Capítulo 2

Estatística Descritiva

"O estatístico, está casado em média com 1,75 esposas, que procuram fazê-lo sair de casa 2,25 noites com 0,5 de sucesso apenas. Possui fronte com 0,02 de inclinação (denotando poder mental), 5/8 de uma conta bancária, e 3,06 filhos; 1,65 dos filhos são do sexo masculino. Apenas 0,07 por cento de todos os estatísticos estão acordados à mesa do café, onde consomem 1,68 xícaras de café (o restante derrama-se em suas camisas). Nas noites de sábado, ele contrata 1/3 de uma baby-sitter para cuidar de suas 3,06 crianças, a não ser que ele tenha 5/8 de uma sogra que more com ele e que faça o serviço pela metade do preço..."

F. Miksch (1950)

Após a definição do problema a ser estudado e o estabelecimento do planejamento da pesquisa (forma pela qual os dados serão coletados, cronograma das atividades, custos envolvidos, exames das informações disponíveis, delineamento da amostra, etc.), o passo seguinte é a **coleta de dados**, que consiste na busca ou aplicação dos dados das variáveis, componentes do fenômeno a ser estudado.

A coleta de dados é **direta** quando os dados são obtidos na fonte originária. Os valores assim compilados são chamados de dados primários, como, por exemplo, dados obtidos em pesquisas de opinião pública, vendas registradas em notas fiscais da empresa, medição de chuva em pluviômetros, contagem do número de carros que passa por dia em um cruzamento, etc.

A coleta de dados é **indireta** quando os dados obtidos provêm de coleta direta. Os valores assim complicados são denominados de dados secundários, como, por exemplo, o cálculo do tempo de vida média, obtido pela pesquisa, nas tabelas demográficas publicadas

pela Fundação Instituto Brasileiro de Geografia e Estatística – IBGE, utilização de dados hidroclimatológicos publicados pela Funceme, dados de vazão publicados pelo DNAEE, etc.

Quando ao tempo, a coleta pode ser classificada em:

- **Contínua:** quando realiza permanentemente,
- **Periódica:** quando é feita em intervalos de tempo, e
- **Ocasional:** quando efetuada sem época preestabelecida.

Objetivando a eliminação de erros capazes de provocar futuros enganos de apresentação e análise, procede-se a uma revisão crítica dos dados, suprimindo os valores estranhos ao levantamento.

Após a crítica, convém organizar os dados de maneira prática e racional, para o melhor entendimento do fenômeno que se está estudando. As etapas seguidas pela Estatística Descritiva pode então pode ser resumida no diagrama da Figura 2.1.

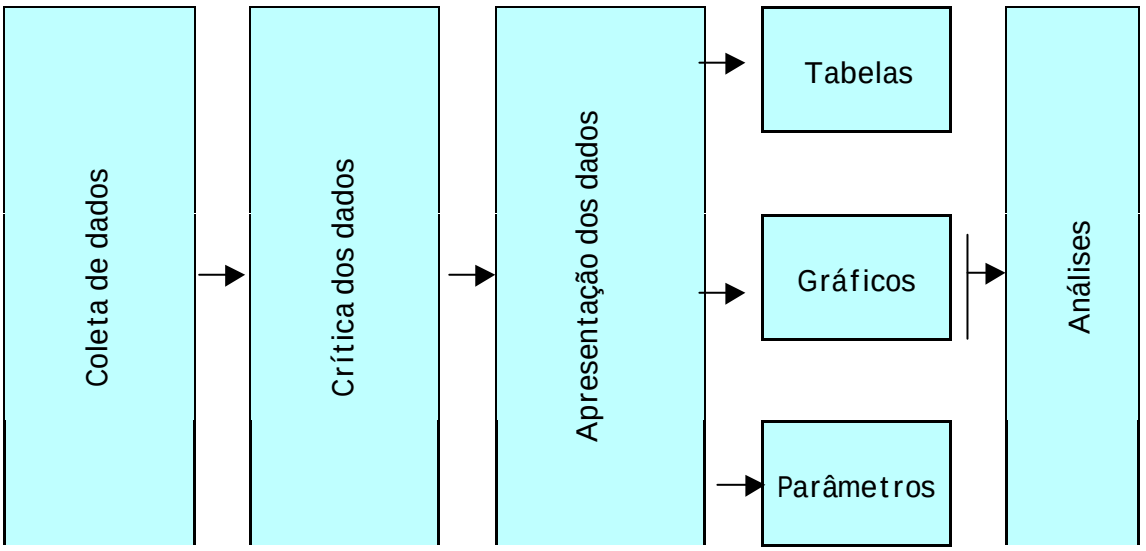


Figura 2.1. Etapas da Estatística Descritiva

1. APRESENTAÇÃO DOS DADOS

Após a crítica, convém organizar os dados de maneira prática e racional, para o melhor entendimento do fenômeno que se está estudando. Só assim os dados podem ser transformados de meros "dados" em "informação. Com os recursos da Estatística

Descritiva, pode-se compreender melhor um conjunto de dados através de suas características. A Estatística Descritiva pode extrair e apresentar a informação contida nos dados coletados apresentando-os de três formas, utilizando a representação **tabular**, a representação através de **parâmetros** – que descreve certas as características numéricas dos dados e a representação **gráfica**, que possibilita uma rápida visão geral do fenômeno estudado.

1.1. TABELAS

Uma tabela deve apresentar cabeçalho, corpo e rodapé. O **cabeçalho** deve conter informação suficiente para que sejam respondidas as seguintes questões:

- **O quê ?** (referente ao fato);
- **Onde ?** (relativo ao lugar);
- **Quando ?** (correspondente à época).

O **corpo** é representado por colunas e sub-colunas dentro das quais serão registrados os dados. O **rodapé** é reservado para as observações pertinentes, bem como a identificação da fonte dos dados.

Conforme o critério de agrupamento, as séries classificam-se em:

a) Série Cronológica, Temporal, Evolutiva ou Histórica. É a série estatística em que os dados são observados segundo a época de ocorrência. Exemplo:

Tabela 2.1. Precipitação anual sobre a Bacia do Rio do Carmo - RN (1911 - 1920)

Ano	Precipitação (mm)
1911	548,0
1912	991,1
1913	824,4
1914	1030,9
1915	155,3
1916	702,1
1917	1093,0
1918	709,4
1919	139,4
1920	661,7

Fonte: Sudene

b) Série Geográfica ou de Localização. É a série estatística em que os dados são observados segundo a localidade de ocorrência. Exemplo:

Tabela 2.2. Chuva observada no Ceará relativa ao mês de maio de 2000

Região	Precipitação (mm)
Litoral Norte	70,8
Litoral do Pecém	96,9
Litoral de Fortaleza	148,1
Maciço de Baturité	148,0
Ibiapaba	29,1
Jaguaribara	76,7
Cariri	28,3
Sertão Central e Inhamuns	43,5

