Curso de Capacitação em Epidemiologia Básica
e Análise da Situação de Saúde
Ministério da Saúde
Secretaria de Vigilância em Saúde

NOÇÕES BÁSICAS DE ESTATÍSTICA

Gleice Margarete de Souza Conceição
Airlane Pereira Alencar
Gizelton Pereira Alencar

NOÇÕES BÁSICAS DE ESTATÍSTICA

Análise Exploratória de Dados

Após a coleta e a digitação de dados em um banco de dados apropriado, o próximo passo é a análise descritiva. Esta etapa é fundamental, pois uma análise descritiva detalhada permite ao pesquisador familiarizar-se com os dados, organizá-los e sintetizá-los de forma a obter as informações necessárias do conjunto de dados para responder as questões que estão sendo investigadas. Tradicionalmente, a análise descritiva limitava-se a calcular algumas medidas de posição e variabilidade. No final da década de 70, Tukey criou uma nova corrente de análise. Utilizando principalmente técnicas visuais, buscando descrever quase sem utilizar cálculos, alguma forma de regularidade ou padrão nos dados, em oposição aos resumos numéricos. Nessa etapa, iremos produzir tabelas, gráficos e medidas resumo que descrevam a tendência dos dados, quantifiquem a sua variabilidade, permitam a detecção de estruturas interessantes e valores atípicos no banco de dados.

Tipo de variáveis

Cada uma das características de interesse observadas ou medidas durante o estudo é denominada de variável. As variáveis que assumem valores numéricos são denominadas *quantitativas*, enquanto que as não numéricas, *qualitativas*.

Uma variável é qualitativa quando seus valores são atributos ou qualidades (por ex: sexo, raça, classe social). Se tais variáveis possuem uma ordenação natural, indicando intensidades crescentes de realização, são classificadas de *qualitativas ordinais* (por ex: classe social - baixa, média ou alta). Se não for possível estabelecer uma ordem natural entre seus valores, são classificadas como *qualitativas nominais* (por ex: Sexo - masculino ou feminino).

As variáveis quantitativas podem ser classificadas ainda em *discretas* ou *contínuas*. Variáveis discretas podem ser vistas como resultantes de contagens, e assumem, em geral, valores inteiros (por ex: Número de filhos). Variáveis contínuas podem assumir qualquer valor dentro de um intervalo especificado e são, geralmente, resultados de uma mensuração (por ex: Peso, em kg; Altura, em metros).

Descrição dos dados

É importante conhecer e saber construir os principais tipos de tabelas, gráficos e medidas resumo para realizar uma boa análise descritiva dos dados. Vamos tentar entender como os dados se distribuem, onde estão centrados, quais observações são mais freqüentes, como é a variabilidade

etc., tendo em vista responder às principais questões do estudo. Cada ferramenta fornece um tipo de informação e o seu uso depende, em geral, do tipo de variável que está sendo investigada. Grosso modo, utilizaremos as duas abordagens sugeridas no quadro:

variável qualitativa*	variável quantitativa
tabela de freqüências	medidas de posição: média, mediana, moda
gráfico de barras	medidas de dispersão: variância, desvio-padrão,
diagrama circular (pizza)	amplitude, coeficiente de variação
	tabela de freqüências
	histograma
	boxplot
	gráfico de linha ou seqüência
	polígono de freqüências

*Esta abordagem também pode ser interessante para as variáveis quantitativas discretas.

Tabela de freqüências

Como o nome indica, conterá os valores da variável e suas respectivas contagens, as quais são denominadas *freqüências absolutas* ou simplesmente, *freqüências.* No caso de variáveis qualitativas ou quantitativas discretas, a tabela de freqüência consiste em listar os valores possíveis da variável, numéricos ou não, e fazer a contagem na tabela de dados brutos do número de suas ocorrências. A freqüência do valor *i* será representada por n_i, a freqüência total por *n* e a *freqüência relativa* por $fi = n_i/n$.

Para variáveis cujos valores possuem ordenação natural (qualitativas ordinais e quantitativas em geral), faz sentido incluirmos também uma coluna contendo as *freqüências acumuladas* f_{ac}, obtidas pela soma das freqüências de todos os valores da variável, menores ou iguais ao valor considerado.

