprete as figuras apresentadas no artigo.

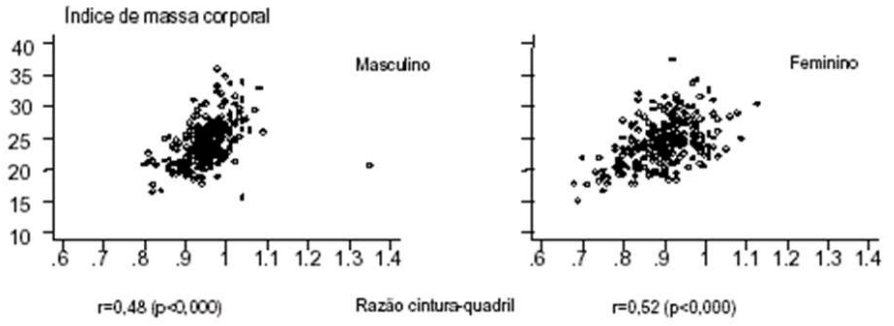


Figura 1 - Correlação entre os valores do índice de massa corporal (Kg/m²) e da razão cintura-quadril de nipo-brasileiros segundo sexo.

Regressão linear simples – estimando a reta de regressão

Admitindo-se y como função linear de x, ajusta-se a "melhor reta" ao conjunto de dados.

Equação da reta: $\hat{y} = a + bx$, onde

$\hat{y}$ = valor ajustado (valor médio predito)

x = valor escolhido de X

$a = \bar{y} - b\bar{x}$; a é denominado intercepto; é o valor predito para x=0

$b = r_{xy} \frac{s_y}{s_x}$; b é denominado coeficiente angular (*slope*). Indica quantas unidades de Y mudam em média, para a mudança de uma unidade de X.

Aplicando-se aos dados do exemplo:

$a = \bar{y} - b\bar{x}$;

$a = 28{,}1 - bx46{,}33$

$$b = r_{xy} \frac{s_y}{s_x};$$

$$r = \frac{1627,53}{\sqrt{2970 x 1421,54}} = 0,79$$

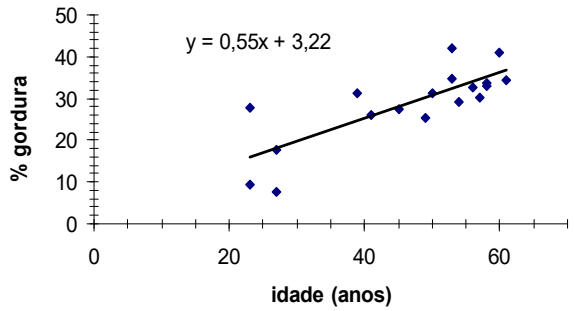
$$b = 0,79 x \frac{9,14}{13,22} = 0,548$$

Substituindo-se o valor b em a, obtém-se a=3,221.

Equação ajustada % gordura= 3,22 + 0,55 (idade)

Com base nesta equação é possível traçar a reta que passa pelos pontos.

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



Interpretação do coeficiente angular da reta: para aumento de 1 ano, o percentual de gordura aumenta 0,55%.

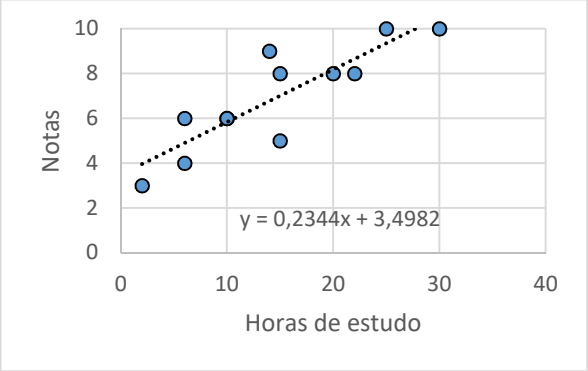
OBS: o coeficiente angular depende das unidades de medida de X e Y. Isto deve ser considerado na decisão da importância do coeficiente angular.
O coeficiente angular da equação de Y=f(X) é diferente do coeficiente angular de X=f(Y), a menos que os desvios padrão de X e Y sejam iguais.

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.

Exemplo

Horas	Nota
30	10
10	6
22	8
14	9
6	4
6	6
25	10
15	5
2	3
10	6
20	8
15	8



=CORREL(B2:B13;C2:C13)
r= 0,867

Medidas de associação

- Razão de incidências
- Odds ratio
- Qui quadrado de Pearson

Razão de riscos (razão de incidências)

Estudo de incidência: estudo de seguimento que permite identificar casos incidentes (casos novos)

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

Desfecho = óbito
Exposição = fumar
Incidência = risco

Risco de morrer (geral) = 472/2789= 0,17 = 17%
Risco de morrer (entre expostos) = r_1=208/1058= 0,20= 20%
Risco de morrer (entre não expostos) = r_0=264/1731= 0,15=15%

Risco relativo = razão de riscos = rr=0,20/0,15= 1,33

Razão de riscos como medida de associação:

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

Razão de riscos como medida de efeito:

Como a razão de riscos (rr) é diferente de 1, e no exemplo, é maior que 1, pode-se dizer que a incidência de mortes parece ser maior entre as pessoas que fumam. Os fumantes apresentam uma incidência 33% maior do que os não fumantes. [(1-1,33)x100=33%]

Uma outra forma de evidenciar a existência de efeito é dizer que a incidência de óbitos entre fumantes é 1,33 vezes a incidência entre os não fumantes.

Risco atribuível:

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

Pela diferença diz-se que 5% dos óbitos excedentes são devido ao fumo.

Se o interesse for investigar fator de proteção:

Seria equivalente a ter interesse nos óbitos entre os não expostos (não fumantes) e assim, o risco relativo seria calculado como

rr = 0,15/0,20= 0,75; que é menor que 1.

[interpretação do RR como medida de associação] Seria necessário testar se o rr calculado vem de população onde o RR é igual a 1. Se estatisticamente 0,75 for diferente de 1, pode-se dizer que existe associação entre as variáveis.

[interpretação como medida de efeito seria] [1-0,75)=0,25; 0,25x100 = 25%. Então, o risco de morte entre não expostos é 25% menor que o risco entre expostos ou o risco de morte entre não expostos é 0,75 vezes o risco entre expostos.

Neste caso diz-se que a exposição é fator de proteção

Exercício 10

Padrão de amamentação de crianças segundo episódios de doenças respiratórias.

Padrão	Um ou mais episódios	Nenhum episódio	Total
Mamadeira e peito	207	238	445
Somente peito	34	72	106
Total	241	310	551

Fonte: Abramson JH e Abramson ZH.

Considerando-se o desfecho: um ou mais episódios de doenças respiratórias e a exposição alimentação com mamadeira e peito,

a) Calcule a incidência de um ou mais episódios de doenças respiratórias, dado que a criança se alimenta de mamadeira e peito;

$I_{desfecho\ \ entre\ \exp ostos} = \text{——} =$

b) Calcule a incidência de um ou mais episódios de doenças respiratórias, dado que a criança se alimenta somente ao seio;

$I_{desfecho\ \ entre\ não\ \exp ostos}$ = —— =

c) Calcule a razão de incidências;

Risco relativo = rr= —— =

d) Calcule a diferença de incidências;

e) Discuta os resultados

Respostas

a) Calcule a incidência de um ou mais episódios de doenças respiratórias, dado que a criança se alimenta de mamadeira e peito;

$$I_{desfecho\ entre\ expostos} = \frac{207}{445} = 0{,}465$$

b) Calcule a incidência de um ou mais episódios de doenças respiratórias, dado que a criança se alimenta somente ao seio;

$$I_{desfecho\ entre\ não\ expostos} = \frac{34}{106} = 0{,}321$$

c) Calcule a razão de incidências;

$$RR = \frac{\frac{207}{445}}{\frac{34}{106}} = \frac{106x207}{34x445} = 1{,}45$$

d) Calcule a diferença de incidências;
RA= 0,465-0,321=0,144

e) Discuta os resultados

Se o RR for estatisticamente diferente de 1 então pode-se dizer que existe associação entre forma de amamentação e doença respiratória. Neste caso, pode-se dizer que a incidência de episódios de doenças respiratórias entre crianças alimentadas na mamadeira e no peito é 45% maior que a incidência entre crianças amamentadas exclusivamente ao seio. Pode-se dizer que 14,4% dos casos de doença respiratória poderiam ser evitados na ausência da exposição.

Estudo do tipo caso-controle

Odds e probabilidade

Probabilidade

Supor que durante um jogo de basquete um jogador acerta a cesta 2 vezes em 5 tentativas.

Chamando p de probabilidade de acerto tem-se que $p = \frac{2}{5}$ = 0,4 ou 40% e a probabilidade de erro, $q = \frac{3}{5}$ = 0,6 ou 60%.

Considerando-se que a probabilidade de acerto ou de erro = p+q = 1; então $q = 1 - p$

Odds ratio

Define-se *odds* como a razão entre a probabilidade de acerto e a probabilidade de erro, ou seja,

Odds= $\frac{p}{1-p}$ [tradução de odds: razão de probabilidades]

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).

Odds ratio [razão de odds]

Exemplo 5:

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 + (expostos)	0-79 (não expostos)	
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: $\frac{96}{109} : \frac{104}{666} = \frac{96x666}{109x104} = 5{,}6$

Razão de odds como medida de associação:

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 do odds ratio (OR) vem de uma população com parâmetro igual a 1.

Razão de odds como medida de efeito:

Se a Odds ratio diferente de 1, e maior que 1, como no exercício, pode-se dizer que 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).

Qui-quadrado de Pearson – indica se há ou não associação. Não mede força de associação.

Duas variáveis qualitativas
X - curso universitário e
Y – sexo do aluno

Questão: sexo do indivíduo influi na escolha do curso?
Situação 1

Curso	Masculino	Feminino	Total
	n	n	n
Economia	24	36	60
Administração	16	24	40
Total	40	60	100

Curso	Masculino		Feminino		Total	
	n	proporção	n	proporção	n	proporção
Economia	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

As proporções de escolha dos cursos não diferem segundo sexo do estudante.

Situação 2

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

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

A distribuição de alunos em cada curso segundo sexo não é a mesma. Sexo e curso podem estar associados.

Se a variável sexo não fosse associada à escolha do curso, quantos indivíduos esperaríamos em Física, entre os homens?

Casela 100 (Física – Masculino)

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

Casela 40 (Ciências Sociais – Masculino)

$\frac{x}{140} = \frac{80}{200}$ ⟶ $x = \frac{140x80}{200} = 56$

Casela 20 (Física – Feminino)

$\frac{x}{60} = \frac{120}{200}$ ⟶ $x = \frac{60x1}{200} = 36$

Casela 40 (Ciências Sociais – Feminino)

⟶

$\frac{x}{60} = \frac{80}{200}$ $x = \frac{60x80}{200} = 24$

Curso	Sexo	Número esperado
Física	Masculino (a)	0,6x140= $\frac{120}{200}x140 = 84$
Física	Feminino (b)	0,6x60= $\frac{120}{200}x60 = 36$
Ciências Sociais	Masculino (c)	0,4x140= $\frac{80}{200}x140 = 56$
Ciências Sociais	Feminino (d)	0,4x60= $\frac{80}{200}x60 = 24$

Tabela de **freqüências 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

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 o quadrado das diferenças entre as frequências observadas e esperadas, divididas pelas freqüências esperadas.

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

Se o Qui-quadrado for igual a zero, então não existe associação entre as variáveis. O Qui-quadrado não mede força de associação.

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

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

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

Exemplo 6

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

Peso ao nascer	Óbito	Sobrevida	Total
Baixo peso (<2500)	24	13	37
Não baixo peso (2500 e mais)	3	10	13
Total	27	23	50

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

O qui-quadrado é diferente de zero. Pode-se suspeitar da existência de associação entre as variáveis.

Calculando-se as porcentagens pode-se entender melhor a associação

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

Peso ao nascer	Óbito		Sobrevida		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.

A tabela sugere que a proporção de óbitos é maior entre os recém-nascidos de baixo peso

Força da associação

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

A associação entre peso ao nascer e condição de sobrevivência é forte.

Exercício 11

Os dados a seguir são de pesquisa que estuda a associação entre amamentação ao seio e Diabetes Mellitus tipo I . Local X. Ano Y.

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.

Com base nos dados apresentados

a) Calcule e apresente o qui-quadrado de Pearson.
b) Os dados sugerem existência de associação entre as variáveis?
c) Se existir associação, calcule o coeficiente de associação para investigar a força da associação.
d) Discuta os resultados

Cálculo das frequências esperadas

Casela 35 (Não – Casos)

$\frac{x}{346} = \frac{52}{692}$ ⟶ $x = \frac{346}{692} = 26$

Casela 311 (Sim – Casos)

$\frac{x}{346} = \frac{640}{692}$ ⟶ $x = \frac{346x640}{692} = 320$

Casela 17 (Não – Controles)

$\frac{x}{346} = \frac{52}{692}$ ⟶ $x = \frac{346}{692} = 26$

Casela 329 (Sim – Controles)

$\frac{x}{346} = \frac{640}{692}$ ⟶ $x = \frac{346x6}{692} = 320$

Cálculo das frequências esperadas

Alimentação ao seio	Grupo (caso/controle)	Número esperado (E)
Não	Casos (a)	$x = \frac{346x640}{692} = 320$
Sim	Casos (b)	$x = \frac{346x640}{692} = 320$
Não	Controles (c)	$x = \frac{346x52}{692} = 26$
Sim	Controles (d)	$x = \frac{346x640}{692} = 320$

Cálculo do qui-quadrado de Pearson

Valores observados O	Valores esperados E	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
35	26	9	81	3,115
311	320	-9	81	0,253
17	26	-9	81	3,115
329	320	9	81	0,253

Qui-quadrado=6,736

Coeficiente de associação de Yule

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

$$Y = \frac{35x329 - 311x17}{35x329 + 311x17} = \frac{6228}{16802} = 0{,}371$$

Os dados a seguir são de pesquisa que estuda a associação entre amamentação ao seio e Diabetes Mellitus tipo I . Local X. Ano Y.

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.

Distribuição de pessoas segundo presença/ausência de Diabetes Mellitus tipo 1 e tipo de amamentação. São Paulo, 1996

Amamentação	Casos		Controles		Total	
Ao seio	n	%	n	%	n	%
Não	35	10,1	17	4,9	52	7,5
Sim	311	89,9	329	95,1	640	92,5
Total	346	100	346	100	692	100

Fonte: Gimeno SGA. Consumo de leite e o Diabetes Mellitus insulino-dependente:um estudo caso-controle. Tese de doutorado, 1996.