Fonte: Funceme

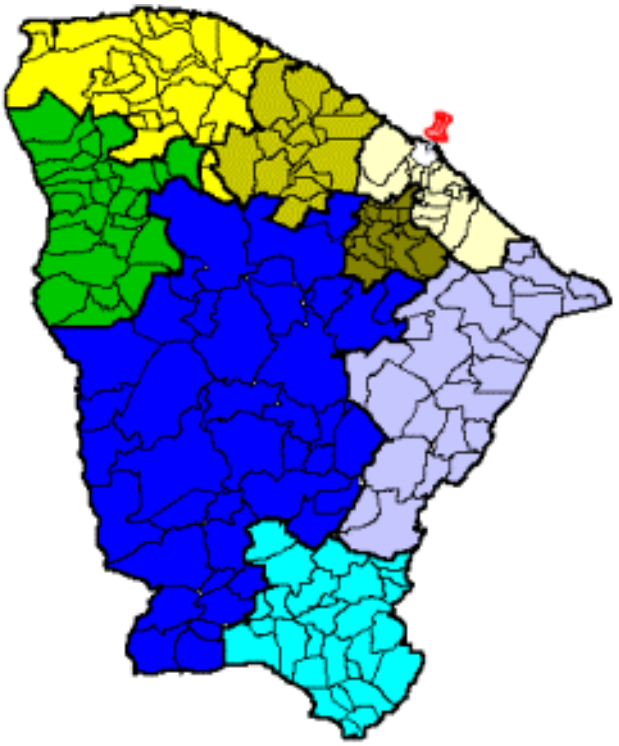


Figura 2.2. Macroregiões climáticas analisadas pela Funceme (Fonte: Funceme)

c) Série Específica. É a série estatística em que os dados são agrupados segundo a modalidade de ocorrência.

Tabela 2.3. Número de instrumentos encontrados nas estações meteorológicas do Nordeste

Instrumento	Quantidade
Pluviógrafo	50
Pluviômetro	500
Tanque Classe A	0
Heliógrafo	33
Psicrômetro	40

Fonte: dados fictícios.

1.1.1. DISTRIBUIÇÃO DE FREQÜÊNCIAS

É a série estatística em que os dados são agrupados com suas respectivas freqüências absolutas. Por constituir-se no tipo de tabela mais importante para a Estatística Descritiva, far-se-á um estudo completo das distribuições de freqüências. A seguir são apresentados os conceitos e os procedimentos usuais na construção dessas tabelas.

Representação da amostra

Como citado anteriormente, a Estatística tem como objetivo encontrar leis de comportamento para todo o conjunto, por meio da sintetização dos dados numéricos, sob a forma de tabelas, gráficos e medidas. A seguir, estão alguns procedimentos comuns para a representação das distribuições de freqüências, que é uma das maneiras de sumarizar os valores de uma variável discreta ou contínua, objeto de estudo.

Dados brutos

O conjunto dos dados numéricos obtidos após a crítica dos valores coletados constitui-se nos dados brutos. Ex:

24 – 23 – 22 – 28 – 35 – 21 – 23 – 33 – 34 – 24 – 21 – 25 – 36 – 26 – 22
30 – 32 – 25 – 26 – 33 – 34 – 21 – 31 – 25 – 31 – 26 – 25 – 35 – 33 – 31

Rol

É o arranjo dos dados brutos em ordem de freqüências crescente ou decrescente.

21 – 21 – 21 – 22 – 22 – 23 – 23 – 24 – 24 - 25 – 25 – 25 – 26 – 26 – 26
28 -30 – 31 – 31 – 31 – 32 – 33 – 33 – 33 – 34 – 34 – 34 – 35 – 35 – 36

Amplitude total ou "range" (R)

É a diferença entre o maior e o menor valor observados. No exemplo anterior,

$R = 36 - 21 = 15.$

Número de classes (K)

De modo a interpretar melhor o que esses números exprimem, intervalos (classes) devem ser criados, preferencialmente, igualmente espaçados. O número deles depende do número de observações (n) e o quão dispersos os dados estão. Não há uma fórmula exata para o cálculo do número de classes. Apresenta-se, a seguir, duas soluções.

- $K = 5$ para $n \leq 25$ e $K = \sqrt{n}$, para $n > 25$.
- Fórmula de Sturges: $K \cong 1 + 3{,}22 \log n$, em que n = tamanho da amostra.

Exemplo: Para n = 49, tem-se que:

$K = \sqrt{49} = 7$ ou $K \cong 1 + 3{,}22 \log 49 \cong 7$.

Amplitude das classes (h)

A especificação da largura do intervalo é uma consideração importante. Intervalos muito grandes resultam em menos classes de intervalo. A amplitude das classes é dada pela relação:

$h \cong R : K$

Assim como no caso do número de classes (K), a amplitude das classes (h) deve ser aproximada para o maior inteiro. Assim, se K ≅ 6,4, arredonda-se para K = 7 ou, se h ≅ 1,7, arredonda-se para h = 2.

Limites das classes

Existem diversas maneiras de expressar os limites das classes. Eis algumas:

a) 20 ⊢⊣ 23: compreende todos os valores entre 20 e 23;
b) 20 ⊢ 23: compreende todos os valores entre 20 e 23, excluindo o 23;
c) 20 ⊣ 23: compreende todos os valores entre 20 e 23; excluindo o 20;

Neste texto, usar-se-á a forma expressa no exemplo b.

Pontos médios das classes (x_i)

É a média aritmética entre o limite superior e o limite inferior da classe. Assim, se a classe for 20 ⊢ 23, tem-se: $x_i = \frac{20 + 23}{2} = 11$, como ponto médio da classe.

Freqüência absoluta (F_i)

É o número de vezes que o elemento aparece na amostra, ou o número de elementos pertencentes a uma classe. No exemplo, F(21)= 3.

Freqüência absoluta acumulada (F_{ac})

É a soma das freqüências dos valores inferiores ou iguais ao valor dado.

Freqüência relativa (f_i)

É a porcentagem daquele valor na amostra. Note que $\Sigma f_i = 1$.A freqüência relativa de um valor é dada por:

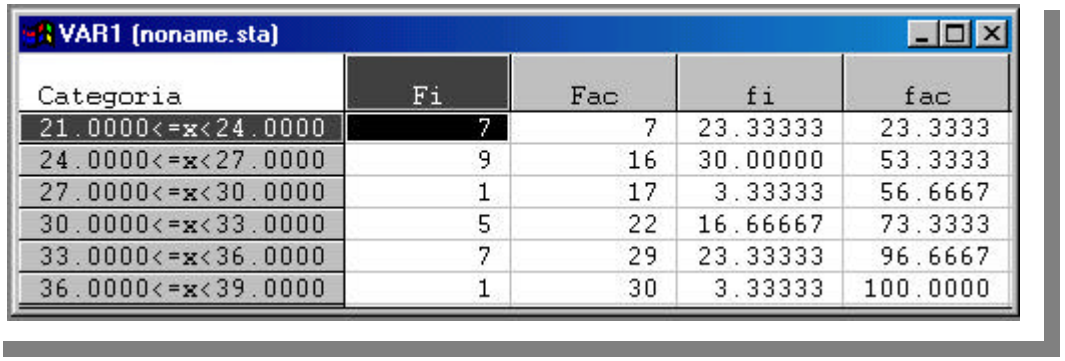
$$f_i = \frac{F_i}{n}$$

Na Tabela 2.4 a coluna da freqüência absoluta (F_i) representa o número de ocorrências de dias chuvosos durante os meses março/abril, no período de 1969 a 1998, em cada respectivo intervalo. A freqüência relativa (f_i) é a informação mais importante, pois

independe do número da amostra. Usando-se o software STATISTICA, obtém-se o resultado observado na Figura 2.3

Tabela 2.4. Número de dias chuvosos no período Março/Abril em Fortaleza (1969 – 1998).

