

Estadística Descritiva

Gilbert Queiroz dos Santos

Rio de Janeiro - RJ
2016

SUMÁRIO

1. Definição de Estatística	3
2. Método Estatístico	3
3. Divisão da Estatística	4
4. População e amostra	5
5. Parâmetro e estatística	5
6. Variável	5
7. Níveis de mensuração dos dados	7
8. Obtenção dos dados	8
9. Amostragem	8
10. Apresentação Tabular	12
11. Séries estatísticas	13
12. Representação gráfica das séries estatísticas	27
13. Características numéricas de uma distribuição de frequências	39
14. Teste de Normalidade	70
15. Correlação	72
16. Associação entre Variáveis Categóricas	74
17 Modelos de Regressão	75
18. Determinação do Tamanho da Amostra	80
Bibliografia	82

1 Definição

Ciência que se preocupa com a organização, descrição, análise e interpretação dos dados experimentais, com base em um conjunto de métodos que se destina a possibilitar a tomada de decisões, face às incertezas (Wallis).

Ou ainda:

É um ramo do conhecimento científico que consta de um conjunto de processos que têm por objeto a observação, a classificação formal e a análise dos fenômenos coletivos ou de massa (finalidade descritiva) e também investigar a possibilidade de fazer inferências indutivas válidas a partir dos dados observados por meio de métodos capazes de permitir esta inferência (finalidade indutiva).

Pode-se dizer que:

“Estatística é a ciência do aprendizado a partir dos dados.”

Mas o que são os dados???

Podemos dizer que dados são coleções de evidências relevantes sobre um fato observado.

Existem várias fontes para se obter dados:

- Governo, indústria ou indivíduos;
- Experiências (experimentos);
- Pesquisa (survey);
- Observações de comportamentos, atitudes etc.

Dados primários: quando são publicados pela própria pessoa ou organização que os haja recolhido. Ex: tabelas do censo demográfico do IBGE.

Dados secundários: quando são publicados ou comunicados por outro pesquisador ou outra organização. Ex: quando determinado jornal publica estatísticas referentes ao censo demográfico extraídas do IBGE.

OBS: É mais seguro trabalhar com fontes primárias. O uso da fonte secundária traz o grande risco de erros de transcrição.

2 Método Estatístico

Tem por finalidade estruturar e a organizar as fases ou etapas que devem ser estabelecidas na abordagem de uma observação estatística.

Suas fases ou etapas principais são:

- Definição do problema;
- Planejamento;
- Coleta de dados;
- Apuração de dados;
- Apresentação de dados;
- Análise e interpretação dos dados.

1º - DEFINIÇÃO DO PROBLEMA: Saber exatamente aquilo que se pretende pesquisar é o mesmo que definir corretamente o problema.

2º - PLANEJAMENTO: Como levantar informações ? Que dados deverão ser obtidos? Qual levantamento a ser utilizado? Censitário? Por amostragem? E o cronograma de atividades? Os custos envolvidos? etc

3º - COLETA DE DADOS: Fase operacional. É o registro sistemático de dados, com um objetivo determinado.

Coleta Direta: quando é obtida diretamente da fonte. Ex: Empresa que realiza uma pesquisa para saber a preferência dos consumidores pela sua marca.

A coleta direta pode ser :

Contínua (registros de nascimento, óbitos, casamentos, etc.),

Periódica (recenseamento demográfico, censo industrial) e

Ocasional (registro de casos de dengue).

Coleta Indireta: É feita por deduções a partir dos elementos conseguidos pela coleta direta, por analogia, por avaliação, indícios ou proporcionalização.

4º - APURAÇÃO DOS DADOS: Resumo dos dados através de sua contagem e agrupamento. É a condensação e tabulação de dados.

5º - APRESENTAÇÃO DOS DADOS: Há duas formas de apresentação, que não se excluem mutuamente. A **apresentação tabular**, ou seja é uma apresentação numérica dos dados em linhas e colunas distribuídas de modo ordenado, segundo regras práticas fixadas pelo Conselho Nacional de Estatística. A **apresentação gráfica** dos dados numéricos constitui uma apresentação geométrica permitindo uma visão rápida e clara do fenômeno.

6º - ANÁLISE E INTERPRETAÇÃO DOS DADOS: A última fase do trabalho estatístico é a mais importante e delicada. Está ligada essencialmente ao cálculo de medidas e coeficientes, cuja finalidade principal é descrever o fenômeno (estatística descritiva). Na estatística indutiva, a interpretação dos dados se fundamenta na Teoria das Probabilidades.

3. Divisão da Estatística

A Estatística divide-se em :

1) Estatística Descritiva:

Que se preocupa com a organização, sumarização e descrição dos dados experimentais. Consiste num conjunto de métodos que ensinam a reduzir uma quantidade de dados bastante numerosa em um número pequeno de medidas, substitutas e representantes daquela massa de dados.

2) Estatística Indutiva:

Que se preocupa com a análise e interpretação dos dados. Consiste em inferir propriedades de um universo a partir de uma amostra com resultados conhecidos.

3) Probabilidade:

Que trata da medição da ocorrência de eventos sujeitos ao aspecto de aleatoriedade.

4. População e Amostra

Objetivando o estudo quantitativo e qualitativo dos dados (ou informações) obtidas nos vários campos da atividade científica, a Estatística manipula dois tipos de conjuntos de dados: a população e a amostra:

a) População- (ou universo) é o conjunto de elementos com pelo menos uma característica comum.

Ex: população de um país, população de um estado, população de município, população de um bairro etc

b) Amostra- é um subconjunto de uma população, necessariamente finito, pois todos os seus elementos serão examinados para efeito da realização do estudo estatístico desejado.

Ex: O Brasil possui 27 unidades federativas (UF), sendo 26 Estados e 1 Distrito Federal. Uma amostra destas unidades poderia ser de 5 UF.

Ou ainda, se estivéssemos interessado em retirar uma amostra de municípios brasileiros, de um total de 5570 municípios, poderíamos escolher 100 municípios, por exemplo.

5. Parâmetro e Estatística

Com relação aos dois tipos de conjuntos de dados : população e amostra, temos os seguintes conceitos na Estatística:

a) **Parâmetro** - é uma medida que se refere à população, ou seja, é obtida com base nos valores da população.

Ex: média (μ), proporção (π), variância (σ^2) e desvio-padrão (σ)

b) **Estatística** – é uma medida que se refere à amostra, ou seja, é obtida com base nos valores da amostra.

Ex: média(\bar{x}), proporção (p), variância (s^2) e desvio-padrão (s).

Na prática, usamos uma estatística para se estimar um parâmetro populacional, que em geral é desconhecido. Ou seja, realizamos um processo de amostragem, que significa retirar uma amostra da população de estudo. Ao fazer isto, estamos cometendo um erro, chamado de erro amostral - ϵ .

O erro amostral (ϵ) é expresso na unidade da variável de estudo. Ele representa a máxima diferença admitida entre o verdadeiro parâmetro populacional (θ) e o seu estimador ($\hat{\theta}$), conhecido como estatística. Então:

$$|\theta - \hat{\theta}| \leq \epsilon$$

6. Variável

Variável é uma característica que pode ser observada ou medida em cada elemento da população ou da amostra, sob as mesmas condições.

Dependendo da pesquisa, uma variável pode ser classificada em variável qualitativa ou variável quantitativa.

Variável qualitativa (categórica): é a que se refere a uma classificação por tipos, categorias ou atributos, ex.: sexo, cor dos olhos, estado civil etc; conseqüentemente, temos as “estatísticas de

atributos”, ou seja, nas variáveis categóricas resumem-se os dados por determinar a frequência de cada uma das categorias observadas e apresentá-las em uma tabela ou gráfico.

Variável quantitativa (numérica): quando seus valores são expressos em números, ex.: idade, peso, altura, renda etc; conseqüentemente, temos as “estatísticas de variáveis”, ou seja, além de verificar frequências, podemos também calcular médias e realizar outras operações.

De acordo com o tipo de variável empregada em uma pesquisa ou estudo, os dados podem ser classificados em:

a) **Dados Nominais ou categóricos:** são aqueles que se referem ao agrupamento e classificação de elementos para a formação de conjuntos distintos (categorias).

Por exemplo: sexo (masculino e feminino)

b) **Dados ordinais:** são aqueles que se referem à avaliação de um fenômeno em termos de sua situação dentro de um conjunto de patamares ordenados, variando desde um patamar mínimo até um patamar máximo.

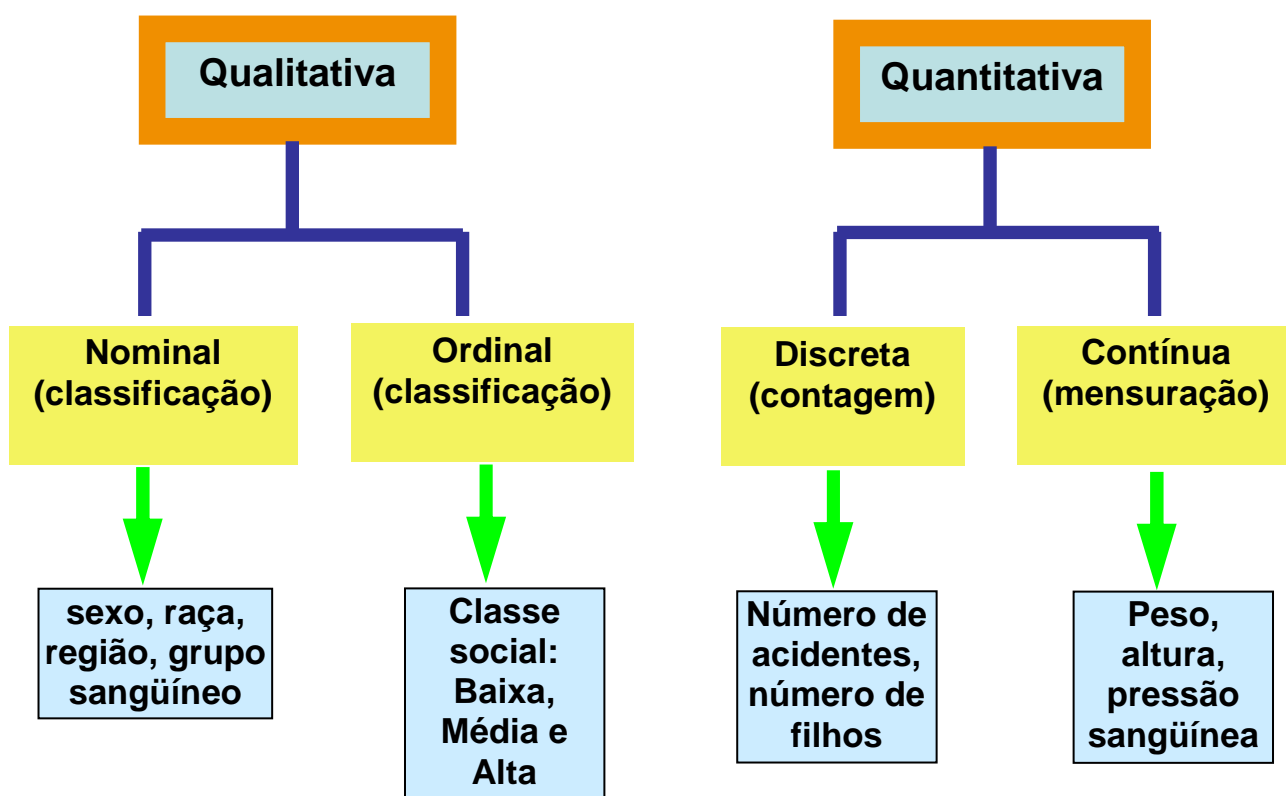
Por exemplo: Nível de escolaridade (fundamental, médio e superior)

c) **Dados discretos :** são aqueles que podem assumir apenas valores pertencentes a um conjunto enumerável, ou seja, a escala numérica se refere ao conjunto dos números inteiros (N).

Por exemplo número de filhos, o ponto obtido em cada jogada, número de defeitos por unidade etc.

d) **Dados contínuos:** são aqueles que assumem quaisquer valores num certo intervalo razoável de variação, ou seja a escala numérica é o conjunto dos números reais (R).

Por exemplo: temperatura, pressão, idade, diâmetro etc.



Quanto à organização, os dados podem ser classificados em:

- a) Dados Brutos - são os dados originais, que ainda não se encontram prontos para análise, pois não foram numericamente organizados
- b) Rol – é um arranjo de dados numéricos em ordem crescente ou decrescente de grandeza.

7. Níveis de Mensuração dos Dados

Na aplicação da Estatística a problemas reais, o nível de mensuração dos dados é um fator de grande importância na determinação de qual procedimento usar, ou seja, quais as possíveis operações aritméticas serão utilizadas e quais técnicas estatísticas serão permitidas para análise.

a) Nível Nominal de Mensuração

É caracterizado pelo ato de nomear ou rotular um objeto, pessoa ou alguma característica. Neste nível de mensuração, não são possíveis operações aritméticas, apenas a contagem de valores.

Ex: sexo (Masculino e Feminino), religião, filiação partidária, estado civil, raça, profissões etc.

b) Nível Ordinal de Mensuração

Neste nível, os dados, além de apresentarem as propriedades inerentes da escala nominal, são postos em ordem do menor ao maior, de forma significativa. A relação $>$ (maior do que) vale para todos os dados, e com isto temos uma escala ordinal.

Ex: status socioeconômico, grau de escolar, hierarquização funcional etc.

c) Nível Intervalar de Mensuração

Neste nível, observam-se que os dados, além de apresentarem as propriedades inerentes da escala ordinal, apresentam intervalos iguais de medição, ou seja, em uma unidade de medida fixa.

d) Nível de Razão

Neste nível, observam-se que os dados, além de apresentarem as propriedades inerentes da escala intervalar, apresentam um quociente significativo entre dois valores, ou seja, uma razão entre os pares de valores no conjunto ordenado.

Temos o seguinte resumo para os níveis de mensuração:

Níveis	Tipo de dados	Operações
Nominal	Numéricos / Não numéricos	Contagem, Proporção
Ordinal	Numéricos/ Não numéricos	Contagem, Proporção
Intervalar	Numéricos	Contagem, proporção, médias
Razão	Numéricos	Contagem, proporção, médias

8. Obtenção dos Dados

Podemos obter os dados da seguinte forma:

- 1) Realizando um censo, ou seja, realizando a coleção de dados obtidos de todos os membros da população. Sua execução, porém, é complexa e envolve muitos recursos e tempo.
- 2) Por meio de uma pesquisa por amostragem (survey), ou seja, realizando o dimensionamento, os critérios para composição e seleção de uma amostra. Sua execução é mais prática.
- 3) Executando um experimento, ou seja, aplicando um determinado tratamento a uma parte da população (amostra) e observando os resultados.
- 4) Por meio de simulação, ou seja, usando um modelo matemático ou físico para reproduzir as condições de uma situação ou processo.

9. Amostragem

Dentre as diversas maneiras de coletar dados, a amostragem é mais freqüente, particularmente nas pesquisas sobre fenômenos sociais e econômicos,

Uma amostra pode ser probabilística, ou seja, quando os elementos amostrais são escolhidos com probabilidades conhecidas.

Uma amostra não-probabilística é aquela em que os elementos amostrais não são escolhidos com probabilidades, ou seja, a escolha dos elementos amostrais é feita de forma deliberada.

9.1 Métodos de Amostragem Probabilística

Os métodos de amostragem probabilísticas mais conhecidos são:

- Amostragem Aleatória Simples (AAS)
- Amostragem Sistemática
- Amostragem Aleatória Estratificada (AAE)
- Amostragem por Conglomerado (em um estágio ou em estágios múltiplos)

Os métodos de amostragem não-probabilísticas são:

- Amostragem de Conveniência
- Amostragem por Cotas

9.2 Determinação Inicial do Tamanho da Amostra

Antes de se escolher qual o método de amostragem a ser utilizado, devemos ter uma noção do tamanho inicial da amostra. Neste caso, teremos como base o erro amostral, dado por:

$$|\theta - \hat{\theta}| \leq \varepsilon$$

E o tamanho N da população alvo do estudo. Usa-se a seguinte expressão para a determinação inicial do tamanho da amostra:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Onde: $n_0 = \frac{1}{\varepsilon^2}$

n_0 – primeira aproximação da amostra

N – tamanho da população

Por exemplo: se $\varepsilon = 0,05$ e $N = 200.000$, temos:

$$n_0 = \frac{1}{\varepsilon^2} = \frac{1}{(0,05)^2} = \frac{1}{0,0025} = 400$$

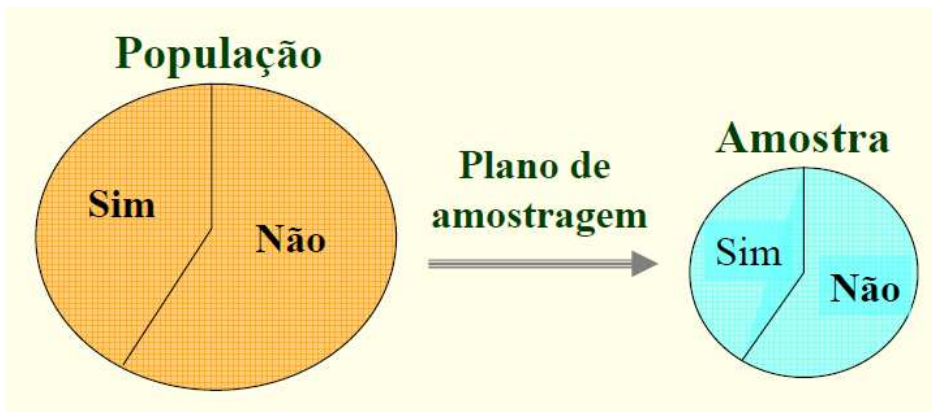
$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{400}{1 + \frac{400}{200000}} = \frac{400}{1,002} = 399,20 \cong 399$$

Se aumentarmos o erro, por exemplo, para $\varepsilon = 0,10$, teremos:

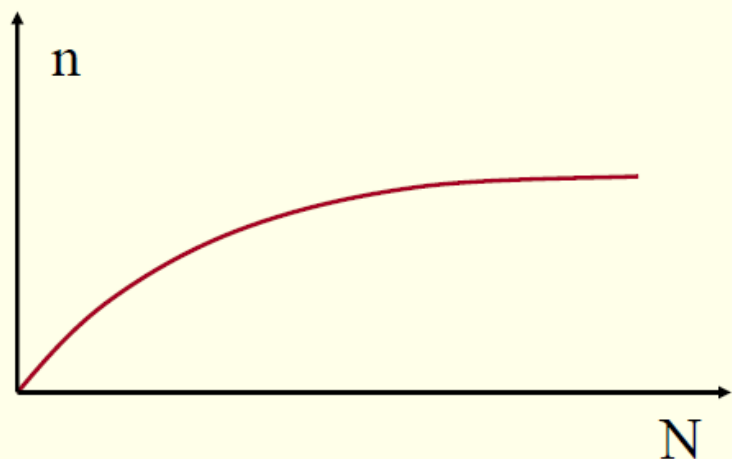
$$n_0 = \frac{1}{\varepsilon^2} = \frac{1}{(0,10)^2} = \frac{1}{0,01} = 100$$

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{100}{1 + \frac{100}{200000}} = \frac{100}{1,0005} = 99,95 \cong 100$$

É necessário considerar que amostra deve ser representativa da população, ou seja:



Logo, é errôneo pensar que o tamanho da amostra deve ser tomado como um percentual do tamanho da população para ser representativa



Por exemplo:

Observe que: $N = 200$ famílias, $E_0 = 4\%$
 $n = 152$ famílias → 76% da população

Observe que: $N = 200.000$ famílias, $E_0 = 4\%$
 $n = 623$ famílias → 0,3% da população

Ou seja, o que influencia o tamanho da amostra é o tamanho da população em estudo e o erro amostral admitido.

9.3 Processo de sorteio dos elementos da amostra

Uma vez determinado o tamanho inicial da amostra, deve-se realizar o sorteio dos elementos que irão compô-la. Este processo depende do Método de Amostragem a ser adotado.

a) Amostragem Aleatória Simples

Neste método, todos os elementos da população têm a mesma chance (probabilidade – $1/n$) de serem selecionados. Atribui-se a cada elemento da população um número distinto. Efetuam-se sucessivos sorteios até completar o tamanho da amostra n . Para realizar o sorteio, utilizar a Tabela de Números Aleatórios - TNA (anexo) que consistem em tabelas que apresentam dígitos de 0 à 9 distribuídos aleatoriamente.

Por exemplo:

Suponha uma população com 500 elementos, que numeramos de 000 a 499 para selecionar uma amostra aleatória de $n=50$ elementos.

O processo termina quando for sorteado o elemento 50. A probabilidade de cada elemento ser selecionado é $p=1/50$

b) Amostragem Sistemática

Conveniente quando a população está ordenada segundo algum critério como fichas, lista telefônica etc.

Procedimento:

1. Definir intervalo de seleção: $K = \frac{N}{n}$

Onde: N=número de elementos da população
n=número de elementos da amostra

2. Determinar o ponto de partida (a_1): sorteio aleatório simples. (de 1 até k)

3. Determinar os elementos da amostra através de uma progressão aritmética.

$$a_n = a_1 + (n-1)K$$

Exemplo:

Se N = 5.000 é o tamanho da população e precisamos de uma amostra de n = 250, dividimos $N/n = 20$. Seleccionamos ao acaso um número de 1 à 20. Suponha que saiu o número 7:

1a unidade a ser seleccionada 7a

2a unidade a ser seleccionada $20 + 7 = 27a$

3a unidade a ser seleccionada $27 + 20 = 47a$

67a, 87a,..., 4987a dando um total de 250 unidades.

c) Amostragem Estratificada

Neste caso, os elementos da população estão agrupados em subpopulações mais ou menos homogêneas denominadas estratos, e distintos entre si. Os estratos são mutuamente exclusivos, ou seja $N_1 + N_2 + \dots + N_k = N$.

Após a determinação dos estratos, selecciona-se uma amostra aleatória simples de cada estrato. Existem dois tipos de amostragem estratificada:

1) De mesmo tamanho ou Uniforme;

2) Proporcional.

No primeiro tipo sorteia-se igual número de elementos em cada estrato. Esse processo é utilizado quando o número de elementos por estrato for aproximadamente o mesmo, ou seja, $n_1 = n_2 = \dots = n_k$ e $n_1 + n_2 + \dots + n_k = n$

No outro caso, utiliza-se proporção para determinar o número de elementos de cada estrato que irão compor a amostra, ou seja, $n_1 \neq n_2 \neq \dots \neq n_k$, mas $n_1 + n_2 + \dots + n_k = n$

As variáveis de estratificação mais comuns são: classe social, idade, sexo, profissão.

Exemplo: Numa localidade com 150 000 habitantes, 45 000 têm menos de 20 anos de idade, 75 000 têm idades entre 30 e 50 anos e 30 000 têm mais de 50 anos de idade. Extrair uma amostra de 30 habitantes desta população pelo processo de amostragem estratificada com partilha proporcional.

$N = 150\,000$, $N_1 = 45\,000$, $N_2 = 75\,000$, $N_3 = 30\,000$ e $n = 30$

$$n_1 = 30 \frac{45\,000}{150\,000} \therefore \boxed{n_1 = 9}; \quad n_2 = 30 \frac{75\,000}{150\,000} \therefore \boxed{n_2 = 15}; \quad n_3 = 30 \frac{30\,000}{150\,000} \therefore \boxed{n_3 = 6}$$

$$\text{Peso 1} = w_1$$

$$\text{Peso 2} = w_2$$

$$\text{Peso 3} = w_3$$

A amostra deverá conter 9 habitantes com menos de 20 anos, 15 com idades entre 20 e 50 anos e 6 com mais de 50 anos.

10. Apresentação Tabular

Um dos métodos usados para a apresentação de dados estatísticos que consegue expor os resultados sobre determinado assunto num só local, sinteticamente, de tal modo que se tenha uma visão mais globalizada daquilo que se vai analisar.

A apresentação tabular dos dados estatísticos se faz mediante tabelas (ou quadros), resultantes da disposição dos respectivos dados em linhas e colunas distribuídas de modo ordenado, seguindo regras práticas adotadas pelos diversos sistemas estatísticos. No Brasil, essas regras foram fixadas pelo Conselho Nacional de Estatística, por meio da Resolução nº 886, de 26 de outubro de 1966.