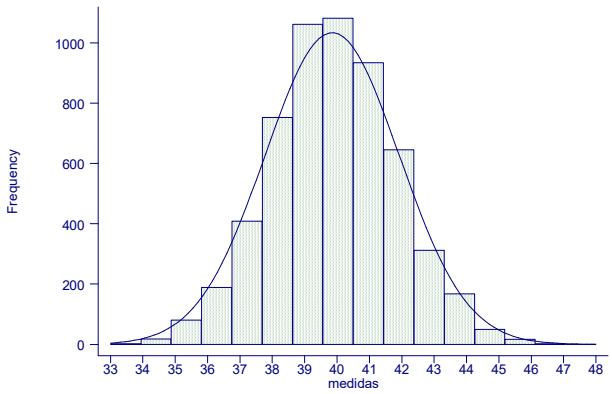
Aula 4

Distribuição normal, 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)

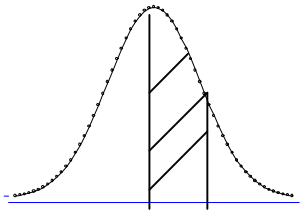
Propriedades:

- 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.

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

Exemplo 7

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.



40 43 X

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

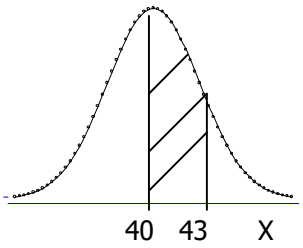
$$P(40 < X < 43) = \int_{40}^{43} \frac{1}{2\sqrt{2\pi}} e^{[-\frac{(x-40)^2}{2x4}]} dx$$

Quantos desvio padrão 43 está em torno da média?

Normal reduzida:

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

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



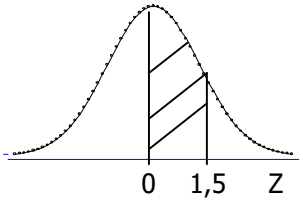
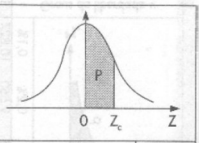


Tabela da Distribuição Normal

Tabela III – Distribuição Normal Padrão

$Z \sim N(0, 1)$

Corpo da tabela dá a probabilidade p, tal que $p = P(0 < Z < Z_c)$



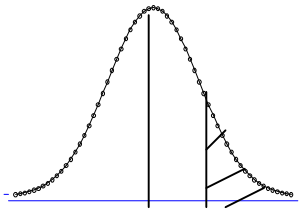
parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983	3,5
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989	3,6
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992	3,7
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995	3,8
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997	3,9
4,0	49997	49997	49997	49997	49997	49997	49998	49998	49998	49998	4,0
4,5	49999	50000	50000	50000	50000	50000	50000	50000	50000	50000	4,5

Utilizando a tabela da curva normal reduzida, P(0<Z<1,5)=0,43319=43,3%.

Exemplo 8

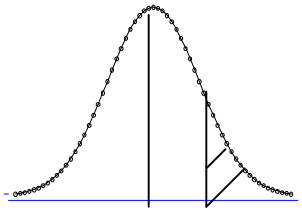
Com base na distribuição de X~N (μ =40, σ =2), calcular:

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



40 43 X

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



0 1,5 Z

Utilizando a tabela da curva normal reduzida, P(Z>1,5)=0,5-0,43319=0,06681= 6,7%

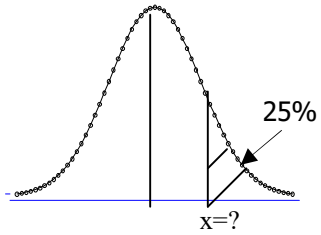
b) a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax entre 35 e 40 polegadas.

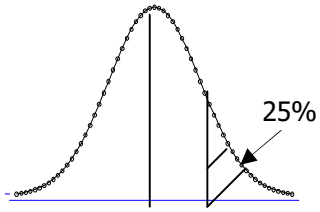
$P(35 \leq X \leq 40) = P\left(\frac{35-40}{2} \leq \frac{X-\mu}{\sigma} \leq \frac{40-4}{2}\right) = P(-2,5 \leq Z \leq 0)$=0,49379 ou 49,4%

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

$P(X \leq 35) = P\left(\frac{X-\mu}{\sigma} \leq \frac{35-4}{2}\right) = P(Z \leq -2,5)$=0,5-0,49379=0,00621 ou 0,6%

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





z=?

Utilizar a transformação Z:

$$Z = \frac{\text{x - } \mu}{\sigma}$$

Para p=0,24857, z=0,67; para 0,25175, z=0,68.
As distâncias são: 0,25-0,24857=0,00143 e 0,25175-0,25=0,00175.
Por meio do cálculo de diferenças observa-se que o valor 0,24857 está mais próximo de 0,25
Portanto será utilizado o valor de z=0,67

$$0{,}67 = \frac{x - 40}{2}$$

$$x = 2x0{,}67 + 40 = 41{,}34\ polegadas$$

Exercício 12

Considerar o imc médio da população idosa do município de São Paulo seguindo uma distribuição normal com média 28 kg/m^2 e desvio padrão 4 kg/m^2. Sorteia-se um indivíduo; qual a probabilidade de que ele tenha

a) imc entre a média e 32 kg/m^2
b) imc entre a média e 24 kg/m^2
c) imc entre 24 kg/m^2 e 32 kg/m^2
d) imc abaixo de 24 kg/m^2
e) imc acima de 24 kg/m^2

Respostas

a) $P(28 \leq X \leq 32) = P\left(\frac{28-28}{4} \leq \frac{X-\mu}{\sigma} \leq \frac{32-28}{4}\right) = P(0 \leq Z \leq 1)$=0,34134 ou 34,1%
b) $P(24 \leq X \leq 28) = P\left(\frac{24-2}{4} \leq \frac{X-\mu}{\sigma} \leq \frac{28-2}{4}\right) = P(-1 \leq Z \leq 0)$= 0,34134 ou 34,1%
c) $P(24 \leq X \leq 32) = P\left(\frac{24-2}{4} \leq \frac{X-\mu}{\sigma} \leq \frac{32-2}{4}\right) = P(-1 \leq Z \leq 1)$=0,34134+0,34134=0,68268 ou 68,3%
d) $(X \leq 24) = P\left(\frac{X-\mu}{\sigma} \leq \frac{24-28}{4}\right) = P(Z \leq -1)$=0,5-0,34134=0,15866 ou 15,9%
e) $(X \geq 24) = P\left(\frac{X-\mu}{\sigma} \geq \frac{24-28}{4}\right) = P(Z \geq -1)$=0,5+0,34134=0,84134 ou 84,1%

Distribuição amostral da média

Considerar a população de idosos do município de São Paulo e que é de interesse estudar o imc deste grupo populacional.

Supor ainda que o imc médio e o desvio padrão da população são conhecidos e iguais a $\mu = 28kg/m^2$ e $\sigma = 4kg/m^2$

Sorteia-se uma amostra de tamanho 1000 e calcula-se o imc médio amostral

amostras	estimativas	
n_1=1000	$\bar{x}_1 = 29{,}4$	Devolve-se os participantes para a população e sorteia-se nova amostra
n_2=1000	$\bar{x}_2 = 27{,}5$	
.	.	.
.	.	.
.	.	.
n_∞=1000	$\bar{x}_\infty = 28{,}7$	

Observa-se que o valor médio obtido para cada amostra não será necessariamente igual aos demais, sendo assim tem-se que a média ($\bar{X}$), antes de ser calculada pode assumir qualquer valor sendo, portanto, uma variável aleatória.

Se a média ($\bar{X}$) é uma variável aleatória então ela terá uma distribuição. Qual é a distribuição da média? É necessário fazer todas as possíveis amostras para saber tal distribuição? A resposta é Não!

Existe um teorema (Teorema Central do Limite) que afirma que

Se X é variável aleatória com média μ e variância σ^2 , então $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

Exemplo 9

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

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

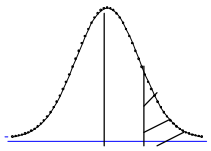
Exercício 13

Considerar o imc médio da população idosa do município de São Paulo seguindo uma distribuição normal com média $\mu = 28kg/m^2$ e desvio padrão $\sigma = 4kg/m^2$. Sorteia-se uma amostra de 25 indivíduos; qual a probabilidade de que o imc médio esteja

a) entre a média e 29 kg/m^2
b) entre a média e 27,5 kg/m^2
c) entre 27,5 kg/m^2 e 29 kg/m^2
d) abaixo de 26 kg/m^2
e) acima de 29 kg/m^2

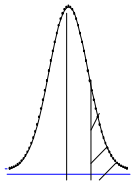
X:IMC

a)



$\mu = 28$ X

$\sigma = 4$



$\bar{X}$

$\mu_{\bar{X}} = 28$

$\sigma_{\bar{x}} = \frac{4}{\sqrt{25}} = 0{,}8$

$P(28 < \bar{X} < 29) = P(\frac{28-28}{\frac{4}{\sqrt{25}}} < Z < \frac{29-28}{\frac{4}{\sqrt{25}}})$ =P(0<Z<1,25); pela tabela da N(0,1), P(0<Z<1,25)=0,39435 ou

39,4%

b)$P(27{,}5 \leq \bar{X} \leq 28) = P\left(\frac{27{,}5-28}{0{,}8} \leq Z_{\bar{X}} \leq \frac{28-28}{0{,}8}\right) = P(-0{,}625 \leq Z_{\bar{X}} \leq 0) = 0{,}23565\ ou\ 23{,}6\%$

c)$P(27{,}5 \leq \bar{X} \leq 29) = P\left(\frac{27{,}5-28}{0{,}8} \leq Z_{\bar{X}} \leq \frac{29-2}{0{,}8}\right) = P(-0{,}625 \leq Z_{\bar{X}} \leq 1{,}25) = 0{,}23565 + 0{,}39435 =$

$0{,}63\ ou\ 63\%$

d)$P(\bar{X} \leq 26) = P\left(Z_{\bar{X}} \leq \frac{26-28}{0{,}8}\right) = P(Z_{\bar{X}} \leq -2{,}5) = 0{,}5 - 0{,}49379 = 0{,}00621\ ou\ 0{,}62\%$

e)$P(\bar{X} \geq 29) = P\left(Z_{\bar{X}} \geq \frac{29-28}{0{,}8}\right) = P(Z_{\bar{X}} \geq 1{,}25) = 0{,}5 - 0{,}39435 = 0{,}10565\ ou\ 10{,}6\%$

Aula 4

Estimativa de parâmetros populacionais por intervalo; Distribuição t de Student

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.

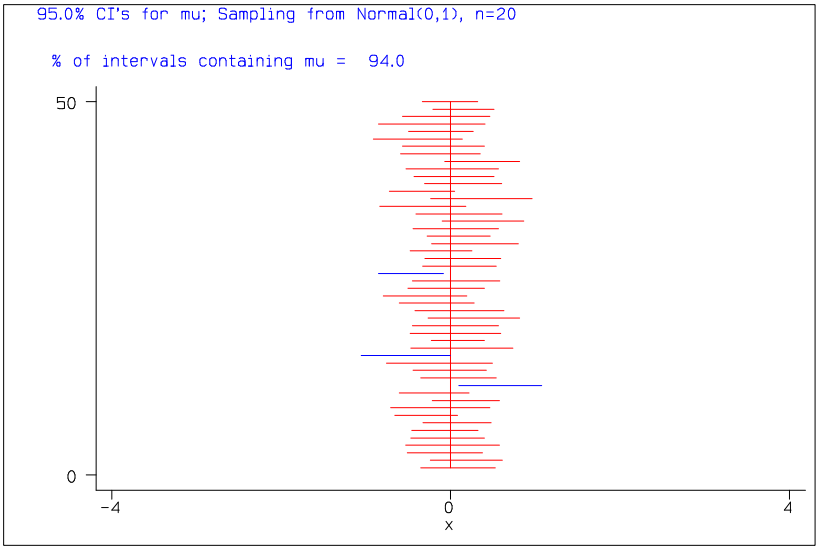
O grau de confiança tem origem na probabilidade associada ao processo de construção do intervalo antes de se obter o resultado amostral.

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 específico intervalo de confiança cobre o verdadeiro valor. Só podemos aplicar o conceito freqüentista 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.

Representação gráfica



A linha vertical representa o parâmetro populacional. O gráfico foi gerado por programa de computacional. 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.

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 gramas) 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.

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

$P(-z \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z) = 1 - \alpha$.

Para $\alpha = 5\%$, $P(-1{,}96 \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +1{,}96) = 0{,}95$

$$P(-1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq +1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

$$P(-\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

Multiplicando tudo por -1

$$P(\overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$$

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 ($\hat{\bar{X}} = \bar{x}$), 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.

$$\text{IC}(95\%) : (\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}};\ \ \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}})$$

Exemplo 17
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 85 mmHg e o desvio padrão populacional (σ) foi 9 mmHg. 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)

Exemplo 18
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.

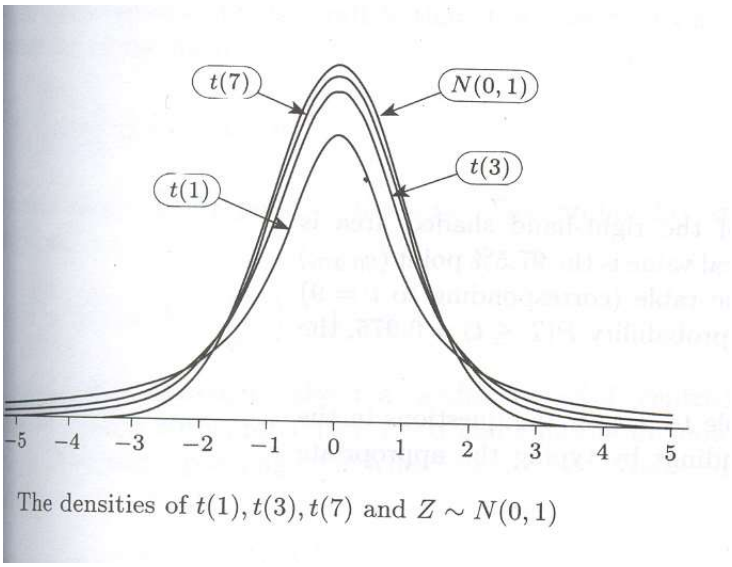
Intervalo de confiança para a média populacional com variância populacional desconhecida

$$\text{IC}(\mu) : \bar{x} - t_{n-1,\alpha/2} \cdot \frac{S_x}{\sqrt{n}}; \bar{x} + t_{n-1,\alpha/2} \cdot \frac{S_x}{\sqrt{n}}$$

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 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.

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 ($\overline{x}$) foi igual a 85 mmHg e o desvio padrão amostral (s) foi 12 mmHg. 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)

Exemplo 19
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.

Aula 5

Teste de hipóteses de uma média populacional (μ) com variância conhecida

Proposta clássica de Neyman e Pearson

Neyman e Pearson propuseram uma abordagem, para a tomada de decisão, que envolve a fixação, antes da realização do experimento, das hipóteses nula e alternativa, e fixação de valores de probabilidade de ocorrência de erros de decisão.