No caso das variáveis quantitativas contínuas, que podem assumir infinitos valores diferentes, é inviável construir a tabela de freqüência nos mesmos moldes do caso anterior, pois obteríamos praticamente os valores originais da tabela de dados brutos. Para resolver este problema, determinamos classes ou faixas de valores e contamos o número de ocorrências em cada faixa. Por ex., no caso da variável peso de adultos, poderíamos adotar as seguintes faixas: 30 |— 40 kg, 40 |— 50 kg, 50 |— 60, 60 |— 70, e assim por diante. Apesar de não adotarmos nenhuma regra formal para estabelecer as faixas, procuraremos utilizar, em geral, de 5 a 8 faixas com mesma amplitude. Eventualmente, faixas de tamanho desigual podem ser convenientes para representar valores nas extremidades da tabela.

Exs.:

Número e Proporção (%) de Óbitos, segundo regiões.
Brasil, 1996 e 1999.

Região	n	%
Norte	16117	4,93
Nordeste	69811	21,37
Sudeste	170050	52,05
Sul	48921	14,97
Centro-Oeste	21830	6,68
BRASIL	326729	100,00

Número e Proporção (%) de Óbitos, segundo sexo e regiões.
Brasil, 1996 e 1999.

Região	masculino		feminino	
	n	%	n	%
Norte	10857	4,85	5260	5,12
Nordeste	46242	20,65	23569	22,93
Sudeste	118774	53,04	51276	49,89
Sul	33113	14,79	15808	15,38
Centro-Oeste	14958	6,68	6872	6,69
BRASIL	223944	100,00	102785	100,00

Gráfico de barras

Para construir um *gráfico de barras*, representamos os valores da variável no eixo das abscissas e suas as freqüências ou porcentagens no eixo das ordenadas. Para cada valor da variável desenhamos uma barra com altura correspondendo à sua freqüência ou porcentagem. Este tipo de gráfico é interessante para as variáveis qualitativas ordinais ou quantitativas discretas, pois permite investigar a presença de tendência nos dados.

Ex.:

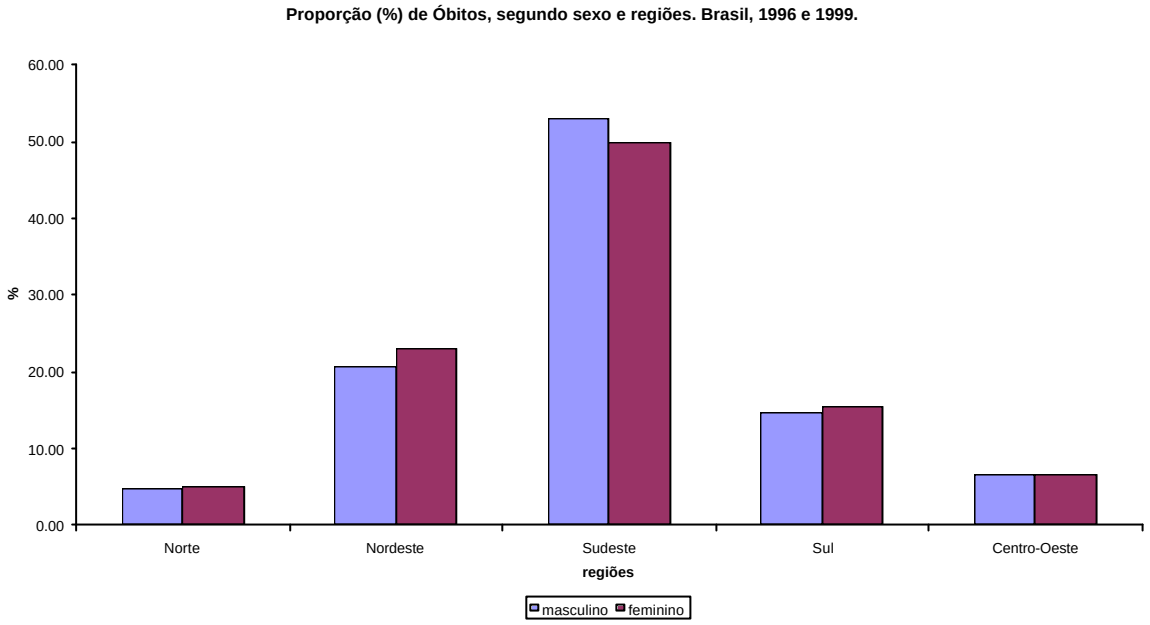
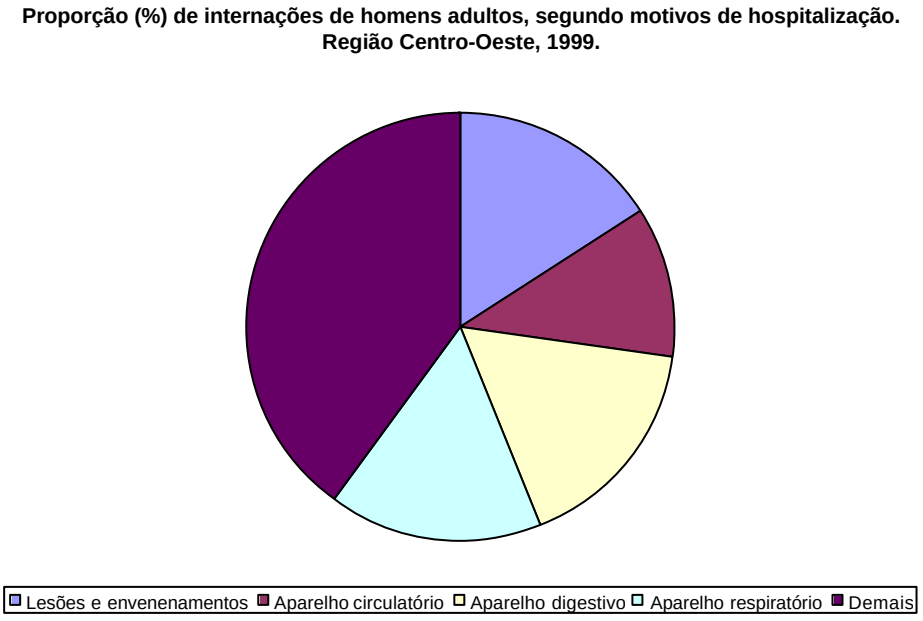