10.1 Tabela

Define-se tabela como um conjunto de dados estatísticos associados a um fenômeno, dispostos em uma ordem de classificação, em uma organização racional e prática de apresentação.

Uma tabela pode ser simples ou de dupla entrada.

10.1.1 Tabela simples

É aquela composta de uma coluna matriz, também chamada coluna indicadora, onde vão inscritos os valores ou modalidades de ordem de classificação e da coluna em que aparecem os valores que representam as ocorrências ou intensidades do fenômeno em causa.

10.1.2 Tabela de dupla entrada

É aquela própria à apresentação das distribuições de dois atributos, qualitativos ou quantitativos, em que existem duas ordens de classificação: uma horizontal e outra em coluna indicadora; nos cruzamentos formados pelas linhas com as colunas encontra-se a frequência dos indivíduos que apresentam conjuntamente as alternativas correspondentes à linha e à coluna que sobre ela se cruzam.

10.2 Elementos de uma Tabela

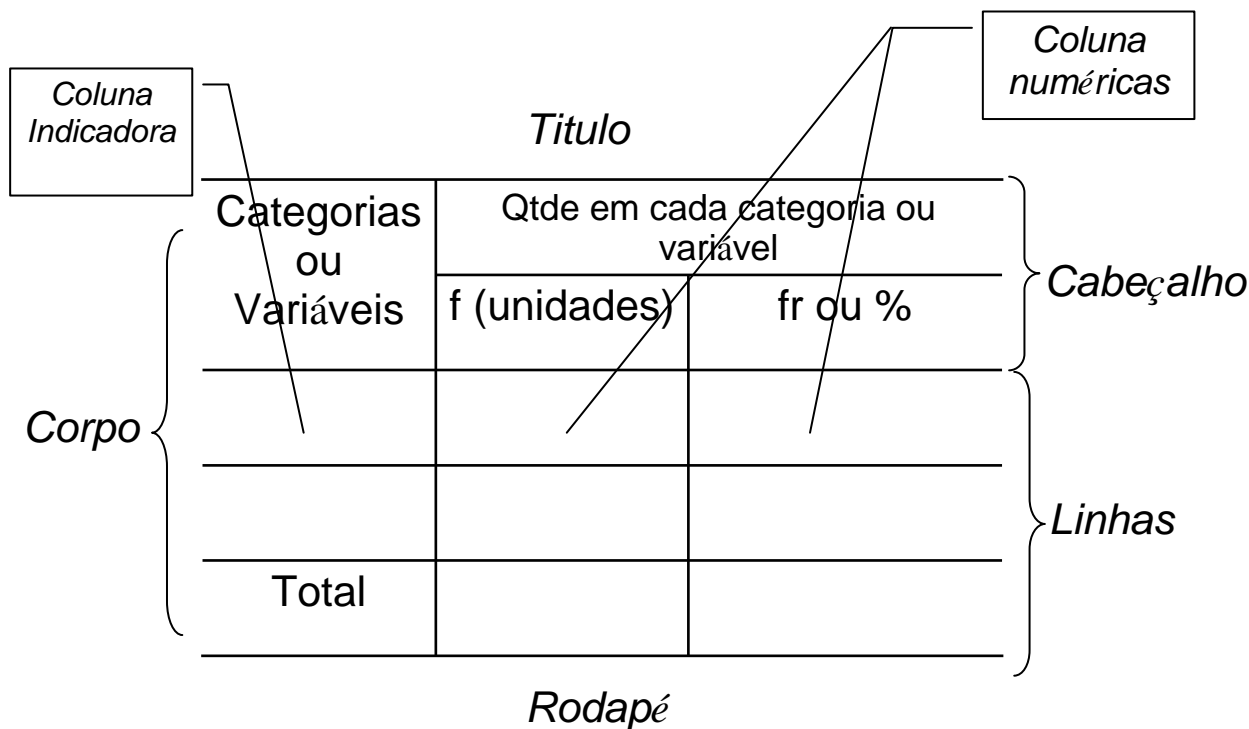
No Brasil, a apresentação tabular é regida pelas Normas de Apresentação Tabular do IBGE (1993)/NBR 14724 da ABNT. As tabelas estatísticas compõem-se de elementos essenciais e elementos complementares.

a) Elementos essenciais:

Os elementos essenciais de uma tabela são: título, corpo, cabeçalho e coluna-indicadora.

b) Elementos complementares:

Os elementos complementares de uma tabela estatística são: fonte, notas e chamadas, todos situados no rodapé da tabela.



11. Séries Estatísticas

Denomina-se série estatística a um conjunto de valores numéricos associados a um fenômeno e que expressa suas variações no tempo, no local e na espécie.

As séries podem ser divididas em dois grupos:

- Séries homógradas;
- Séries heterógradas.

11.1 Séries homógradas

Aplicadas no caso em que a variável é discreta.

As séries temporais, geográficas e específicas formam as principais séries homógradas.

a) Séries temporais (cronológicas, evolutivas, históricas ou marchas)

São séries em que a variável de estudo varia em função da época ou do tempo, permanecendo fixos a região ou o local e o fenômeno.

Exemplo:

Produção de Petróleo Bruto – Brasil (1000 m³)

Anos	Produção
1976	9.702
1977	9.332
1978	9.304
1979	9.608
1980	10.562

Fonte: Conjuntura Econômica, fev/83

b) Séries geográficas (espaciais ou de localização, territoriais)

São séries em que a variável de estudo varia em função da região, do local ou do espaço, permanecendo fixos a época ou o tempo e o fenômeno.

População Estimada por Estado - 2007

Estados	População
Rio de Janeiro	15.420.375
São Paulo	39.827.570
Ceará	8.185.286
Amazonas	3.221.939
Minas Gerais	19.273.506

Fonte: IBGE

c) Séries específica (categóricas, qualitativas)

São séries em que a variável de estudo varia em função do fenômeno, permanecendo fixos a época ou o tempo e a região ou local.

Produção Agrícola no Brasil – 1974
(Produtos Seleccionados)

Especificações	Produção em 1.000 t
Algodão em caroço	1.959
Cacau	165
Café	3.220
Cana de açúcar	96.412
Soja	7.876

Fonte: Revista Comércio e Mercado, mar/76

Freqüentemente, são usadas séries estatísticas conjugadas, onde são cruzados dois ou mais tipos de séries; pode-se ter as conjugações geográfico-temporal (ou espaço-temporal), geográfico-especificativa, especificativo-temporal, especificativo-geográfico-temporal etc.

Exemplos:

a) Série geográfico-temporal (espaço-temporal)

Agências do Banco do Brasil - 2011 a 2012

Estados	2011	2012
Rio de Janeiro	10	15
Ceará	12	20
Amazonas	5	10
Minas Gerais	20	30

Fonte: IBGE

b) Série geográfico-específica

Produção das principais lavouras do Nordeste

Estados	Produção (1.000 t)	
	Arroz	Arroz
Maranhão	11	16
Ceará	13	21
Bahia	6	12
Pernambuco	21	32

Fonte: IBGE

c) Série específico-temporal

Evolução do corpo docente do Sistema Educacional (2010 – 2011)

Nível	Anos	
	2010	2011
Básico	10.000	15.000
Fundamental	12.000	20.000
Superior	20.000	30.000

Fonte: INEP

11.2 Séries heterógradas

Mais comumente chamadas de Distribuições de Frequências, mantendo fixos a época, a região e o fenômeno.

11.2.1 Distribuições de Frequência (série de frequências)

É uma série em que o fenômeno, a época e a região permanecem fixos, porém o fenômeno pode ser subdividido em grupos de classes que têm a finalidade de tornar mais cômodo o estudo.

Defini-se **frequência** (ou **frequência simples**) de um dado valor de uma variável (qualitativa ou quantitativa) como o número de vezes que esse valor foi observado.

Denota-se a frequência do i -ésimo valor observado por f_i .

Define-se **frequência total** f_t como a soma de todos os elementos observados nas frequências simples. Sendo o n o número total de valores observados, verifica-se imediatamente que:

$$f_t = \sum_{i=1}^n f_i = n$$

Define-se **frequência relativa** fr_i (ou frequência relativa simples), ou proporção, de um dado valor de uma variável (qualitativa ou quantitativa), como o quociente de sua frequência pelo número total de elementos observados, da seguinte forma:

$$fr_i = \frac{f_i}{n}$$

Lembrando que: $\sum_{i=1}^n f_i = n$

Define-se **freqüência absoluta acumulada** F (ou F_{ac}) como a soma das freqüências simples das classes inferiores com a da classe considerada, da seguinte forma:

$$F_j = \sum_{i=1}^{j \leq k} f_i$$

Define-se **freqüência relativa acumulada** F_{ri} (ou F_{ra}) como o quociente da freqüência absoluta acumulada (F) pelo total de dados observados (n), ou seja:

$$F_{ri} = \frac{F_i}{n}$$

Estas freqüências são condensadas em uma única tabela, de fácil manejo, denominada Tabela de Distribuição de Freqüências.

Dependendo da variável de estudo (qualitativa ou quantitativa), as tabelas de distribuição de freqüências serão classificadas em:

- Tabelas de Freqüência para Dados não Agrupados ou não Tabulados em Classe;
- Tabelas de Freqüência de Dupla Entrada para Dados não Agrupados ou não Tabulados em Classe;
- Tabelas de Freqüência para Dados Agrupados ou Tabulados em Classe.

1) Tabela de Freqüência para Dados não Agrupados ou não Tabulados em Classe

a) Variável qualitativa

Neste caso, usamos uma tabela simples, onde em uma coluna são apresentadas as categorias, em outra as freqüências e em uma terceira as freqüências relativas, conforme exemplo abaixo:

Título:

Categorias	Freqüências			
	f (unidades)	F	fr ou (%)	Fra ou (%)
Total	$\sum_{i=1}^n f_i = n$			

Fonte:

Exemplo:

Distribuição dos fundos relativos por Estados

Categorias	Freqüências			
	f (unidades)	F	fr ou (%)	Fra ou (%)
São Paulo	38	38	0,281 ou 28,1%	0,281 ou 28,1%
Rio de Janeiro	30	68	0,222 ou 22,2%	0,503 ou 50,3%
Rio Grande do Sul	35	103	0,259 ou 25,9%	0,762 ou 76,2%
Minas Gerais	15	118	0,111 ou 11,1%	0,873 ou 87,3%
Demais Estados	17	135	0,127 ou 12,7%	1 ou 100%
Total	135		1 ou 100%	

b) Variável quantitativa discreta

Neste caso, usamos uma tabela simples, onde em uma coluna são apresentados os valores da variável, e nas outras as frequências, conforme exemplo abaixo:

Variável Discreta	Título			
	Frequências			
	f (unidades)	F	fr ou (%)	Fra ou (%)
Valor 1				
Valor n				
Total	$\sum_{i=1}^n f_i = n$			

Fonte:

Exemplo:

a) Seja a variável qualitativa (categórica) Sexo, dada abaixo:

> Sexo

[1] "MASCULINO" "FEMININO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO"
 [7] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO"
 [13] "MASCULINO" "MASCULINO" "MASCULINO" "FEMININO" "MASCULINO" "FEMININO"
 [19] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [25] "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [31] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "FEMININO"
 [37] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [43] "MASCULINO" "MASCULINO" "MASCULINO"

Vamos montar uma tabela de distribuição de frequências para esta variável. A tabela seria assim:

Distribuição por Sexo				
Sexo	f	F	Fr	Fra
Feminino	11	11	24,44	24,44
Masculino	34	45	75,56	75,56
Total	45	-	100,00	-

b) Seja o número de defeitos por unidade, obtidos a partir de aparelhos retirados de uma linha de montagem:

2, 4, 2, 1, 2, 3, 1, 0, 5, 1, 0, 1, 1, 2, 0, 1, 3, 0, 1, 2

A tabela seria:

Distribuição no N° de Defeitos por Unidade				
N° de defeitos	f	F	Fr	Fra
0	4	4	0,20	0,20
1	7	11	0,35	0,55
2	5	16	0,25	0,80
3	2	18	0,10	0,90
4	1	19	0,05	0,95
5	1	20	0,05	1,00
Total	20		1,00	

2) Tabela de Frequência Dupla Entrada para Dados não Agrupados ou não Tabulados em Classe

Este tipo de tabela se aplica quando estamos trabalhando com duas ou mais variáveis. Neste caso, estaremos interessados em realizar uma análise conjunta das variáveis escolhidas. A tabela de dupla entrada, tem seguinte forma:

Título:

Variável 1	Variável 2			
	Catg. 1	...	Catg n	Total
Catg. 1				
...				
Catg. n				
Total				

Fonte:

Por exemplo:

Seja a variável Sexo, com os seguintes dados:

[1] "MASCULINO" "FEMININO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO"
 [7] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO"
 [13] "MASCULINO" "MASCULINO" "MASCULINO" "FEMININO" "MASCULINO" "FEMININO"
 [19] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [25] "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [31] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "FEMININO"
 [37] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
 [43] "MASCULINO" "MASCULINO" "MASCULINO"

Seja a variável Cor dos Olhos, com os seguintes dados:

[1] "CASTANHOS" "CASTANHOS" "VERDES"
 [4] "CASTANHOS" "AZUIS" "CASTANHOS"
 [7] "CASTANHOS" "CASTANHOS" "CASTANHOS"
 [10] "CASTANHOS ESCUROS" "CASTANHOS" "CASTANHOS CLAROS"
 [13] "VERDES" "CASTANHOS ESCUROS" "CASTANHOS ESCUROS"
 [16] "CASTANHOS" "CASTANHOS" "CASTANHOS"
 [19] "CASTANHOS ESCUROS" "CASTANHOS" "CASTANHOS"
 [22] "CASTANHOS" "CASTANHOS" "CASTANHOS ESCUROS"
 [25] "CASTANHOS ESCUROS" "CASTANHOS ESCUROS" "VERDES"
 [28] "CASTANHOS" "CASTANHOS ESCUROS" "CASTANHOS"
 [31] "CASTANHOS" "CASTANHOS" "CASTANHOS"
 [34] "CASTANHOS" "CASTANHOS ESCUROS" "CASTANHOS"
 [37] "CASTANHOS" "CASTANHOS" "CASTANHOS ESCUROS"
 [40] "CASTANHOS" "CASTANHOS ESCUROS" "CASTANHOS"
 [43] "CASTANHOS" "CASTANHOS" "CASTANHOS"

Colocando as duas variáveis, lado a lado, temos:

	Sexo	Cor dos Olhos
[1,]	"MASCULINO"	"CASTANHOS"
[2,]	"FEMININO"	"CASTANHOS"
[3,]	"FEMININO"	"VERDES"
[4,]	"FEMININO"	"CASTANHOS"
[5,]	"MASCULINO"	"AZUIS"
[6,]	"MASCULINO"	"CASTANHOS"
[7,]	"MASCULINO"	"CASTANHOS"
[8,]	"FEMININO"	"CASTANHOS"
[9,]	"FEMININO"	"CASTANHOS"
[10,]	"MASCULINO"	"CASTANHOS ESCUROS"

: : :

Desta forma, temos que contar os pares (Sexo, Cor dos Olhos). Assim, vamos poder construir a seguinte tabela:

Título: Distribuição por Sexo e Cor dos Olhos				
Sexo	Cor dos Olhos			
	Azuis	Castanhos	Verdes	Total
Feminino	0	10	1	11
Masculino	1	31	2	34
Total	1	41	3	45

As tabelas anteriores podem ser elaboradas com o auxílio o programa R, da seguinte forma:

a) para a variável categórica "Sexo":

> Sexo

```
[1] "MASCULINO" "FEMININO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO"
[7] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO"
[13] "MASCULINO" "MASCULINO" "MASCULINO" "FEMININO" "MASCULINO" "FEMININO"
[19] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
[25] "FEMININO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
[31] "MASCULINO" "FEMININO" "FEMININO" "MASCULINO" "MASCULINO" "FEMININO"
[37] "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO" "MASCULINO"
[43] "MASCULINO" "MASCULINO" "MASCULINO"
```

Comandos necessários:

```
> tab.sexo<-table(Sexo) # cria a tabela só com as frequências simples
```

```
> tab.sexo # verificando a tabela
```

Sexo

FEMININO MASCULINO

11 34

```
> F<-cumsum(tab.sexo) # faz a soma acumulada das frequências
```

```
> tab.sexo.p<-prop.table(tab.sexo)*100 # cria as frequências relativas
```

```
> Fra<-cumsum(tab.sexo.p) # faz a soma acumulada das frequências relativas
```

```
> dist<-cbind(tab.sexo,F, tab.sexo.p, Fra) # cria a tabela completa
```

```
> dist # mostra a tabela completa
```

```
      tab.sexo  F tab.sexo.p      Fra
FEMININO      11 11   24.44444  24.44444
MASCULINO      34 45   75.55556 100.00000
```

Falta adicionar a linha "total" na tabela acima. Os comandos são:

```
Total<-c(sum(tab.sexo), NA, sum(tab.sexo.p), NA) # cria a linha "Total"
```

```
Total # verifica o conteúdo
```

```
dist<-rbind(dist, Total) # adiciona a linha na tabela
```

```
dist # verifica o resultado final
```

```
      tab.sexo  F tab.sexo.p      Fra
FEMININO      11 11   24.44444  24.44444
MASCULINO      34 45   75.55556 100.00000
Total          45 NA  100.00000      NA
```

Manualmente, você pode incluir o nome "Sexo" para coluna onde aparece as categorias, alterar o nome da coluna "tab.sexo" por "f" e da coluna "tab.sexo.p" por "fr". Com isto, o resultado será:

Sexo	f	F	fr	Fra
FEMININO	11	11	24.44444	24.44444
MASCULINO	34	45	75.55556	100.00000
Total	45	NA	100.00000	NA

O R possui um pacote para fazer esta tabela no Word. Para isto, é necessário instalar alguns pacotes e fazer alguns ajustes.

b) Para a variável numérica "Número de defeitos":

```
> def
[1] 2 4 2 1 2 3 1 0 5 1 0 1 1 2 0 1 3 0 1 2
```

Comandos necessários:

```
> tab.def<-table(def) # cria a tabela
> F1<-cumsum(tab.def) # faz a soma acumulada das frequências simples
> tab.def.p<-prop.table(tab.def)*100# cria as frequências relativas
> Fra1<-cumsum(tab.def.p) # faz a soma acumulada das frequências relativas
> dist2<-cbind(tab.def,F1, tab.def.p, Fra1) # cria a tabela completa
> dist2 # mostra a tabela completa
  tab.def F1 tab.def.p Fra1
0      4  4      20    20
1      7 11      35    55
2      5 16      25    80
3      2 18      10    90
4      1 19       5    95
5      1 20       5   100
```

Falta adicionar a linha "total" na tabela acima. Os comandos são:

```
Total2<-c(sum(tab.def), NA, sum(tab.def.p), NA) # cria a linha total
Total2 # verifica o conteúdo
dist2<-rbind(dist2, Total2) # adiciona a linha "total" na tabela
dist2 # verifica o resultado final
```

	tab.def	F1	tab.def.p	Fra1
0	4	4	20	20
1	7	11	35	55
2	5	16	25	80
3	2	18	10	90
4	1	19	5	95
5	1	20	5	100
Total2	20	NA	100	NA

Da mesma forma, manualmente, você pode incluir o nome "Nº de defeitos" para coluna onde aparece os valores dos defeitos, alterar o nome da coluna "tab.def" por "f" e da coluna "tab.def.p" por "fr". Com isto, o resultado será:

Nº de defeitos	f	F	fr	Fra
0	4	4	20	20
1	7	11	35	55
2	5	16	25	80
3	2	18	10	90

4	1	19	5	95
5	1	20	5	100
Total	20	NA	100	NA

c) Para duas variáveis categóricas - tabela de dupla entrada

Trabalhando com duas variáveis categóricas, por exemplo:

```

Sexo    Cor dos Olhos
[1,] "MASCULINO" "CASTANHOS"
[2,] "FEMININO" "CASTANHOS"
[3,] "FEMININO" "VERDES"
[4,] "FEMININO" "CASTANHOS"
[5,] "MASCULINO" "AZUIS"
[6,] "MASCULINO" "CASTANHOS"
[7,] "MASCULINO" "CASTANHOS"
[8,] "FEMININO" "CASTANHOS"
[9,] "FEMININO" "CASTANHOS"
[10,] "MASCULINO" "CASTANHOS ESCUROS"
:
:
:

```

Comandos necessários:

```

> tab.dupla<-table(Sexo, cor.olhos) # cria a tabela de dupla entrada
> tab.dupla # verifica os resultados

```

	cor.olhos					
Sexo	AZUIS	CASTANHOS	CASTANHOS	CASTANHOS CLAROS	CASTANHOS ESCUROS	VERDES
FEMININO	0	9	0	0	1	1
MASCULINO	1	19	1	1	10	2

Falta adicionar os totais por linha e coluna. Então fazemos o seguinte:

```

> rowtot <- apply(tab.dupla,1,sum) # faz a soma das linhas
> rowtot # verifica os resultados
  FEMININO MASCULINO
      11       34
> tab.dupla<-cbind(tab.dupla, rowtot) # adiciona na tabela a soma
das linhas

```

```

> tab.dupla # verifica os resultados

```

	AZUIS	CASTANHOS	CASTANHOS	CASTANHOS CLAROS	CASTANHOS ESCUROS	VERDES	rowtot
FEMININO	0	9	0	0	1	1	11
MASCULINO	1	19	1	1	10	2	34

```

> coltot <- apply(tab.dupla,2,sum) # faz a soma das colunas
> coltot # verifica os resultados

```

AZUIS	CASTANHOS	CASTANHOS	CASTANHOS CLAROS	CASTANHOS ESCUROS	VERDES	rowtot
1	28	1	1	11	3	45

```

> tab.dupla<-rbind(tab.dupla, coltot) # adiciona na tabela a soma
das colunas

```

```

> tab.dupla # verifica os resultados finais

```

	AZUIS	CASTANHOS	CASTANHOS	CASTANHOS	CLAROS	CASTANHOS	ESCUROS	VERDES	rowtot
FEMININO	0	9	0		0		1	1	11
MASCULINO	1	19	1		1		10	2	34
coltot	1	28	1		1		11	3	45

3) Tabela de Freqüências para Dados Agrupados ou Tabulados em Classe

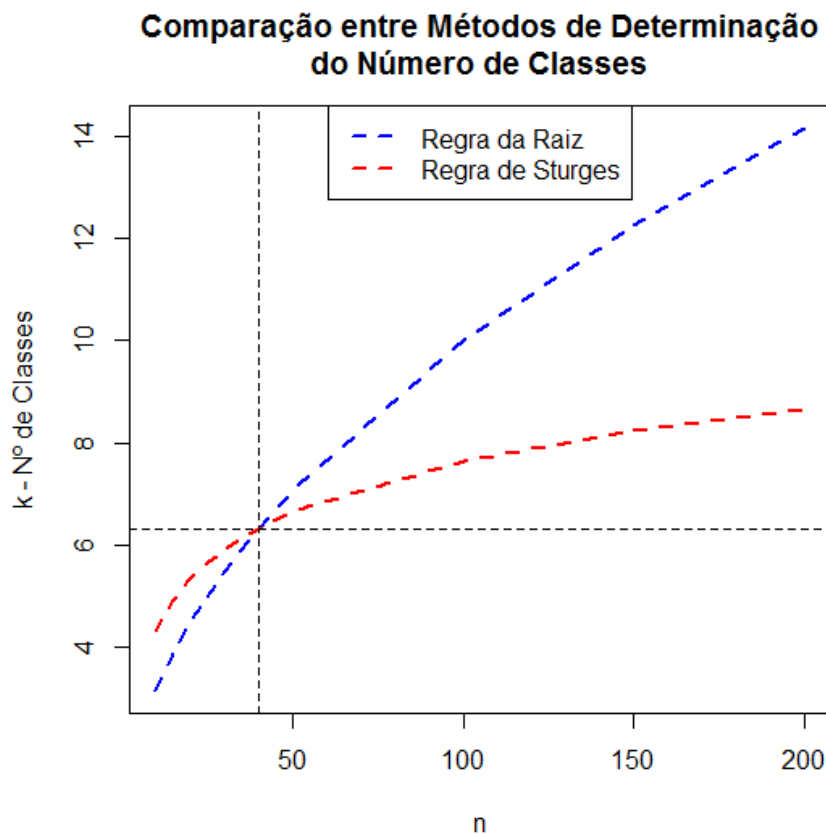
É utilizada quando temos uma variável quantitativa (contínua ou discreta em grande quantidade – $n \geq 25$). Neste caso, a Tabela de Distribuição de Freqüências é composta por intervalo de classes, freqüências, freqüências relativas, freqüências acumuladas e freqüências relativas acumuladas. Sua construção é bastante simples e segue o roteiro abaixo:

1. Determina-se o maior e o menor número dos dados brutos;
2. Calcula-se a Amplitude Total AT, dado por $AT = X_{\text{maior}} - X_{\text{menor}}$;
3. Determina-se o nº de intervalos de classe k, dado por $k = 1 + 3,322 (\log_{10} n)$ (**Fórmula de Sturges**);
4. Determina-se a amplitude do intervalo de classe h, dado por: $h = \frac{AT}{k}$
5. Determina-se os limites dos intervalos de classe;
6. Determina-se o número de observações que caem dentro de cada intervalo, para com isto, determinar as freqüências de classe.