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

$$H_0 : \mu_{SDIG} = \mu_{Sadia} \qquad \text{ou} \qquad H_0 : \mu_{SDIG} = 3000$$
$$H_a : \mu_{SDIG} < \mu_{Sadia} \qquad\qquad H_a : \mu_{SDIG} < 3000$$

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

$\alpha = \Pr obabilidad e(erro\ tipoI)$ = Probabilidade (Rejeitar H_0 e H_0 é verdade)

$\beta = \Pr obabilidad e(erro\ tipoII)$ = 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

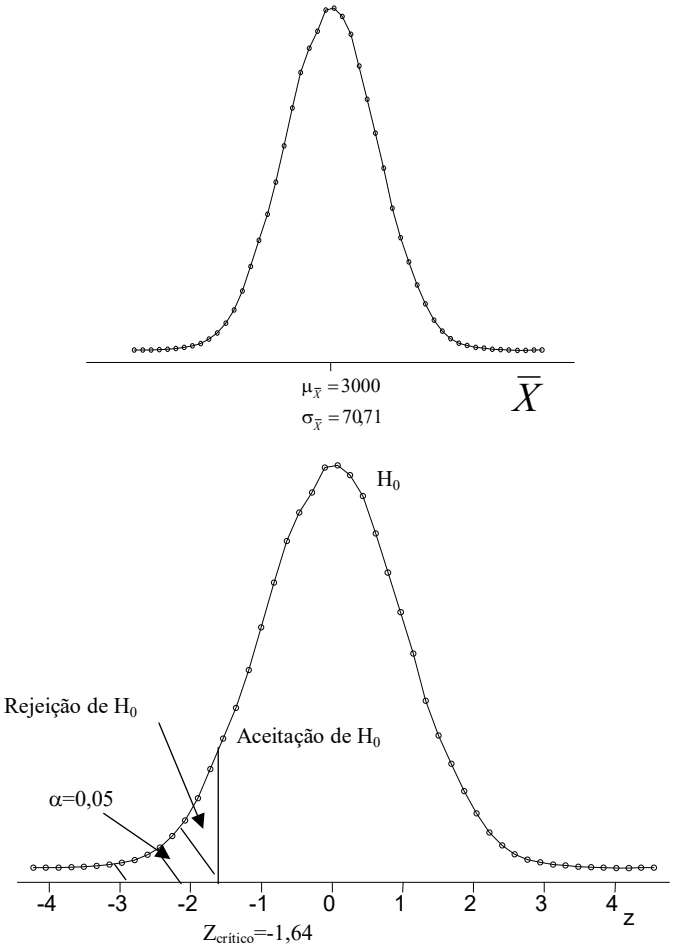
Distribuição de probabilidade

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, tem-se: $\overline{X} \sim N(\mu_{\overline{X}} = 3000; \sigma_{\overline{X}} = \frac{500}{\sqrt{50}})$

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.

É possível realizar o teste comparando a média observada na amostra $(\bar{x}_{obs} = 2800)$ 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 $\bar{x}_{critico}$.

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 $H_a : \mu_{SDIG} \neq \mu_{Sadias}$

Cálculo do tamanho mínimo da amostra

Para uma hipótese monocaudal, onde $\begin{cases} H_0 : \mu_{SDIG} = 3000 \\ H_a : \mu_{SDIG} < 3000 \end{cases}$

$n = \frac{(Z_\alpha + Z_\beta)^2}{d^2}$, 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.

Teste de hipóteses de uma média populacional (μ) (com variância conhecida)

Abordagem de Fisher

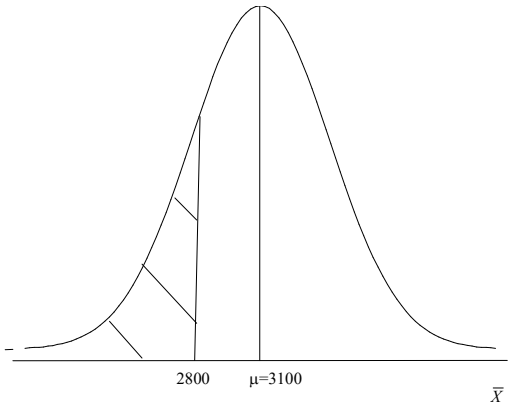
Situação:
Estudos mostram que crianças sadias possuem peso médio (m) ao nascer igual a 3100 gramas e desvio padrão $\sigma = 610 gramas$. Suspeita-se que crianças que nascem com síndrome de desconforto idiopático grave possuem peso ao nascer abaixo do peso ao nascer da população de crianças sadias.

Proposição (equivalente à H_0 de Neynman e Pearson): 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.

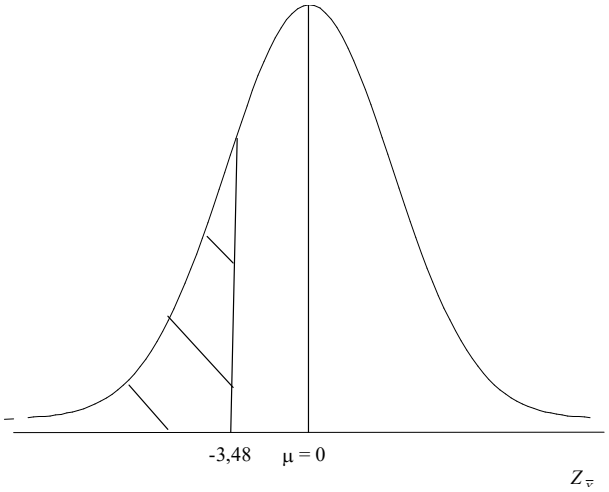
Supondo-se que as crianças da amostra (com síndrome) vêm de uma população com mesma dispersão do peso ao nascer de crianças sadias, teste a hipótese de que crianças com síndrome de desconforto idiopático grave possuem peso médio ao nascer igual ao peso médio ao nascer de crianças sadias.

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.

$$P(\overline{X} \le 2800) = P(\frac{\overline{X} - m}{\sigma_{\overline{X}}} \le \frac{2800 - 3100}{\frac{610}{\sqrt{50}}}) = P(Z_{\overline{X}} \le \frac{-300}{86{,}27}) = P(Z_{\overline{X}} \le -3{,}48)$$



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 gramas. Possivelmente a amostra vem de uma população com média menor que 3100 gramas. 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).

Exemplo 20

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? Realize um teste de hipóteses segundo a abordagem de Fisher para responder a pergunta.

Exemplo 21

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 desconfiou que o mesmo estava sendo pouco consumido 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.

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.

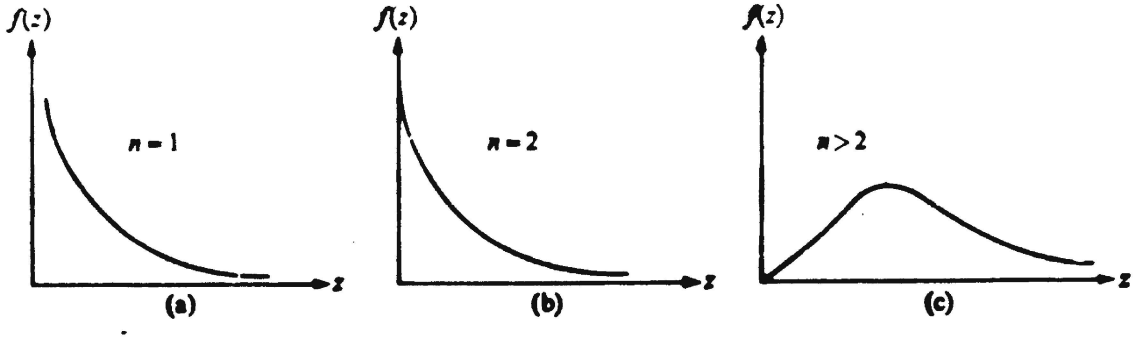
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

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.

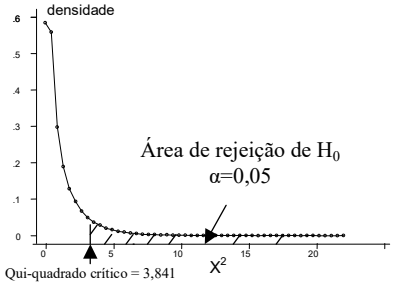
Distribuição qui-quadrado ($\chi^2_{(n-1)}$) com (n-1) graus de liberdade

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.

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.

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
				$\chi^2 = 6{,}762$	$\chi^2_{corrigido} = 5{,}19$

Exemplo
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.

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.

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.

Exemplo 22
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]

Gabarito

Aula 1

Exemplo 1 - Classificar quanto à natureza, as seguintes variáveis:

Solução:

Variável	**Tipo (natureza)**
Condição de saúde (doente, não doente)	Qualitativa nominal
Tipo de parto (normal, cesário)	Qualitativa nominal
Nível de colesterol sérico (mg/100cc)	Quantitativa contínua
Tempo de um procedimento cirúrgico (minutos)	Quantitativa contínua
Número de praias consideradas poluídas	Quantitativa discreta

Exemplo 2 –
São fornecidos dados de altura de uma amostra de 351 mulheres idosas selecionadas aleatoriamente de uma comunidade para um estudo de osteoporose.
Solução:

a) Distribuição de mulheres idosas segundo a altura, Local X. Ano Y.

Altura (cm)	Nº	%
140|--145	1	0,3
145|--150	11	3,1
150|--155	52	14,8
155|--160	109	31,1
160|--165	106	30,2
165|--170	50	14,3
170|--175	18	5,1
175|--180	4	1,1
Total	351	100

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

b) Pode-se observar na tabela que aproximadamente 60% das mulheres idosas têm a estatura entre 155cm e 164,9cm.

Exemplo 3
Solução:

a) Calculando-se o percentual "fixando" o hábito de fumar e investigando a distribuição dos níveis de β-caroteno entre fumantes e não fumantes; em outras palavras, comparando-se fumantes e não fumantes quanto aos níveis de β-caroteno.

Distribuição de gestantes segundo níveis de β-caroteno (mg/L) e hábito de fumar.

β-caroteno (mg/L)	Fumante		Não fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46	79,3	74	56,1	120	63,2
Normal (0,214 – 1,00)	12	20,7	58	43,9	70	36,8
Total	58	100	132	100	190	100

b) Calculando-se o percentual "fixando" o nível de β-caroteno e investigando a distribuição do hábito de fumar entre gestantes com nível baixo e normal de β-caroteno.

Distribuição de gestantes segundo níveis de β-caroteno (mg/L) e hábito de fumar.

β-caroteno (mg/L)	Fumante		Não fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46	38,3	74	61,7	120	100
Normal (0,214 – 1,00)	12	17,1	58	82,9	70	100
Total	58	30,5	132	69,5	190	100

c) interpretando-se a tabela do item **a**:
Do total de fumantes, 79,3% apresentam nível baixo de β-caroteno. Entre não fumantes este percentual é de 56,1%. Parece existir associação; a proporção de pessoas com nível baixo de β-caroteno parece maior entre fumantes.
Interpretando-se a tabela do item **b**: Entre as gestantes com nível baixo de β-caroteno, 38,3% eram fumantes enquanto que entre as que tinham nível normal de β-caroteno, este percentual era de 17,1%. Pode ser que exista associação; a proporção de fumantes parece maior entre as gestantes com nível baixo de β-caroteno.

Exemplo 6

$\overline{x}_B =$ 210,3mg/100ml

Exemplo 7

$\overline{x}_{Meninos} =$ 2042,2 kcal.

$\overline{x}_{Meninas} =$ 1690 kcal

Exemplo 8

Ordenando-se os valores:

137	153	175	185	194	212	224	242	250	263
148	168	184	188	202	213	226	246	252	344

Mediana: 207 kcal

Exemplo 9
Com os dados do exemplo 7, calcule a quantidade mediana de energia para os meninos e para as meninas:

Meninos: mediana = 1866 kcal; meninas: mediana = 1553 kcal

Exemplo 10

Variância: $s^2 =$ 2336,7 mg/100ml^2
Desvio padrão s=48,3 mg/100ml
Coeficiente de Variação de Pearson CV=23,0%

Exemplo 12

a) 207/445 = 0,46
b) 34/106 = 0,32
c) $\frac{207}{445} \Big/ \frac{34}{106} = \frac{106x207}{445x34} = 1,45$
d) 0,46-0,32=0,14
e) A incidência de um ou mais episódios de doença respiratória parece ser maior entre as crianças com padrão de amamentação mamadeira e peito. As crianças com este padrão de amamentação apresentam uma incidência 44% maior que aquelas com padrão de amamentação somente peito. Pela diferença das incidências pode-se dizer que 14% de episódios de doença respiratória poderiam ser evitadas se fosse adotado o padrão de "somente peito". É possível que exista associação entre padrão de amamentação e episódio de doença respiratória uma vez que a razão de incidência é diferente de 1,0. Somente após o teste de hipótese será possível averiguar se estas medidas, do ponto de vista estatístico, são iguais ou diferentes de 1,0.

Exemplo 13

Solução:

a) $\frac{38}{790} = 0{,}048$ ou 4,8%

b) $\frac{39}{928} = 0{,}042$ ou 4,2%

c) $\frac{\frac{38}{790}}{\frac{39}{928}} = \frac{38x928}{39x790} = 1{,}14$

d) 0,048-0,042 = 0,006 ou 0,6%

e) A incidência de doença coronariana entre homens com alto consumo de café é 1,14 vezes a incidência entre os que consomem moderadamente.

Aula 3

Exemplo 14

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

b) $P(35 < X < 40) = P(\frac{35-40}{2} < \frac{X-\mu}{\sigma} < \frac{40-40}{2}) = P(-2{,}5 < Z < 0)$

Utilizando a tabela da curva normal reduzida: P(-2,5<Z<0)=0,49379 = 49,4%

c) $P(X < 35) = P(\frac{X-\mu}{\sigma} < \frac{35-40}{2}) = P(Z < -2{,}5)$

Utilizando a tabela da curva normal reduzida: P(Z<-2,5)=0,5-0,49379=0,0062 ou 0,6%

d) 0,25 – olhar tabela da normal

Z = 0,675

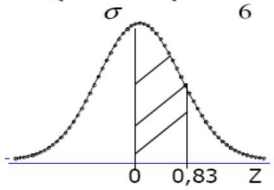
$Z = \frac{x-m}{\sigma} = \frac{x-40}{2}$

$0{,}675 = \frac{x-40}{2}$ = 1,35 = x - 40 = x = 41,35 polegadas

Exemplo 15

a) X: altura; X~N(160,6)

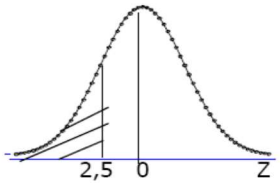
$$P(160 < X < 165) = P(\frac{160-160}{6} < \frac{X-\mu}{\sigma} < \frac{165-160}{6}) = P(0 < Z < 0{,}83)$$



Utilizando a tabela da curva normal reduzida, P(0<Z<0,83)=0,29673 ou 29,7%

b) X: altura; X~N(160,6)

$$P(X < 145) = P(\frac{X - \mu}{\sigma} < \frac{145 - 160}{6}) = P(Z < -2,5)$$



Utilizando a tabela da curva normal reduzida, P(Z<-2,5)=0,5-0,49379=0,0062 ou 0,6%

c) $P(X > 170) = P(\frac{170 - 160}{6} > \frac{X - \mu}{\sigma}) = P(1,66 > Z)$

Utilizando a tabela da curva normal reduzida, P(1,66 > Z)=0,5-0,45154=0,048 = 4,8%.