Número de dias chuvosos	F_i	F_{ac}	f_i	f_{ac}
$21 \leq x \leq 24$	7	7	7/30	7/30
$24 \leq x \leq 27$	9	16	9/30	16/30
$27 \leq x \leq 30$	1	17	1/30	17/30
$30 \leq x \leq 33$	5	22	5/30	22/30
$33 \leq x \leq 36$	7	29	7/30	29/30
$36 \leq x \leq 39$	1	30	1/30	30/30
Σ	30	-	1	-



Categoria	Fi	Fac	fi	fac
21.0000<=x<24.0000	7	7	23.33333	23.3333
24.0000<=x<27.0000	9	16	30.00000	53.3333
27.0000<=x<30.0000	1	17	3.33333	56.6667
30.0000<=x<33.0000	5	22	16.66667	73.3333
33.0000<=x<36.0000	7	29	23.33333	96.6667
36.0000<=x<39.0000	1	30	3.33333	100.0000

Figura 2.3. Tabela de freqüência utilizando-se o software STATISTICA

1.2. PARÂMETROS

Já vimos como sintetizar uma amostra sob a forma de tabelas e distribuições de freqüências. A amostra, no entanto, também pode ser representada por parâmetros, que podem ser as medidas de tendência central, de dispersão, de assimetria e de curtose.

1.2.1. MEDIDAS DE POSIÇÃO

Tais medidas orientam-nos quanto à posição da distribuição no eixo x (eixo dos números reais), possibilitando que comparemos séries de dados entre si pelo confronto desses números. São chamadas de medidas da tendência central, pois representam os fenômenos pelo seus valores médios, em torno dos quais tendem a concentrar-se os dados.

a) Média Aritmética – Dados Não Agrupados

Sejam $x_1, x_2, x_3, ..., x_n$, portanto, "n" valores da variável X. A média aritmética simples de X representada por $\bar{x}$ é definida por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \text{ ou simplesmente } \boxed{\bar{x} = \frac{\sum x}{n}}$$

em que "n" é o número de elementos do conjunto.

Exercício 2.1

Determinar a média aritmética simples dos valores: 3, 7, 8, 10 e 11.

$$\bar{x} = \frac{\sum x}{n} = \frac{3 + 7 + 8 + 10 + 11}{5}$$

$$\bar{x} = 7{,}8$$

b) Média Aritmética – Dados Agrupados

Quando os dados estiverem agrupados numa distribuição de freqüência usaremos a média dos valores $x_1, x_2, ..., x_n$ - pontos médios de cada classe - ponderados pelas respectivas freqüências absolutas: $F_1, F_2, F_3, ..., F_n$. Assim:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i F_i}{n} \quad \text{ou} \quad \boxed{\bar{x} = \frac{\sum x_i F_i}{n}}$$

Exercício 2.2

Seja a distribuição de freqüências abaixo. Calcular a média aritmética.

Tabela 2.5. Número de dias chuvosos no período Março/Abril em Fortaleza no período 1969 - 1999.

Número de dias chuvosos	F_i	x_i	x_iF_i
$20 \le x \le 23$	5	21,5	107,5
$23 \le x \le 26$	8	24,5	196,0
$26 \le x \le 29$	4	27,5	110,0
$29 \le x \le 32$	4	30,5	122,0
$32 \le x \le 35$	6	33,5	201,0
$35 \le x \le 38$	3	36,5	109,5
Σ	30		846

$\bar{x}$ = 846/30 = 28,2

c) Mediana ($\tilde{x}$)

Colocados os valores em ordem crescente, Mediana é o elemento que ocupa a posição central.

Vamos considerar, em primeiro lugar, a determinação da mediana para o caso de variável **discreta**, isto é, para distribuição de freqüência simples. Neste caso precisamos considerar duas situações: para "n" (número de elementos da amostra) ímpar e para "n" par.:

Caso 1: Se n for ímpar, a mediana será o elemento central (de ordem $\frac{n+1}{2}$).

Exercício 2.3

Calcular a mediana da amostra

5 7 8 10 14

$\tilde{x}$ = 8, pois é o elemento central, ou seja , o de ordem 3.

Caso 2: Se n par, a mediana será a média entre os elementos centrais (de ordens $\frac{n}{2}$ e $\frac{n}{2}+1$).

Exercício 2.4

Calcular a mediana da amostra:

5 7 8 10 14 15

$\tilde{x}$ = 9, pois a média dos elementos centrais 8 (ordem 3) e 10 (ordem 4) é igual a 9.

Vamos considerar, agora, a determinação da mediana para o caso de variável contínua (dados agrupados em classes). O procedimento a ser seguido é:

1º Passo: Calcula-se a ordem $\frac{n}{2}$ .Como a variável é contínua, não se preocupe se n é par ou ímpar.

2º Passo: Pela F_{ac} identifica-se a classe que contém a mediana (classe M_d).

3º Passo: Utiliza-se a fórmula:

$$\tilde{x} = \ell_{MD} + \frac{\left(\frac{n}{2} - \sum F\right).h}{F_{MD}}$$

onde:

ℓ_{MD} = limite inferior de classe Md.

n = tamanho da amostra ou número de elementos.

$\sum F$ = soma das freqüências anteriores à classe Md.

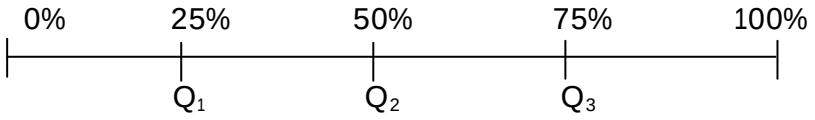
h = amplitude de classe Md.

F_{MD} = freqüência da classe Md.

---

d) Quartis

Os quartis dividem um conjunto de dados em quatro partes iguais. Assim:



Q_1 = 1º quartil, deixa abaixo e si 25% dos elementos .

Q_2 = 2º quartil, coincide com a mediana, deixa abaixo de si, 50% dos elementos.

Q_3 = 3º quartil, deixa abaixo de si, 75% dos elementos.

Utilizamos os quartis apenas para dados agrupados em classes. As fórmulas para a determinação dos quartis Q_1 e Q_3 são semelhantes utilizada para o cálculo da Md.

Determinação de Q_1:

1º Passo: Calcula-se n/4.

2º Passo: Identifica-se a classe Q_1 pela Fac.

3º Passo: Aplica-se a fórmula:

$$Q_1 = \ell_{Q_1} + \frac{\left(\frac{n}{4} - \sum F\right).h}{F_{Q_1}}$$

Determinação de Q_3:

1º Passo: Calcula-se 3n/4.