Diagrama Circular

Para construir um *diagrama circular* ou gráfico de *pizza*, repartimos um disco em setores circulares correspondentes às porcentagens de cada valor (calculadas multiplicando-se a freqüência relativa por 100). Este tipo de gráfico adapta-se muito bem para as variáveis qualitativas nominais.

Ex.:

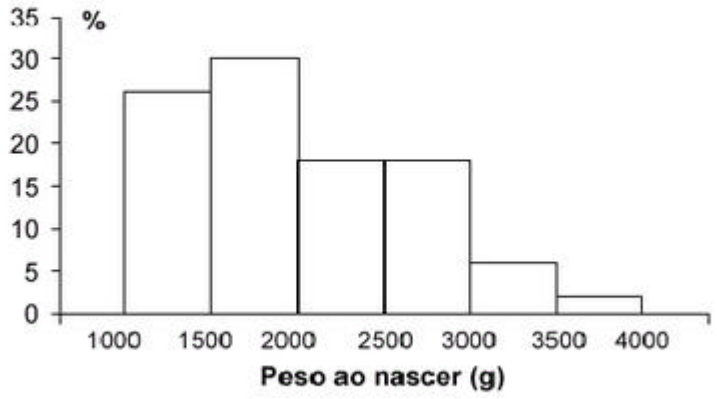


Histograma

O *histograma* consiste em retângulos contíguos com base nas faixas de valores da variável e com área igual à freqüência relativa da respectiva faixa. Desta forma, a altura de cada retângulo é denominada densidade de freqüência ou simplesmente densidade definida pelo quociente da área pela amplitude da faixa. Alguns autores utilizam a freqüência absoluta ou a porcentagem na construção do histograma, o que pode ocasionar distorções (e, conseqüentemente, más interpretações) quando amplitudes diferentes são utilizadas nas faixas.

Ex.:

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



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

Medidas de posição (tendência central)

São medidas que visam localizar o centro de um conjunto de dados, isto é, identificar um valor em torno do qual os dados tendem a se agrupar. As medidas de posição ou de tendência central mais utilizadas são: *média aritmética, mediana e moda.*

média aritmética: é a soma de todas as observações dividida pelo número de observações.

Ex.: média aritmética de 3, 4, 7, 8 e 8.

$$\overline{x} = \frac{3+4+7+8+8}{5} \Rightarrow \overline{x} = \frac{30}{5} \Rightarrow \overline{x} = 6$$

mediana: valor que ocupa a posição central dos dados ordenados; é o valor que deixa metade dos dados abaixo e metade acima dele. Se o número de observações for par, a mediana será a média aritmética dos dois valores centrais.

Ex.: mediana de

a) 3, 4, 7, 8 e 8 ? Md=7

b) 3, 4, 7, 8, 8 e 9 ? $\text{Md} = \frac{7+8}{2} \Rightarrow \text{Md} = \frac{15}{2} \Rightarrow \text{Md} = 7{,}5$

moda: é o valor mais freqüente no conjunto de dados.