Aula 4

Exemplo 18
Solução:

$\left(5,3 - 1,96\frac{0,6}{\sqrt{16}}; 5,3 + 1,96\frac{0,6}{16}\right)$; IC 95%: (5,006mg; 5,594mg)

Exemplo 19
Solução:

a) $\left(56 - 2,064\frac{8}{\sqrt{25}}; 56 + 2,064\frac{8}{\sqrt{25}}\right)$; IC 95%: (52,7; 59,3)

b) Deposita-se neste intervalo 95% de confiança de que cobrirá o verdadeiro valor do peso médio.

Aula 5

Exemplo 20

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 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.

$$\begin{cases} H_0 : \mu_{consumo} = 8kg \\ H_a : \mu_{consumo} < 8kg \end{cases}$$

Estatística do teste: $Z_{\bar{X}obs} = \frac{\bar{x}_{obs} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{7,2 - 8}{\frac{2}{\sqrt{25}}} = -2$

Teste monocaudal à esquerda. Para nível de significância de 5%, obtém-se os valores de $Z_{crítico}$: -1,64.

Decisão

Como Z_{obs} está na área de rejeição de H_0, decide-se rejeitar H_0.

Conclusão

Foi encontrada diferença estatisticamente significante entre o consumo médio mensal per capita do produto para nível de significância $\alpha = 0{,}05$. Portanto, a diretoria deve tirar este item de produção.

Exemplo 21

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? Realize um teste de hipóteses segundo a abordagem de Fisher para responder a pergunta.

Proposição inicial: O nível médio populacional de pessoas com deficiência de vitamina k (μ_k)=20mg/100ml; $Z_{observado}$ = -2,38; teste bicaudal; p=0,8%
Conclusão: os dados não são compatíveis com uma distribuição que tem nível de protrombina igual a 20mg/dl. Pode-se dizer que pacientes com deficiência de vitamina K vêm de uma população com nível médio de protrombina menor que pessoas da população sem deficiência (p= 0,008).

Exemplo 22

H_0: não existe associação
H_a: existe associação

Abordagem de Neyman e Pearson

Cálculo do qui-quadrado

Valores observados (O)	Valores esperados (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$	$\frac{(\lvert O-E\rvert-0{,}5)^2}{E}$
56	47,48	8,52	72,5904	1,528863	1,35468408
22	30,52	-8,52	72,5904	2,378453	2,10748362
84	92,52	-8,52	72,5904	0,78459	0,69520536
68	59,48	8,52	72,5904	1,220417	1,08137861
		Qui-quadrado (χ^2)=		5,912325	$\chi^2_{corrigido} = 5{,}23875$

Obs: a resposta está com muitas casas decimais mas 2 seriam suficientes desde que sejam feitas aproximações corretamente.

Decisão:
O valor do qui-quadrado calculado é maior do que o valor do qui-quadrado crítico para 1 grau de liberdade ($\chi^2_{crítico}$ = 3,841) e nível de significância de 5%, portanto, rejeita-se H_0.

Conclusão: Pode-se dizer que existe associação entre níveis de betacaroteno (mg/L) e hábito de fumar.

Abordagem de Fisher

Calculando-se o valor de p pelo Excel, para 1 gl, o valor de $2{,}0\% \leq p_{\text{ corrigido}} \leq 2{,}5\%$

Existe evidência contrária à independência entre as variáveis. A associação observada ocorre não devido ao acaso. Pode-se dizer que os dados são compatíveis com existência de associação entre níveis de β-caroteno (mg/L) e hábito de fumar ($2{,}0\% \leq p_{\text{ corrigido}} \leq 2{,}5\%$).

Exemplos complementares

Tabelas

Observe como os dados são apresentados na tabela abaixo.

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.]*

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.

Gráficos

Observe e analise o gráfico abaixo.

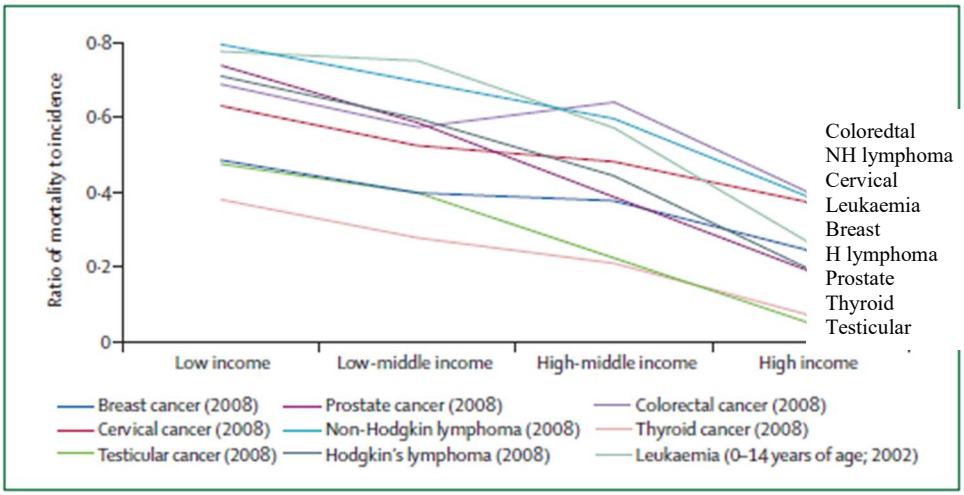


Figure: **Ratio of mortality to incidence in a specific year by cancer type and country income**
Case fatality (calculated by approximation from the ratio of mortality to incidence in a specific year) is much lower in high-income countries than in low-income countries for cancers that are treatable, such as childhood leukaemia (0·26 *vs* 0·78) and testicular cancer (0·05 *vs* 0·47), treatable if detected early, such as breast cancer (0·24 *vs* 0·48), or preventable, such as cervical cancer (0·37 *vs* 0·63). Estimates are based on International Agency for Research on Cancer GLOBOCAN data for 2002 and 2008 (http://globocan.iarc.fr).[3,4]

Farmer P et al. Expansion of cancer care and control in countries of low and middle income: a call to action. The Lancet. 2010; v367:1186-1193.

Medidas de tendência central e dispersão

Analise as tabelas a seguir:

Tabela 1 *- Valores mínimo e máximo e médias dos parâmetros dietéticos obtidos através de dois recordatórios de 24-horas.*

Dietetic Variables	Crude values			
	Mean	Standard Deviation	Minimum	Maximum
Energy (kcal)	2,326.18	883.50	1,045.20	5,938.42
Fat (g)	89.03	38.30	35.42	253.08
Carbohydrate (g)	305.31	121.36	117.68	744.61
Protein (g)	82.15	32.84	28.89	202.29

***Tabela 2** - Valores de variabilidade intra- (S_w) e interpessoal (S_b), coeficientes de variância (CV%) e razões de variância (λ).*

Dietetic Variables	S_w	CV_w (%)	S_b	CV_b (%)	λ
Energy (kcal)	832.34	35.78	658.91	28.33	1.60
Fat (g)	39.98	44.90	25.85	29.03	2.39
Carbohydrate (g)	108.95	35.69	93.76	30.71	1.35
Protein (g)	34.96	42.56	21.61	26.31	2.62

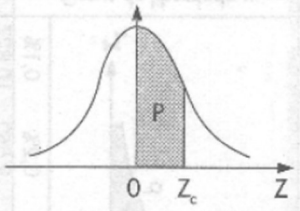
Costa MMF, Takeyama L, Voci SM, Slater B, Silva MV. Within- and between-person variations as determinant factors to calculate the number of observations to estimate usual dietary intake of adolescents. Rev. bras. epidemiol. 2008; 11(4):541-548.

Tabela da Distribuição Normal

Tabela III — Distribuição Normal Padrão

$Z \sim N(0, 1)$

Corpo da tabela dá a probabilidade p, tal que $p = P(0 < Z < Z_c)$



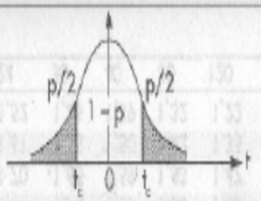
parte inteira e primeira decimal de Z_c	Segunda decimal de Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062	1,8
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670	1,9
2,0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169	2,0
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574	2,1
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899	2,2
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158	2,3
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361	2,4
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520	2,5
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643	2,6
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736	2,7
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807	2,8
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861	2,9
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900	3,0
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929	3,1
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950	3,2
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965	3,3
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976	3,4
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983	3,5
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989	3,6
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992	3,7
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995	3,8
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997	3,9
4,0	49997	49997	49997	49997	49997	49997	49998	49998	49998	49998	4,0
4,5	49999	50000	50000	50000	50000	50000	50000	50000	50000	50000	4,5

Tabela da Distribuição t de *Student*

Tabela V – Distribuição *t* de Student

Corpo da tabela dá os valores t_c tais que $P(-t_c < t < t_c) = 1 - p$.

Para $\nu > 120$, usar a aproximação normal.



Graus de liberdade ν	p = 90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	Graus de liberdade ν
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	15,894	31,821	63,657	318,309	636,619	1
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	4,849	6,965	9,925	22,327	31,598	2
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	3,482	4,541	5,841	10,214	12,924	3
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	2,998	3,747	4,604	7,173	8,610	4
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	2,756	3,365	4,032	5,893	6,869	5
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	2,612	3,143	3,707	5,208	5,959	6
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,517	2,998	3,499	4,785	5,408	7
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041	8
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,398	2,821	3,250	4,297	4,781	9
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,359	2,764	3,169	4,144	4,587	10
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,328	2,718	3,106	3,025	4,437	11
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,303	2,681	3,055	3,930	4,318	12
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,282	2,650	3,012	3,852	4,221	13
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,264	2,624	2,977	3,787	4,140	14
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,248	2,602	2,947	3,733	4,073	15
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,235	2,583	2,921	3,686	4,015	16
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,224	2,567	2,898	3,646	3,965	17
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,214	2,552	2,878	3,610	3,922	18
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,205	2,539	2,861	3,579	3,883	19
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,197	2,528	2,845	3,552	3,850	20
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,189	2,518	2,831	3,527	3,819	21
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,183	2,508	2,819	3,505	3,792	22
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,177	2,500	2,807	3,485	3,768	23
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,172	2,492	2,797	3,467	3,745	24
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,166	2,485	2,787	3,450	3,725	25
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,162	2,479	2,779	3,435	3,707	26
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,158	2,473	2,771	3,421	3,690	27
28	0,127	0,256	0,389	0,530	0,684	0,855	1,056	1,313	1,701	2,048	2,154	2,467	2,763	3,408	3,674	28
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,150	2,462	2,756	3,396	3,659	29
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,147	2,457	2,750	3,385	3,646	30
35	0,126	0,255	0,388	0,529	0,682	0,852	1,052	1,306	1,690	2,030	2,133	2,438	2,724	3,340	3,591	35
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,123	2,423	2,704	3,307	3,551	40
50	0,126	0,254	0,387	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,109	2,403	2,678	3,261	3,496	50
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,099	2,390	2,660	3,232	3,460	60
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,076	2,358	2,617	3,160	3,373	120
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,054	2,326	2,576	3,090	3,291	∞
	p = 90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	4%	2%	1%	0,2%	0,1%	

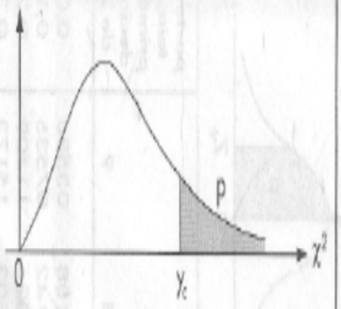
Tabela da Distribuição Qui-quadrado

Tabela IV – Distribuição Qui-quadrado

$Y \sim \chi^2(\nu)$

Corpo da tabela dá os valores y_c tais que $P(Y > y_c) = p$.

Para valores $\nu > 30$, use a aproximação normal dada no texto.



Graus de liberdade ν	p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	Graus de liberdade ν
1	0,0³16	0,0⁴63	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	4,218	5,024	5,412	6,635	9,550	10,827	1
2	0,020	0,040	0,051	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	6,438	7,378	7,824	9,210	12,429	13,815	2
3	0,115	0,185	0,216	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	8,311	9,348	9,837	11,345	14,796	16,266	3
4	0,297	0,429	0,484	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	10,026	11,143	11,668	13,277	16,924	18,467	4
5	0,554	0,752	0,831	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	11,644	12,832	13,388	15,086	18,907	20,515	5
6	0,872	1,134	1,237	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	13,198	14,449	15,033	16,812	20,791	22,457	6
7	1,239	1,564	1,690	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	14,703	16,013	16,622	18,475	22,601	24,322	7
8	1,646	2,032	2,180	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	16,171	17,534	18,168	20,090	24,352	26,125	8
9	2,088	2,532	2,700	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	17,608	19,023	19,679	21,666	26,056	27,877	9
10	2,558	3,059	3,247	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	19,021	20,483	21,161	23,209	27,722	29,588	10
11	3,053	3,609	3,816	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	20,412	21,920	22,618	24,725	29,354	31,264	11
12	3,571	4,178	4,404	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	21,785	23,337	24,054	26,217	30,957	32,909	12
13	4,107	4,765	5,009	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	23,142	24,736	25,472	27,688	32,535	34,528	13
14	4,660	5,368	5,629	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	24,485	26,119	26,873	29,141	34,091	36,123	14
15	5,229	5,985	6,262	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	25,816	27,488	28,259	30,578	35,628	37,697	15
16	5,812	6,614	6,908	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	27,136	28,845	29,633	32,000	37,146	39,252	16
17	6,408	7,255	7,564	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	28,445	30,191	30,995	33,409	38,648	40,790	17
18	7,015	7,906	8,231	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	29,745	31,526	32,346	34,805	40,136	42,312	18
19	7,633	8,567	8,906	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	31,037	32,852	33,687	36,191	41,610	43,820	19
20	8,260	9,237	9,591	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	32,321	34,170	35,020	37,566	43,072	45,315	20
21	8,897	9,915	10,283	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	33,597	35,479	36,343	38,932	44,522	46,797	21
22	9,542	10,600	10,982	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	34,867	36,781	37,659	40,289	45,962	48,268	22
23	10,196	11,293	11,688	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	36,131	38,076	38,968	41,638	47,391	49,728	23
24	10,856	11,992	12,401	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	37,389	39,364	40,270	42,980	48,812	51,179	24
25	11,524	12,697	13,120	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	38,642	40,646	41,566	44,314	50,223	52,620	25
26	12,198	13,409	13,844	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	39,889	41,923	42,856	45,642	51,627	54,052	26
27	12,879	14,125	14,573	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	41,132	43,194	44,140	46,963	53,022	55,476	27
28	13,565	14,847	15,308	16,928	18,939	21,588	23,647	27,336	31,319	34,027	37,916	41,337	42,370	44,461	45,419	48,278	54,411	56,893	28
29	14,258	15,574	16,047	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	43,604	45,722	46,693	49,588	55,792	58,302	29
30	14,953	16,306	16,791	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	44,834	46,979	47,962	50,892	57,167	59,703	30
	p = 99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	

Excel

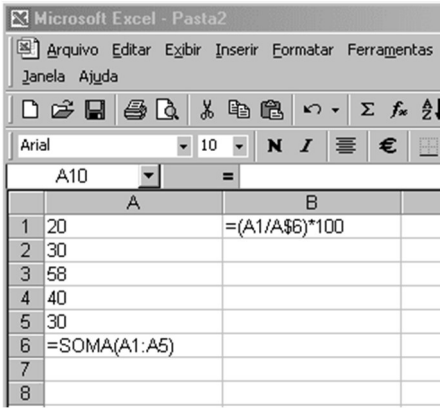
Assuntos que serão apresentados

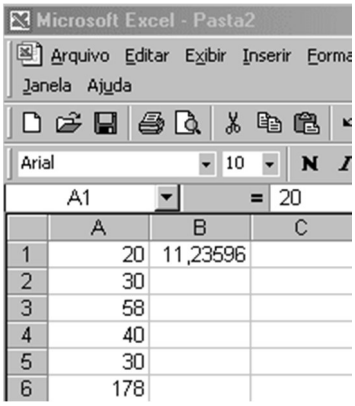
1- Cálculo de porcentagem simples e acumulada; construção de tabelas
2- Construção de gráficos
   2.1 – Diagrama de barras (uma variável)
   2.2 – Diagrama linear (uma e duas variáveis); escala aritmética e logarítmica
   2.3- Histograma – intervalos de classe iguais
   2.4 – Polígono de freqüências – intervalos de classes iguais
   2.5 – Polígono de freqüências – intervalos de classe diferentes
   2.6 – Diagrama de barras (duas variáveis)
   2.7 – Diagrama de freqüências acumuladas
   2.8 – Diagrama de dispersão, coeficiente de correlação de Pearson
   2.9 – Equação da reta de regressão linear simples
3- Cálculo de estatísticas: média, mediana, desvio padrão
4- Cálculo de probabilidades
   4.1 – Distribuição normal
   4.2 – Distribuição t de Student
   4.3 – Distribuição qui-quadrado

1 - Cálculo de porcentagem simples e acumulada; construção de tabelas

O Excel é uma planilha estruturada em linhas numeradas e colunas identificáveis por letras (A, B, C, ...) assim é possível se referir a cada célula ou casela, por exemplo, a célula A5 é a quinta célula na primeira linha. O Excel é utilizado para elaboração de planilhas que envolvem cálculos, para desenhar gráficos e também como banco de dados.