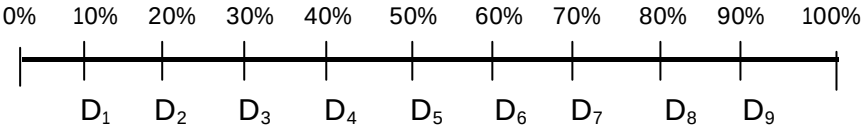
2º Passo: Identifica-se a classe Q_3 pela Fac.

3º Passo: Aplica-se a fórmula:

$$Q_3 = \ell_{Q_3} + \frac{\left(\frac{3n}{4} - \sum F\right).h}{F_{Q_3}}$$

e) Decis

Continuando o estudo das medidas separatrizes mediana e quartis, temos os decis. São os valores que dividem a série em 10 partes iguais.



Como você já deve ter percebido, a fórmula neste caso também é semelhante às separatrizes anteriores. Ei-la:

1º Passo: Calcula-se $\frac{i \cdot n}{10}$, em que i = 1, 2, 3, 4, 5, 6, 7, 8 e 9.

2º Passo: Identifica-se a classe D_i pela Fac.

3º Passo: Aplica-se a fórmula:

$$D_i = \ell_{D_i} + \frac{\left(\frac{in}{10} - \sum F\right).h}{F_{D_i}}$$

em que :

ℓ_{D_i} = limite inferior da classe D_i, i = 1, 2, 3, ..., 9

n = tamanho da amostra

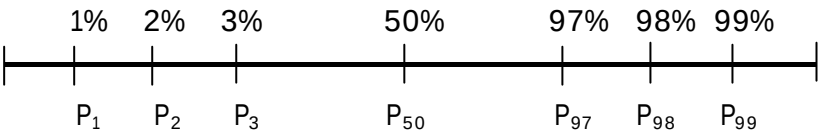
h = amplitude da classe

F_{D_i} = freqüência da classe D_i

$\sum F$ = soma das freqüências anteriores à classe D_i

e) Percentis

São as medidas que dividem a amostra em 100 artes iguais. Assim:



Seu cálculo é dado por:

1º Passo: Calcula-se $\frac{in}{100}$, em que i = 1, 2, 3, ..., 98, 99.

2º Passo: Pela Fac identifica-se a classe P_i .

3º Passo: Usa-se a fórmula:

$$P_i = \ell_{P_i} + \frac{\left(\frac{in}{100} - \sum F\right).h}{F_{P_i}}$$

em que:

ℓ_{P_i} = limite da classe P_i , em que i = 1, 2, 3, ..., 99

n = tamanho da amostra

$\sum F$ = soma das freqüências anteriores à classe P_i

h = amplitude da classe

F_{P_i} = freqüência da classe P_i

Para a determinação das separatrizes (mediana, quartis, decis e percentis) pode-se utilizar o gráfico da freqüência acumulada. Assim:

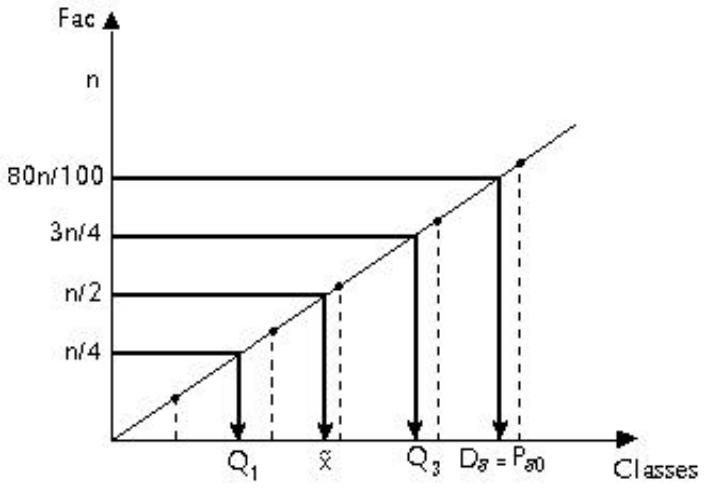


Figura 2.4. Freqüências acumuladas

f) Moda

Dentre as principais medidas de posição, destaca-se a Moda. É o valor mais freqüente da distribuição. Para distribuições simples (sem agrupamento em classes) a identificação da Moda é facilitada pela simples observação do elemento que apresenta maior freqüência. Assim, para a distribuição abaixo a Moda será 248. Indica-se Mo = 248. Notem que esse número é o mais comum nesta distribuição (aparece mais vezes).

Tabela 2.6. Dados não agrupados

xi	243	245	248	251	307
Fi	7	17	23	20	8

Para dados agrupados em classe, temos diversas fórmulas para o cálculo da Moda. Apresentaremos dois processos:

1º processo: fórmula de Czuber

1º Passo: Identifica-se a classe modal (aquela que possuir maior freqüência).

2º Passo: Aplica-se a fórmula:

$$Mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2}.h$$

em que:

ℓ = limite inferior da classe modal.

Δ_1= diferença entre a freqüência da classe modal e a imediatamente anterior.

Δ_2= diferença entre a freqüência da classe modal e a imediatamente posterior.

h = amplitude da classe.

2º processo: determinação gráfica da moda

É preciso construir o histograma da distribuição, identificar a classe modal (aquela com maior altura) e fazer a construção indicada abaixo. Assim:

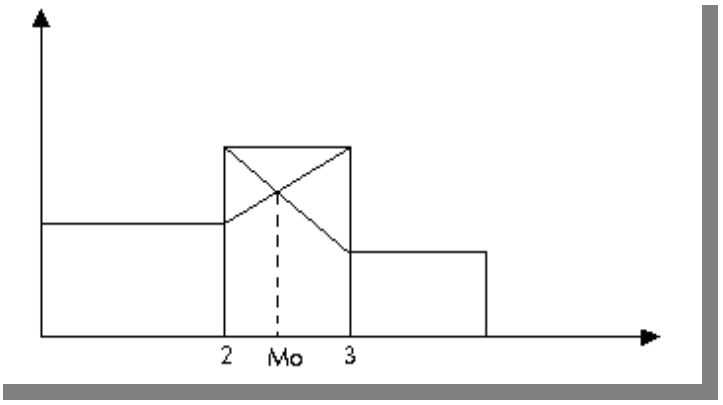


Figura 2.5. Método gráfico para o cálculo da Moda

g) Relação entre média, mediana e moda

Em uma distribuição simétrica, observa-se que a Média = Mediana = Moda (Figura 2.6).