Observação:

Pode-se usar, também, para determinar o número de classe k, a **regra da raiz** dada por $k = \sqrt{n}$, sendo o n a quantidade de dados.

Observe a comparação entre os dois métodos:



Seu formato é o seguinte:

Título:

Ordem da Classe i	Intervalos de classe	X_i (Ponto médio)	f_i Frequência	F_i Frequência Acumulada	fr_i Frequência Relativa	Fr_i Frequência Relativa Acumulada
1	$L_{inf} - L_{Sup}$	$\frac{l + L}{2}$	f_1	F_1	fr_1	Fr_1
...						
...						
K						
Total			$\sum_{i=1}^n f_i = n$		$\sum_{i=1}^n fr_i$	

Fonte:

Onde:

a) Ordem dos Intervalos de Classes:

São representadas simbolicamente por i , sendo $i = 1, 2, 3, \dots, k$, onde k é o número total de classes.

b) Intervalos de classes

Os intervalos são compostos pelo extremos de cada classe e pela amplitude dos intervalos de classe. Para determinada classe i , limite inferior é simbolizado por li e o limite superior por Li . De acordo com o IBGE, as classes devem ser escritas empregando-se os símbolos "|---", "---" ou "---|", conforme o caso.

A amplitude dos intervalos de classes h_i é o tamanho do intervalo que define a classe. Para cada classe i , a amplitude do intervalo é simbolizado por hi e é obtido pela diferença entre os seus limites, ou seja

$$h_i = L_i - l_i$$

d) Ponto médio de uma classe

É o ponto que divide a classe no meio. O ponto médio da classe i é simbolizado por X_i e calculado por

$$X_i = \frac{l_i + L_i}{2}$$

O ponto médio é o valor representativo da classe.

e) Frequências

Frequência simples ou absoluta (f_i)

Frequência relativa (fr_i)

Frequência acumulada (F_i)

Frequência relativa acumulada (Fr_i)

Obs:

1) Na construção da tabela de distribuição de frequências, podem ser usado os seguintes símbolos:

a) "|---": indica que o intervalo irá conter o valor que se encontra à esquerda deste símbolo (limite inferior - l), mas não irá conter o valor que está à direita (limite superior - L), equivalente ao seguinte intervalo $[a, b[$;

b) "---": que indica que tanto o valor da sua esquerda (l) quando o valor da sua direita (L) não serão incluídos no intervalo, equivalente ao intervalo $]a, b[$; e
 c) "|---|": que indica que ambos valores (l e L) serão incluídos no intervalo, equivalente ao intervalo $[a, b]$.

- 2) **O valor de k deve ser sempre arredondado**, independente do tipo de variável numérica (discreta ou contínua), para o número inteiro imediatamente superior a k ou o número inteiro imediatamente inferior a k , conforme as regras de arredondamento;
- 3) **O valor de h deve ser arredondado**, seguindo as regras de arredondamento, **quando a variável numérica for discreta** (exemplo: idade, notas de teste, ou outra qualquer); **quando for contínua não é necessário** fazer este arredondamento, pois são permitidos, aqui, valores fracionados, desde que, no final do processo, todos os dados sejam distribuídos na tabela;
- 4) **Verifique após os cálculos de AT , k e h se $AT < k.h$** , e nesse caso empregue o símbolo "|---", na tabela; caso $AT = k.h$, pode-se usar o símbolo "|---|" na construção da tabela, deste que todos os dados sejam distribuídos na mesma; caso isso não aconteça, o valor de h **deve ser arredondado para o maior inteiro, no caso discreto, e para o número fracionado maior do que o valor de h calculado anteriormente, no caso contínuo, obedecendo o número de casas decimais da variável em estudo.**;
- 5) **O primeiro valor a ser inserido na tabela** de distribuição de frequências deve ser o menor valor do rol de dados, ou seja, o **X_{menor}** ;

Exemplo: sejam 25 valores da variável diâmetro de peças produzidas por uma máquina em milímetros:

21,5	21,4	21,8	21,5	21,6
21,7	21,6	21,4	21,2	21,7
21,3	21,5	21,7	21,4	21,4
21,5	21,9	21,6	21,3	21,5
21,4	21,5	21,6	21,9	21,5

Seguindo o roteiro, temos:

1. $n = 25$
2. Maior: 21,9; Menor: 21,2
3. $AT = 21,9 - 21,2 = 0,70$
4. $K = 1 + 3,322 \log n = 1 + 3,322 \log(25) = 5,61 \cong 6$
5. $h = AT / k \rightarrow h = 0,70 / 6 = 0,12$
6. $AT < k.h \rightarrow 0,70 < (6 * 0,12 = 0,72) \rightarrow \text{Ok}$

Para montar a tabela de distribuição de frequências, devemos antes fazer o ordenamento dos dados, da seguinte forma:

21,2	21,3	21,3	21,4	21,4
21,4	21,4	21,4	21,5	21,5
21,5	21,5	21,5	21,5	21,5
21,6	21,6	21,6	21,6	21,7
21,7	21,7	21,8	21,9	21,9

A tabela de Distribuição de Frequências fica assim:

"Distribuição de Frequências do diâmetro de peças produzidas"

Classe	Intervalos de classe		X	f	F	Fr	Fra
1	21,20	-- 21,32	21,26	3	3	0,12	0,12
2	21,32	-- 21,44	21,38	5	8	0,20	0,32
3	21,44	-- 21,56	21,50	7	15	0,28	0,60
4	21,56	-- 21,68	21,62	4	19	0,16	0,76
5	21,68	-- 21,80	21,74	3	22	0,12	0,88
6	21,80	-- 21,92	21,86	3	25	0,12	1,00
				25			

Os cálculos anteriores poderiam ser obtidos com o auxílio do R, por meio dos seguintes comandos:

criando a variável "diam" e colocando os valores observados nela

```
diam<-c(21.5, 21.4, 21.8, 21.5, 21.6, 21.7, 21.6, 21.4, 21.2, 21.7, 21.3, 21.5, 21.7, 21.4,
        21.4, 21.5, 21.9, 21.6, 21.3, 21.5, 21.4, 21.5, 21.6, 21.9, 21.5)
```

diam # verificando os resultados

```
[1] 21.5 21.4 21.8 21.5 21.6 21.7 21.6 21.4 21.2 21.7 21.3 21.5 21.7 21.4 21.4
```

```
[16] 21.5 21.9 21.6 21.3 21.5 21.4 21.5 21.6 21.9 21.5
```

x.min<-min(diam) # achando o menor valor de "diam"

```
> x.min
```

```
[1] 21.2
```

x.max<-max(diam) # achando o maior valor de "diam"

```
> x.max
```

```
[1] 21.9
```

AT<-x.max - x.min # calculando AT

```
> AT
```

```
[1] 0.7
```

k<-1+3.322*log(length(diam), 10) # calculando o valor de k

```
> k
```

```
[1] 5.643957
```

k<-round(k, 0) # arredondando k

```
> k
```

```
[1] 6
```

h<-AT/k # calculando h

```
> h
```

```
[1] 0.1166667
```

h<-round(h, 2) # arredondando para 2 casas

```
> h
```

```
[1] 0.12
```

AT < k*h # verificando AT < k*h

```
[1] TRUE # resposta verdadeira
```

```
#calculando os limites dos intervalos
ini<-x.min
x.br<-0
for (i in 1:(k+1)){
  x.br[i]<-ini
  ini<-ini+h
}

> x.br # vendo os resultados
[1] 21.20 21.32 21.44 21.56 21.68 21.80 21.92

#Para construir a tabela de frequências

install.packages("fdth") # instala pacote "fdth"
require(fdth) # carrega o pacote

x.dist2=fdt(diam,start=x.min,end=x.max+h,h=h) # cria a tabela com o pacote "fdth"
x.dist2 # verifica o resultado
```

```
Class limits f    rf rf(%) cf cf(%)
[21.2,21.32) 3 0.12    12  3    12
[21.32,21.44) 5 0.20    20  8    32
[21.44,21.56) 7 0.28    28 15    60
[21.56,21.68) 4 0.16    16 19    76
[21.68,21.8)  3 0.12    12 22    88
[21.8,21.92) 3 0.12    12 25   100
```

Esta tabela pode ser feita no Word, usando o R, com os seguintes comandos:

```
install.packages("ReporteRs") # instala o pacote "ReportRs"
library(ReporteRs) # carrega o pacote no R

mydoc4 = docx() # define o documento no Word
mydoc4 = addFlexTable( mydoc4, FlexTable(print(x.dist2)) ) # cria a tabela em "mydoc4"
writeDoc( mydoc4, file = "x_dist2.docx") # escreve a tabela e salva em "Meus documentos" com o
                                         nome "x_dist2.docx"
```

O resultado ao abrir o arquivo "x_dist2.docx" é:

Class limits	f	Rf	rf(%)	cf	cf(%)
[21.2,21.32)	3	0.12	12	3	12
[21.32,21.44)	5	0.20	20	8	32
[21.44,21.56)	7	0.28	28	15	60
[21.56,21.68)	4	0.16	16	19	76
[21.68,21.8)	3	0.12	12	22	88
[21.8,21.92)	3	0.12	12	25	100

Falta adicionar a linha "Total" e retirar as linhas da esquerda e da direita. Isto pode ser feito manualmente no Word.

O resultado final fica:

Class limits	f	Rf	rf(%)	cf	cf(%)
[21.2,21.32)	3	0.12	12	3	12
[21.32,21.44)	5	0.20	20	8	32
[21.44,21.56)	7	0.28	28	15	60
[21.56,21.68)	4	0.16	16	19	76
[21.68,21.8)	3	0.12	12	22	88
[21.8,21.92)	3	0.12	12	25	100
Total	25	-	100	-	-

12. Representação gráfica das séries estatísticas

Uma vez montada a tabela com as devidas frequências, os dados podem ser representados de diversas formas. Toda representação gráfica deve obedecer aos seguintes requisitos:

- Simplicidade;
- Clareza; e
- Veracidade.

Os principais tipos de representação gráfica são:

- Diagramas;
- Estereogramas;
- Cartogramas;
- Pictogramas.

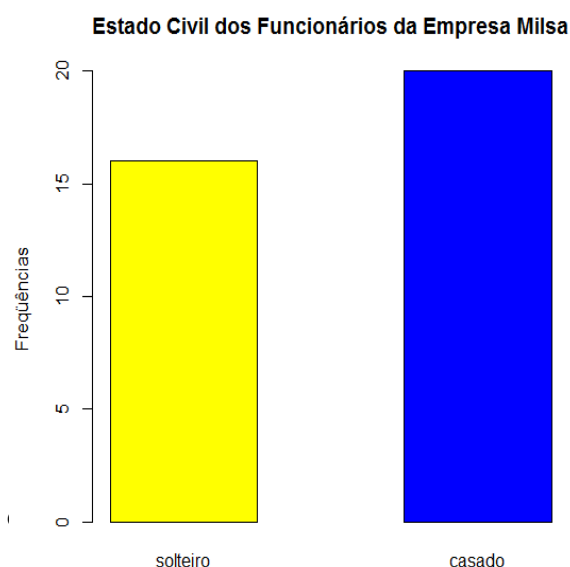
12.1 Diagramas

São representações geométricas no espaço bidimensional. Os principais diagramas são:

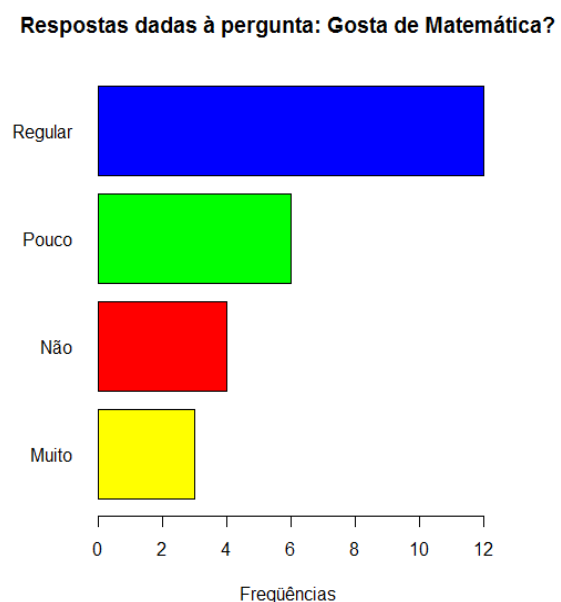
- Gráfico em colunas;
- Gráfico em barras;
- Gráfico em setores;
- Gráfico de porcentagens complementares;
- Gráfico polar;
- Diagrama de ramo-e-folhas;
- Diagrama de pontos;
- Histograma
- Polígono de frequências;
- Gráficos lineares ou de linhas.

Gráficos em colunas e em barras

a) Gráfico em colunas



b) Gráfico em Barras



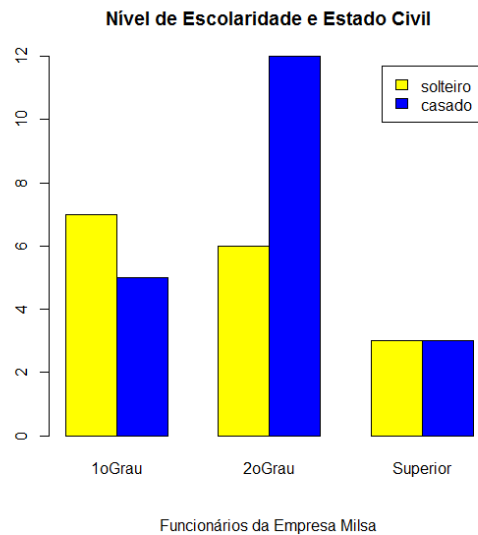
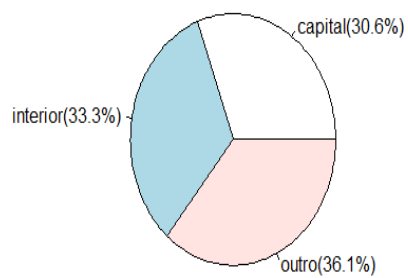


Gráfico em setores e porcentagens complementares

a) Gráfico em setores

Região de procedência dos funcionários da Empresa Milsa



b) Porcentagens complementares

Nível de Escolaridade e Estado Civil

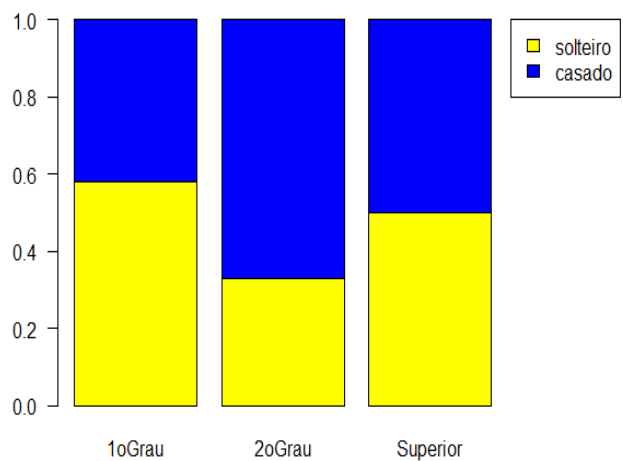


Gráfico polar

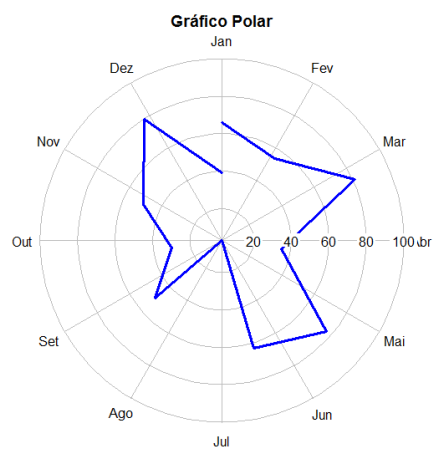


Diagrama de ramo-e-folha

Representa um conjunto de dados quantitativos separando cada valor em duas partes: ramo (como o dígito mais à esquerda) e a folha(como o dígito mais à direita).

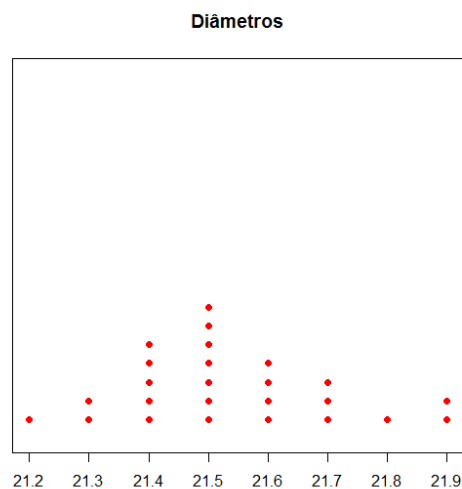
Ex: sejam os seguintes valores:

4.00 4.56 5.25 5.73 6.26 6.66 6.86 7.39 7.44 7.59 8.12 8.46
8.74 8.95 9.13 9.35 9.77 9.80 10.53 10.76 11.06 11.59 12.00 12.79
13.23 13.60 13.85 14.69 14.71 15.99 16.22 16.61 17.26 18.75 19.40 23.30

Ramo	Folha
4	06
5	37
6	379
7	446
8	157
9	01488
10	58
11	16
12	08
13	269
14	77
15	
16	026
17	3
18	8
19	4
20	
21	
22	
23	3

Diagrama de Pontos

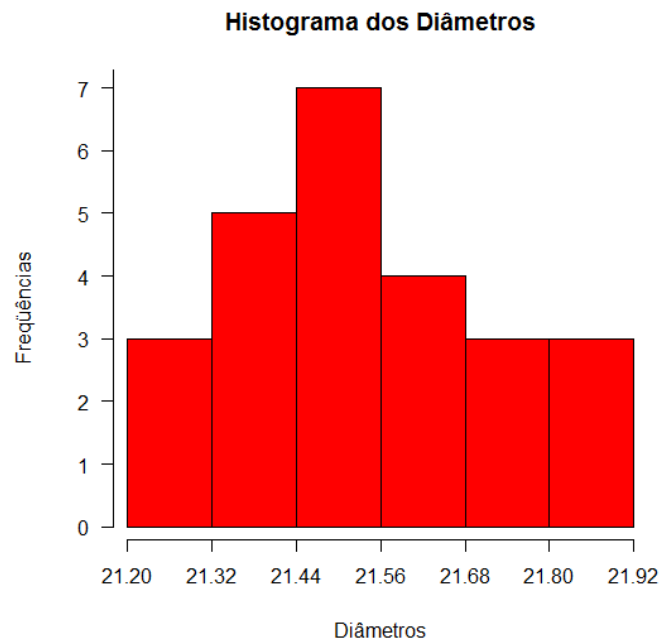
É útil para avaliar se há ou parece haver alguma estrutura no processo de observação dos dados.



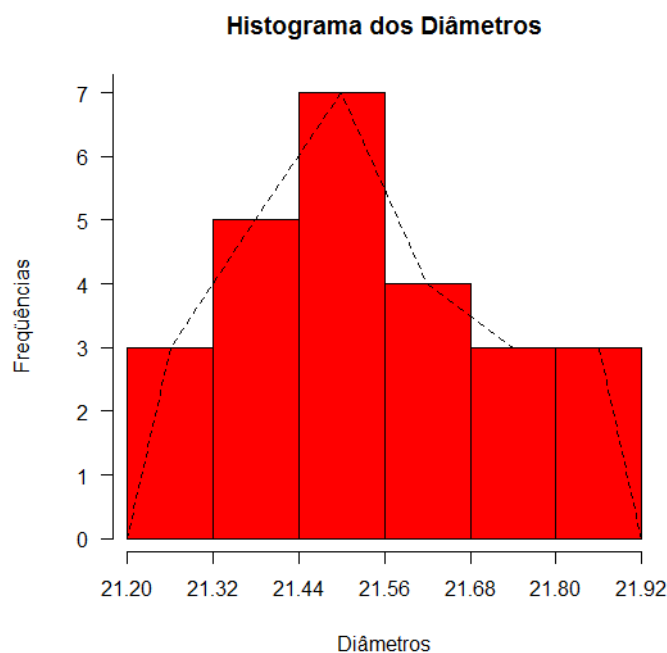
Histograma

A representação gráfica da Tabela de Distribuição de Frequências é o histograma, que é formado por um conjunto de retângulos justapostos cujas bases se localizam no eixo horizontal, de tal modo que seus pontos médios coincidam com os pontos médios dos intervalos de classe e seus limites coincidam com os limites das classes.

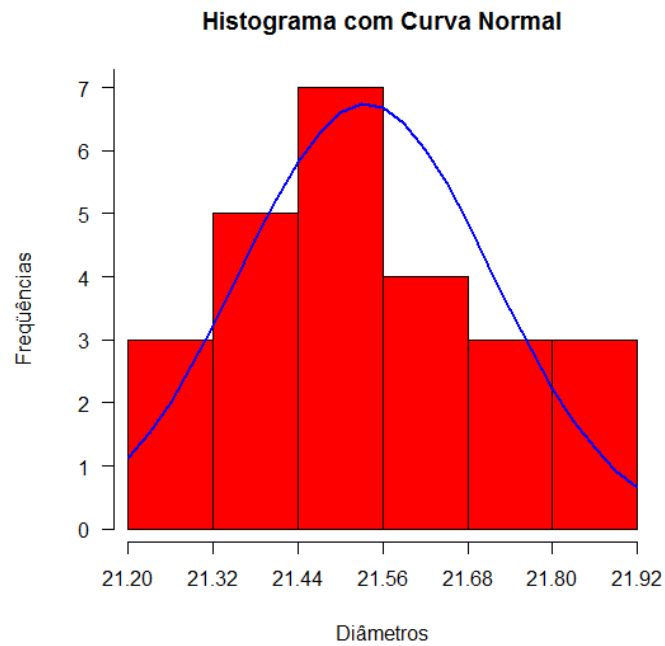
Por exemplo:



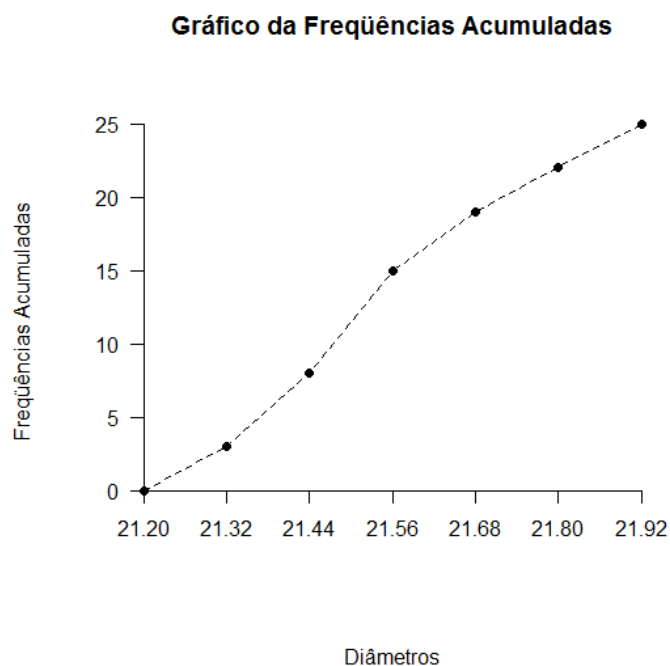
Juntamente com o histograma, também é apresentado o Polígono de Frequências, conforme o gráfico abaixo:



Alguns programas estatísticos apresentam o histograma com a curva da Distribuição Normal, conforme o gráfico abaixo:



E para as frequências absolutas acumuladas, podemos construir o Gráfico de Frequências Acumuladas:



Além dos gráficos vistos anteriormente, são usados também os seguintes gráficos para representar dados estatísticos:

Gráfico de linha

São amplamente empregados para representar fenômenos contínuos no tempo (série temporal). Neste gráfico, temos no eixo x a variável tempo, que é a principal característica de uma série temporal.