Cálculo de percentual
Digitar na coluna A, linhas 1, 2, 3, 4 e 5 os valores 20, 30, 58, 40 e 30. Na célula A6 digitar a fórmula =SOMA(A1:A5) ou clicar sobre o ícone ∑ e pressionar a tecla Enter. Na casela B1, digitar a fórmula =(A1/A$6)*100 para calcular o percentual de 20 em relação ao total e, em seguida, pressionar Enter. OBS: O $ fixa a linha. Também é possível usar $ pela tecla <F4>



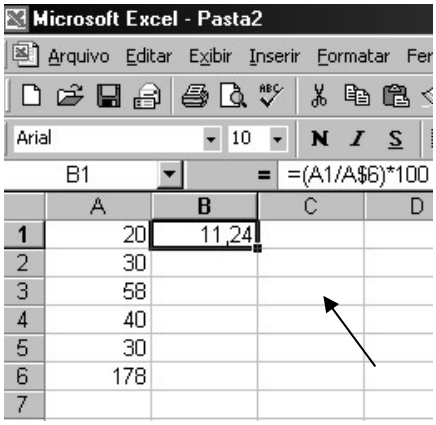


No lugar da fórmula irá aparecer o resultado 11,23596 que pode ser formatado para duas casas decimais utilizando a seguinte seqüência de comandos: formatar, célula, número, escolher número de casas decimais, por exemplo 2. Clicar em OK para que o Excel execute o comando.

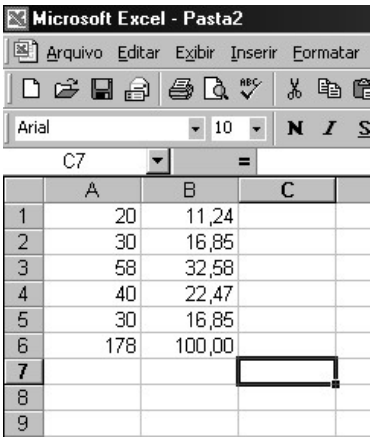
Cópia da fórmula para as outras caselas

- clique o mouse sobre a célula que será copiada;
- segure o mouse sobre o quadradinho do lado direito na base do retângulo;

- segure e arraste o mouse até a célula B5. Solte o botão do mouse;
- percorra, utilizando a seta para cima, cada casela e confira as fórmulas;
- posicione o cursor na célula B6 e clique no ícone ∑ , pressione Enter.



Resultado final:



2 - Construção de gráficos

Atenção: Se a versão do Excel for em inglês, utilizar para vírgula o ponto. Se a versão for em português, utilizar para representar casas decimais, a vírgula.

2.1. Diagrama de barras (uma variável)

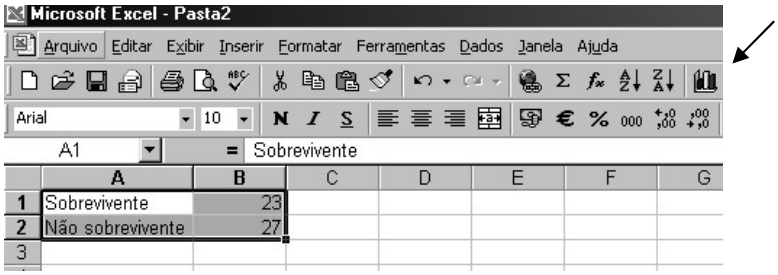
Lista de procedimentos para apresentar os dados da tabela em um gráfico apropriado:

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo condição de sobrevivência

Condição do recém-nascido	Nº	%
Sobrevivente	23	46,0
Não sobrevivente	27	54,0
Total	50	100

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

Digitar em uma coluna as categorias da variável (sobrevivente e não sobrevivente) e em outra coluna, os valores da freqüência ou do percentual. Marque as duas colunas e clique sobre o ícone de gráficos.

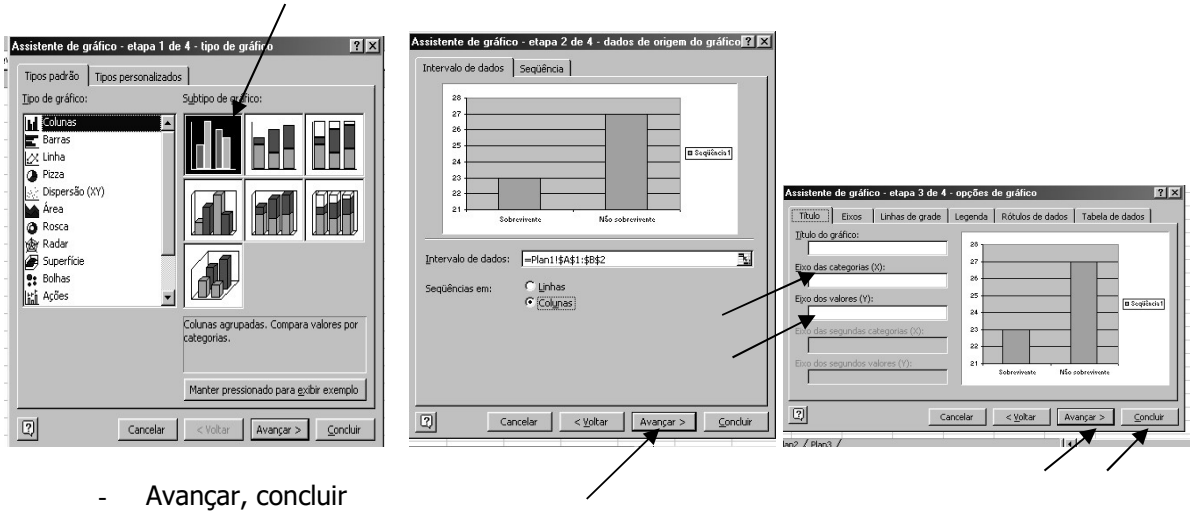


Escolher o gráfico de colunas e clicar sobre o primeiro sub-tipo de gráfico. Notar os demais subtipos.

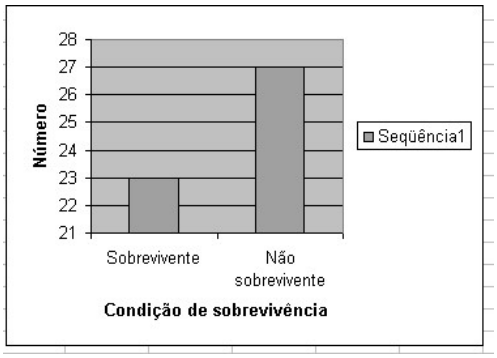
Clicar sobre Avançar. Pode-se visualizar o diagrama de barras.

Notar a origem "escolhida" pelo Excel. É possível alterar a origem, caso seja de interesse, após a conclusão do gráfico.

Clicar em avançar e no **menu assistente de gráfico** inserir os títulos dos eixos X e Y. O título do gráfico pode ser digitado depois de levar o gráfico para o *Word for Windows*.

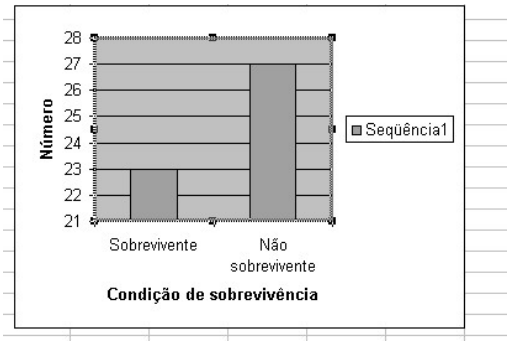


- Avançar, concluir



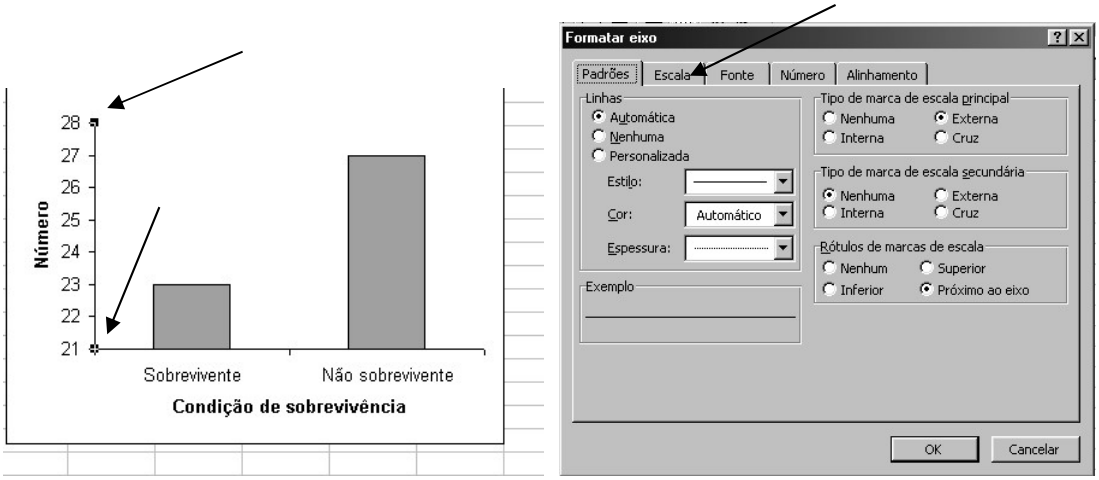
Para editar (melhorar a apresentação) do gráfico

- clicar sobre a caixa seqüência 1 e pressionar a tecla Del. Também é possível configurar no assistente, antes de concluir.
- dar duplo clique sobre a área cinza do gráfico para escolher a cor do fundo do gráfico.
- para retirar as linhas de grade clique sobre uma linha e automaticamente todas serão selecionadas. Clicar em Del. Também é possível marcar a área do gráfico clicando sobre este uma vez. A área ficará constrita em um retângulo limitado por quadradinhos.

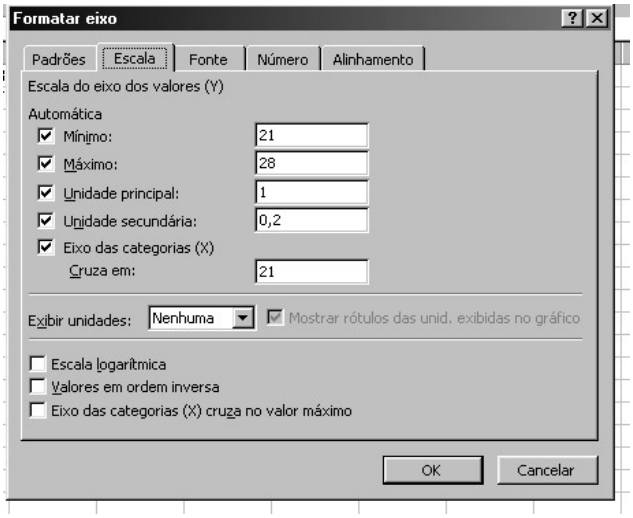


- clicar sobre a área marcada, com o botão direito do mouse e escolher Opções de gráfico. Neste menu é possível alterar os eixos, as linhas de grade, decidir sobre a legenda, rótulos de dados e decidir se a tabela de dados será ou não incluída. OBS: normalmente deve-se apresentar o gráfico ou a tabela, mas não ambos.

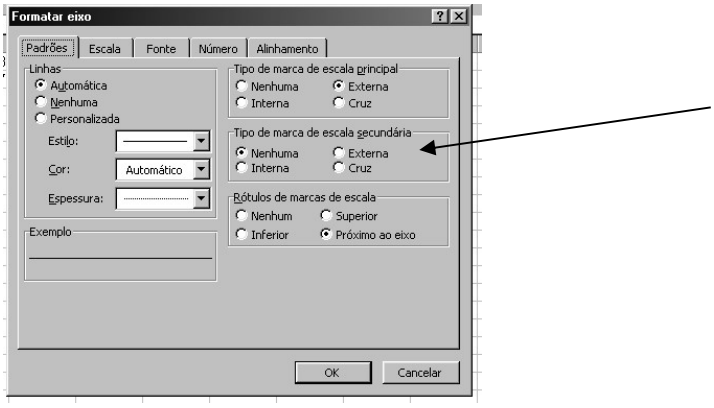
Para mudar a escala é necessário clicar uma vez sobre o eixo. Este ficará marcado. Clicar duas vezes sobre o eixo já marcado ou simplesmente clique com o botão da direita sobre o eixo, mesmo sem estar marcado, e escolha **formatar eixo**.



escolha Escala

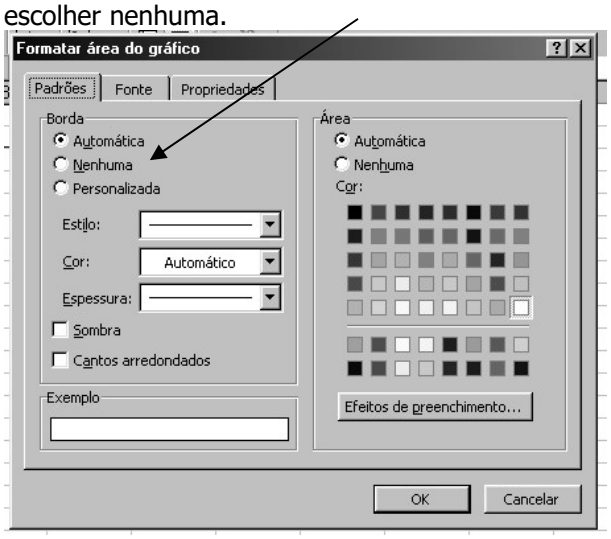


- digitar no campo mínimo o valor zero;
- o valor máximo também pode ser alterado (neste exemplo não é necessário);
- a unidade principal também pode ser alterada. Deixe 5 e veja o resultado; depois mude para 8 e veja o resultado. A unidade secundária só aparecerá se no menu formatar eixo for escolhido algum tipo de marca, por exemplo, externa (o *default* é nenhuma)



Voltando para Formatar eixo (clicando sobre o eixo e escolhendo escala), notar que a escala pode ser **logarítmica**.