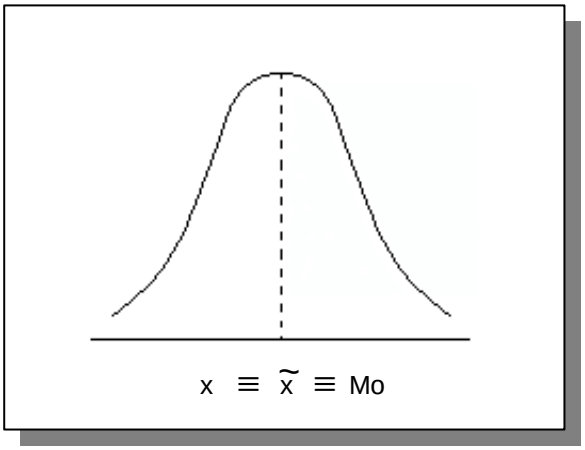


Figura 2.5. Relação entre média, mediana e moda em uma distribuição simétrica

Em uma distribuição com assimetria negativa, observa-se que a Média < Mediana < Moda (Figura 2.7 a). Em uma distribuição assimétrica positiva, observa-se que a Média > Mediana > Moda (Figura 2.7 b).

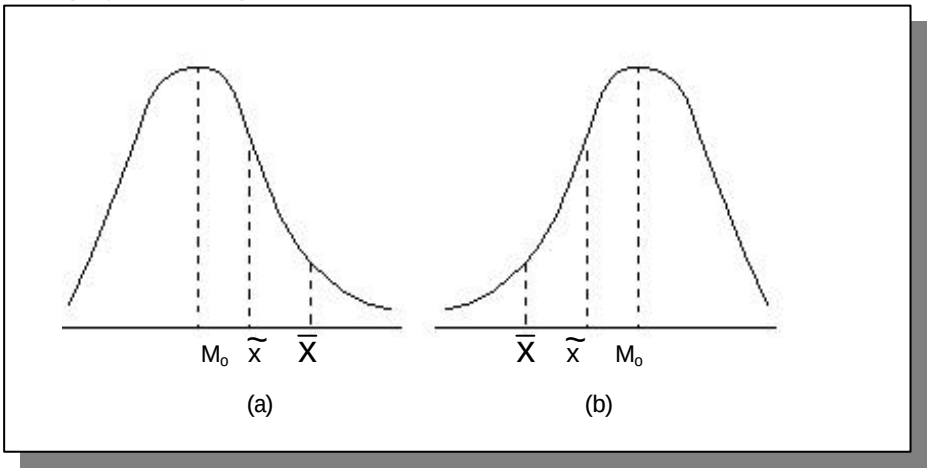


Figura 2.7. Relação entre média, mediana e moda em distribuições assimétrica

1.2.2. MEDIDAS DE DISPERSÃO

Servem para medir o grau de dispersão dos valores individuais em torno da média, ou ainda, para verificar o grau de representatividade da média. Sejam as duas séries a seguir:

a) 20, 20, 20, 20, 20

b) 15, 10, 20, 25, 30

Para ambas as séries temos média igual a 20. Nota-se, entretanto, que os valores da série "a" se concentram totalmente na média 20, enquanto os valores da série "b" se

dispersam em torno do mesmo valor. Ou seja, a série "a" não apresenta dispersão e os valores da série "b" estão dispersos em torno de 20. Entre as medidas de dispersão, destacam-se a amplitude total, a variância, o desvio-padrão e o coeficiente de variação.

a) Amplitude total (ou Range)

É a diferença entre o maior e menor dos valores da série. Ou seja:

$$R = x_{max} - x_{mín}$$

A utilização da amplitude total como medida de dispersão é muito limitada, pois é uma medida que depende apenas dos valores extremos, não sendo afetada pela variabilidade interna dos valores da série. Sejam as duas series a seguir:

a) 1, 1, 1, 1, 1, 100

b) 1, 30, 32, 45, 75, 100

Ambos tem R = 100 - 1= 99 !!!

b) Desvio Médio (D_M)

Considerando nosso propósito de medir a dispersão ou o grau de variabilidade dos valores em torno da média, nada mais interessante do que estudarmos o comportamento dos desvios de cada valor individual da série em relação à media, ou seja, o desvio individual, dado por:

$$d_i = (x_i - \bar{x})$$

Entretanto, por uma das propriedades da média tem-se que $\Sigma (x_i - \bar{x}) = 0$. Temos então que solucionar o problema: queremos calcular a média dos desvios, porém sua soma é nula.

O **Desvio Médio** considera o módulo de cada desvio (xi - $\overline{x}$), evitando com isso que Σ di = 0. Assim sendo, o Desvio Médio é dado por:

$$D_M = \frac{\sum |x_i - \overline{x}| . F_i}{n} = \frac{\sum |d_i| . F_i}{n}$$

Trata-se, pois, da média aritmética dos desvios individuais considerados em módulos (valor absoluto).

c) Variância

A Variância também procura solucionar o problema de Σ di = 0. Para isso considera o quadrado de cada desvio $(x_i - \overline{x})^2$, evitando com isso que o somatório seja nulo. Assim, a variância é dada por:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 . F_i}{n} = \frac{\sum d_i^2 . F_i}{n}$$

Trata-se, pois, da média aritmética dos quadrados dos desvios. O símbolo σ^2 indica variância e lê-se "sigma ao quadrado" e $\overline{x}$ é a média da população. Para o caso do cálculo da variância de valores amostrais é conveniente usarmos a seguinte fórmula:

$$S^2 = \frac{\sum (x_i - \overline{x})^2 . F_i}{n - 1}$$

Como você deve ter notado, as diferenças entre as fórmulas são: para o caso da variância populacional s^2 tendo como denominador o tamanho da amostra (n). Para o cálculo da variância amostral (S^2), utiliza-se a média amostral $(\overline{x})$, tendo como denominador o tamanho da amostra menos um (n - 1). Para o cálculo da variância, é mais interessante o uso das seguintes fórmulas:

$$ó^2 = \frac{1}{n}\left[\sum x_i^2 F_i - \frac{\left(\sum x_i F_i\right)^2}{n}\right]$$

$$S^2 = \frac{1}{n-1}\left[\sum x_i^2 F_i - \frac{\left(\sum x_i F_i\right)^2}{n}\right]$$

que são obtidas por transformações nas respectivas fórmulas originais.

d) Desvio-padrão

Observando a fórmula para o cálculo da variância, notamos tratar-se de uma soma de quadrados. Dessa forma, a unidade de variável for, por exemplo, variável original, necessitamos definir outra medida de dispersão, que é a raiz quadrada da variância – o desvio-padrão. Assim:

$\sigma = \sqrt{\sigma^2}$ é o desvio-padrão populacional

$S = \sqrt{S^2}$ é o desvio-padrão amostral

Resumindo: para o cálculo do desvio-padrão deve-se primeiramente determinar o valor da variância e, em seguida, extrair a raiz quadrada desse resultado.

e) Coeficiente de variação

Trata-se de uma medida relativa de dispersão, útil para a comparação em termos relativos do grau de concentração em torno da média de séries distintas. É dado por:

$$C.V = \frac{\sigma}{\overline{x}} \quad \text{ou} \quad C.V = \frac{S}{\overline{x}}$$

1.2.3. MEDIDAS DE ASSIMETRIA

Já foi observado que, em uma distribuição simétrica, a média, a moda e a mediana coincidem e que os quartis ficam eqüidistantes da mediana, o que não ocorre numa distribuição assimétrica.