Ex.: Número de filhos por funcionário de uma certa empresa:

Nº de filhos	0	1	2	3	5	Total
Freqüência	4	5	7	3	1	20

Medidas de dispersão

As medidas de tendência central fornecem informações valiosas mas, em geral, não são suficientes para descrever e discriminar diferentes conjuntos de dados. As *medidas de dispersão* ou *variabilidade* permitem visualizar a maneira como os dados espalham-se (ou concentram-se) em torno do valor central. Para mensurarmos esta variabilidade podemos utilizar as seguintes estatísticas: *amplitude total; distância interquartílica; desvio médio; variância; desvio padrão e coeficiente de variação.*

Amplitude total: é a diferença entre o maior e o menor valor do conjunto de dados.

Ex.: dados: 3, 4, 7, 8 e 8.

amplitude total = 8 – 3 = 5

Distância interquartílica: é a diferença entre o terceiro e o primeiro quartil de um conjunto de dados. O primeiro quartil é o valor que deixa um quarto dos valores abaixo e três quartos acima dele. O terceiro quartil é o valor que deixa três quartos dos dados abaixo e um quarto acima dele. O segundo quartil é a mediana. (O primeiro e o terceiro quartis fazem o mesmo que a mediana para as duas metades demarcadas pela mediana.) Ex.: quando se discutir o boxplot.

Desvio médio: é a diferença entre o valor observado e a medida de tendência central do conjunto de dados.

Variância: é uma medida que expressa um desvio quadrático médio do conjunto de dados, e sua unidade é o quadrado da unidade dos dados.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Desvio Padrão: é raiz quadrada da variância e sua unidade de medida é a mesma que a do conjunto de dados.

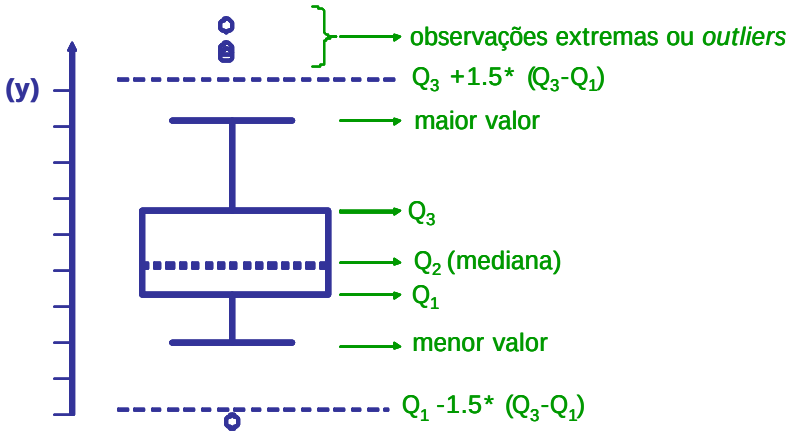
$$s = \sqrt{s^2}$$

Coeficiente de variação: é uma medida de variabilidade relativa, definida como a razão percentual entre o desvio padrão e a média, e assim sendo uma medida adimensional expressa em percentual.

$$cv = \frac{s}{\bar{x}}$$

Boxplot

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de valores, uma vez que são afetados, de forma exagerada, por valores extremos. Além disso, apenas com estas duas medidas não temos idéia da assimetria da distribuição dos valores. Para solucionar esses problemas, podemos utilizar o *Boxplot*. Para construí-lo, desenhamos uma "caixa" com o nível superior dado pelo terceiro quartil (Q_3) e o nível inferior pelo primeiro quartil (Q_1). A mediana (Q_2) é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo, que não sejam observações discrepantes. O critério para decidir se uma observação é discrepante pode variar; por ora, chamaremos de discrepante os valores maiores do que $Q_3+1.5*(Q_3-Q_1)$ ou menores do que $Q_1-1.5*(Q_3-Q_1)$.