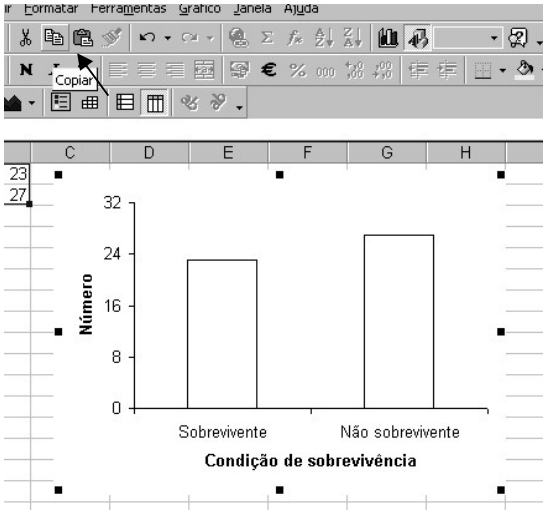
O gráfico está pronto. Para quem quiser tirar a borda do gráfico, é necessário clicar sobre a borda externa, e clicar com o botão direito do mouse, escolhendo formatar área do gráfico, e em Borda, escolher nenhuma.



Alterando cores

As cores do gráfico podem ser alteradas utilizando duplo clique sobre as barras. No menu formatar seqüência de dados escolher a cor da área. Notar que existe a opção de efeitos de preenchimento.

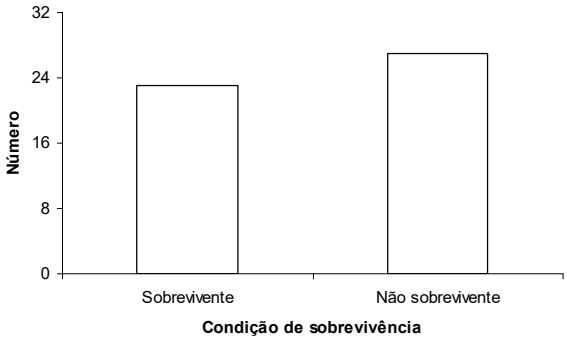
Uma vez que o gráfico esteja terminado, este pode ser copiado para o Word. Para tanto, selecione o gráfico, clique sobre o ícone copiar, abra o Word, deixe algumas linhas para o título e clique no ícone colar.



Resultado final no Word

OBS: digitando-se o título no documento Word e copiando-se o gráfico (como Figura ou Objeto)

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo condição de sobrevivência



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

2.2 - Diagrama linear com uma e duas variáveis (escala aritmética e logarítmica),

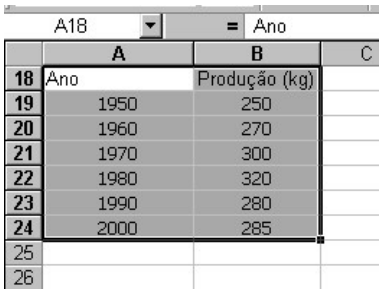
Exemplo
Os dados são relativos à produção mundial de grãos por pessoa ano no período de 1950 a 2000.

Distribuição da produção mundial de grãos por pessoa/ano segundo ano.

Ano	Produção (kg)
1950	250
1960	270
1970	300
1980	320
1990	280
2000	285

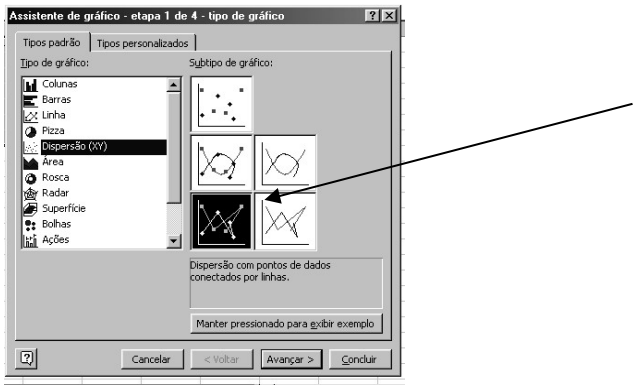
Fonte: State of the World, 2001. The Worldwatch Institute

- digitar os dados em qualquer lugar da planilha;
- marcar as colunas posicionando o mouse sobre a primeira casela, segundo o botão esquerdo e arrastando o mouse até a última casela.

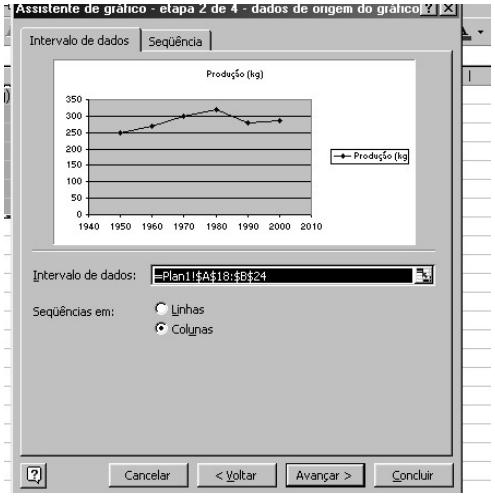


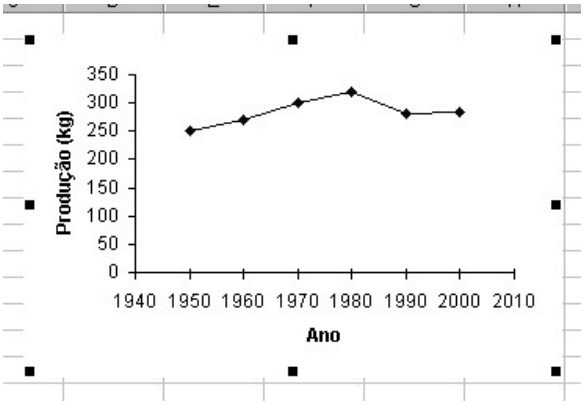
	A	B	C
18	Ano	Produção (kg)	
19	1950	250	
20	1960	270	
21	1970	300	
22	1980	320	
23	1990	280	
24	2000	285	
25			
26			

- clicar sobre o ícone de gráficos e escolher dispersão;
- escolher como subtipo, o terceiro gráfico da coluna com 3 opções;
- Clicar em avançar.

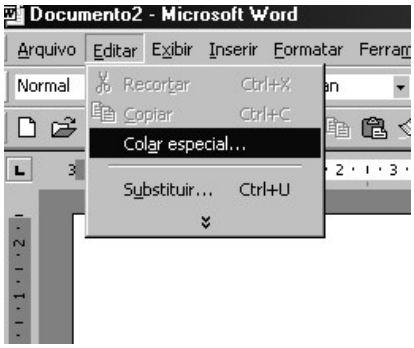


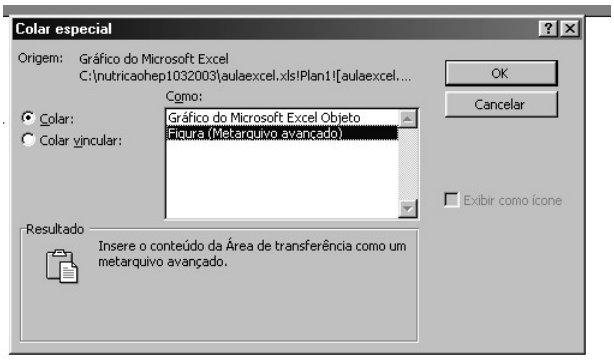
- clicar em avançar, e no assistente de gráficos, escrever os títulos dos eixos X e Y;
- retirar as linhas de grade e a cor do fundo;
- retirar a legenda e o título do gráfico.



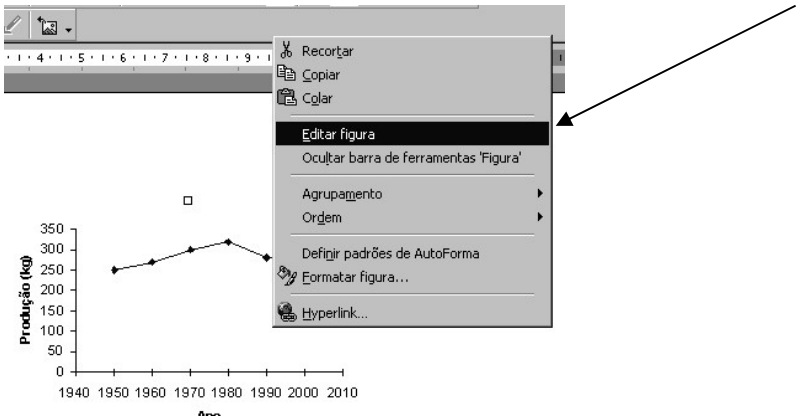


- Concluir o gráfico;
- Selecionar o gráfico para copiá-lo no Word;
- Clicar no ícone copiar, abrir um documento no Word e colar. Escolher colar especial , figura



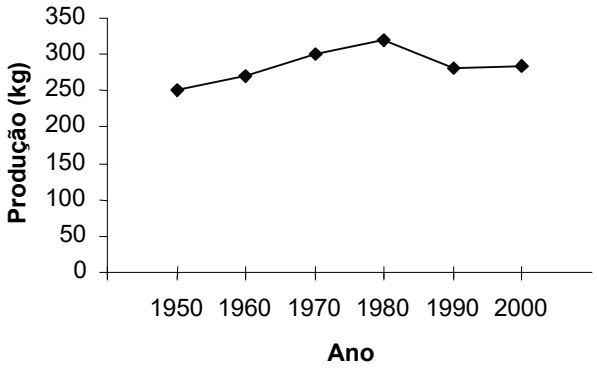


Abrir a figura, clicando sobre o gráfico, com o botão direito do mouse. Escolher editar figura



- editar a figura, retirando os anos 1940 e 2010.

Resultado final (no *Word*)



Exercício

Apresente os dados da tabela em um gráfico apropriado. Construir o gráfico nas escalas aritmética e logarítmica. Decida qual escala é mais apropriada.

Coeficientes de mortalidade por câncer de esôfago (por 100.000 hab.).
Município de São Paulo, 1968-1998.

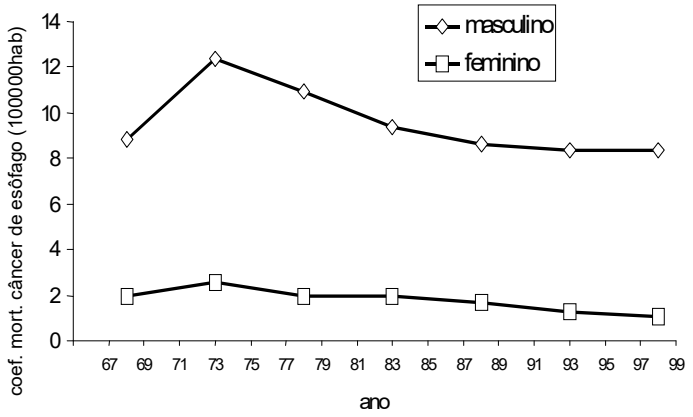
Ano	Masculino	Feminino
1968	8,81	2,00
1973	12,38	2,61
1978	10,93	1,98
1983	9,41	2,00
1988	8,60	1,67
1993	8,33	1,27
1998	8,37	1,12

Fonte: Incidência de câncer no Município de São Paulo, 1997-1998. Registro de Câncer de São Paulo. FSP/USP.

a) digitar os dados e selecionar a área desde o título da primeira coluna até o último valor da terceira coluna;
b) clicar no botão de gráfico e escolher dispersão
c) selecionar a opção unir pontos (último gráfico da primeira coluna)
d) clicar em avançar; digitar o nome do eixo X e do eixo Y; concluir.
e) Retirar as grades e o fundo de cor cinza; retirar a borda do gráfico.
f) Selecionar o gráfico e salvá-lo no Word como figura.

Resultado final (no *Word*):

Coeficientes de mortalidade por câncer de esôfago (por 100.000 hab.).
Município de São Paulo, 1968-1998.



Fonte: Incidência de câncer no Município de São Paulo, 1997-1998. Registro de Câncer de São Paulo. FSP/USP.

Para mudar a escala do eixo Y de escala aritmética para logarítmica:

- No Excel, clicar sobre o eixo Y e escolher Escala;
- Mudar a escala de aritmética para logarítmica;
- Selecionar o gráfico e copiá-lo como figura, no Word.

2.3 - Histograma - intervalos com mesma amplitude

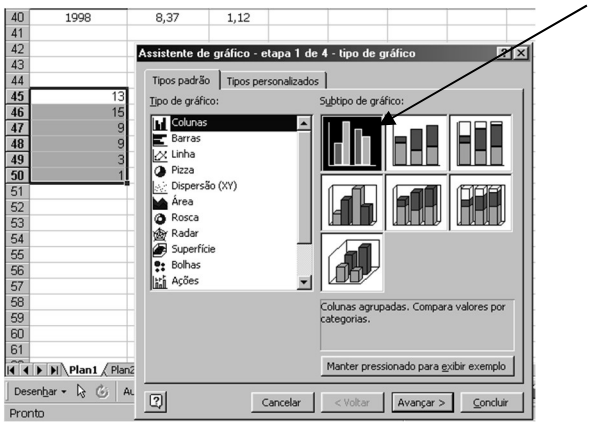
Considerar os dados apresentados na tabela.

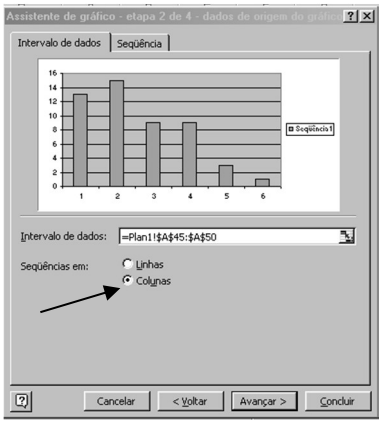
Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g)

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

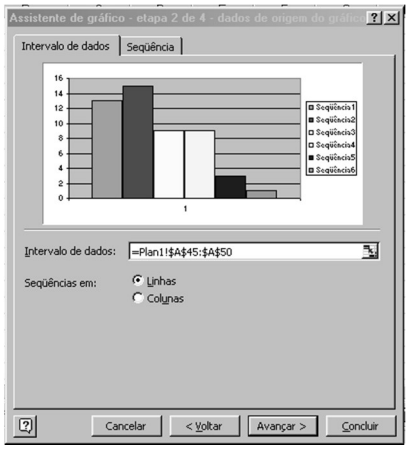
Fonte: Hand DJ et al. A handbook of small data sets. Chapman&Hall, 1994.