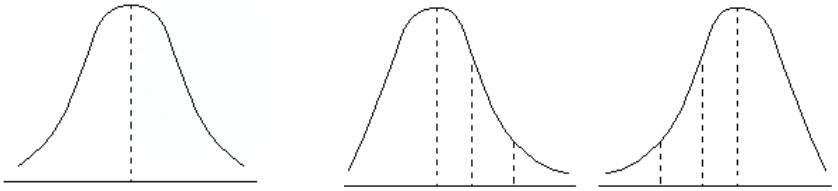


Figura 2.8. Distribuições simétrica, assimétrica negativa e assimétrica positiva

O grau de assimetria de uma distribuição é medido pelo **coeficiente de assimetria.**

a. Primeiro Coeficiente de Pearson - quando se dispõe de valores da média e do desvio-padrão.

$$A_S = \frac{\bar{x} - Mo}{S} \quad \text{ou} \quad A_S = \frac{\bar{x} - Mo}{\sigma}$$

Quando não se tem condições de calcular a média e o desvio-padrão utiliza-se:

$$\bar{x} - Mo = 3(\bar{x} - Md)$$

$$A_S = \frac{3(\bar{x} - Md)}{\sigma}$$

a.2. Segundo Coeficiente de Pearson

$$A_S = \frac{Q_3 + Q_1 - 2\tilde{x}}{Q_3 - Q_1}$$

Se $A_S = 0$ a distribuição é simétrica.

Se $A_S > 0$ a distribuição é assimétrica positiva.

Se $A_S < 0$ a distribuição é assimétrica negativa.

1.2.4. MEDIDAS DE CURTOSE

Entende-se por curtose o grau de achatamento de uma distribuição. Com referência do grau de achatamento, pode-se ter:

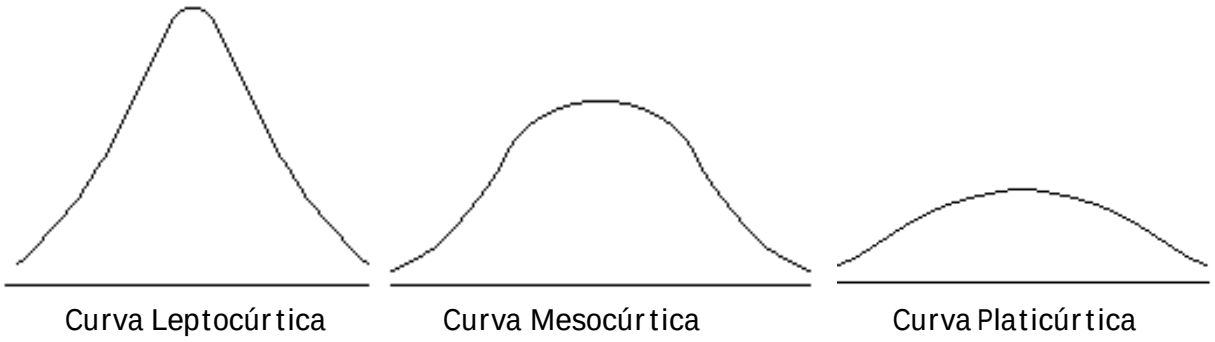


Figura 2.9. Distribuições leptocúrtica, mesocúrtica e platicúrtica

Para medir o grau de achatamento da distribuição utiliza-se o coeficiente de curtose:

$$K = \frac{Q_3 - Q_1}{2\left(P_{90} - P_{10}\right)}$$

1.3. GRÁFICOS

A representação gráfica das séries estatísticas tem por finalidade dar uma idéia, a mais imediata possível, dos resultados obtidos, permitindo chegar-se a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores da série. Não há apenas uma maneira de representar graficamente uma série estatística. A escolha do gráfico mais apropriado ficará a critério do analista. Contudo, os elementos simplicidade, clareza e veracidade devem ser considerados quando da elaboração de um gráfico. Encontram-se a seguir os principais tipos de gráficos.

a) Gráfico em Colunas

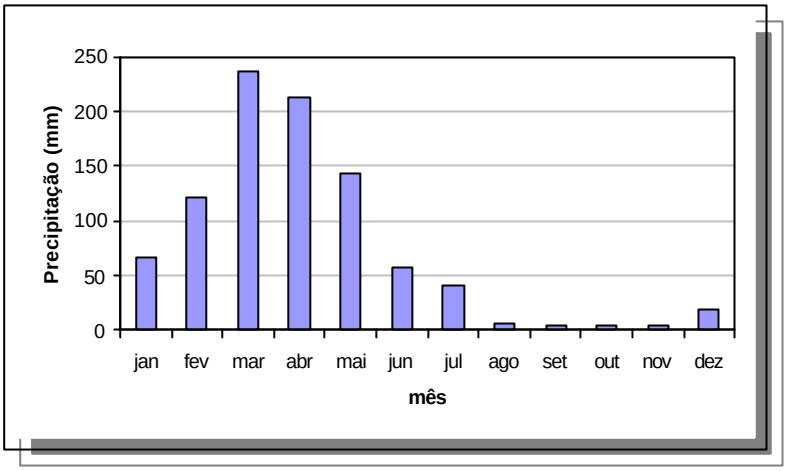


Figura 2.10. Gráfico em colunas

b) Gráfico em Barras

É semelhante ao gráfico em colunas, porém os retângulos são dispostos horizontalmente. Sua configuração é mostrada na Figura 2.10.

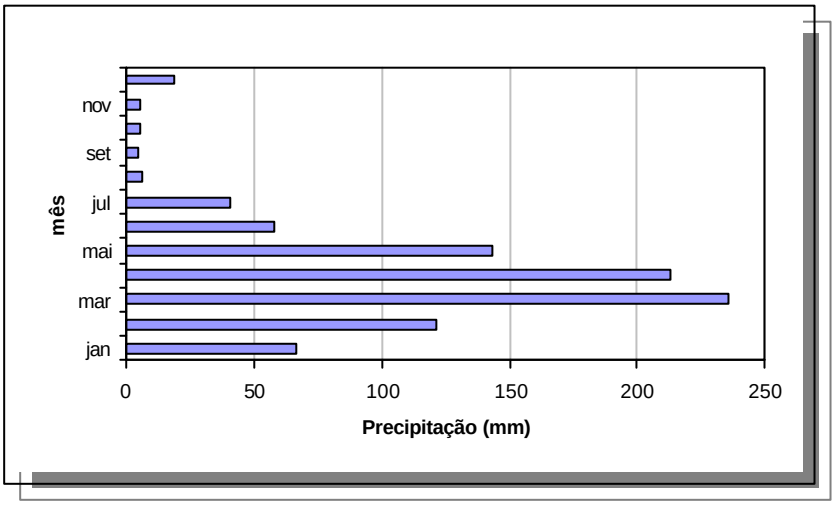


Figura 2.11. Gráfico em barras

c) Gráfico em Setores

É a representação gráfica de uma série estatística, em um círculo, por meio de setores. É utilizado principalmente quando se pretende comparar cada valor da série com o total. Para construí-lo, divide-se o círculo em setores, cujas áreas serão proporcionais aos valores da série. Essa divisão poderá ser obtida pela solução da regra de três.