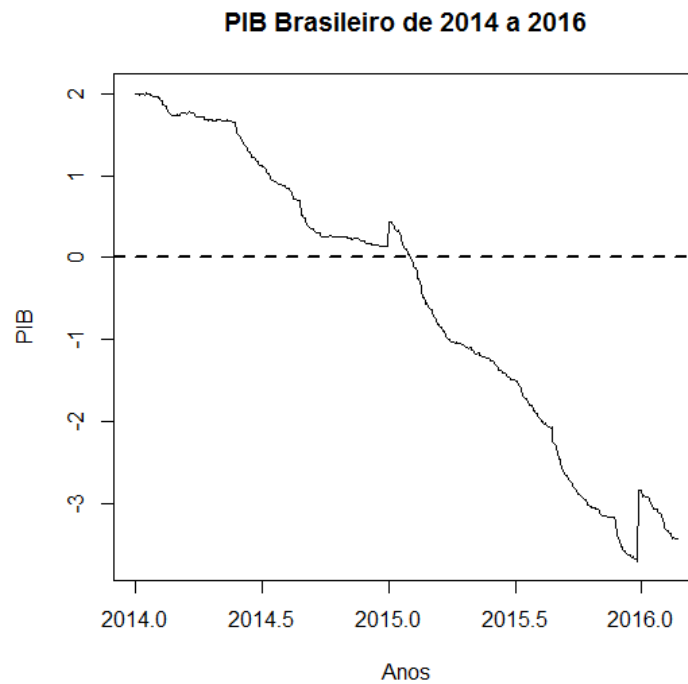
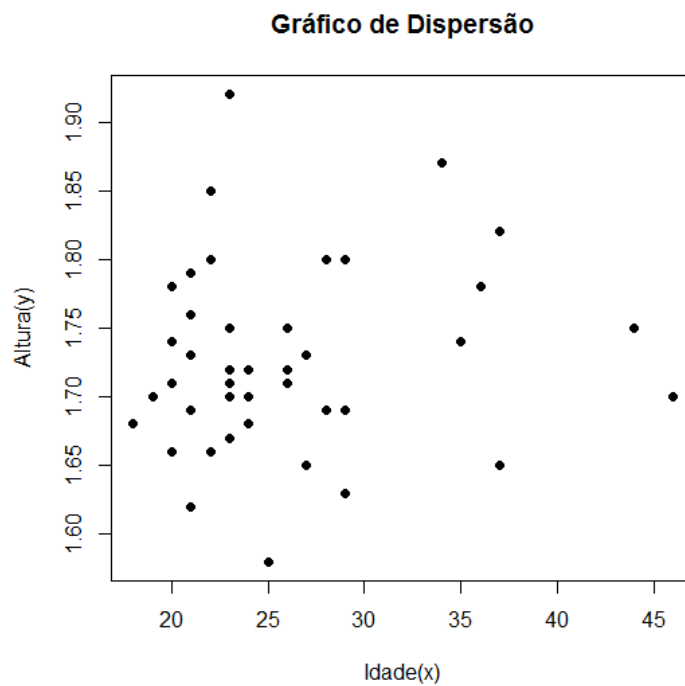


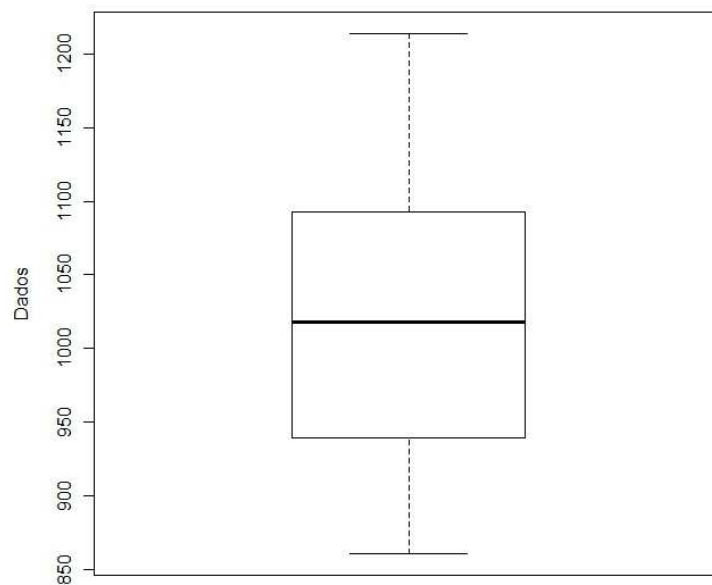
Gráfico de dispersão

Onde representamos o comportamento da relação entre a variável x e a variável y .



Boxplot

Este gráfico mostra como está o comportamento da distribuição dos dados. É utilizado para avaliar a distribuição empírica dos dados. Ele será detalhado mais adiante.

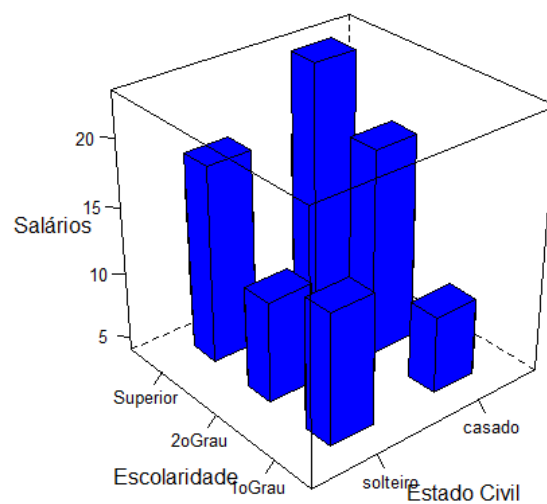


12.2 Estereogramas

São representações geométricas no espaço tridimensional. Os volumes dos sólidos geométricos devem ser proporcionais aos valores da série que procura representar.

Ex.:

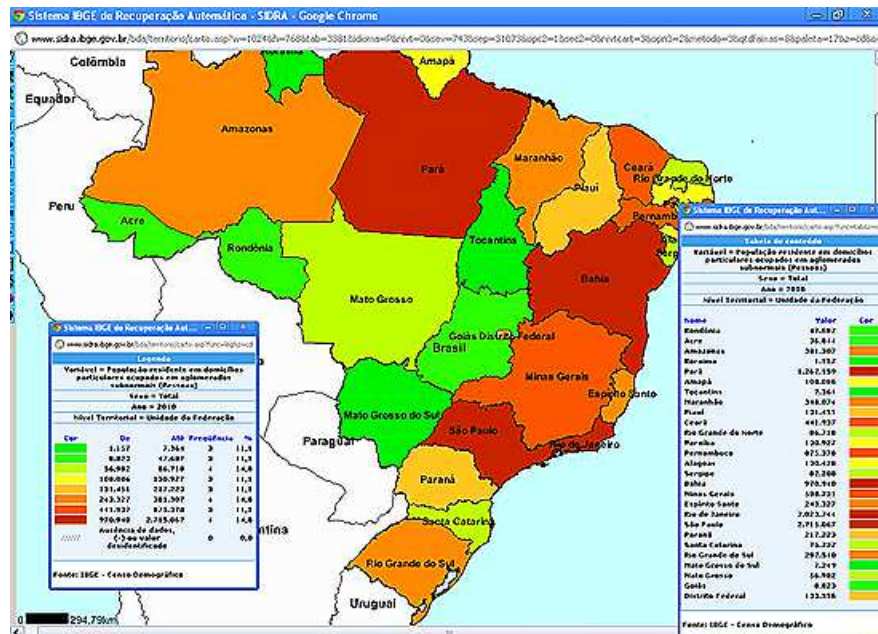
Distribuição de Salários por Escolaridade e Estado Civil



12.3 Cartogramas

São ilustrações em cartas geográficas. Neste tipo de representação se relacionam os valores da série (que é sempre geográfica ou espacial) com seus respectivos locais de ocorrência.

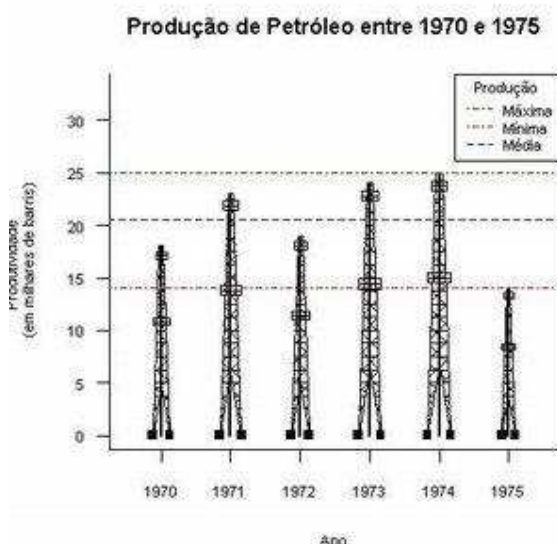
Ex.:



12.4 Pictogramas

São gráficos construídos a partir de figuras ou conjunto de figuras representativas da intensidade do fenômeno. Têm a vantagem de despertar a atenção do público leitor.

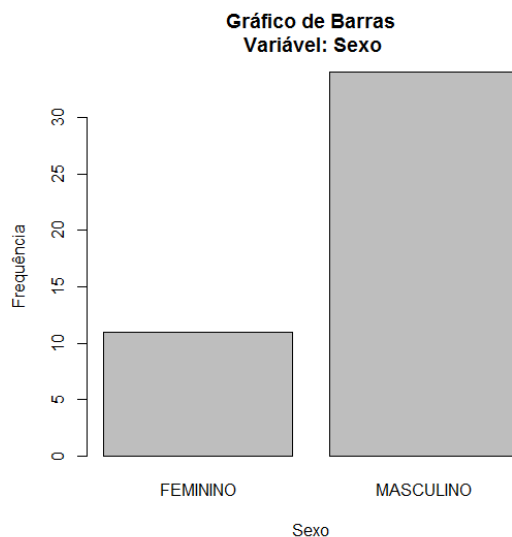
Ex.:



Para fazer os gráficos no R, vamos utilizar os seguintes comandos:

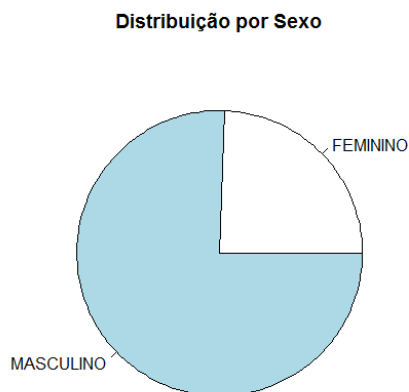
a) Para o gráfico de colunas

```
tab1<-table(dados$SEXO) # cria a tabela e coloca em "tab1"
tab1
FEMININO MASCULINO # resultado de "tab1"
      11      34
# comando para o gráfico
barplot(tab1,main="Gráfico de Barras\n Variável: Sexo", ylab= "Frequência", xlab="Sexo")
```



b) Para o Gráfico de setores

```
tab1<-table(dados$SEXO) # cria a tabela e coloca em "tab1"
tab1
FEMININO MASCULINO # resultado de "tab1"
      11      34
pie(tab1, main="Distribuição por Sexo") # comando básico para o gráfico
```

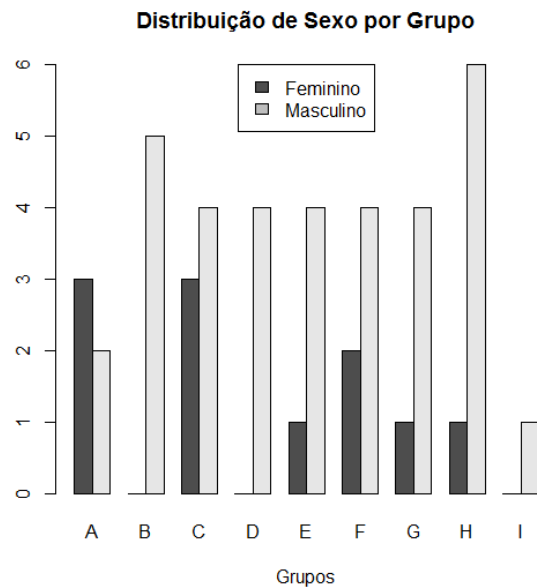


c) Para o gráfico de colunas duplas

```
tab2<-table(dados$SEXO, dados$GRUPO) # cria a tabela de dupla entrada
tab2 # verifica o resultado da tabela
```

	A	B	C	D	E	F	G	H	I
FEMININO	3	0	3	0	1	2	1	1	0
MASCULINO	2	5	4	4	4	4	4	6	1

```
barplot(tab2, beside=T, main="Distribuição de Sexo por Grupo", xlab="Grupos") # cria o gráfico
legend("top", legend=c("Feminino", "Masculino"), fill = c("gray30", "gray")) # cria a legenda
```



d) Para o gráfico de colunas complementares

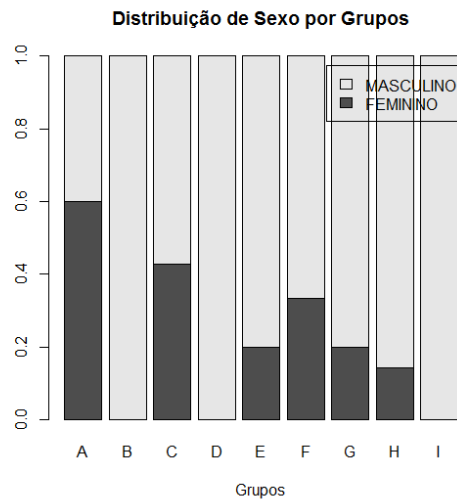
```
tab2<-table(dados$SEXO, dados$GRUPO) # cria a tabela de dupla entrada com valores absolutos
p.tab2<-prop.table(tab2,2) # transforma a tabela "tab2" em porcentagens
p.tab2 # verifica os resultados
```

	A	B	C	D	E	F
FEMININO	0.6000000	0.0000000	0.4285714	0.0000000	0.2000000	0.3333333
MASCULINO	0.4000000	1.0000000	0.5714286	1.0000000	0.8000000	0.6666667

	G	H	I
FEMININO	0.2000000	0.1428571	0.0000000
MASCULINO	0.8000000	0.8571429	1.0000000

cria o gráfico

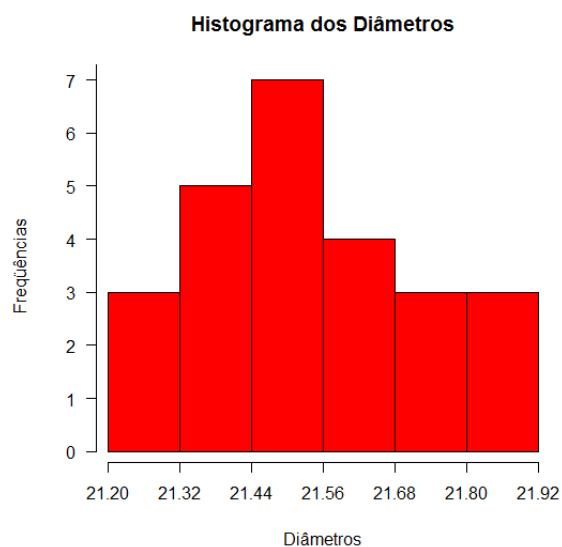
```
barplot(p.tab2, legend = T, main="Distribuição de Sexo por Grupos", xlab="Grupos")
```



e) Para o histograma

Após termos calculado AT, k, h e termos definidos os limites dos intervalos de classe, ainda termos definidas as frequências de cada intervalo, podemos usar os seguintes comandos:

```
#comando para o histograma
h.x<-hist(diam, breaks=x.br, right=FALSE, axes=FALSE, col="red", main="Histograma dos
Diâmetros", xlab="Diâmetros", ylab="Frequências")
# definindo os valores do "eixo x" com os valores dos limites dos intervalos
axis(1, at=c(x.br), pos=-0.2)
# definindo os valores do "eixo y"
ini.h=x.min-0.02
ini.h
axis(2, at=seq(0, 10, by=1), pos=c(ini.h))
```



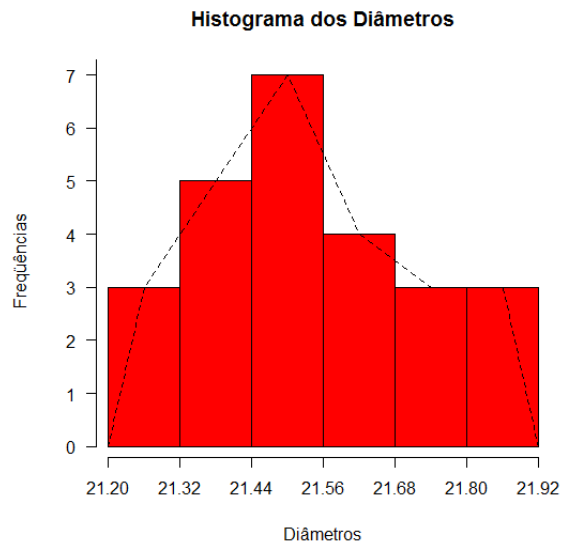
Para adicionar o "Polígono de Frequências" no gráfico acima, os comandos são (obs: o gráfico deve estar ativo no R):

```
mids<-(x.br[-1]+x.br[-length(x.br)])*0.5 # cria os pontos médios dos intervalos
mids
mids1<-c(x.br[1], mids, x.br[length(x.br)]) # cria um vetor que começa com o menor valor de X e
# vai até o maior valor de x
```

```

mids1
c<-c(0, h.x$counts, 0) # cria um vetor que começa em 0 para pelas frequências dos intervalos e
                        # termina em 0
c
lines(mids1, c, lwd=1, lty=2) # plota a linha no histograma

```



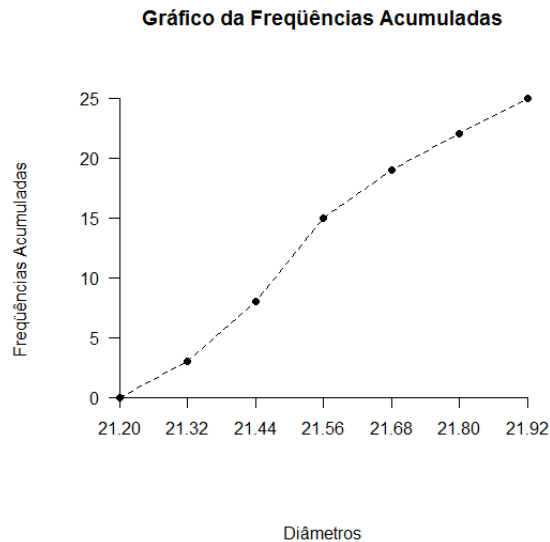
Para fazer o gráfico de "Frequências Acumuladas", temos:

```

fac<-cumsum(h.x$counts) # faz a frequência acumulada
fac
fac1<-c(0, fac) #cria um vetor que começa em 0 e pelos valores acumulados
fac1

# plota a linha formada por "x.br" (limites dos intervalos) e "fac1" (valores acumulados)
plot(x.br, fac1, type="l", lty=2, xlim=c(min(x.br)-0.02, max(x.br)+0.02), ylim=c(-4, max(fac)+2),
     main="Gráfico da Frequências Acumuladas" , xlab="Diâmetros", ylab="Frequências
Acumuladas" , axes=F)
axis(1, at=x.br, pos=0) # define os valores do "eixo x"
axis(2, at=seq(0, max(fac1), by=5), pos=min(x.br)) # define os valores do "eixo y"
points(x.br, fac1, pch=16) # inclui pontos no gráfico

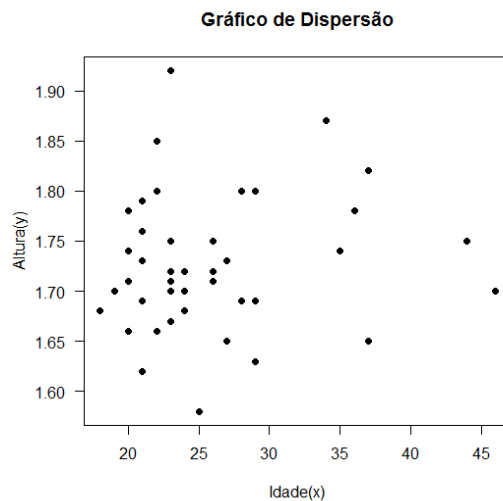
```



f) Para o Gráfico de Dispersão

a Idade está no "eixo x" e a altura no "eixo y"

```
plot(dados$IDADE, dados$ALTURA, main="Gráfico de Dispersão", xlab="Idade(x)",
ylab="Altura(y)", pch=16 )
```



13. Características numéricas de uma distribuição de freqüências

13.1 Medidas de Posição ou Localização

Essas medidas fornecem valores que caracterizam o comportamento de uma série de dados, indicando a posição ou a localização dos dados em relação ao eixo dos valores assumidos pela variável ou característica em estudo.

As medidas de posição ou localização são subdivididas em medidas de tendência central (média, mediana e moda) e medidas separatrizes (quartis, quintis, decis e percentis).

13.1.1 Medidas de Tendência Central

São indicadores que permitem que se tenha uma primeira idéia ou resumo, do modo como se distribuem os dados de uma variável aleatória.

Sevem para localizar a distribuição de freqüências sobre o eixo de variação da variável em questão.

13.1.1.1 Média Aritmética ou simplesmente Média

E o valor representativo de um conjunto de valores que corresponde ao centro de gravidade da distribuição de freqüências.

Sendo x_i , com $i = 1, 2, \dots, n$, o conjunto de dados sem freqüências ou não agrupados, definimos sua média por:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Se os dados estiverem dispostos em uma tabela de freqüências formadas por k linhas, poderemos obter a média por:

$$\bar{X} = \frac{\sum_{i=1}^k X_i f_i}{\sum_{i=1}^k f_i}$$

No caso em que os dados foram distribuídos em intervalos de classes de freqüências, podemos calcular a média utilizando a expressão acima, mas substituindo os x_i pelos pontos médios das classes.

Propriedades da média

a) a soma algébrica dos desvios tomados em relação à média é nula.

$$\sum (x_i - \bar{X}) = 0.$$

b) a soma algébrica dos quadrados dos desvios (em relação à média) é mínima.

$$\sum (x_i - \bar{X})^2 \leq \sum (x_i - y_i)^2, \text{ onde } \bar{X} \neq y_i.$$

c) somando ou subtraindo uma constante a todos valores de uma variável, a média ficará acrescida ou subtraída dessa constante.

$$\frac{\sum (x_i + k)}{n} = \frac{\sum x_i + \sum k}{n} = \frac{\sum x_i + nk}{n} = \bar{X} + k.$$

d) multiplicando (ou dividindo) todos os valores de uma variável por uma constante, a média ficará multiplicada ou dividida por essa constante.

$$\frac{\sum kx_i}{n} = \frac{k \sum x_i}{n} = k\bar{X}.$$

Obs: além da média aritmética, temos outras médias a saber:

13.1.1.2 Média Geométrica:

Quando os dados crescem de forma exponencial, a média aritmética pode não representar bem os dados. Neste caso, utiliza-se a média geométrica.