- No Excel, digitar os valores 13, 15, 9, 9, 3, 1 (ou os percentuais) em uma coluna;
- Selecionar os valore e escolher gráfico de colunas.





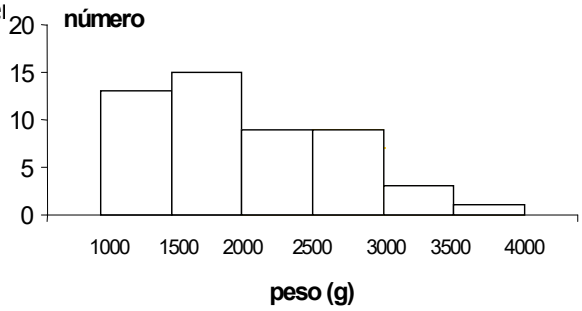
- Mude, em "Seqüências em", de colunas para linhas;



- Avançar; inserir títulos dos eixos; apagar a legenda, as grades e o fundo cinza. Como as faixas etárias fazem parte de uma única variável, sugere-se deixar todas as barras com a mesma cor. Clicar em um retângulo e alterar a cor deste para cor específica. Clicar no próximo retângulo e pressionar a tecla de função <F4>, que repete o último comando. Formatar área do gráfico retirando a borda;
- Selecionar o gráfico, clicar no ícone de copiar, salvá-lo no Word como figura;
- No Word, abrir figura, abrir caixa de diálogo sob as barras e digitar valores.

Resultado final (no *Word*):

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nasce

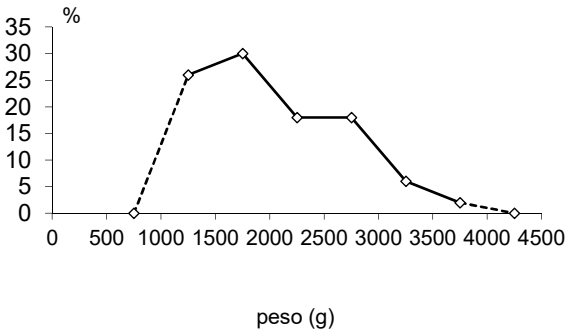


Fonte: Hand DJ et al*. A handbook of small data sets.* Chapman&Hall, 1994.

2.4 - Polígono de freqüências (intervalos de classe iguais)

- em uma coluna digitar os pontos médios começando pelo ponto médio de um intervalo hipotético anterior e depois do ponto médio referente ao último intervalo, digitar o ponto médio de um intervalo hipotético posterior ao último;
- digitar na outra coluna o número (ou percentual);
- selecionar os dados e clicar no ícone de gráficos. Escolher dispersão com a opção de ligar os pontos (último gráfico da primeira coluna). Avançar;
- inserir nome nos eixos X e Y; retirar a legenda, as grades e a cor do fundo;
- clicar sobre os pontos uma vez e sobre o primeiro segmento mais uma vez. Clicar o botão direito do mouse, escolher formatar ponto de dados. Escolher no menu padrão, linha, estilo tracejado;
- formatar o primeiro segmento. Clicar sobre o último segmento e clicar sobre a tecla <F4> para repetir o último comando;
- selecionar o gráfico e formatar a área retirando a borda. Copiar e colar no Word como figura.

Resultado final (no Word):



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

2.5 - Polígono de freqüência com intervalos de classe diferentes

Considere os dados apresentados na tabela a seguir.

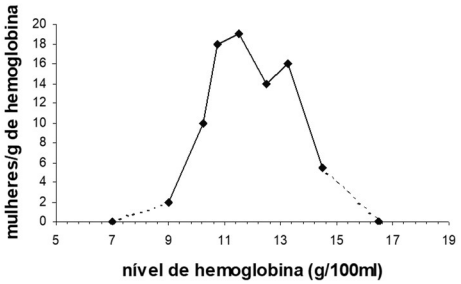
Distribuição de mulheres segundo nível de hemoglobina (g/100ml).

Nível de hemoglobina (g/100ml)	n⁰	%
8,0|--10,0	4	5,7
10,0|--10,5	5	7,1
10,5|--11,0	9	12,9
11,0|--12,0	19	27,1
12,0|--13,0	14	20,0
13,0|--13,5	8	11,4
13,5|--15,5	11	15,7
Total	70	100

Fonte: Kirkwood BR. Essentials of Medical Statistics.1988.

- em uma coluna digitar os pontos médios começando pelo ponto médio de um intervalo hipotético anterior e depois do ponto médio referente ao último intervalo, digitar o ponto médio de um intervalo hipotético posterior ao último (considerar para o primeiro e últimos intervalos, amplitudes iguais à primeira e à última respectivamente);
- digitar nas outras colunas o número e a amplitude de classe;
- fazer os ajuste – número de pessoas dividido pela amplitude de classe;
- selecionar a coluna dos pontos médios e a coluna do ajuste. Para selecionar colunas não adjacentes, selecione os pontos médios, pressione a tecla control (Ctrl) e, com o mouse, selecione os valores do ajuste. No ícone de gráficos, escolher dispersão com a opção de ligar os pontos (último gráfico da primeira coluna). Avançar;
- Inserir nome nos eixos X e Y; retirar a legenda, as grades e a cor do fundo;
- clicar sobre os pontos uma vez e sobre o primeiro segmento mais uma vez. Clicar o botão direito do mouse, escolher formatar ponto de dados. Escolher no menu padrão, linha, estilo tracejado;
- Formatar o primeiro segmento. Clicar sobre o último segmento e clicar sobre a tecla <F4> para repetir o último comando;
- Selecionar o gráfico e formatar a área retirando a borda. Copiar e colar no Word como figura.

Resultado final:



Distribuição de mulheres segundo concentração de hemoglobina (g/100ml)

2.6 - Diagrama de barras com duas variáveis

Considere os dados apresentados na tabela a seguir

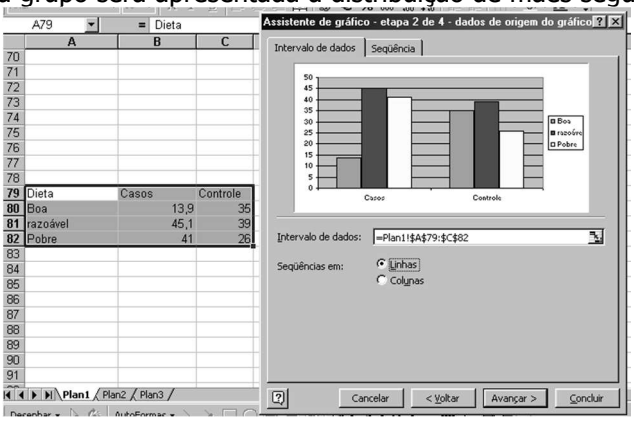
Distribuição de recém-nascidos segundo condição caso - com defeitos do tubo neural; controle – recém-nascidos que não tinham defeitos do tubo neural e dieta materna.

Dieta	Casos		Controles		Total	
	N	%	n	%	n	%
Boa	34	13,9	43	35,0	77	21,0
Razoável	110	45,1	48	39,0	158	43,0
Pobre	100	41,0	32	26,0	132	36,0
Total	244	100	123	100	367	100

Representação gráfica:

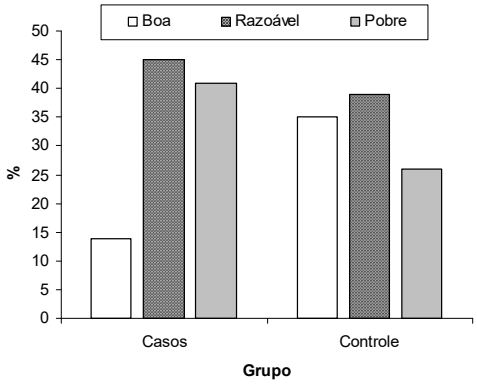
- digitar as categorias da variável dieta em uma coluna; na coluna seguinte digitar o percentual de casos e o de controles. Digitar os "títulos" das colunas;

- selecionar valores incluindo nomes das colunas. Escolher no ícone de gráficos, gráfico colunas; mudar de colunas para linhas, uma vez que deve somar 100% em casos e 100% em controles e dentro de cada grupo será apresentada a distribuição de mães segundo tipo de dieta.



- Avançar; inserir título nos eixos, concluir;
- Clicar no gráfico, retirar grades e cor de fundo, posicionar a legenda;
- Para separar as barras (variável dieta é qualitativa) clicar com o botão direito do mouse dentro da primeira barra e selecionar formatar seqüência de dados. Escolher opções e em sobreposição, deixar o valor –30;
- Alterar as cores das barras deixando em tons de cinza. Para tanto, clique dentro da primeira barra e escolha em padrão, a cor branca para a primeira série de dados. Repita o procedimento para as demais barras;
- Retirar a borda, copiar o gráfico e salvá-lo no Word como figura (salvar especial).

Resultado final (no *Word*):



Distribuição de recém-nascidos segundo condição caso - com defeitos do tubo neural; controle – recém-nascidos que não tinham defeitos do tubo neural e dieta materna.

2. 7 - Diagrama de freqüências acumuladas

Utilizando os dados a seguir, calcule o percentual acumulado de recém-nascidos segundo peso ao nascer. Construa o gráfico de freqüências acumuladas. Diga qual é o valor da variável que deixa 50% dos valores abaixo dele.

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g).

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. A handbook of small data sets. Chapman&Hall, 1994.

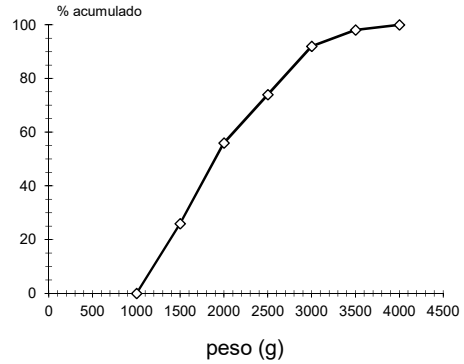
- digitar em uma coluna os valores da variável peso correspondentes aos limites superiores dos intervalos. Começar com o limite inferior do primeiro intervalo;
- na coluna à direita digitar o número de pessoas de cada intervalo; na primeira classe (corresponde ao valor do limite inferior da primeira classe), digitar zero;
- Calcular os percentuais e na coluna a seguir calcular os percentuais acumulados.

	n	%	% acum
1000	0	=(B97/B$104)*100	=C97
1500	13	=(B98/B$104)*100	=D97+C98
2000	15	=(B99/B$104)*100	=D98+C99
2500	9	=(B100/B$104)*100	=D99+C100
3000	9	=(B101/B$104)*100	=D100+C101
3500	3	=(B102/B$104)*100	=D101+C102
4000	1	=(B103/B$104)*100	=D102+C103
	=SOMA(B97:B103)		

- selecionar as colunas dos valores de peso e da porcentagem acumulada; escolher o gráfico de dispersão com opção de unir pontos. Dar nome para os eixos X e Y.

Resultado final (no *Word*):

Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g).



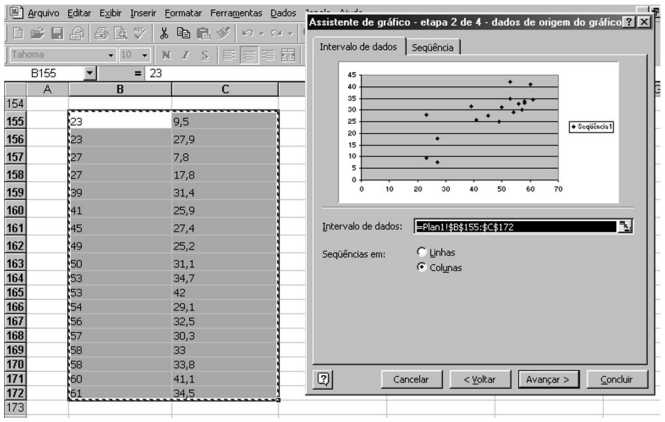
Fonte: Hand DJ et al. *A handbook of small data sets.* Chapman&Hall, 1994.

2.8 - Diagrama de dispersão

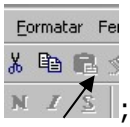
Utilizando os dados a seguir, construa o diagrama de dispersão entre as variáveis porcentagem de gordura e idade. Calcule o coeficiente de correlação de Pearson.

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

- digitar em uma coluna os valores da idade e em uma coluna adjacente, os valores da variável % de gordura;
- digitar o nome das variáveis;
- marcar os valores, clicar sobre o ícone de gráficos e escolher o gráfico de dispersão; escolher primeiro gráfico;
- avançar, inserir títulos dos eixos X e Y; retirar legenda, linhas de grade e cor cinza do fundo; concluir.



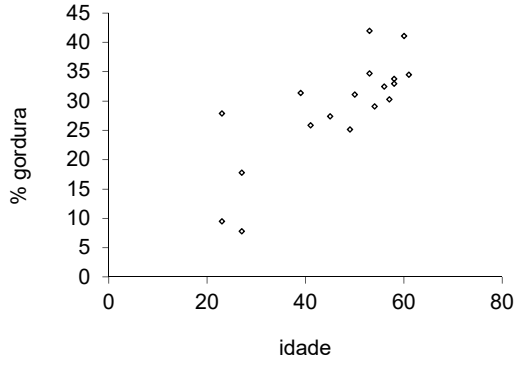
- clicar sobre o gráfico e formatar área, retirando a linha da borda;



- selecionar gráfico, clicar sobre o ícone de copiar ; no Word, clicar sobre Editar; escolher colar especial, opte por Figura.

Resultado final (no *Word*):

Distribuição de pacientes segundo idade e gordura corporal



Cálculo do coeficiente de correlação de Pearson

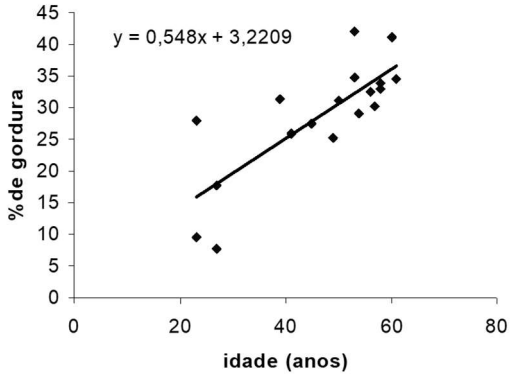
- em uma casela abaixo do último valor digitado, escreva a fórmula para o cálculo do coeficiente de correlação de Pearson: =correl(b155:b172;c155:c172). Dependendo da versão do Excel, o ponto e vírgula da fórmula deverão ser substituídos por vírgula.

Coeficiente de correlação de Pearson®= +0,79

2.9 – Equação da reta de regressão linear simples
Para inserir a equação da reta de regressão linear simples:

- clicar sobre os pontos do diagrama de dispersão com o botão do lado esquerdo do mouse;
- clicar novamente sobre um dos pontos marcados, porém com o botão direito do mouse;
- escolher adicionar linha de tendência;
- clicar em opções e na base do menu escolher exibir equação no gráfico;
- clicar sobre a caixa da equação e posicioná-la em um lugar adequado no gráfico.