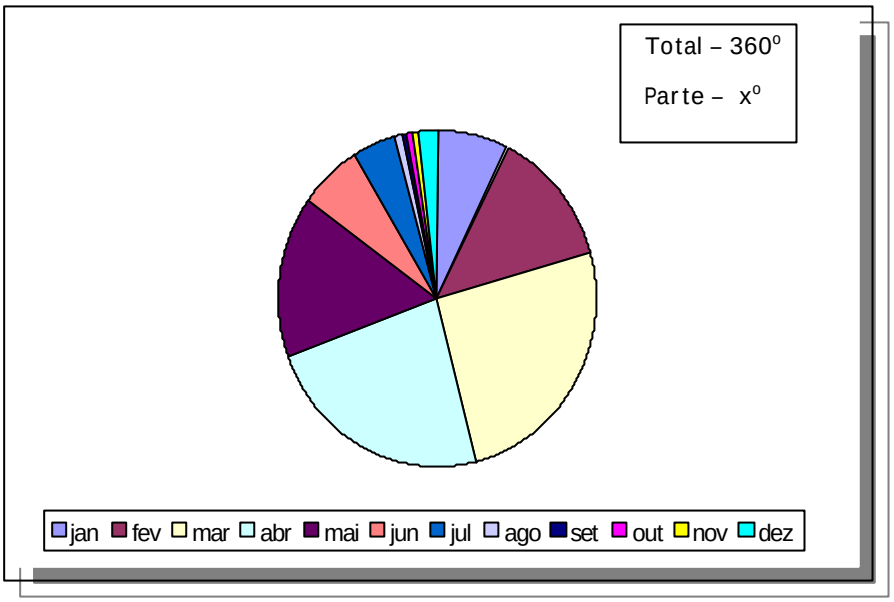


Figura 2.12. Gráfico em setores

d) Gráfico Polar

É representação de uma série por meio de um polígono. Geralmente presta-se para apresentação de séries temporais. Para construí-lo, divide-se uma circunferência em tantos arcos iguais quantos forem os dados a representar. Pelos pontos de divisas traçam-se raios.

Em cada raio é representado um valor da série, marcando-se um ponto cuja distância ao centro é diretamente proporcional a esse valor. A seguir unem-se os pontos (linha em vermelho). Exemplo:

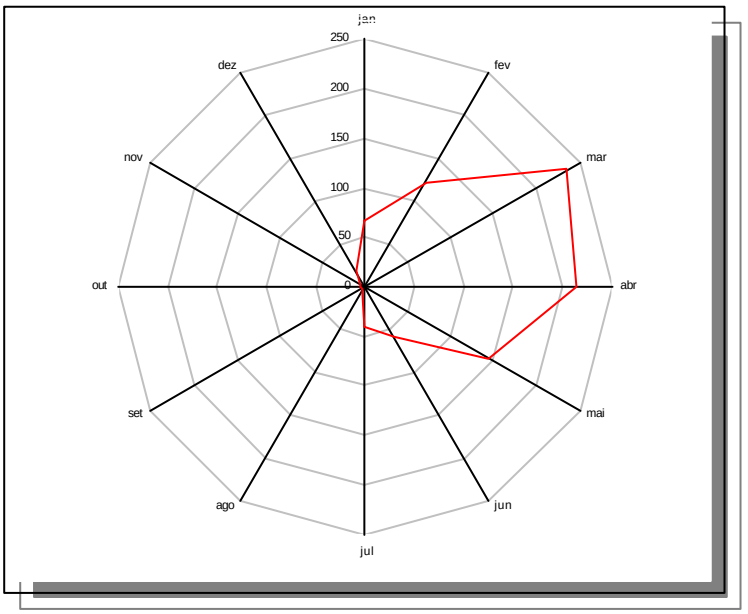


Figura 2.13. Gráfico polar

e) Histograma

É a representação gráfica de uma distribuição de freqüência por meio de retângulos justapostos. O exemplo a seguir refere-se às chuvas máximas anuais no posto pluviométrico de Cedro/Ceará. Os histogramas a seguir foram elaborados com o software STATISTICA.