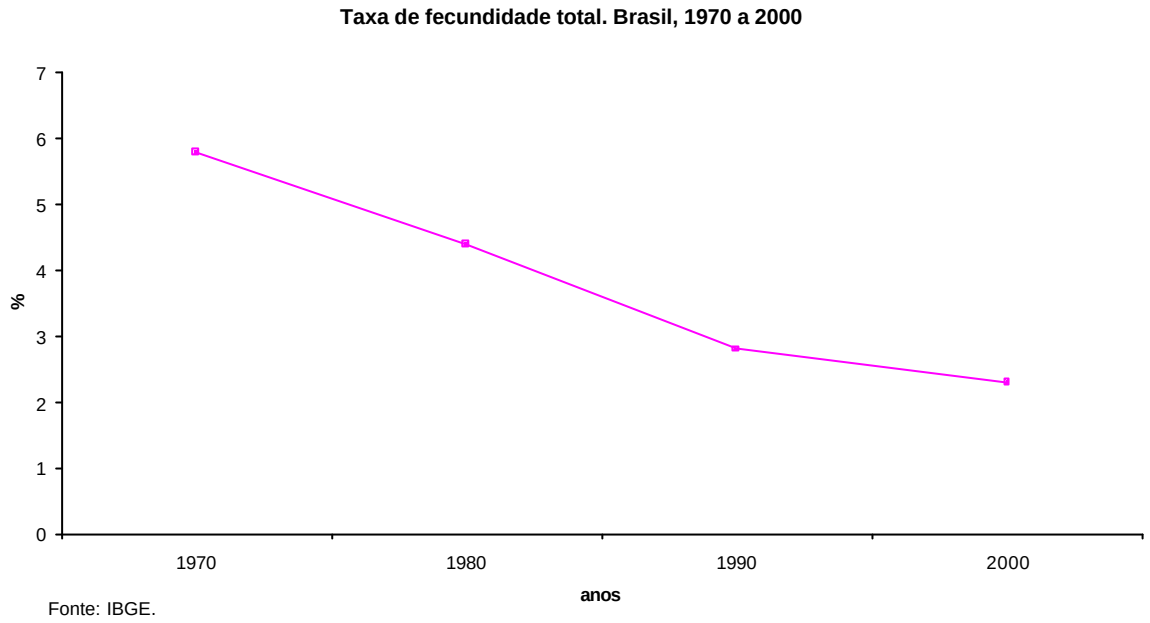


O Boxplot fornece informações sobre posição, dispersão, assimetria, caudas e valores discrepantes.

Gráfico de linha ou seqüência

Adequados para apresentar observações medidas ao longo do tempo, enfatizando sua tendência ou periodicidade.

Ex.:

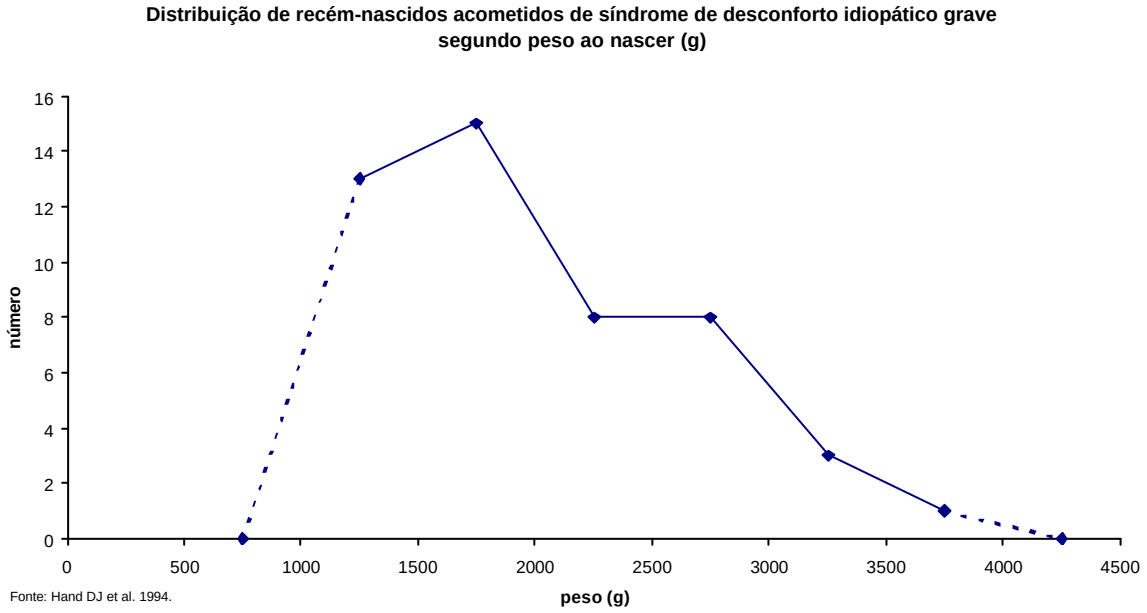


Fonte: IBGE.

Polígono de freqüências

Semelhante ao histograma, mas construído a partir dos pontos médios das classes.