É a raiz de ordem n do produto do conjunto de dados x_i , sem frequências ou não agrupados, é dada por:

$$G = \sqrt[n]{\prod x_i}$$

No caso em que os dados estão em tabelas de frequência ou agrupados em intervalos de classe, em que é possível identificar suas frequências, temos:

$$G = \sqrt[n]{\prod X_i^{f_i}}, \text{ onde } n = \sum f_i$$

Quando o número de observações for muito grande, é aconselhável o emprego de logaritmos (decimal ou neperiano)

$$G = \sqrt[n]{\prod x_i} = \prod x_i^{\frac{1}{n}}$$

Aplicando o logaritmo decimal em G , temos:

$$\log G = \log \left[\prod x_i^{\frac{1}{n}} \right] = \frac{1}{n} \log [\prod x_i] = \frac{1}{n} \sum \log x_i = \frac{\sum \log x_i}{n}$$

Para obter G , temos que calcular o antilog da seguinte maneira:

$$G = \text{anti log} \left[\frac{\sum \log x_i}{n} \right]$$

Quando os dados estiverem em uma tabela de frequências, o cálculo será seguinte maneira:

$$G = \text{anti log} \left[\frac{\sum f_i \log x_i}{n} \right]$$

13.1.1.3 Média harmônica

É utilizada quando estamos trabalhando com grandezas inversamente proporcionais ou quando temos situações em que a média de taxas é desejada.

A média harmônica de um conjunto de dados sem frequências ou não agrupados é a recíproca da média aritmética das recíprocas dos números, ou seja:

$$H = \frac{n}{\sum \frac{1}{x_i}}$$

No caso em que os dados estão em tabelas de frequência ou agrupados em intervalos de classe, em que é possível identificar suas frequências, temos:

$$H = \frac{n}{\sum \frac{f_i}{X_i}}$$

Relação entre a média aritmética, geométrica e harmônica

A média geométrica é menor do que ou igual à média aritmética, mas é maior do que ou igual à média harmônica, ou seja,

$$H \leq G \leq \bar{X}$$

13.1.1.4 Média Quadrática ou Raiz Média Quadrática (RMQ)

É um tipo de média que é calculada com base nos valores de x elevados ao quadrado. É definida por:

$$RMQ = \sqrt{\overline{X^2}} = \sqrt{\frac{\sum X^2}{n}} \text{ - Para dados sem frequência ou não agrupados}$$

$$RMQ = \sqrt{\overline{X^2}} = \sqrt{\frac{\sum X^2 f}{n}} \text{ - Para dados com frequência}$$

Exemplo:

a) Valores sem frequência ou não agrupados

x = 2, 2, 3, 5, 6, 8, 8, 8, 10

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{2+2+3+5+6+8+8+8+10}{9} = \frac{52}{9} = 5,78$$

$$G = \sqrt[n]{\prod x_i} = \sqrt[9]{2*2*2*5*6*8*8*8*10} = \sqrt[9]{18432000} = 4,97$$

$$G = \text{anti log} \left[\frac{\sum \log x_i}{n} \right] = \text{anti log} \left[\frac{\log 2 + \log 2 + \log 3 + \log 5 + \log 6 + \log 8 + \log 8 + \log 8 + \log 10}{9} \right] =$$

$$= \text{anti log} \left[\frac{0.301 + 0.301 + 0.477 + 0.699 + 0.778 + 0.903 + 0.903 + 0.903 + 1.000}{9} \right] =$$

$$= \text{anti log} \left[\frac{6.266}{9} \right] = \text{anti log}(0,696) = 10^{0,696} = 4,97$$

$$H = \frac{n}{\sum \frac{1}{x_i}} = \frac{9}{\frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{6} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{10}} =$$

$$= \frac{9}{0.50 + 0.50 + 0.333 + 0.200 + 0.167 + 0.125 + 0.125 + 0.125 + 0.100} = \frac{9}{2,175} = 4,14$$

Verificamos que : $H \leq G \leq \bar{X}$, pois $4,14 < 4,97 < 5,78$

b) Para dados com frequência, mas não agrupados

Nr de defeitos	f
0	4
1	7
2	5
3	2
4	1
5	1
Total	20

Vamos usar a fórmula:

$$\bar{X} = \frac{\sum_{i=1}^k X_i f_i}{\sum_{i=1}^k f_i}$$

Para isto, vamos criar uma coluna na tabela e chamá-la de "Xf". Nesta coluna, vamos incluir os valores de X*f de cada linha, da seguinte forma:

Nr de defeitos	f	Xf
0	4	0
1	7	7
2	5	10
3	2	6
4	1	4
5	1	5
Total	20	32

$$\bar{X} = \frac{\sum_{i=1}^k X_i f_i}{\sum_{i=1}^k f_i} = \frac{32}{20} = 1,60$$

Para o cálculo da Média Geométrica(G) e da Média Harmônica (H), devemos adicionar mais coluna na tabela acima:

Nr de defeitos	f	Xf	log x	f*log x	$\frac{f}{x}$
0	4	0	-	-	-
1	7	7	0	0	7
2	5	10	0,301	1,505	2,500
3	2	6	0,477	0,954	0,667
4	1	4	0,602	0,602	0,250
5	1	5	0,699	0,699	0,200
Total	20	32	-	3,76	10,617

Usando as fórmulas abaixo, temos:

$$G = \text{anti log} \left[\frac{\sum f_i \log x_i}{n} \right] = \text{anti log} \left[\frac{3,76}{20} \right] = \text{anti log}(0,188) = 10^{0,188} = 1,54$$

$$H = \frac{n}{\sum \frac{f_i}{X_i}} = \frac{16}{10,617} = 1,51$$

(Obs: aqui tivemos que considerar como valores de f, os valores 7, 5, 2, 1,1, que dá um total de 16, isto porque na primeira linha não foi possível calcular o log e a divisão f/x, em virtude do valor x = 0; isto não acontece quando todos os valores de x são válidos, ou seja, $x_i \neq 0$)

Verificamos que : $H \leq G \leq \bar{X}$, pois $1,51 < 1,54 < 1,60$

c) Para dados agrupados

Vamos usar como exemplo, a tabela de distribuição de frequências abaixo:

Class limits	f	Rf	rf(%)	cf	cf(%)
[21.2,21.32)	3	0.12	12	3	12
[21.32,21.44)	5	0.20	20	8	32
[21.44,21.56)	7	0.28	28	15	60
[21.56,21.68)	4	0.16	16	19	76
[21.68,21.8)	3	0.12	12	22	88
[21.8,21.92)	3	0.12	12	25	100
Total	25	-	100	-	-

Para calcular as médias (Média Aritmética, Geométrica e Harmônica) precisamos incluir na tabela acima os Pontos Médios de cada intervalo, e as mesmas colunas do exemplo anterior:

Class limits	PM	f	Rf	rf(%)	cf	cf(%)	Xf	log x	f*log x	$\frac{f}{x}$
[21.2,21.32)	21.26	3	0.12	12	3	12	63,78	1,328	3,984	0,141
[21.32,21.44)	21.38	5	0.20	20	8	32	106,90	1,330	6,650	0,234
[21.44,21.56)	21.50	7	0.28	28	15	60	150,50	1,332	9,324	0,326
[21.56,21.68)	21.62	4	0.16	16	19	76	86,48	1,335	5,340	0,185
[21.68,21.8)	21.74	3	0.12	12	22	88	65,22	1,337	4,011	0,138
[21.8,21.92)	21.86	3	0.12	12	25	100	65,58	1,340	4,020	0,137
Total	-	25	-	100	-	-	538,46		33,329	1,161

Vamos usar as seguintes fórmulas:

$$\bar{X} = \frac{\sum_{i=1}^k X_i f_i}{\sum_{i=1}^k f_i} = \frac{538,46}{25} = 21,54$$

$$G = \text{anti log} \left[\frac{\sum f_i \log x_i}{n} \right] = \text{anti log} \left[\frac{33,329}{25} \right] = \text{anti log}(1,333) = 10^{1,333} = 21,53$$

$$H = \frac{n}{\sum \frac{f_i}{X_i}} = \frac{25}{1,161} = 21,53$$

Verificamos que : $H \leq G \leq \bar{X}$, pois $21,53 \leq 21,53 < 21,54$

13.1.1.5 Mediana

É o valor que divide a distribuição de frequências em duas partes iguais.

a) Para dados simples:

I) Se n for ímpar, a mediana será o elemento de ordem $PM_d = \frac{(n+1)}{2}$

II) Se n for par, a mediana será o valor médio entre os elementos de ordem $P1M_d = \frac{n}{2}$ e $P2M_d = \frac{n}{2} + 1$

Exemplo:

n ímpar: 2, 2, 3, 5, 6, 8, 8, 8, 10 \Rightarrow aqui o $n = 9$, usando $PM_d = \frac{(n+1)}{2}$ temos, $PM_d = \frac{(9+1)}{2} = 5$

logo a mediana será o valor de ordem 5, que no caso é $Md = 6$, pois ocupa a 5ª posição no rol.

n par: 5, 5, 7, 9, 11, 12, 15, 18 \Rightarrow aqui o $n = 8$, usando $P1M_d = \frac{n}{2}$ e $P2M_d = \frac{n}{2} + 1$ temos:

$$P1Md = \frac{8}{2} = 4 \text{ e } P2Md = \frac{8}{2} + 1 = 5$$

ou seja, vamos usar o valor que está na posição 4 e o valor que está na posição 5, que correspondem aos valores 9 e 11. Daí, tiramos uma média destes valores que é igual a $Md = \frac{P1Md + P2Md}{2}$

$$\frac{9+11}{2} = 10, \text{ esta é a mediana.}$$

Observe que os dados devem estar em rol, ou seja devem estar ordenados em ordem de grandeza.

b) Para dados que apresentam frequência, mas não estão agrupados em intervalos de classe, deve-se seguir a idéia acima.

Exemplo:

Nr de defeitos	f
0	4
1	7
2	5
3	2
4	1
5	1
Total	20

Como n é par, temos: $P1Md = \frac{20}{2} = 10$ e $P2Md = \frac{20}{2} + 1 = 11$

Aqui, é necessário determinar a frequência acumulada. Logo:

Nr de defeitos	f	F
0	4	4
1	7	11
2	5	16
3	2	18
4	1	19
5	1	20
Total	20	

Verificando na tabela, percebemos que o valor 0 se repete 4 vezes, e o valor 1 se repete 7 vezes. Juntos temos 11 valores acumulados. Ou seja, a décima posição é ocupada pelo valor "1", e a décima primeira é ocupada pelo valor "1", também.

Então, a mediana será:

$$Md = \frac{P1Md + P2Md}{2} = \frac{1+1}{2} = 1$$

c) Para os dados agrupados em intervalos de classe devemos usar a seguinte fórmula:

$$Md = l_{\inf Md} + \left(\frac{(n/2) - F_{antMd}}{f_{Md}} \right) h_{Md}$$

onde:

$l_{\inf Md}$ - Limite Inferior da classe que contém a Md

$F_{ant Md} = \Sigma f_{ant Md}$ - Soma das freq. anterior à classe da Md (Frequência acumulada anterior à classe da Md)

f_{Md} - Frequência da classe da Md

h_{Md} - Amplitude da classe da Md

Exemplo:

Classes	f	F
2 --- 4	3	3
4 --- 6	5	8
6 --- 8	7	15
8 --- 10	4	19
10 --- 12	1	20
Total	20	

Neste caso, vamos começar por $\frac{n}{2} = \frac{20}{2} = 10$ e $\frac{n}{2} + 1 = \frac{20}{2} + 1 = 11$,

ou seja, a mediana é o valor que ocupa a 10ª e a 11ª posição. Estes valores estão no intervalo de 6 a 8, porque até o anterior temos 8 elementos. Identificamos assim a classe da mediana. Vamos agora usar a fórmula:

$$Md = l_{\inf Md} + \left(\frac{(n/2) - F_{ant Md}}{f_{Md}} \right) h_{Md} = 6 + \left(\frac{10 - 8}{7} \right) 2 = 6 + \frac{4}{7} = 6,57. \text{ Esta}$$

é a mediana para os dados da tabelados.

Existe uma forma mais rápida para este cálculo, que usa a seguinte relação:

$$\frac{h_{Md}}{f_{Md}} = \frac{X_{Md}}{Dif_{Md}}$$

$$Md = l_{\inf Md} + X_{Md}$$

Onde:

X_{Md} – é o valor que se quer achar

h_{Md} – é a amplitude do intervalo da mediana

f_{Md} – é a frequência do intervalo da mediana

Dif_{Md} – é a diferença entre $\frac{n}{2}$ e a soma das frequências anteriores ao intervalo da mediana.

Então, pelo exemplo temos:

$$\frac{h_{Md}}{f_{Md}} = \frac{X_{Md}}{Dif_{Md}} \Rightarrow \frac{2}{7} = \frac{X_{Md}}{\left(\frac{20}{2} - 8\right)} \Rightarrow \frac{2}{7} = \frac{X_{Md}}{2} \Rightarrow X_{Md} = \frac{4}{7} = 0,57$$

$$Md = l_{inf\ Md} + X_{Md} \Rightarrow Md = 6 + 0,57 = 6,57$$

13.1.1.6 Moda

É o valor que ocorre com a maior frequência, ou de máxima frequência.

Para dados simples: valor (ou valores) de máxima frequência.

Para dados agrupados:

1) Moda Bruta:

Neste caso, verifica-se o intervalo com a maior frequência. A moda bruta será o ponto médio deste intervalo.

2) Método de King

Neste método, usa-se a seguinte fórmula:

$$Mo = l_{inf\ Mo} + \left(\frac{f_{post}}{f_{ant} + f_{post}} \right) h_{Mo}$$

Onde:

$l_{inf\ Mo}$: é o limite inferior do intervalo onde está a moda;

f_{ant} : é a frequência do intervalo anterior ao da moda

f_{post} : é a frequência do intervalo posterior ao da moda

h_{Mo} : é a amplitude do intervalo da moda

3) Método de Czuber

É o método considerado mais preciso. É definido por: $Mo = l_{inf\ Mo} + \left(\frac{d_1}{d_1 + d_2} \right) h_{Mo}$

Onde:

$l_{inf\ Mo}$: é o limite inferior do intervalo onde está a moda;

d_1 - diferença entre a frequência da classe modal e a imediatamente anterior.

d_2 - diferença entre a classe modal e a imediatamente posterior.

h_{Mo} : é a amplitude do intervalo da moda

Exemplos:

Dados simples (não tabelados) : 2, 2, 3, 5, 6, 8, 8, 8, 10 $M_o = 8$

Dados tabelados discretos:

Nr de defeitos	f	
0	4	
1	7	
2	5	
3	2	
4	1	
5	1	
Total	20	

→ $M_o = 1$

Moda bruta

Classes	f	F
2 --- 4	3	3
4 --- 6	5	8
6 --- 8	7	15
8 --- 10	4	19
10 --- 12	1	20
Total	20	

Neste caso, verifica-se a maior freqüência, no caso o intervalo de 6 a 8. A moda será o ponto médio deste intervalo, ou seja

$$\frac{6+8}{2} = 7$$

Método de King

Classes	f	F
2 --- 4	3	3
4 --- 6	5	8
6 --- 8	7	15
8 --- 10	4	19
10 --- 12	1	20
Total	20	

Neste caso, usa-se a fórmula $M_o = l_{\inf Mo} + \left(\frac{f_{post}}{f_{ant} + f_{post}} \right) h_{Mo}$, então

$$M_o = l_{\inf Mo} + \left(\frac{f_{post}}{f_{ant} + f_{post}} \right) h_{Mo} = 6 + \left(\frac{4}{5+4} \right) 2 = 6 + \frac{8}{9} = 6,89$$

Método de Czuber

Classes	f	F
2 --- 4	3	3
4 --- 6	5	8
6 --- 8	7	15
8 --- 10	4	19
10 --- 12	1	20
Total	20	

Neste caso, usa-se a fórmula $M_o = l_{\inf Mo} + \left(\frac{d_1}{d_1 + d_2} \right) h_{Mo}$, então

$$M_o = l_{\inf Mo} + \left(\frac{d_1}{d_1 + d_2} \right) h_{Mo} = 6 + \left(\frac{(7-5)}{(7-5) + (7-4)} \right) 2 = 6 + \left(\frac{2}{5} \right) 2 = 6 + \frac{4}{5} = 6,80$$

Relação entre a média, a mediana e a moda

A relação entre a média, a mediana e a moda é a seguinte:

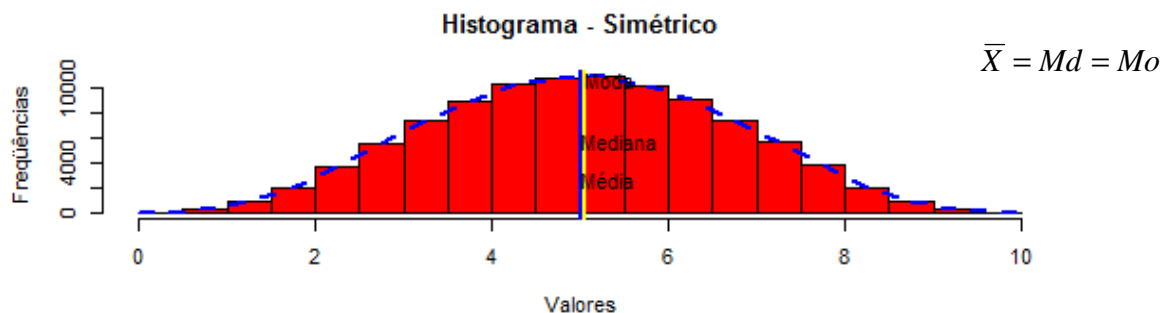
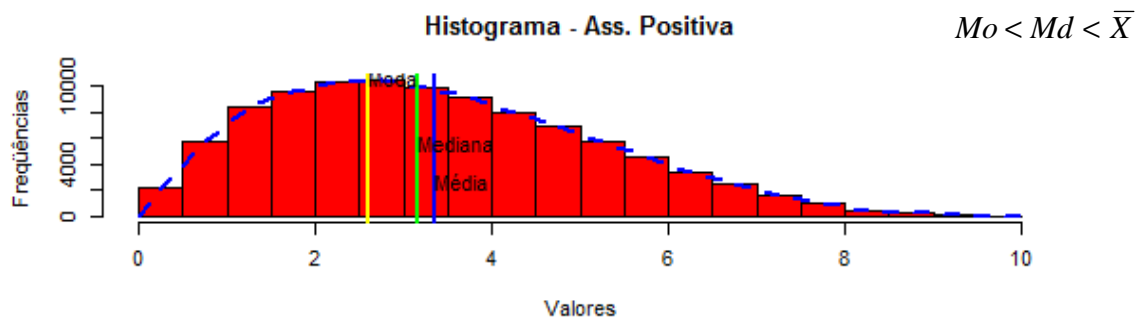
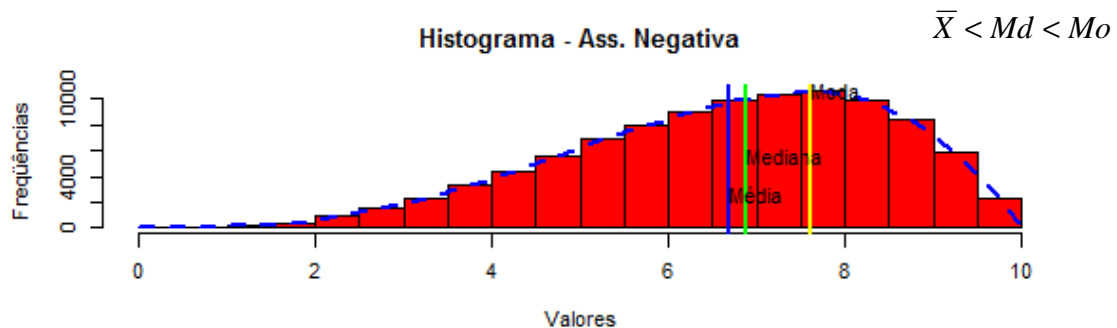
$$\bar{x} - m_o \cong 3(\bar{x} - m_d)$$

Por meio dela, é possível ter uma noção inicial de como está a distribuição dos dados, com relação à assimetria, ou seja:

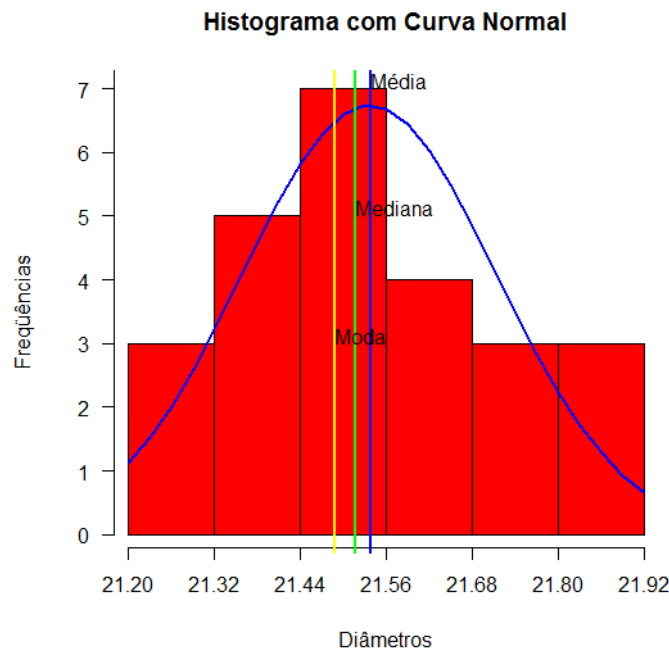
$\bar{X} < Md < Mo$, indica que a Assimetria é Negativa;

$\bar{X} = Md = Mo$, que indica que existe simetria na distribuição.

$Mo < Md < \bar{X}$, que indica que a Assimetria é positiva.



Para os dados “Diâmetro”, temos:



$$Mo < Md < \bar{X}$$

Ass. Positiva

13.1.2 Medidas de Posição: Separatrizes

São valores que dividem uma série ordenada de dados ou uma distribuição de frequência em partes iguais.

Principais separatrizes:

QUARTIL (Q_i) : divide a série ou a distribuição em quatro partes iguais.

QUINTIL (K_i): divide a série ou distribuição em cinco partes iguais.

DECIL (D_i) : divide a série ou a distribuição em dez partes iguais.

PERCENTIL (P_i) ou CENTIL (C_i) : divide a série ou a distribuição em cem partes iguais.

13.1.2.1 Quartil

a) Para dados simples:

Neste caso usamos:

$$PQ_1 = \frac{(n)}{4}$$

$$PQ_2 = \frac{(n+1)}{2}$$

$$PQ_3 = \frac{3(n)}{4}$$

Onde n é o tamanho do conjunto de dados

Uma vez encontrada a posição, utilizar:

$$Q_i = X_{ant\ pq} + (PQ_i - P_{ant\ pq})(X_{post\ pq} - X_{ant\ pq})$$

O resultado dado pela expressão acima indica a posição que estará o valor do conjunto de dados que representa o quartil considerado.