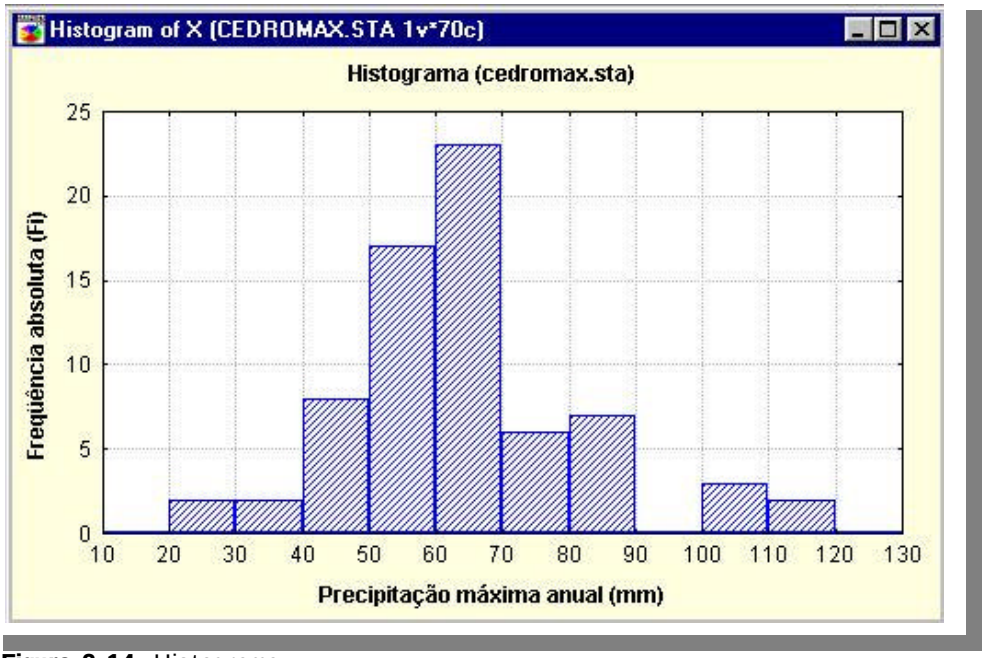


Figura 2.14. Histograma

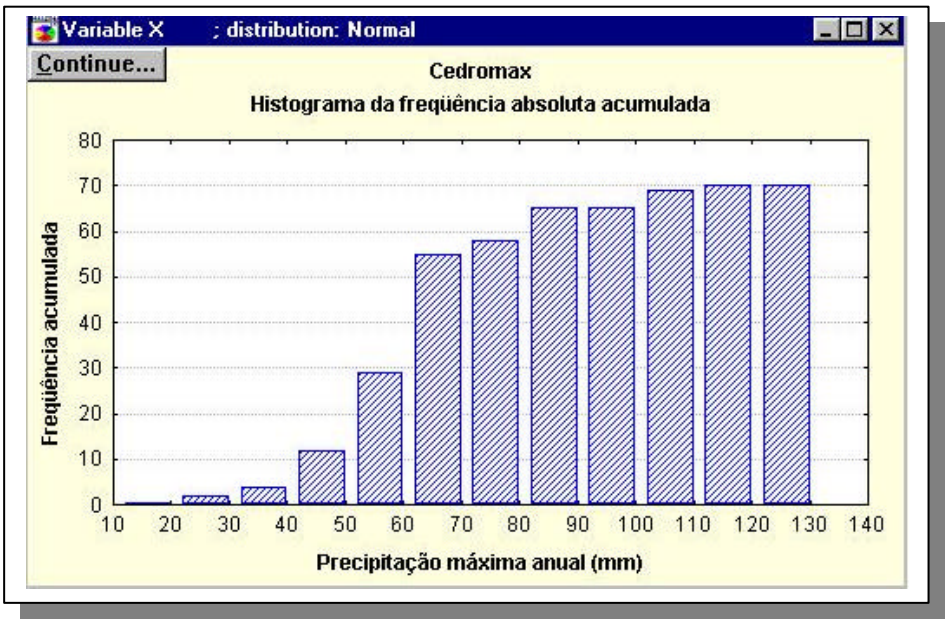


Figura 2.15. Histograma da freqüência absoluta acumulada

f) Polígono de freqüências

É a representação gráfica de uma distribuição por meio de um polígono, unindo-se os pontos médios das classes.

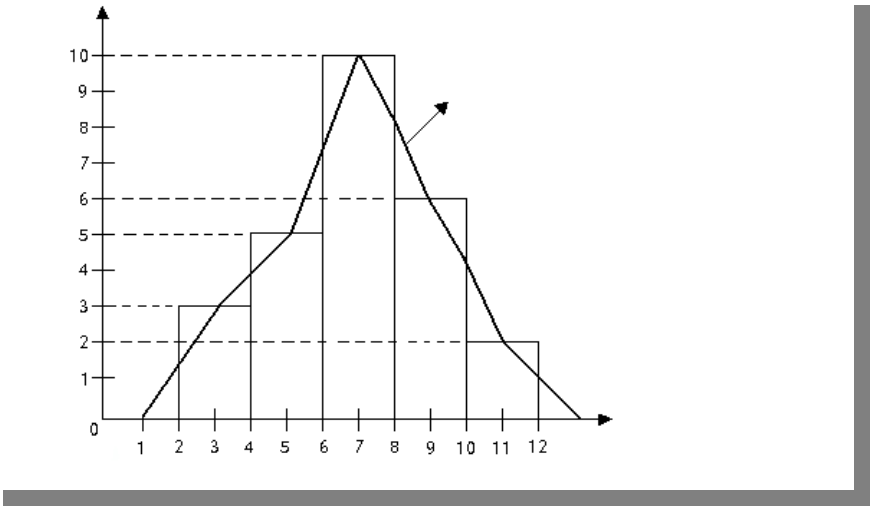


Figura 2.16. Polígono de freqüências

g) Diagrama de caixas (Boxplots)

O diagrama de caixas (boxplot) é um gráfico que consiste em uma reta que se prolonga do menor ao maior valor, e um retângulo com retas traçadas no primeiro quartil (Q_1), na mediana e no terceiro quartil (Q_3).

Dado o conjunto de dados abaixo, traçar o diagrama de caixas.

52 – 52 –60 –60 –60 –60 –63 –63 –66 –67 – 68
69 –71 –72 –73 –75 – 78 -80 – 82 – 83 – 88 – 90

Utilizando o STATISTICA, observa-se pelo boxplot que a distribuição é ligeiramente assimétrica positiva, com mínimo em 52, máximo em 90, Q_1 igual a 60, Q_3 igual 76,5 e mediana igual a 68,5.

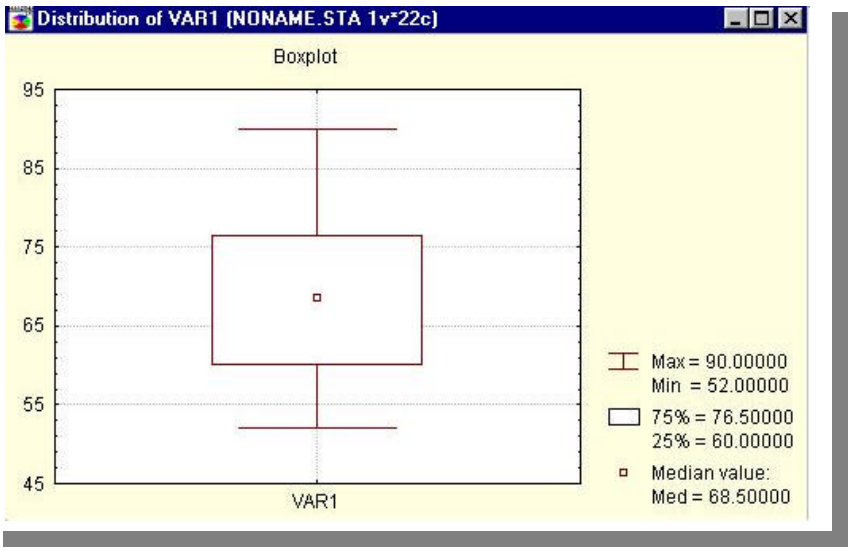


Figura 2.17 Boxplot

Os boxplots são especialmente úteis para a identificação de **outliers**. Caso no conjunto de dados acima tivesse ocorrido um erro ao se digitar 90 (colocou-se 900, por exemplo):

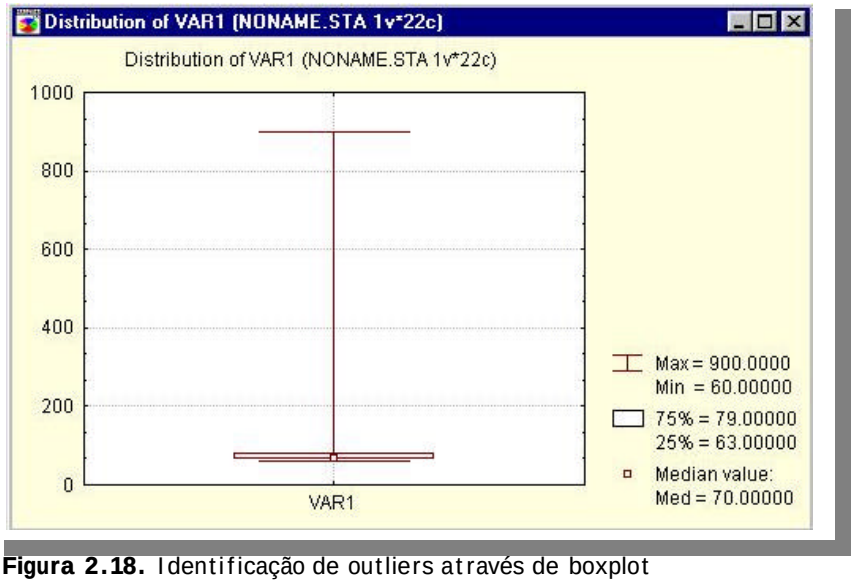


Figura 2.18. Identificação de outliers através de boxplot