Ex.:

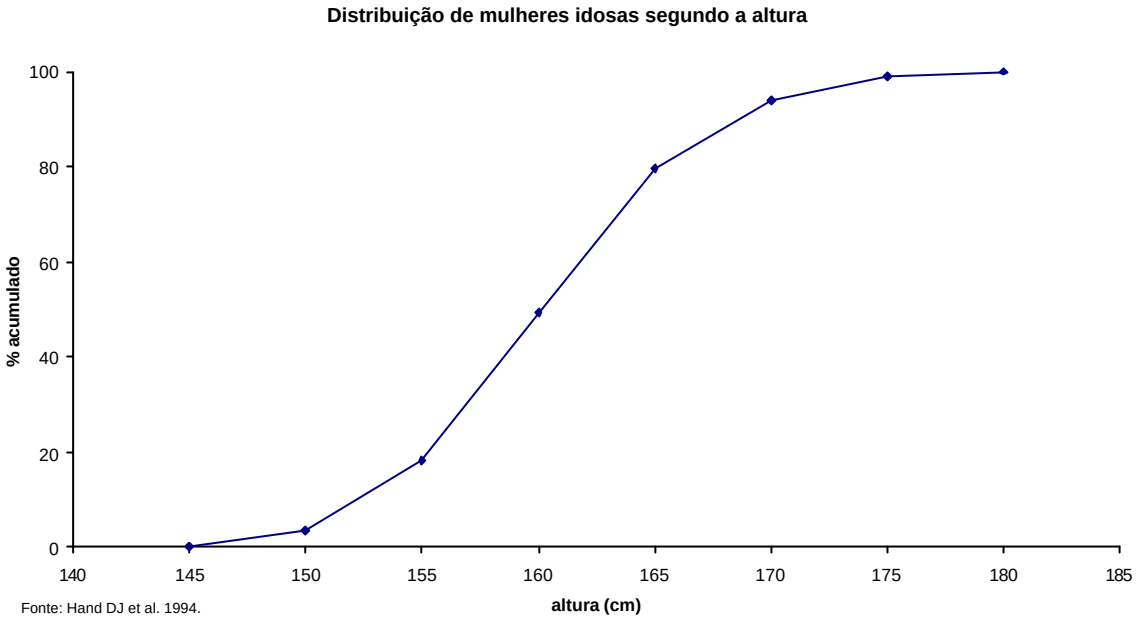


Fonte: Hand DJ et al. 1994.

Gráfico de ogiva

Apresenta uma distribuição de freqüências acumuladas, utiliza uma poligonal ascendente utilizando os pontos extremos.

Ex.:

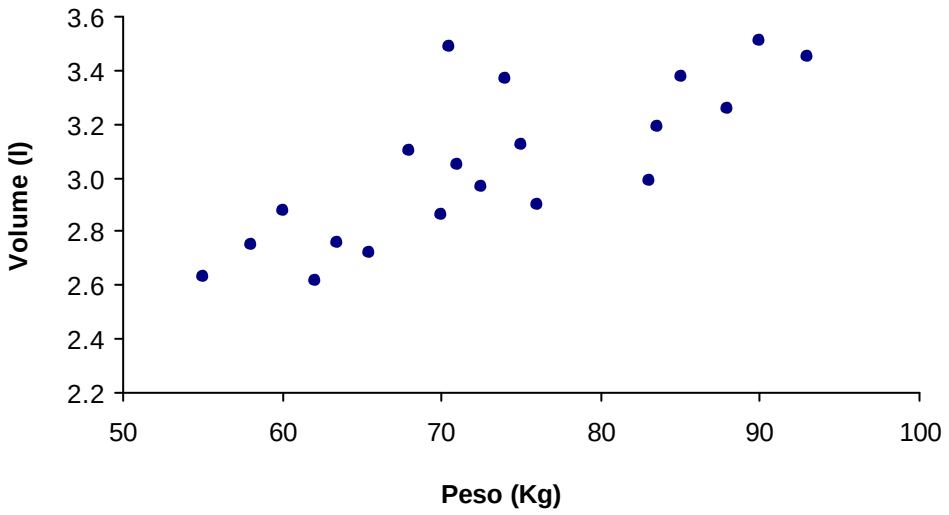


Fonte: Hand DJ et al. 1994.

Diagrama de dispersão

Adequado para descrever o comportamento conjunto de duas variáveis quantitativas. Cada ponto do gráfico representa um par de valores observados.

Ex:



BIBLIOGRAFIA BÁSICA

BUSSAB WO, MORETTIN PA (2002). Estatística Básica. 5ª ed. São Paulo: Saraiva Editora.

CALLEGARI-JACQUES SM (2003). Bioestaística – princípios e aplicações. Porto Alegre: Artmed. 255p.

MAGALHÃES MN, LIMA ACP (2004). Noções de probabilidade e estatística. 6ª ed. São Paulo: Edusp. 392 p.