Por exemplo: seja o seguinte rol de dados 1, 2, 5, 5, 5, 8, 10, 11, 12, 12, 13, 15

Para determinar o Q_1 , sabendo que $n = 12$, faremos $PQ_i = \frac{i(n)}{4} \Rightarrow PQ_1 = \frac{1(12)}{4} = 3$. Então o Q_1 é o terceiro elemento do rol, no caso 5.

$$Q_i = X_{ant\ pq} + (PQ_i - P_{ant\ pq})(X_{post\ pq} - X_{ant\ pq}) \Rightarrow Q_1 = 2 + (3 - 2)(5 - 2) = 2 + 3 = 5$$

Para determinar o Q3: $PQ_i = \frac{i(n)}{4} \Rightarrow PQ_3 = \frac{3(12)}{4} = 9$

$$Q_3 = 11 + (9 - 8)(12 - 11) = 11 + 1 = 12$$

Para dados agrupados em intervalos de classe, temos:

$$Q_i = l_{\inf Q_i} + \left(\frac{i(25\%)n - F_{ant Q_i}}{f_{Q_i}} \right) h_{Q_i}$$

Para cada $i = 1, 2, 3$, temos:

$$Q_1 = l_{\inf Q_1} + \left(\frac{25\%n - F_{ant Q_1}}{f_{Q_1}} \right) h_{Q_1}$$

$$Q_2 = l_{\inf Q_2} + \left(\frac{50\%n - F_{ant Q_2}}{f_{Q_2}} \right) h_{Q_2}$$

$$Q_3 = l_{\inf Q_3} + \left(\frac{75\%n - F_{ant Q_3}}{f_{Q_3}} \right) h_{Q_3}$$

Onde:

$L_{\inf Q}$ – Limite inferior da classe de Q

$F_{ant Q} = \sum f_{ant Q}$ – Soma das frequências anteriores a classe de Q (Frequência Acumulada anterior)

f_Q – Frequência da classe de Q

h_Q – Amplitude do intervalo de classe de Q

Exemplo:

Vamos calcular o quartil Q1 e Q3, da distribuição abaixo:

Class limits	f	Rf	rf(%)	cf	cf(%)
[21.2,21.32)	3	0.12	12	3	12
[21.32,21.44)	5	0.20	20	8	32
[21.44,21.56)	7	0.28	28	15	60
[21.56,21.68)	4	0.16	16	19	76
[21.68,21.8)	3	0.12	12	22	88
[21.8,21.92)	3	0.12	12	25	100
Total	25	-	100	-	-

Para achar Q1, primeiro temos que achar $PQ_1 = \frac{(n)}{4} = \frac{25}{4} = 6,25$. Isto significa que Q1 está na posição 6,25, que se encontra no segundo intervalo, que vai de 21,32 a 21,44. A frequência acumulada anterior $F_{ant} = 3$, a frequência deste intervalo $f = 5$, e $h = 0,12$, pois $21,44 - 21,32 = 0,12$. Então, temos:

$$Q_1 = l_{\inf Q_1} + \left(\frac{25\%n - F_{ant Q_1}}{f_{Q_1}} \right) h_{Q_1} = 21,32 + \left(\frac{6,25 - 3}{5} \right) * 0,12 = 21,32 + 0,078 = 21,398 \approx 21,40$$

Para achar Q3, primeiro temos que achar $PQ_3 = \frac{3(n)}{4} = \frac{3(25)}{4} = 18,75$. Isto significa que Q3 está na posição 18,75, que se encontra no quarto intervalo, que vai de 21,56 a 21,68. A frequência

acumulada anterior $F_{ant} = 15$, a frequência deste intervalo $f = 4$, e $h = 0,12$, pois $21,68 - 21,56 = 0,12$. Então, temos:

$$Q_3 = l_{\inf Q_3} + \left(\frac{75\%n - F_{ant Q_3}}{f_{Q_3}} \right) h_{Q_3} = 21,56 + \left(\frac{18,75 - 15}{4} \right) * 0,12 = 21,56 + 0,1125 = 21,6725 \cong 21,67$$

13.1.2.3 Quintil (K)

Para dados simples:

Encontrar a posição do quintil utilizando: $Pk_i = \frac{i(n)}{5}$ onde $i = 1, \dots, 4$ e n é o tamanho do conjunto de dados.

Uma vez encontrada a posição, utilizar:

$$k_i = X_{ant} + (Pk_i - P_{ant})(X_{post} - X_{ant})$$

Por exemplo: seja o seguinte rol de dados 1, 2, 5, 5, 5, 8, 10, 11, 12, 12, 13, 15

Para determinar o K1, sabendo que $n = 12$, faremos $PK_i = \frac{i(n)}{5} \Rightarrow PK_1 = \frac{1(12)}{5} = 2,4$. Então o K₁ é o elemento do rol, entre o da posição 2 e o da posição 3. Vamos determiná-lo:

$$K_i = X_{ant pk} + (PK_i - P_{ant pk})(X_{post pk} - X_{ant pk}) \Rightarrow K_1 = 2 + (2,4 - 2)(5 - 2) = 2 + 1,2 = 3,2$$

Para determinar o K4, sabendo que $n = 12$, faremos $PK_i = \frac{i(n)}{5} \Rightarrow PK_4 = \frac{4(12)}{5} = 9,6$. Então o K₄ é o elemento do rol, entre o da posição 9 e o da posição 10. Vamos determiná-lo:

$$K_4 = 12 + (9,6 - 9)(12 - 12) = 12 + 0 = 12$$

Para dados agrupados:

$$k_i = l_{\inf k_i} + \left(\frac{(20i)\%n - F_{ant k_i}}{f_{k_i}} \right) h_{k_i}$$

Onde:

$Lin f k$ – Limite inferior da classe de k

$F_{ant k}$ – Frequência acumulada anterior a classe de k

$f k$ – Frequência da classe de k

$h k$ – Amplitude do intervalo de classe de k

Exemplo: vamos calcular K1 para a distribuição abaixo

Class limits	f	Rf	rf(%)	Cf	cf(%)
[21.2,21.32)	3	0.12	12	3	12
[21.32,21.44)	5	0.20	20	8	32
[21.44,21.56)	7	0.28	28	15	60
[21.56,21.68)	4	0.16	16	19	76
[21.68,21.8)	3	0.12	12	22	88
[21.8,21.92)	3	0.12	12	25	100
Total	25	-	100	-	-

Para achar K1, primeiro temos que achar $Pk_i = \frac{i(n)}{5} \Rightarrow PK_1 = \frac{1(25)}{5} = 5$. Isto significa que K1 está na posição 5, que se encontra no segundo intervalo, que vai de 21,32 a 21,44. A frequência acumulada anterior $F_{ant} = 3$, a frequência deste intervalo $f = 5$, e $h = 0,12$, pois $21,44 - 21,32 = 0,12$. Então, temos:

$$K_1 = l_{inf\ k1} + \left(\frac{PK_1 - F_{ant\ k1}}{f_{k1}} \right) h_{k1} = 21,32 + \left(\frac{5 - 3}{5} \right) * 0,12 = 21,32 + 0,048 = 21,368 \cong 21,37$$

13.1.2.4 Decil

Para dados simples:

Neste caso usamos $PD_i = \frac{i(n)}{10}$, onde $i = 1$ a 4 e de 5 a 9 e n é o tamanho do conjunto de dados.

Para $i = 5$, utilizar $PD_5 = \frac{10(n+1)}{2}$

Uma vez encontrada a posição, utilizar:

$$D_i = X_{ant} + (PD_i - P_{ant})(X_{post} - X_{ant})$$

Para dados agrupados em intervalos de classe, temos:

$$D_i = l_{inf\ D_i} + \left(\frac{(10i)\%n - F_{ant\ D_i}}{f_{D_i}} \right) h_{D_i}$$

Onde:

$L_{inf\ D}$ – Limite inferior da classe de D

$F_{ant\ D} = \sum f_{ant\ D}$ – Soma das frequências anteriores a classe de D (Frequência Acumulada anterior)

f_D – Frequência da classe de D

h_D – Amplitude do intervalo de classe de D

Exemplo: vamos calcular D1 para a distribuição abaixo

Class limits	f	Rf	rf(%)	Cf	cf(%)
[21,2,21,32)	3	0,12	12	3	12
[21,32,21,44)	5	0,20	20	8	32
[21,44,21,56)	7	0,28	28	15	60
[21,56,21,68)	4	0,16	16	19	76
[21,68,21,8)	3	0,12	12	22	88
[21,8,21,92)	3	0,12	12	25	100
Total	25	-	100	-	-

Para achar D1, primeiro temos que achar $PD_i = \frac{i(n)}{10} \Rightarrow PD_1 = \frac{1(25)}{10} = 2,5$. Isto significa que D1 está na posição 2,5, que se encontra no primeiro intervalo, que vai de 21,2 a 21,32. A frequência acumulada anterior $F_{ant} = 0$, a frequência deste intervalo $f = 3$, e $h = 0,12$, pois $21,32 - 21,2 = 0,12$. Então, temos:

$$D_1 = l_{inf\ D1} + \left(\frac{PD_1 - F_{ant\ D1}}{f_{D1}} \right) h_{D1} = 21,2 + \left(\frac{2,5 - 0}{3} \right) * 0,12 = 21,2 + 0,10 = 21,30$$

13.1.2.5 Percentil ou Centil

Para dados simples:

Neste caso usamos $PP_i = \frac{i(n)}{100}$, onde $i = 1$ a 49 e de 51 a 99 e n é o tamanho do conjunto de dados.

Para $i = 50$, utilizar:

$$PP_{50} = \frac{(n+1)}{2}$$

Uma vez encontrada a posição, utilizar:

$$P_i = X_{ant} + (PP_i - P_{ant})(X_{post} - X_{ant})$$

Para dados agrupados em intervalos de classe, temos:

$$P_i = l_{\inf P_i} + \left(\frac{i\%n - F_{ant P_i}}{f_{P_i}} \right) h_{P_i}$$

Onde:

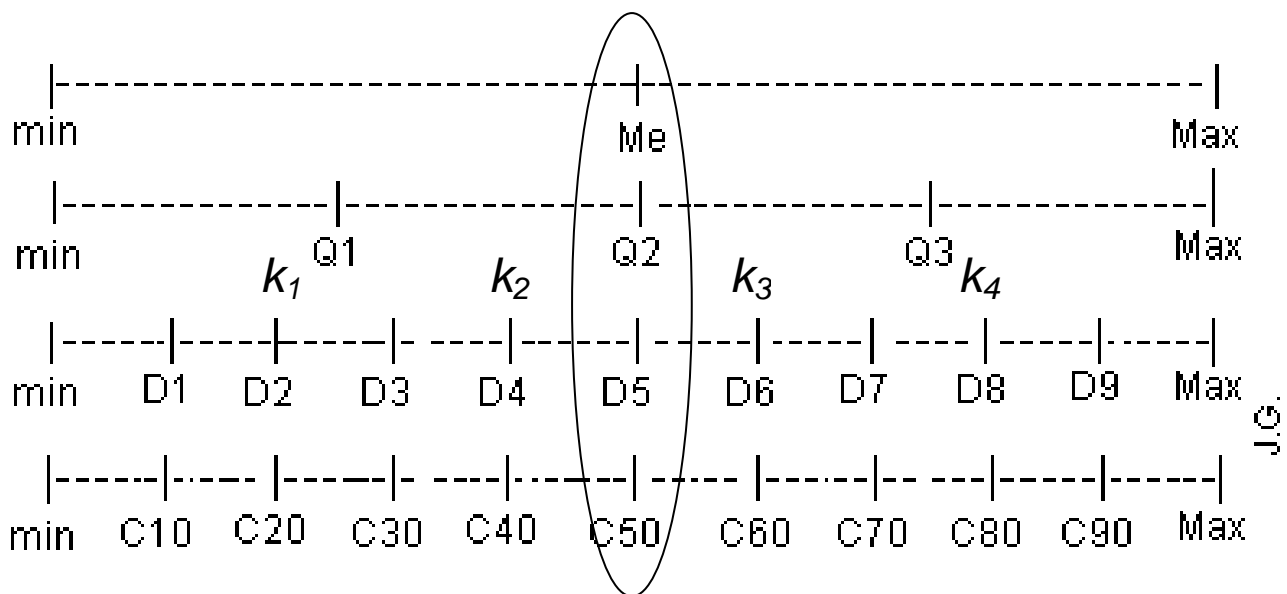
$l_{\inf P}$ – Limite inferior da classe de P

$F_{ant P}$ – Frequência acumulada anterior a classe de P

f_P – Frequência da classe de P

h_P – Amplitude do intervalo de classe de P

Graficamente, temos:



$$Md = Q_2 = D_5 = P_{50}$$

Exemplo: Calcule o valor do vigésimo percentil (P_{20}).

Nºtel.	f	F	fr	Fr
7 --12	3	3	10,00%	10,00%
12 --17	10	13	33,33%	43,33%
17 --22	8	21	26,67%	70,00%
22 --27	5	26	16,67%	86,67%
27 --32	2	28	6,67%	93,34%
32 --37	2	30	6,66%	100%
	30			

$$P_i = l_{\inf P_i} + \left(\frac{i\%n - F_{ant P_i}}{f_{P_i}} \right) h_{P_i}$$

$$P_{20} = 12 + \left(\frac{20\%(30) - 3}{10} \right) 5$$

$$P_{20} = 12 + \left(\frac{6 - 3}{10} \right) 5$$

$$P_{20} = 12 + 1,50$$

$$P_{20} = 13,50$$

13.2 Esquema de Cinco Números e Boxplot (ou Gráfico Box-and-Whisker)

Após estudar as principais medidas de posição dos dados numéricos, é importante identificar e descrevê-los em um formato resumido.

1) Esquema de Cinco Números

Um esquema de cinco números consiste em determinar:

Xmenor Q1 Mediana Q3 Xmaior

Se os dados são perfeitamente simétricos, temos:

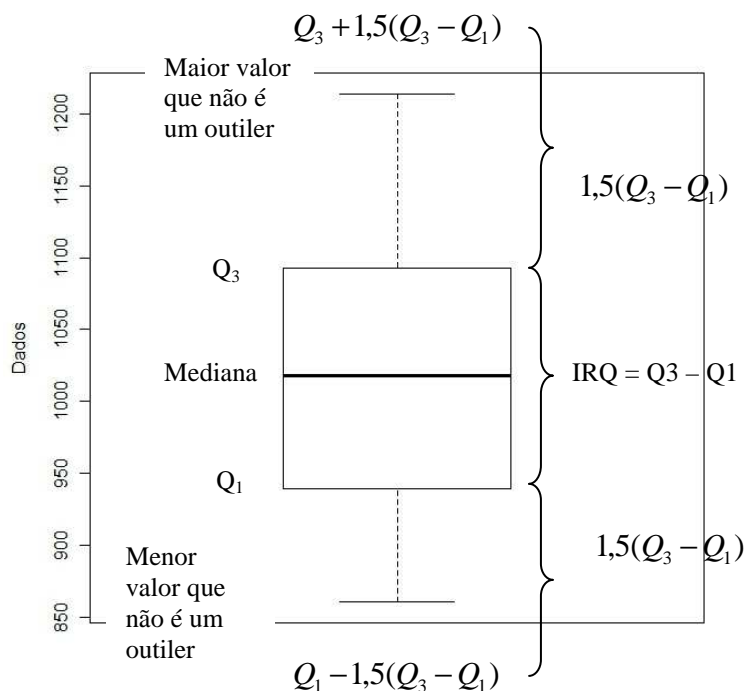
- A distância de Q1 até a mediana é igual à distância da mediana até Q3;
- A distância de Xmenor até a Q1 é igual à distância de Q3 até Xmaior;

2) Boxplot

O boxplot (gráfico de caixa) é um gráfico utilizado para avaliar a distribuição empírica dos dados. Ele serve como representação gráfica do esquema de Cinco Números. a utilização do gráfico permite avaliar a simetria e distribuição dos dados. O boxplot é formado pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite superior. Os limites são calculados da forma abaixo:

$$\text{Limite inferior: } Q_1 - 1,5(Q_3 - Q_1)$$

$$\text{Limite superior: } Q_3 + 1,5(Q_3 - Q_1)$$



Uma outra utilização do boxplot refere-se a identificação de pontos de discrepância ou observações discrepantes, os famosos "outliers". Estes valores podem afetar de forma substancial o resultado das análises estatísticas. A existência destas observações discrepantes estão relacionadas com erros de medição, erros de execução e variabilidade inerente aos elementos da população. Para identificação de um outlier, fazemos o seguinte: seja x^o um valor da variável de estudo; compara-se este valor x^o com $Q_1 - 1,5(Q_3 - Q_1)$ e com $Q_1 + 1,5(Q_3 - Q_1)$. O valor x^o será um outlier se:

$$x^o < Q_1 - 1,5(Q_3 - Q_1)$$

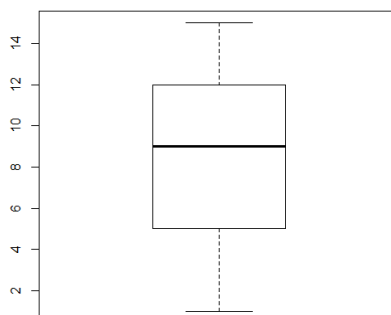
ou

$$x^o > Q_1 + 1,5(Q_3 - Q_1)$$

Uma vez identificado o outlier, o pesquisador poderá eliminá-lo, caso seja apenas um valor, ou trocá-los pela média da variável de estudo, calculada sem os referidos valores. Porém, deve-se investigar as razões que levaram ao surgimento destes valores.

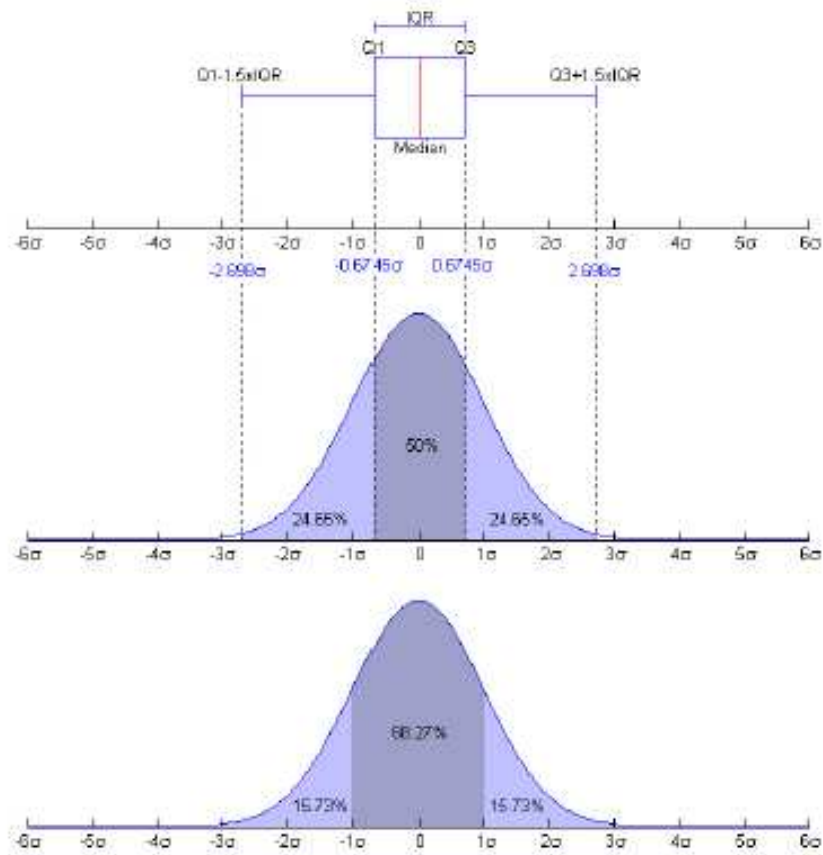
Como exemplo, temos os mesmos valores apresentados para o cálculo de Q₁ e Q₃, ou seja:

1, 2, 5, 5, 5, 8, 10, 11, 12, 12, 13, 15. O valor de Q₁ para estes dados foi igual a 5 e de Q₃ foi igual a 12. A mediana foi igual a 9. O boxplot para estes dados fica assim:



Percebe-se que este caso, não há valores discrepantes.

É possível fazer a comparação do Boxplot com a distribuição Normal. A seguir temos esta comparação:



13.3 Medidas de dispersão

São medidas que traduzem a variação de um conjunto de dados em torno da média, ou seja, da maior ou menor variabilidade dos resultados obtidos. Permitem identificar até que ponto os resultados se concentram ou não ao redor da tendência central de um conjunto de observações. Quanto maior for a dispersão, menor é a concentração e vice-versa. As medidas de dispersão podem ser absolutas e relativas.

13.3.1 Medidas de Dispersão Absolutas

13.3.1.1 Amplitude Total

É a diferença entre o maior e o menor valores do conjunto de dados.

$$R = X_{\max} - X_{\min}$$

13.3.1.2 Amplitude Interquartílica - AI

É a diferença entre o terceiro quartil Q3 e o primeiro quartil Q1

$$AI = Q_3 - Q_1$$

13.3.1.3 Amplitude entre os Percentis 10-90 – AP10-90

É a diferença entre o Percentil 90 e o Percentil 10.

$$AP_{10-90} = P_{90} - P_{10}$$

13.3.1.4 Variância

É a média dos quadrados dos desvios em relação a média.

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad \text{- Para dados simples}$$

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{n-1} \quad \text{- Para dados agrupados em intervalos de classe}$$

É possível calcular a variância de outra maneira.

Sabendo que:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \quad \text{- Para dados simples}$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 f_i = \sum_{i=1}^n X_i^2 f_i - \frac{\left(\sum_{i=1}^n X_i f_i\right)^2}{n} \quad \text{- Para dados com frequência}$$

Temos:

$$s^2 = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1} \quad \text{- Para dados simples}$$

$$s^2 = \frac{\sum_{i=1}^n X_i^2 f_i - \frac{\left(\sum_{i=1}^n X_i f_i\right)^2}{n}}{n-1} \quad \text{- Para dados com frequência}$$

Propriedades da variância

- 1) Se $X = k$, onde k é uma constante, $\text{Var}(X) = 0$

- 2) Se $Y = X + k$, onde k é uma constante, $\text{Var}(Y) = \text{Var}(X)$
- 3) Se $Y = kX$, onde k é uma constante, $\text{Var}(Y) = k^2\text{Var}(X)$.

A variância é uma medida de dispersão extremamente importante na Teoria Estatística. Do ponto de vista prático, ela tem o inconveniente de se expressar numa unidade quadrática em relação à variável em questão. Esse inconveniente é sanado com a definição do desvio-padrão.

13.3.1.5 Desvio-padrão

Define-se desvio-padrão como a raiz quadrada da variância.

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

O desvio-padrão se expressa na mesma unidade da variável, sendo, por isso, de maior interesse que a variância nas aplicações práticas.

É comum apresentar a média e o desvio-padrão para indicar a amplitude da dispersão da amostra, da seguinte forma:

$$\bar{X} \pm s$$

13.3.1.6 Desvio Absoluto Médio - DAM

É a média dos valores absolutos das diferenças entre as observações e a média.

$$DAM = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} \text{ - Para dados simples}$$

$$DAM = \frac{\sum_{i=1}^n |X_i - \bar{X}| f_i}{n} \text{ - Para dados agrupados em intervalos de classe}$$

13.3.1.7 Desvio Quartílico – DQ

$$DQ = \frac{Q_3 - Q_1}{2}$$

13.3.1.8 Relações empíricas entre as medidas de dispersão absolutas:

$$DMA = \frac{4}{5} s$$

$$DQ = \frac{2}{3} s$$

$$DMA = \frac{6}{5} DQ$$

13.3.1.9 Desvio Absoluto ao redor da Mediana (MAD)

O MAD é uma medida robusta da variabilidade de uma variável. É calculado por:

$$MAD = \text{mediana}(|x_i - \text{mediana}(x_i)|)$$

O MAD é um consistente estimador do desvio-padrão. Então, se os dados possuem uma distribuição Normal, temos que:

$$\hat{\sigma} = 1,4826MAD$$

13.3.2 Medidas de Dispersão Relativas

13.3.2.1 Coeficiente de variação

Define-se coeficiente de variação como o quociente entre o desvio-padrão e a média.

$$CV = \left(\frac{s}{\bar{X}} \right) 100$$

Sua vantagem é caracterizar a dispersão dos dados em valores relativos a seu valor médio. Assim, uma pequena dispersão absoluta pode ser, na verdade, considerável quando comparada com a ordem de grandeza dos valores da variável e vice-versa. Quando consideramos o coeficiente de variação, enganos de interpretação desse tipo são evitados.

Existe uma escala para a verificação do grau de dispersão, em função do coeficiente de variação:

$CV \leq 10\%$, grau de dispersão baixo;
 $10\% < CV \leq 20\%$, grau de dispersão médio;
 $20\% < CV \leq 30\%$, grau de dispersão alto;
 $CV > 30\%$, grau de dispersão muito alto.