Resultado final no Word



Distribuição de pacientes segundo idade e gordura corporal

3 - Cálculo de estatísticas: média, mediana, variância e desvio padrão (construindo fórmulas e utilizando funções)

Supor os valores

166	158	202	162	135	82	150	86	121

- digitar os valores em uma coluna;
- no final da coluna digitar as fórmulas para cada medida.
- indicar na coluna anterior qual medida está sendo calculada.

175		
176		
177		166
178		158
179		202
180		162
181		135
182		82
183		150
184		86
185		121
186	média	=MÉDIA(C177:C185)
187	mediana	=MED(C177:C185)
188	variância(n)	=VARP(C177:C185)
189	variância(n-1)	=VARA(C177:C185)
190	desvio padrão (n)	=DESVPADP(C177:C185)
191	desvio padrão (n-1)	=DESVPADA(C177:C185)
192		

É possível calcular a média, variância e desvio padrão desenvolvendo as fórmulas

valor x	x-média	(x-média)^2		
166	=C177-C$187	=D177^2	variância (n)=	=E186/9
158	=C178-C$187	=D178^2	variância (n-1)=	=E186/8
202	=C179-C$187	=D179^2		
162	=C180-C$187	=D180^2	desvio padrão (n)=	=RAIZ(G177)
135	=C181-C$187	=D181^2	desvio padrão (n-1)=	=RAIZ(G178)
82	=C182-C$187	=D182^2		
150	=C183-C$187	=D183^2		
86	=C184-C$187	=D184^2		
121	=C185-C$187	=D185^2		
		=SOMA(E177:E185)		

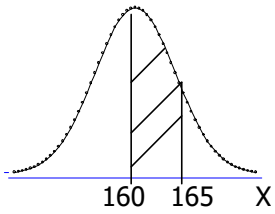
4 - Cálculo de probabilidade

4.1 - Distribuição Normal

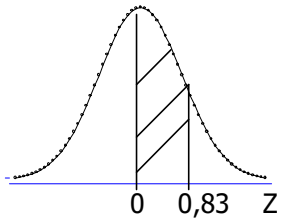
Considerar a altura de 351 mulheres idosas como seguindo uma distribuição normal com média 160 cm e desvio padrão 6 cm. Sorteia-se uma mulher; qual a probabilidade de que ela tenha

f) altura entre 160 cm e 165 cm?

X: altura; X~N(160,6)



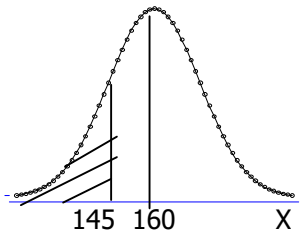
$$P(160 < X < 165) = P(\frac{160-160}{6} < \frac{X-\mu}{\sigma} < \frac{165-160}{6}) = P(0 < Z < 0{,}83)$$



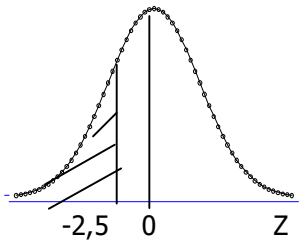
Utilizando a tabela da curva normal reduzida, $P(0<Z<0{,}83)=0{,}29673$ ou 29,7%

Fórmula no Excel: DIST.NORMP(0,83)-0,5=0,29673

g) altura menor do que 145 cm?
X: altura; X~N(160,6)



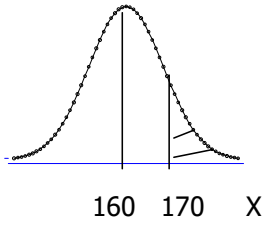
$$P(X<145)=P(\frac{X-\mu}{\sigma}<\frac{145-160}{6})=P(Z<-2{,}5)$$



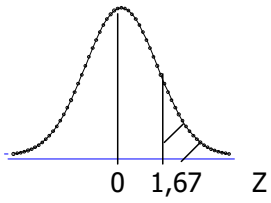
Utilizando a tabela da curva normal reduzida, P(Z<-2,5)=0,5-0,49379=0,0062 ou 0,6%

Fórmula no Excel: DIST.NORMP(-2,5)=0,00620968

h) altura maior do que 170 cm?



$$P(X>170)=P(\frac{X-\mu}{\sigma}>\frac{170-160}{6})=P(Z>1{,}67)$$

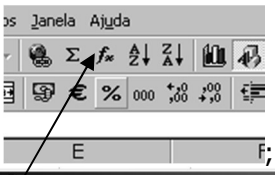


Utilizando a tabela da curva normal reduzida, P(Z>1,67)=0,5 – 0,45254=0,0475 ou 4,7%

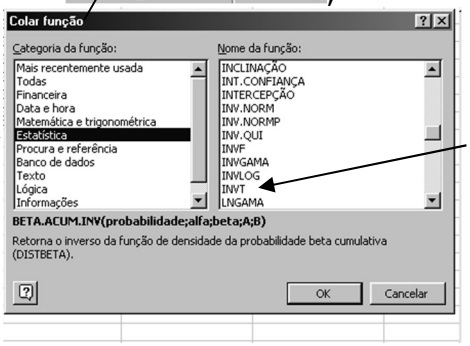
Fórmula no Excel: 1-DIST.NORMP(1,67)=0,0474597

4.2 – Distribuição t de Student
Valor de t crítico para uma área de 5% e 10 graus de liberdade e teste bicaudal:
Fórmula no Excel: INVT(0,05;10)

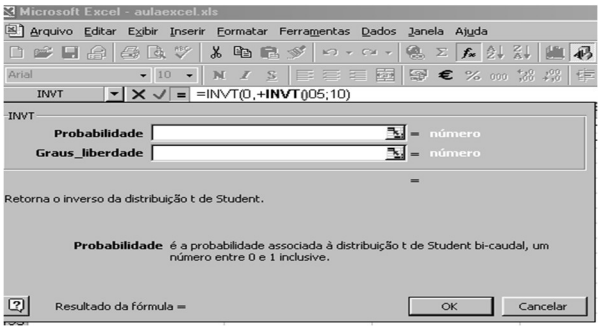
Ou seguir a seqüência: Clicar em Inserir função



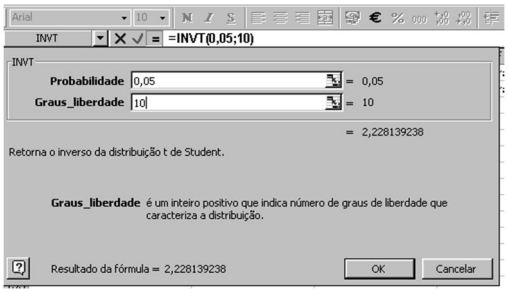
;



Escolher Estatística e INVT

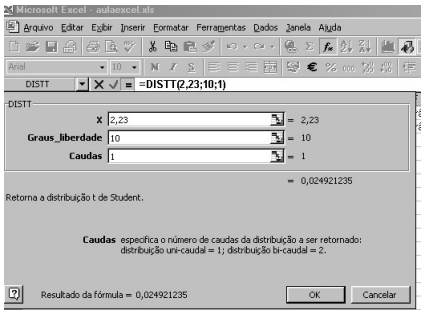


- Notar que o valor de t é para um teste bi-caudal



- digitar o valor da área e o número de graus de liberdade

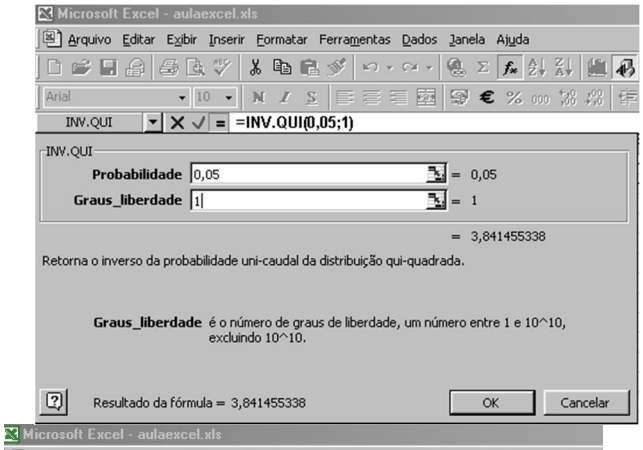
Valor da área para t observado igual a 2,23 e teste monocaudal: 0,024921

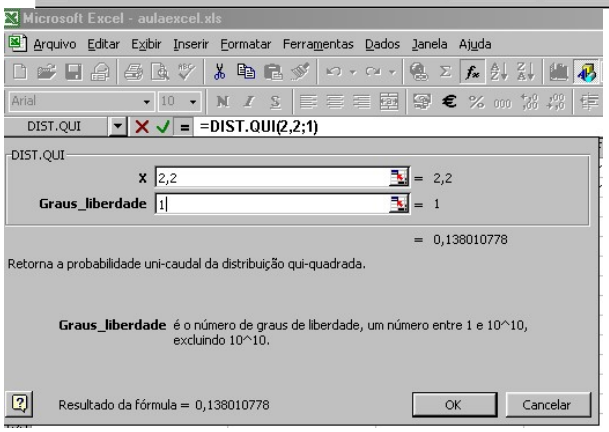


4.3 – Distribuição qui-quadrado

- clicar em função; escolher Estatística; e no sub-tipo, INV.QUI.
- Digitar a probabilidade e o número de graus de liberdade.

Se o valor do qui-quadrado for conhecido e o que se deseja saber á a área que fica à direita do número então utilizar no sub-tipo a DIST.QUI





Resposta

Exercício 2 - Classificar quanto à natureza, as seguintes variáveis:

Variável	**Tipo (natureza)**
Condição de saúde (doente, não doente)	Qualitativa nominal
Tipo de parto (normal, cesário)	Qualitativa nominal
Nível de colesterol sérico (mg/100cc)	Quantitativa contínua
Tempo de um procedimento cirúrgico (minutos)	Quantitativa contínua
Número de praias consideradas poluídas	Quantitativa discreta

Exercício 3

Variável sexo

Distribuição de idosos segundo sexo. Município de São Paulo, 2013

Sexo	n	%
Feminino	34	68
Masculino	16	32
Total	50	100

Interpretação: Observa-se que 68% dos idosos é do sexo feminino

Variável número de doenças crônicas

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	

Distribuição de idosos segundo número de doenças crônicas. Município de São Paulo, 2013

Número de doenças crônicas	n	%
0	7	14,3
1	13	26,5
2	12	24,5
3	13	26,5
4	3	6,1
6	1	2,0
Total	49	100

Interpretação: Observa-se que 77,5% dos idosos apresentam de 1 a 3 doenças crônicas.

Variável 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	

Distribuição de idosos segundo idade. Município de São Paulo, 2013

Idade (anos)	n	%
60 |-- 65	5	10
65 |-- 70	4	8
70 |-- 75	10	20
75 |-- 80	5	10
80 |-- 85	10	20
85 |-- 90	9	18
90 |-- 95	7	14
Total	50	100

Interpretação: observa-se que 50% dos idosos se encontram em idades de 70 a 84 anos

Distribuição de idosos segundo idade. Município de São Paulo, 2013

Idade (anos)	n	%
60 |-- 70	9	18
70 |-- 80	15	30
80 |-- 90	19	38
90 |--100	7	14
Total	50	100

Interpretação: observa-se que 68% dos idosos se encontram em idades de 70 a 89 anos

Exercício 4

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 freqüências relativas. Fixando o 100% no total de fumantes e não fumantes.
b) Calcule as freqüências 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.

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

a)

Distribuição de gestantes segundo níveis de β-caroteno (mg/L) e hábito de fumar.

β-caroteno (mg/L)	Fumante		Não Fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46	79,3	74	56,1	120	63,2
Normal (0,214 – 1,00)	12	20,7	58	43,9	70	36,8
Total	58	100	132	100	190	100

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

Cálculo dos percentuais

$$\frac{46}{58} = (0{,}793) * 100 = 79{,}3$$

$$\frac{12}{58} = (0{,}2069) * 100 = 20{,}7$$

$$\frac{74}{132} = (0{,}561) * 100 = 56{,}1$$

$$\frac{58}{132} = (0{,}439) * 100 = 43{,}9$$

Interpretação:

Independente do hábito de fumar, 63,2% das gestantes apresentam nível baixo de beta caroteno. As variáveis podem estar associadas pois entre as fumantes esta porcentagem é de 79,3% contra 56,% entre as não fumantes.

b)

Distribuição de gestantes segundo níveis de β-caroteno (mg/L) e hábito de fumar.

β-caroteno (mg/L)	Fumante		Não Fumante		Total	
	n	%	n	%	n	%
Baixo (0 – 0,213)	46	38,3	74	61,7	120	100
Normal (0,214 – 1,00)	12	17,1	58	82,9	70	100
Total	58	30,5	132	69,5	190	100

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

Cálculo dos percentuais

$$\frac{46}{120} = (0{,}383) * 100 = 38{,}3$$

$$\frac{74}{120} = (0{,}6167) * 100 = 61{,}7$$

$$\frac{12}{70} = (0{,}171) * 100 = 17{,}1$$

$$\frac{58}{70} = (0{,}829) * 100 = 82{,}9$$

Interpretação:

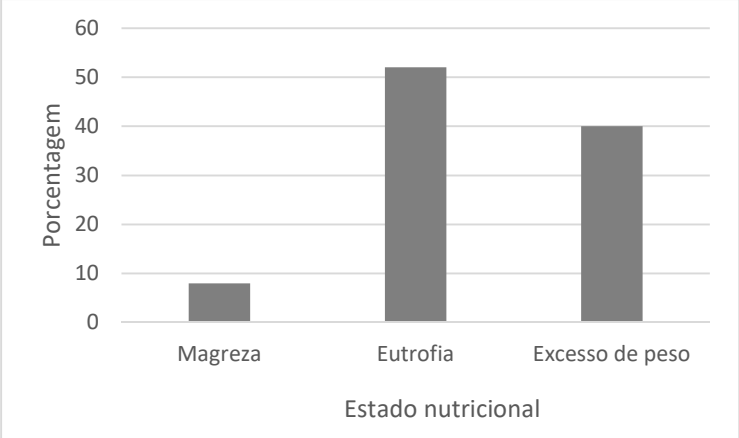
Independentemente do nível de beta caroteno, 30,5% das gestantes são fumantes. Entre as que apresentam nível baixo de betacaroteno, este percentual é de 38,3% e entre as com nível normal, este percentual é de 17,1%. A análise indica possível associação entre as variáveis.

Exercício 5 – Apresente o diagrama de barras para a variável imc em três categorias

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:≤ 21 kg/m^2; eutrofia: 22-27 kg/m^2; excesso de peso ≥28 kg/m^2



(2) magreza:≤ 21 kg/m^2; eutrofia: 22-27 kg/m^2; excesso de peso ≥28 kg/m^2

Distribuição de idosos segundo classificação nutricional(2). Município de São Paulo, 2013

Interpretação:

Pode-se observar que a avaliação do estado nutricional indica a presença de excesso de peso em 40% e magreza em 8% dos idosos.