SOARES JF, SIQUEIRA AL (1999). Introdução à estatística médica. Belo Horizonte, UFMG: Coopmed Editora Médica. 300p.

Exercícios

Um questionário foi aplicado aos alunos do primeiro ano de uma escola fornecendo as seguintes informações:

Id: identificação do aluno
Turma: turma a que o aluno foi colocado (A ou B)
Sexo: F se feminino, M se masculino
Idade: idade, em anos
Alt: altura em metros
Peso: peso em quilogramas
Filhos: número de filhos na família
Fuma: hábito de fumar, sim ou não
Toler: tolerância ao cigarro:
(I) Indiferente, (P) Incomoda pouco e (M) Incomoda muito
Exerc: horas de atividade física, por semana
Cine: número de vezes que vai ao cinema, por semana
OpCine: opinião a respeito das salas de cinema na cidade:
(B) regular a boa e (M) muito boa
TV : horas gastas assistindo TV, por semana
OpTV: opinião a respeito da qualidade da programação na TV:
(R) ruim, (M) média, (B) boa e (N) não sabe

O conjunto de informações disponíveis, após a tabulação do questionário ou pesquisa de campo, é denominado tabela de dados brutos e contém os dados da maneira que foram coletados inicialmente. Os valores obtidos para cada uma dessas informações estão apresentados na Tabela 1.1.

Tabela 1.1. Informações de questionário estudantil – dados brutos

Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Toler	Exerc	Cine	OpCine	TV	OpTV
1	A	F	17	1,60	60,5	2	NAO	P	0	1	B	16	R
2	A	F	18	1,69	55,0	1	NAO	M	0	1	B	7	R
3	A	M	18	1,85	72,8	2	NAO	P	5	2	M	15	R
4	A	M	25	1,85	80,9	2	NAO	P	5	2	B	20	R
5	A	F	19	1,58	55,0	1	NAO	M	2	2	B	5	R
6	A	M	19	1,76	60,0	3	NAO	M	2	1	B	2	R
7	A	F	20	1,60	58,0	1	NAO	P	3	1	B	7	R
8	A	F	18	1,64	47,0	1	SIM	I	2	2	M	10	R
9	A	F	18	1,62	57,8	3	NAO	M	3	3	M	12	R
10	A	F	17	1,64	58,0	2	NAO	M	2	2	M	10	R
11	A	F	18	1,72	70,0	1	SIM	I	10	2	B	8	N
12	A	F	18	1,66	54,0	3	NAO	M	0	2	B	0	R
13	A	F	21	1,70	58,0	2	NAO	M	6	1	M	30	R
14	A	M	19	1,78	68,5	1	SIM	I	5	1	M	2	N
15	A	F	18	1,65	63,5	1	NAO	I	4	1	B	10	R
16	A	F	19	1,63	47,4	3	NAO	P	0	1	B	18	R
17	A	F	17	1,82	66,0	1	NAO	P	3	1	B	10	N
18	A	M	18	1,80	85,2	2	NAO	P	3	4	B	10	R
19	A	F	20	1,60	54,5	1	NAO	P	3	2	B	5	R
20	A	F	18	1,68	52,5	3	NAO	M	7	2	B	14	M
21	A	F	21	1,70	60,0	2	NAO	P	8	2	B	5	R
22	A	F	18	1,65	58,5	1	NAO	M	0	3	B	5	R
23	A	F	18	1,57	49,2	1	SIM	I	5	4	B	10	R
24	A	F	20	1,55	48,0	1	SIM	I	0	1	M	28	R
25	A	F	20	1,69	51,6	2	NAO	P	8	5	M	4	N
26	A	F	19	1,54	57,0	2	NAO	I	6	2	B	5	R
27	B	F	23	1,62	63,0	2	NAO	M	8	2	M	5	R
28	B	F	18	1,62	52,0	1	NAO	P	1	1	M	10	R
29	B	F	18	1,57	49,0	2	NAO	P	3	1	B	12	R
30	B	F	25	1,65	59,0	4	NAO	M	1	2	M	2	R
31	B	F	18	1,61	52,0	1	NAO	P	2	2	M	6	N
32	B	M	17	1,71	73,0	1	NAO	P	1	1	B	20	R
33	B	F	17	1,65	56,0	3	NAO	M	2	1	B	14	R
34	B	F	17	1,67	58,0	1	NAO	M	4	2	B	10	R
35	B	M	18	1,73	87,0	1	NAO	M	7	1	B	25	B
36	B	F	18	1,60	47,0	1	NAO	P	5	1	M	14	R
37	B	M	17	1,70	95,0	1	NAO	P	10	2	M	12	N
38	B	M	21	1,85	84,0	1	SIM	I	6	4	B	10	R
39	B	F	18	1,70	60,0	1	NAO	P	5	2	B	12	R
40	B	M	18	1,73	73,0	1	NAO	M	4	1	B	2	R
41	B	F	17	1,70	55,0	1	NAO	I	5	4	B	10	B
42	B	F	23	1,45	44,0	2	NAO	M	2	2	B	25	R
43	B	M	24	1,76	75,0	2	NAO	I	7	0	M	14	N
44	B	F	18	1,68	55,0	1	NAO	P	5	1	B	8	R
45	B	F	18	1,55	49,0	1	NAO	M	0	1	M	10	R
46	B	F	19	1,70	50,0	7	NAO	M	0	1	B	8	R
47	B	F	19	1,55	54,5	2	NAO	M	4	3	B	3	R
48	B	F	18	1,60	50,0	1	NAO	P	2	1	B	5	R
49	B	M	17	1,80	71,0	1	NAO	P	7	0	M	14	R
50	B	M	18	1,83	86,0	1	NAO	P	7	0	M	20	B