13.3.2.2 Coeficiente de Variação de Thorndike

$$CV_t = \frac{s}{Md} \cdot 100$$

13.3.2.3 Coeficiente de Variação pelo Intervalo Quartil ou Coeficiente de Variação Quartílico

$$CV_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \cdot 100$$

13.3.2.4 Desvio Quartil Reduzido (DSR)

Por definição é a amplitude semi-interquartílica sobre a mediana.

$$DQR = \frac{\frac{Q_3 - Q_1}{2}}{Md} \cdot 100 = \frac{Q_3 - Q_1}{2Md} \cdot 100$$

Como exemplo, temos:

a) para valores não tabelados

dist

	x	desvio	desv.abs	desvio2
	2	-3.78	3.78	14.2884
	2	-3.78	3.78	14.2884
	3	-2.78	2.78	7.7284
	5	-0.78	0.78	0.6084
	6	0.22	0.22	0.0484
	8	2.22	2.22	4.9284
	8	2.22	2.22	4.9284
	8	2.22	2.22	4.9284
	10	4.22	4.22	17.8084
Total	52	NA	22.22	69.5600

$$DAM = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n} = \frac{22.22}{9} = 2,469$$

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{69,56}{9-1} = 8,695$$

$$s = \sqrt{s^2} = \sqrt{8.695} = 2,949$$

$$DQ = \frac{Q_3 - Q_1}{2} = \frac{8 - 2,25}{2} = 2,875$$

$$CV = \left(\frac{s}{\bar{X}} \right) 100 = \frac{2,949}{5,78} 100 = 51,016\%$$

$$CV_t = \frac{s}{Md} \cdot 100 = \frac{2,949}{6} 100 = 49,15\%$$

$$CV_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \cdot 100 = \frac{8 - 2,25}{8 + 2,25} 100 = 56,098\%$$

$$DQR = \frac{\frac{Q_3 - Q_1}{2}}{Md} \cdot 100 = \frac{Q_3 - Q_1}{2Md} \cdot 100 = \frac{8 - 2,25}{2 * 6} 100 = 47,92\%$$

b) para valores tabelados

Com base nos cálculos anteriores para médias, moda e mediana, usando a tabela abaixo, temos:

Class limits	PM	f	Rf	rf(%)	cf	cf(%)	Xf	log x	f*log x	$\frac{f}{x}$
[21.2,21.32)	21.26	3	0.12	12	3	12	63,78	1,328	3,984	0,141
[21.32,21.44)	21.38	5	0.20	20	8	32	106,90	1,330	6,650	0,234
[21.44,21.56)	21.50	7	0.28	28	15	60	150,50	1,332	9,324	0,326
[21.56,21.68)	21.62	4	0.16	16	19	76	86,48	1,335	5,340	0,185
[21.68,21.8)	21.74	3	0.12	12	22	88	65,22	1,337	4,011	0,138
[21.8,21.92)	21.86	3	0.12	12	25	100	65,58	1,340	4,020	0,137
Total	-	25	-	100	-	-	538,46		33,329	1,161

Devemos acrescentar algumas colunas após a última coluna f/x. Para efeito didático, vamos retirar as colunas Xf, log x, f*log x e f/x, e em seu lugar vamos digitar outras colunas. Então temos:

Class limits	PM	f	Rf	rf(%)	cf	cf(%)	Desvio	Desvio- \bar{X} *f	(Desvio- \bar{X}) ² f
[21.2,21.32)	21.26	3	0.12	12	3	12			
[21.32,21.44)	21.38	5	0.20	20	8	32			
[21.44,21.56)	21.50	7	0.28	28	15	60			
[21.56,21.68)	21.62	4	0.16	16	19	76			
[21.68,21.8)	21.74	3	0.12	12	22	88			
[21.8,21.92)	21.86	3	0.12	12	25	100			
Total	-	25	-	100	-	-			

Fazendo os cálculos, temos:

Class limits	PM	f	Rf	rf(%)	cf	cf(%)	Desvio	Desvio- \bar{X} *f	(Desvio- \bar{X}) ² f
[21.2,21.32)	21.26	3	0.12	12	3	12	-0,2784	0,8352	0,23251968
[21.32,21.44)	21.38	5	0.20	20	8	32	-0,1584	0,7920	0,12545280
[21.44,21.56)	21.50	7	0.28	28	15	60	-0,0384	0,2688	0,01032192
[21.56,21.68)	21.62	4	0.16	16	19	76	0,0816	0,3264	0,02663424
[21.68,21.8)	21.74	3	0.12	12	22	88	0,2016	0,6048	0,12192768
[21.8,21.92)	21.86	3	0.12	12	25	100	0,3216	0,9648	0,31027968
Total	-	25	-	100	-	-	-	3,792	0,827136

$$DAM = \frac{\sum_{i=1}^n |X_i - \bar{X}| * f}{n} = \frac{3,792}{25} = 0,16$$

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 * f}{n-1} = \frac{0,827}{25-1} = 0,042$$

$$s = \sqrt{s^2} = \sqrt{0,042} = 0,204$$

$$DQ = \frac{Q_3 - Q_1}{2} = \frac{21,67 - 21,40}{2} = 0,137$$

$$CV = \left(\frac{s}{\bar{X}} \right) 100 = \frac{0,204}{21,54} 100 = 0,95\%$$

$$CV_r = \frac{s}{Md} \cdot 100 = \frac{0,204}{21,52} 100 = 0,95\%$$

$$CV_q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \cdot 100 = \frac{21,67 - 21,40}{21,67 + 21,40} 100 = 0,64\%$$

$$DQR = \frac{\frac{Q_3 - Q_1}{2}}{Md} \cdot 100 = \frac{Q_3 - Q_1}{2Md} \cdot 100 = \frac{21,67 - 21,40}{2 \cdot 21,52} 100 = 0,64\%$$

13.4 Momentos

São quantidades numéricas ou valores de uma distribuição de uma variável X, usadas para a caracterização de determinadas medidas, tais como a média aritmética e a variância, além de medidas do formato da distribuição como a assimetria e a curtose.

São determinados por meio do valor esperado (média) das potências de X. As esperanças das sucessivas potências de X constituem o conceito de momentos dessa variável aleatória.

Momento de ordem r

$$m_r = \frac{\sum_{i=1}^n X_i^r}{n} \quad \text{- Para dados simples}$$

$$m_r = \frac{\sum_{i=1}^n X_i^r f_i}{n} \quad \text{- Para dados agrupados em intervalos de classe}$$

Momento de ordem r centrado na média

$$\mu_r = \frac{\sum_{i=1}^n (X_i - \bar{X})^r}{n} \quad \text{- Para dados simples}$$

$$\mu_r = \frac{\sum_{i=1}^n (X_i - \bar{X})^r f_i}{n} \quad \text{- Para dados agrupados em intervalos de classe}$$

Momentos Importantes de uma Distribuição

Momento de ordem r = 1 : média

Para dados simples:

$$m_1 = \frac{\sum_{i=1}^n X_i}{n}$$

Para dados agrupados:

$$m_1 = \frac{\sum_{i=1}^n X_i f_i}{n}$$

Momento de ordem $r = 2$ centrado na média - σ^2

Para dados simples:

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Para dados agrupados:

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f_i}{n}$$

Momento de ordem $r = 3$ centrado na média

Para dados simples:

$$\mu_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{n}$$

Para dados agrupados:

$$\mu_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3 f_i}{n}$$

Momento de ordem $r = 4$ centrado na média

Para dados simples:

$$\mu_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}$$

Para dados agrupados:

$$\mu_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 f_i}{n}$$

Relação entre os momentos

$$\mu_2 = m_2 - m_1^2$$

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3$$

$$\mu_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$$

13.5 Assimetria

É o grau de desvio, ou afastamento da simetria, de uma distribuição.

13.5.1 Critério de Pearson

Pelo critério de Pearson, à medida que a distribuição deixa de ser simétrica, a média, a moda e a mediana vão se afastando, aumentando cada vez mais a diferença existente entre elas.

Seu cálculo pode ser definido por:

$$A_s = \frac{\bar{X} - Mo}{s} \text{ Primeiro coeficiente de assimetria de Pearson}$$

$$A_s = \frac{3(\bar{X} - Md)}{s} \text{ Segundo coeficiente de assimetria de Pearson}$$

Podemos verificar a assimetria de uma distribuição comparando os resultados com:

Se $AS < 0$ - ass. Negativa

Se $AS = 0$ - simétrica

Se $AS > 0$ - ass. Positiva

13.5.2 Critério de Bowley

Pelo critério de Bowley, à medida que a distribuição deixa de ser simétrica, os quartis deixam de serem equidistantes da mediana.

Seu cálculo pode ser definido por:

$$A_S = \frac{Q_3 - 2Md + Q_1}{Q_3 - Q_1} \text{ Coeficiente quartílico de assimetria}$$

13.5.3 Critério de Kelley

O critério de Bowley despreza 50% das ocorrências. Para evitar isso, Kelley aconselha o uso de percentis equidistantes da mediana, tais como o P10 e P90, surgindo daí a seguinte fórmula:

$$A_S = \frac{P_{90} - 2Md + P_{10}}{P_{90} - P_{10}} \text{ Coeficiente percentílico de assimetria}$$

13.5.4 Critério de Fisher

Esta medida de assimetria utiliza o 2º e o 3º momento centrado na média, ou seja:

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \text{ - 2º momento centrado na média para dados não tabelados}$$

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f}{n} \text{ - 2º momento centrado na média para dados tabelados}$$


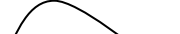

$$\mu_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{n} \text{ - 3º momento centrado na média para dados não tabelados}$$

$$\mu_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3 f_i}{n} \text{ - 3º momento centrado na média para dados tabelados}$$

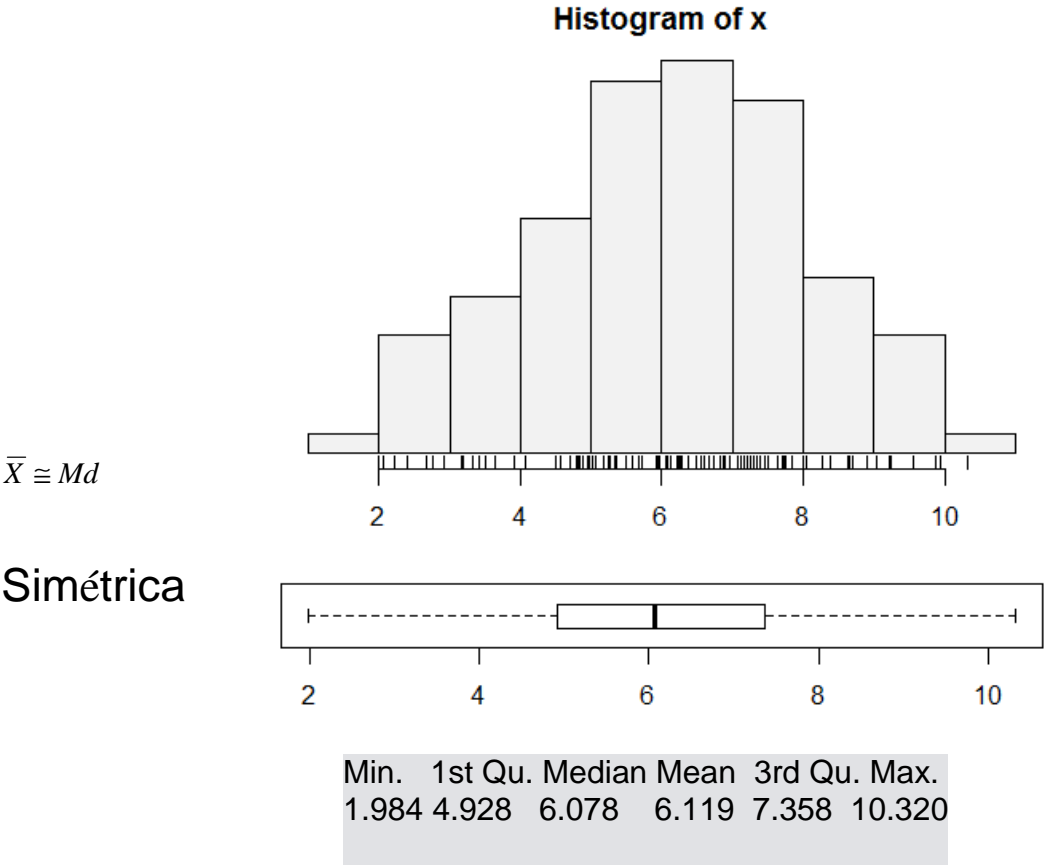
É calculado da seguinte forma:

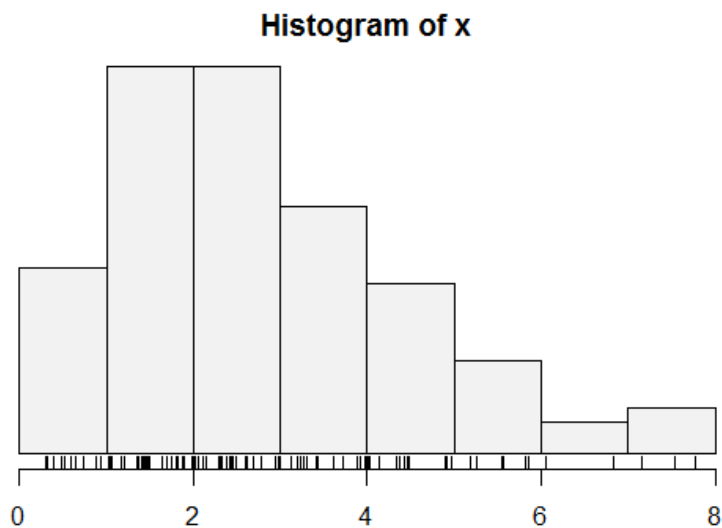
$$g_1 = \frac{n}{(n-1)(n-2)} \frac{\mu_3}{(\sqrt{\mu_2})^3} \text{ quando } n < 25$$

$$g_1 = \frac{\mu_3}{(\sqrt{\mu_2})^3}, \text{ quando } n \geq 25$$

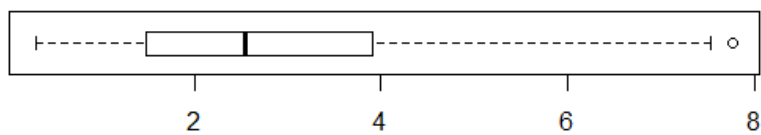
Se $g_1 < 0$ - ass. Negativa - 
 Se $g_1 = 0$ - simétrica 
 Se $g_1 > 0$ - ass. Positiva - 

Ex: Para dados simulados temos:



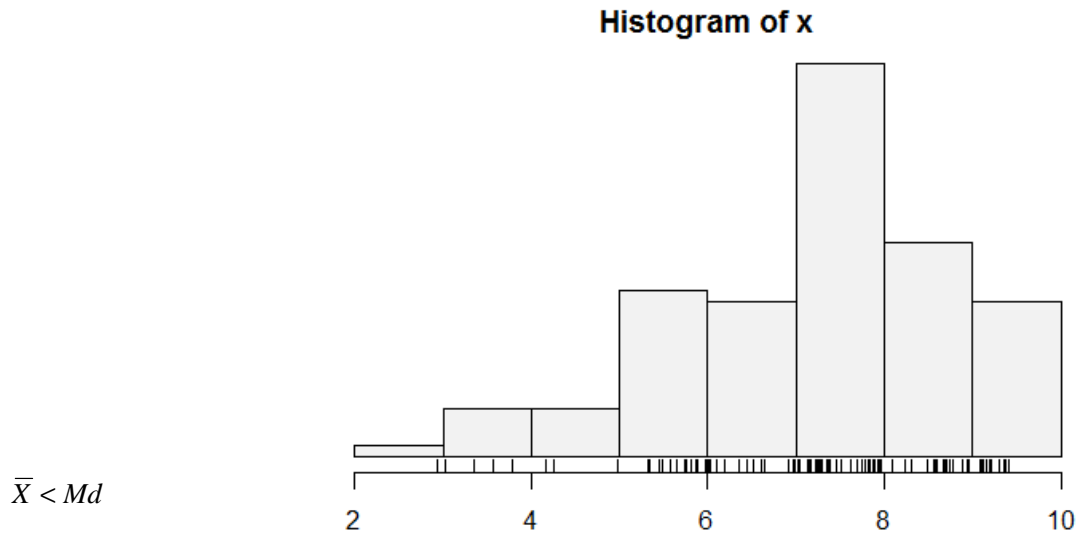


$$Md < \bar{X}$$



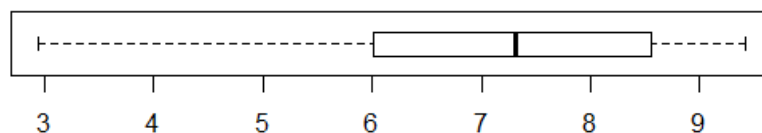
Ass. Positiva

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3108	1.4840	2.5430	2.8080	3.8950	7.7600



$$\bar{X} < Md$$

Ass. Negativa

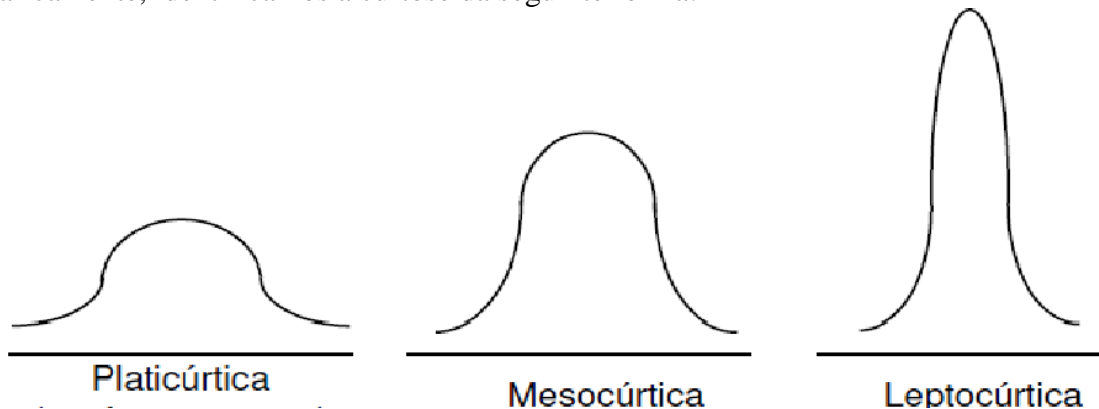


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.093	7.411	8.266	7.939	9.020	9.904

13.6 Curtose

É o grau de achatamento de uma distribuição, considerado usualmente em relação a uma distribuição Normal.

Graficamente, identificamos a curtose da seguinte forma:



13.6.1 Coeficiente Percentílico de Curtose

Um dos coeficientes mais utilizados para medir o grau de achatamento ou curtose de uma distribuição. É calculado a partir do intervalo interquartil, além dos percentis P10 e P90, da seguinte forma:

$$k = \frac{\frac{Q_3 - Q_1}{2}}{P_{90} - P_{10}} = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$

Se $k > 0,263$ – dizemos que a distribuição é platicúrtica

Se $k = 0,263$ – dizemos que a distribuição é mesocúrtica

Se $k < 0,263$ – dizemos que a distribuição é leptocúrtica

Obs: a fórmula acima também pode ser escrita em função dos Decis 1 e 9, uma vez que $D1 = P10$ e $D9 = P90$. Então a expressão fica:

$$k = \frac{\frac{Q_3 - Q_1}{2}}{D_9 - D_1} = \frac{Q_3 - Q_1}{2(D_9 - D_1)}$$

13.6.2 Coeficiente de Curtose de Fisher

Esta medida de curtose utiliza o 2º e o 4º momento centrado na média, ou seja:

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad - 2^\circ \text{ momento centrado na média para dados não tabelados}$$

$$\mu_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 f}{n} \quad - 2^\circ \text{ momento centrado na média para dados tabelados}$$

$$\mu_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}$$

- 4º momento centrado na média para dados não tabelados

$$\mu_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 f_i}{n} \quad \text{- 4º momento centrado na média para dados tabelados}$$

É calculado da seguinte forma:

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\mu_4}{\mu_2^2} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \text{quando } n < 25$$

$$g_2 = \frac{\mu_4}{\mu_2^2} - 3, \quad \text{quando } n \geq 25$$

Podemos verificar a curtose de uma distribuição comparando os resultados com:

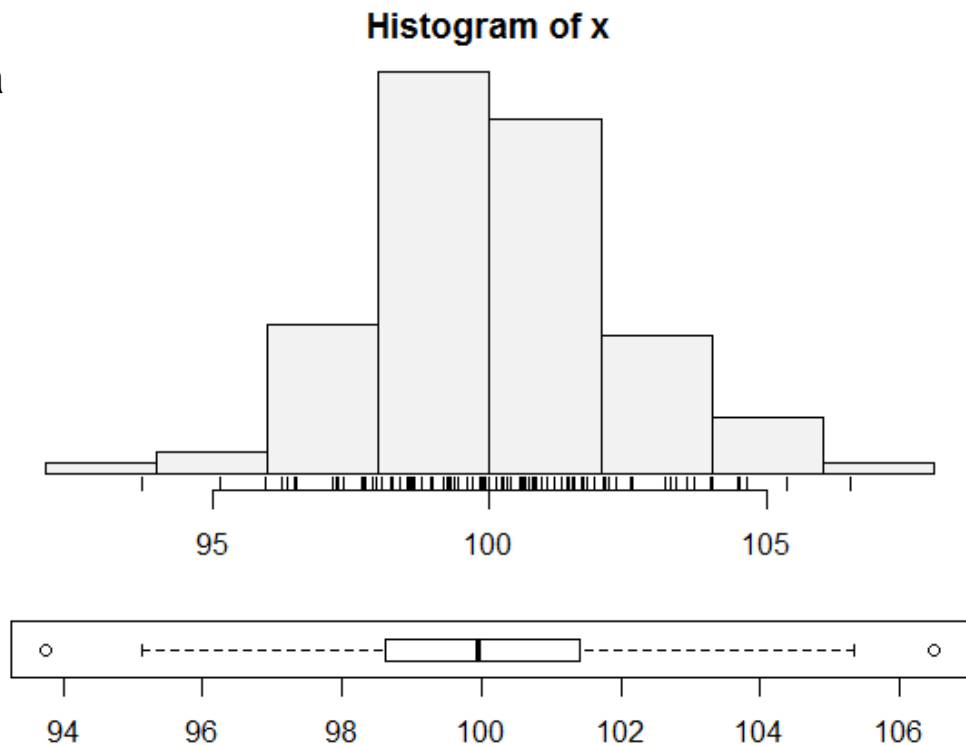
Se $g_2 < 0$ – dizemos que a distribuição é platicúrtica

Se $g_2 = 0$ – dizemos que a distribuição é mesocúrtica

Se $g_2 > 0$ – dizemos que a distribuição é leptocúrtica

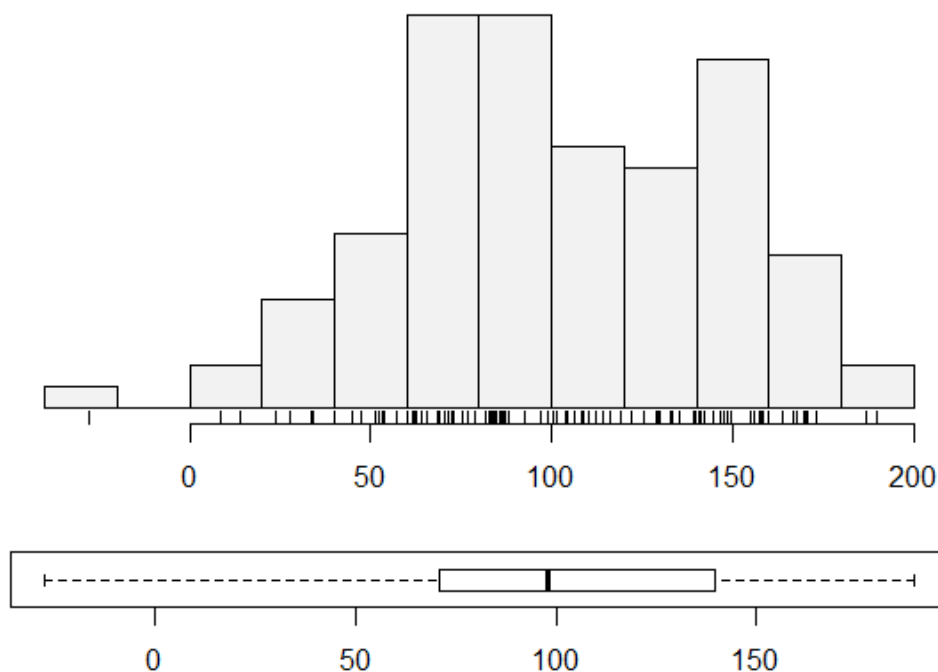
Exemplos para dados simulados:

Leptocúrtica



Histogram of x

Platicúrtica



Como exemplo, temos:

- a) para dados não tabelados
- b) para dados tabelados

14. Teste de Normalidade

Teste de Normalidade é um teste apropriado para verificar a hipótese dos valores apresentarem uma distribuição normal (H_0) contra a hipótese de não apresentarem (H_1).