1. Construa a tabela de freqüências para a variável sexo e interprete.

Sexo	n_i	f_i
F		
M		
total	n=50	

n_i = freqüência do valor i

$f_i = n_i / n$

2. Construa a tabela de freqüências para as demais variáveis qualitativas e interprete.

3. Calcule medidas descritivas (de posição e dispersão) para a idade dos estudantes do sexo masculino. Interprete.

4. Construa o boxplot da variável peso para os dois sexos. Interprete.

Feminino

Ordem	Peso
37	70,0
36	66,0
35	63,5
34	63,0
33	60,5
32	60,0
31	60,0
30	59,0
29	58,5
28	58,0
27	58,0
26	58,0
25	58,0
24	57,8
23	57,0
22	56,0
21	55,0
20	55,0
19	55,0
18	55,0
17	54,5
16	54,5
15	54,0
14	52,5
13	52,0
12	52,0
11	51,6
10	50,0
9	50,0
8	49,2
7	49,0
6	49,0
5	48,0
4	47,4
3	47,0
2	47,0
1	44,0

Masculino

Ordem	Peso
13	95,0
12	87,0
11	86,0
10	85,2
9	84,0
8	80,9
7	75,0
6	73,0
5	73,0
4	72,8
3	71,0
2	68,5
1	60,0

5. Uma nova ração foi fornecida a suínos recém desmamados e deseja-se avaliar sua eficiência. A ração tradicional dava um ganho de peso ao redor de 3,5 kg em um mês. A seguir, apresentamos os dados referentes ao ganho, em quilos, para essa nova ração, aplicada durante um mês em 200 animais nas condições acima.

a. Construa o histograma
b. Determine o 1º , 2º e 3º quartis.
c. Você acha que a nova ração é mais eficiente que a tradicional? Justifique.

Ganho de peso (kg)	n_i	f_i	d_i
1.0+ - - - 2.0	45		
2.0+ - - - 3.0	83		
3.0+ - - - 4.0	52		
4.0+ - - - 5.0	15		
5.0+ - - - 6.0	4		
6.0+ - - - 7.0	1		
Total			

6. Como parte de uma avaliação médica em uma empresa, foi medida a freqüência cardíaca dos funcionários de um determinado setor.

Freqüência cardíaca (bpm)	n_i	f_i	d_i
60+ - - - 65	11		
65+ - - - 70	35		
70+ - - - 85	68		
75+ - - - 80	20		
80+ - - - 85	12		
85+ - - - 90	10		
90+ - - - 95	1		
95+ - - - 100	3		
Total			

a. Obtenha o histograma.
b. Freqüências cardíacas que estejam abaixo de 62 ou acima de 92 requerem acompanhamento médico. Qual é a porcentagem de funcionários nestas condições?
c. Uma freqüência ao redor de 72 batidas por minuto é considerada padrão. Você acha que de modo geral esses funcionários se encaixam nesse caso?

8. O que acontece com a média e o desvio padrão:

a. Se um mesmo número é somado a todos os elementos de um conjunto de dados?

b. Se cada elemento de um conjunto de dados for multiplicado por um valor constante.

9. Comente as seguintes afirmativas:

c. Sempre a metade dos dados está abaixo da média.

d. A média é o valor típico de um conjunto de dados.

e. Enquanto tivermos alunos com rendimento abaixo da média, não poderemos descansar.