Existem vários testes para se verificar a normalidade. Um deles é o Teste de Jarque-Bera. Neste teste, são utilizados o Coeficiente Momento de Assimetria (g_1) e o Coeficiente Momento de Curtose (g_2). É calculado por:

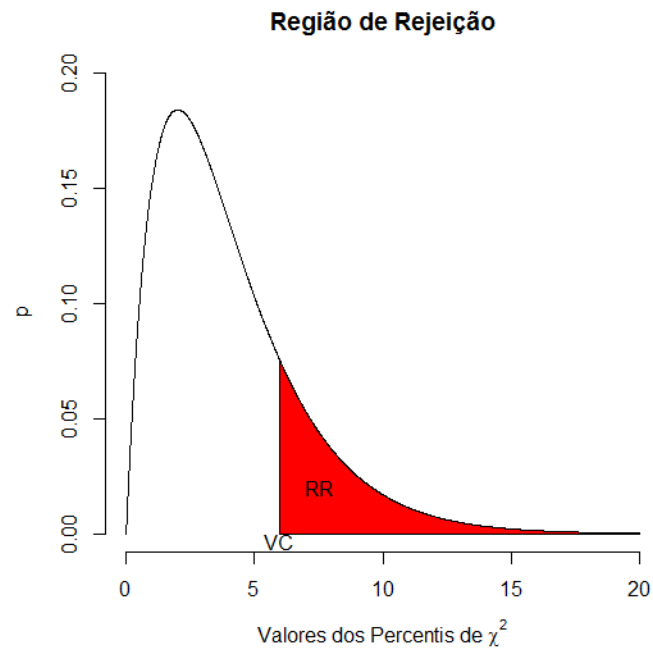
$$JB = \frac{n}{6} \left(g_1^2 + \frac{1}{4} g_2^2 \right)$$

A estatística JB tem distribuição assintótica χ^2 com 2 graus de liberdade, sob a hipótese nula.

Rejeita-se a hipótese nula (H_0), ou seja a hipótese da normalidade se o valor de JB, comparado em uma tabela de Qui-quadrado (χ^2), com um nível de significância α e (2) graus de liberdade, apresentar um valor-p menor do que o valor admitido, que por ser:

Assim, se o p-valor for menor do que 5% (ou 10%), $p < 0,05$ ($p < 0,10$), então rejeita-se a normalidade. Já se $p > 0,05$, não se rejeita a normalidade.

Uma forma mais fácil de saber se rejeitaremos ou não H_0 , é comparando com os valores tabelados de Qui-quadrado, com uma nível de significância α , com o valor calculado de JB. Para tanto, usamos o seguinte esquema:



Os valores de VC para os valores de α mais usuais são:

Alfa (α)	VC
1%	9,210
5%	5,991
10%	4,605

Então, rejeita-se H_0 se $JB > VC$.

15. Correlação

O coeficiente de correlação de Pearson (r) ou coeficiente de correlação produto-momento ou o r de Pearson mede o grau da correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados entre -1,0 e 1,0 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados.

Este coeficiente, normalmente representado pela letra " r " assume apenas valores entre -1 e 1.

$r = 1$ Significa uma correlação perfeita positiva entre as duas variáveis.

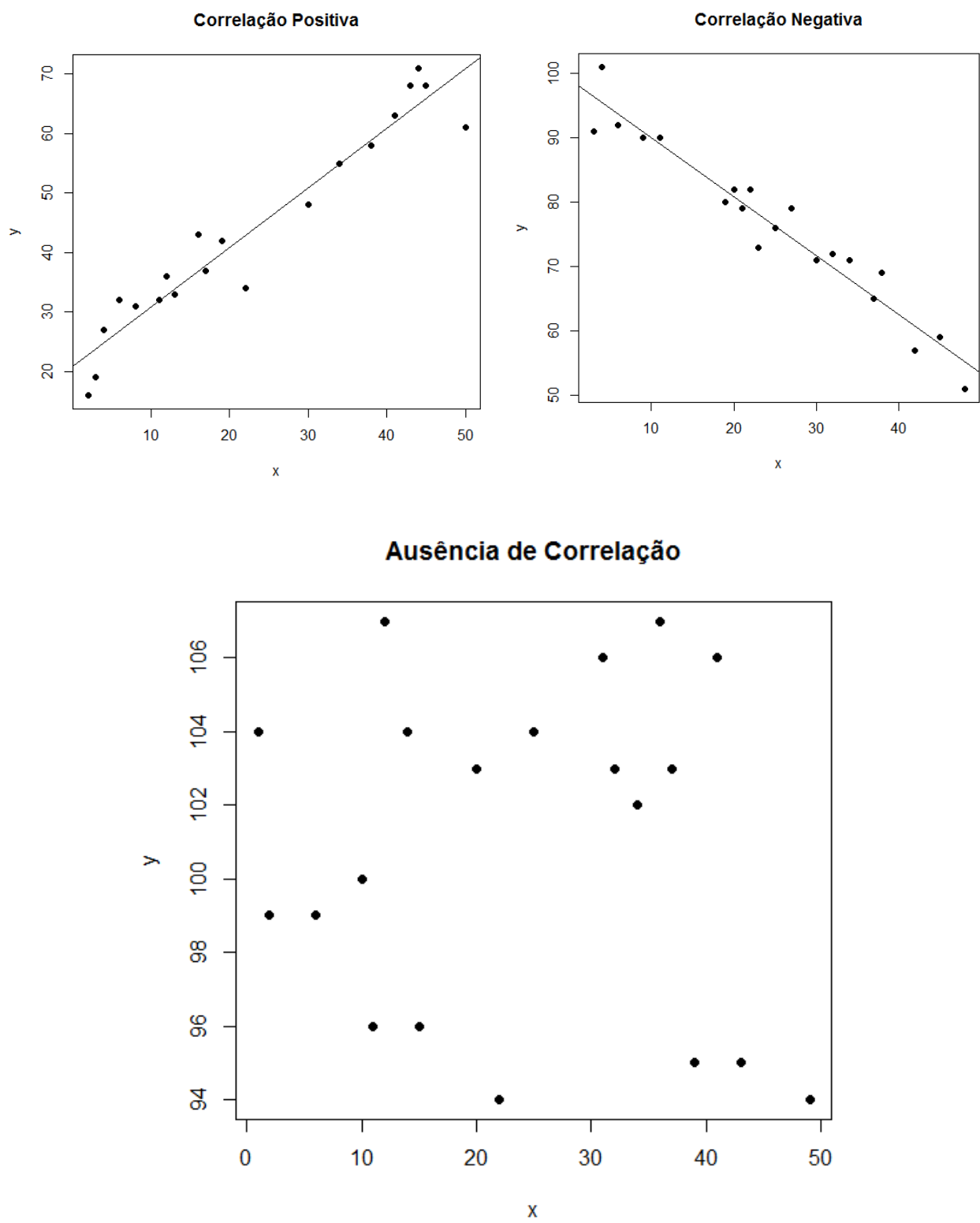
$r = -1$ Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.

$r = 0$ Significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma outra dependência que seja "não linear". Assim, o resultado $r=0$ deve ser investigado por outros meios.

Interpretação:

Coeficiente de correlação	Correlação
$r = 1$	Perfeita positiva
$0,8 \leq r < 1$	Forte positiva
$0,5 \leq r < 0,8$	Moderada positiva
$0,1 \leq r < 0,5$	Fraca positiva
$0 < r < 0,1$	Ínfima positiva
0	Nula
$-0,1 < r < 0$	Ínfima negativa
$-0,5 < r \leq -0,1$	Fraca negativa
$-0,8 < r \leq -0,5$	Moderada negativa
$-1 < r \leq -0,8$	Forte negativa
$r = -1$	Perfeita negativa

Graficamente:



Observação 1: Não se verificar correlação linear, **não significa que não se verifique outro tipo de correlação**, por exemplo, exponencial.

Observação 2: Qualquer que seja a correlação verificada, **correlação não significa causalidade**.

Para o cálculo do coeficiente de correlação, usamos a seguinte expressão:

$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2]} \sqrt{[n \sum Y^2 - (\sum Y)^2]}}$$

Onde n é o número de pares de dados amostrais.

Os cálculos ficam facilitados com o auxílio da tabela abaixo:

X	Y	XY	X ²	Y ²
...
...
...
ΣX	ΣY	ΣXY	ΣX²	ΣY²

16. Associação entre Variáveis Categóricas

Quando estamos estudando a relação entre duas variáveis categóricas, não usamos o termo "correlação". Neste caso, fala-se em “medida de associação”. Usa-se, então, o Coeficiente de Contigência C , dado por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Onde: $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ é o valor de Qui-quadrado, calculado a partir de uma tabela de dupla entrada.

$E_{ij} = \frac{\sum_{i=1}^r O_i \sum_{j=1}^c O_j}{n}$ é o nº esperado da linha i da coluna j

O_{ij} é o nº de observações classificados na linha i da coluna j

Estes cálculos são feitos a partir de uma tabela de dupla entrada abaixo:

Variável 1 (Linhas)	Variável 2(Colunas)				
	X1	X2	...	X _c	Total
Y1	O11	O12		O1 _c	ΣY1
Y2					ΣY2
...	
Y _r	O _{r1}	O _{r2}		O _{rc}	ΣY _r
Total	ΣX1	ΣX2		ΣX _c	N

Observações:

- para o caso 2x2 (gl=1), quando $N > 40$, utilizar no cálculo de χ^2 a correção de continuidade, ou seja:

$$\chi^2 = \frac{N \left(\left| AD - BC \right| - \frac{N}{2} \right)^2}{(A+B)(C+D)(A+C)(B+D)}$$

- quando $20 \leq N \leq 40$, a prova de χ^2 , pode ser empregada com a correção de continuidade, desde que nenhuma frequência esperada seja inferior a 5
- se a menor frequência esperada for inferior a 5, utilizar a prova de Fisher
- quando $N < 20$, utilizar a prova de Fisher em qualquer caso.

Para $gl > 1$ ($c > 2$ e $r > 2$), a prova pode ser aplicada somente se o número de células com frequência esperada inferior a 5 é inferior a 20% do total de células e se nenhuma célula tem frequência esperada inferior a 1. As frequências esperadas podem ser aumentadas combinando-se as categorias adjacentes.

Porém, o coeficiente descrito acima não varia entre 0 e 1. O valor máximo de C depende do número de linhas e colunas da tabela de dupla entrada. Para evitar este inconveniente, costuma-se empregar o Coeficiente de Contingência Modificado, dado por:

$$C' = \sqrt{\frac{(j\chi^2)}{[(j-1)(\chi^2 + n)]}}$$

Onde $j = \min(r, c)$, sendo "r" o número de linhas e "c" o número de colunas da tabela

O Coeficiente de Contingência Modificado satisfaz $0 \leq C' \leq 1$.

17 Modelos de Regressão

Conjunto de métodos e técnicas para o estabelecimento de fórmulas empíricas que interpretem a relação funcional entre variáveis com boa aproximação.

Deseja-se encontrar alguma forma de medir a relação entre as variáveis de cada conjunto, de tal modo que essa medida pudesse mostrar:

- a) se há relação entre as variáveis e, caso afirmativo, se é fraca ou forte;
- b) que, se essa relação existir, estabeleceremos um modelo que interprete a relação funcional existente entre as variáveis;
- c) que construindo o modelo, usá-lo-emos para fins de predição.

Suponhamos que Y seja uma variável que nos interessa estudar e prever o seu comportamento. É de se esperar que os valores da variável Y (dependente) sofram influências dos valores de um número infinito de variáveis X_1, X_2, \dots, X_N (independentes) e que exista uma função g que expresse tal dependência, ou seja

$$Y = g(X_1, X_2, \dots, X_N)$$

É impraticável a utilização das N variáveis ou por desconhecimento dos valores de algumas ou pela dificuldade de mensuração e tratamento de outras, logo se usa um número menor de variáveis (k) e o modelo fica

$$Y = f(X_1, X_2, \dots, X_k) + h(X_{k+1}, X_{k+2}, \dots, X_N)$$

Todas as influências das variáveis $X_{k+1}, X_{k+2}, \dots, X_N$, sobre as quais não exercemos controle, serão consideradas como casuais, e associaremos uma variável aleatória U , obtendo o seguinte modelo:

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

onde $f(X_1, X_2, \dots, X_k)$ é a componente funcional do modelo e ε a parte aleatória.

Problemas na análise dos modelos de regressão:

- o problema da especificação do modelo
Consiste em determinar qual o tipo de função f que melhor explique a relação entre Y e X_1, X_2, \dots, X_k
- o problema da estimação dos parâmetros
Consiste em estimar o valor dos diversos parâmetros que aparecem na especificação adotada.
- o problema da adaptação e significância do modelo adotado
Consiste em verificar se a especificação adotada na primeira etapa se adapta convenientemente aos dados observados.

17.1 Modelo de regressão linear simples

Quando a função f que relaciona X e Y é da seguinte forma:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

onde: - $\alpha + \beta X_i$ é a componente funcional, que representa a influência da variável independente X sobre o valor de Y e define o eixo da nuvem de pontos, que nesse caso será uma reta;

- ε_i é a componente aleatória, que representa a influência de outros fatores.

Sobre ε_i temos:

- tem distribuição Normal;
- é uma variável aleatória com média igual a 0 e variância igual a σ^2 , ou seja
 $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$, logo $\varepsilon_i \approx N(0; \sigma^2)$
- a $Cov(\varepsilon_i; \varepsilon_j) = \sigma^2$ para $i = j$ e $Cov(\varepsilon_i; \varepsilon_j) = 0$ para $i \neq j$

17.1.1 O modelo matemático

Seja $\hat{Y}_i = a + bX_i$ uma estimativa de $Y_i = \alpha + \beta X_i + U_i$, onde a e b são os estimadores de α e β e seja $e_i = (Y_i - \hat{Y}_i)$ o erro de estimação ou desvio.

Deseja-se minimizar a soma dos desvios ao quadrado, ou seja minimizar $\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$.

Chamando de S a soma dos desvios ao quadrado e substituindo

$$\hat{Y}_i$$

$$S = \sum e_i^2 = \sum [Y_i - (a + bX_i)]^2$$

Devemos encontrar os valores de a e b que minimizam S, ou seja, a soma dos desvios ao quadrado. Para isto, vamos utilizar o Método dos Mínimos Quadrados (MMQ). Mas para que isto aconteça, vamos calcular as derivadas parciais de S, em relação a a e b, e igualar estas derivadas parciais a zero, da seguinte forma:

$$\frac{\partial S}{\partial a} = 0 \Rightarrow -2 \sum_{i=1}^n (Y_i - a - bX_i) = 0$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow -2 \sum_{i=1}^n X_i (Y_i - a - bX_i) = 0$$

Estas expressões vão resultar no seguinte sistema:

$$\begin{cases} na + b \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \\ \left(\sum_{i=1}^n X_i \right) a + \left(\sum_{i=1}^n X_i^2 \right) b = \sum_{i=1}^n X_i Y_i \end{cases}$$

A primeira expressão do sistema resulta em: $a = \frac{\sum_{i=1}^n Y_i - b \sum_{i=1}^n X_i}{n}$

Substituindo na segunda, temos: $\left(\sum_{i=1}^n X_i \right) \left(\frac{\sum_{i=1}^n Y_i - b \sum_{i=1}^n X_i}{n} \right) + \left(\sum_{i=1}^n X_i^2 \right) b = \sum_{i=1}^n X_i Y_i$

Que nos dá: $b = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$

Retornando à primeira expressão:

$$a = \frac{\sum_{i=1}^n Y_i - b \sum_{i=1}^n X_i}{n} = \frac{\sum_{i=1}^n Y_i}{n} - \left(\frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \right) \frac{\sum_{i=1}^n X_i}{n}$$

Que resulta em:

$$a = \frac{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i Y_i \sum_{i=1}^n X_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$$

E:

$$b = \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$$

A qualidade do ajuste linear pode ser verificada pelo Coeficiente de Determinação R^2 , dado por:

$$R^2 = \frac{\sum_{i=1}^n (a + bX_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Substituindo no numerador $\bar{Y} = a + b\bar{X}$ e no denominador

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

temos:

$$R^2 = \frac{b^2 \left[n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 \right]}{n \sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i \right)^2}$$

O valor de R^2 varia de 0 a 1, e quanto mais próximo de 1, melhor é o ajuste.

Para a realização dos cálculos, os dados devem ser dispostos conforme a tabela abaixo:

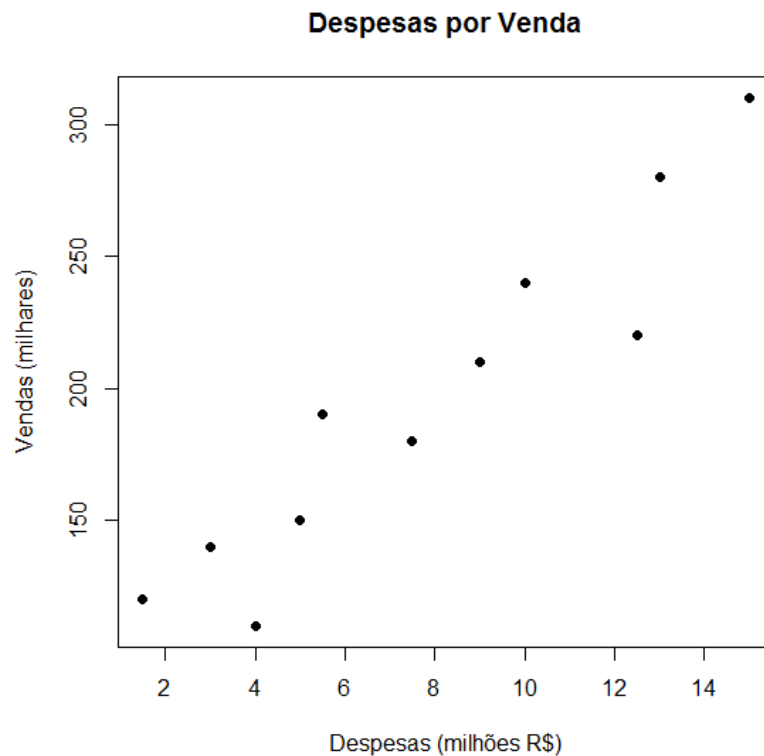
X	Y	X ²	XY	Y ²
ΣX	ΣY	ΣX ²	ΣXY	ΣY ²

Exemplo:

Suponha que exista uma relação linear entre as variáveis X = despesas com propaganda e Y = vendas de certo produto. Considerando os dados abaixo, determine a reta de mínimos quadrados, os testes e o coeficiente de explicação:

X (milhões de reais)	Y (milhares de unidades)
1,5	120
5,5	190
10,0	240
3,0	140
7,5	180
5,0	150
13,0	280
4,0	110
9,0	210
12,5	220
15,0	310

Graficamente, temos:



Primeiramente, devemos fazer o seguinte:

X (milhões de reais)	Y (milhares de unidades)	XY	X ²	Y ²
1,5	120	180	2,25	14400
5,5	190	1045	30,25	36100
10,0	240	2400	100	57600
3,0	140	420	9	19600
7,5	180	1350	56,25	32400
5,0	150	750	25	22500
13,0	280	3640	169	78400
4,0	110	440	16	12100
9,0	210	1890	81	44100
12,5	220	2750	156,25	48400
15,0	310	4650	225	96100
86	2150	19515	870	461700

Usando as fórmulas dadas, temos:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{2150}{11} = 195,45 \quad \bar{X} = \frac{\sum X}{n} = \frac{86}{11} = 7,82$$

$$S_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 19515 - \frac{86(2150)}{11} = 2705,91$$

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n} = 870 - \frac{(86)^2}{11} = 197,64$$

$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 461700 - \frac{(2150)^2}{11} = 41472,73$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{2705,91}{197,64} = 13,69$$

$$a = \bar{Y} - b\bar{X} = 195,45 - 13,69(7,82) = 88,39$$

Então, o modelo $\hat{Y}_i = a + bX_i$, fica $\hat{Y}_i = 88,39 + 13,69X_i$

O Coeficiente de explicação é dado por: $R^2 = \frac{VE}{VT} = \frac{bS_{xy}}{S_{yy}} = \frac{(13,69)(2705,91)}{41472,73} = 0,89$ ou 89%.

Este resultado indica que o modelo explica 89% da variação total de Y

Saída de um Pacote Estatístico - R

Call:

`lm(formula = dados$Y ~ dados$X)`

Residuals:

Min	1Q	Median	3Q	Max
-39.555	-8.984	10.513	14.136	26.284

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	88.413	14.027	6.303	0.00014 ***
dados\$X	13.691	1.577	8.680	1.15e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.17 on 9 degrees of freedom

Multiple R-squared: 0.8933, Adjusted R-squared: 0.8814

F-statistic: 75.35 on 1 and 9 DF, p-value: 1.147e-05

18. Determinação do Tamanho da Amostra

Para determinar o tamanho da amostra, devemos saber qual é a dimensão da população que servirá de base para o estudo, ou seja, o valor de N.

Uma população é dita finita quando se consegue enumerar todos os elementos que a formam.

Refere-se a um universo limitado em uma dada unidade de tempo. Exemplificando pode-se dizer que a quantidade de automóveis produzidos por uma fábrica em um mês, a população de uma cidade e o número de alunos de uma sala de aula são exemplos de uma população finita.

Uma população é dita infinita quando os elementos não podem ser contados. Refere-se a um universo não delimitado. Os resultados (cara ou coroa) obtidos em sucessivos lances de uma moeda, o conjunto dos números inteiros, reais ou naturais são exemplos de populações infinitas.

Então, temos o seguinte:

Para média:

População Finita

$$n = \frac{Z^2 \cdot \sigma^2 \cdot N}{\varepsilon^2 (N-1) + Z^2 \sigma^2}$$

População Infinita

$$n = \left(\frac{Z \cdot \sigma}{\varepsilon} \right)^2$$

Para a proporção:

População Finita

$$n = \frac{Z^2 \cdot P \cdot Q \cdot N}{\varepsilon^2 (N-1) + Z^2 \cdot P \cdot Q}$$

População Infinita

$$n = \frac{Z^2 \cdot P \cdot Q}{\varepsilon^2}$$

Onde:

Z = abscissa da distribuição normal padrão, fixado um nível de $(1 - \alpha)\%$ de confiança para a construção de um intervalo de confiança; Z pode assumir os seguintes valores:

Se o nível for de 95,5%, $Z = 2$

Se o nível for de 95%, $Z = 1,96$

Se o nível for de 99%, $Z = 2,57$

σ = desvio padrão da população; quando não sabemos este valor, substituímos por s, ou seja, o desvio padrão amostral

ε = é o erro amostral admitido

N = tamanho da população

P = proporção populacional; quando não sabemos este valor, substituímos por p, ou seja, o valor da proporção amostral

Q = 1 – P; quando não temos este valor, substituímos por q = 1 – p

Quando não se conhecem os valores populacionais σ^2 , P e Q, utilizam-se os valores amostrais s^2 , p e q, nas fórmulas acima.

Bibliografia:

Bussab, Wilton de O., Morettin, Pedro A. Estatística Básica. 8. Ed. São Paulo: Saraiva, 2013.

Morettin, Luiz Gonzaga. Estatística Básica: Probabilidade e Inferência. Volume único. São Paulo: Ed. Pearson, 2011.

Belfiore, Patrícia, Estatística Aplicada a Administração, Contabilidade e Economia com Excel e SPSS. 1. Ed. Rio de Janeiro: Elsevier, 2015.

Pinheiro, João Ismael D. et al. Estatística Básica: a arte de trabalhar com dados. 2. Ed. Rio de Janeiro: Elsevier, 2015

Martins, Gilberto de Andrade. Estatística Geral e Aplicada. 3. Ed. São Paulo: Atlas, 2008.