Estatística Descritiva

Departamento de Matemática
Escola Superior de Tecnologia de Viseu

Gestão de Empresas
Marketing
Contabilidade e Administração

Conceitos Básicos

- População ou Universo Estatístico: conjunto de elementos sobre o qual incide o estudo estatístico
- Característica Estatística ou Atributo: a característica que se observa nos elementos da população
- Modalidades (incompatíveis e exaustivas): as diversas formas em que se apresenta a característica estatística
- Amostra: subconjunto finito da população
  Razões para a recolha de uma amostra: dimensão excessiva da população, estudo de natureza destrutiva, economia e tempo

Conceitos Básicos

Exemplo

- O Gestor de produção de uma fábrica pretende ter uma ideia da percentagem de peças defeituosas que a fábrica produziu em determinado período de tempo

  *A **população** em estudo é constituída por todas as peças produzidas pela fábrica durante aquele período de tempo*

  *A **característica estatística** tem apenas duas modalidades: peça defeituosa e peça não defeituosa*

Conceitos Básicos

Exemplo

- Num estudo de mercado para construção de um centro comercial, interessa estudar o rendimento familiar mensal dos habitantes de uma determinada cidade

  *A **população** é constituída pelas famílias daquela cidade e a **característica estatística** é o rendimento familiar mensal*

  *As **modalidades** do rendimento familiar mensal não se podem enumerar; são todos os valores desde, por exemplo, 50 contos até 1000 contos*

Conceitos Básicos

Exemplo

- Uma determinada empresa pretende realizar um inquérito aos seus trabalhadores, onde lhes é pedido para classificarem a qualidade do serviço do bar/refeitório segundo a seguinte escala: fraco, razoável, bom ou muito bom

  *Os trabalhadores da fábrica constituem a **população** em estudo, e a **característica estatística** é a opinião acerca da qualidade do serviço do bar/refeitório*

  *Neste estudo o atributo pode manifestar-se nas seguintes **modalidades**: fraco, razoável, bom ou muito bom*

Tipos de Dados Estatísticos

- Quantitativos
  por exemplo: nº diário de nascimentos no hospital de Viseu
  altura dos alunos da ESTV

  - Discretos - nº finito ou infinito numerável de modalidades
    por exemplo: nº diário de nascimentos no hospital de Viseu

  - Contínuos - pode assumir qualquer valor num intervalo de números reais
    por exemplo: altura de um aluno da ESTV

- Qualitativos
  por exemplo: cor dos cabelos
  estado civil

Escalas de Medida de Dados Estatísticos

Escala Nominal (dados qualitativos)

Apresentam-se em diferentes categorias ou classes, não ordenáveis

Exemplos

- Estado civil dos empregados de uma empresa
- Religião
- Cor de cabelos
- Os números das camisolas dos futebolistas
- Sexo de um indivíduo (***característica dicotómica*** ou ***binária***)
- Numa sondagem de opinião, a resposta à pergunta ”É a favor da despenalização do aborto?”(***característica dicotómica*** ou ***binária***)

Para lidar com este tipo de dados é frequente atribuir um código numérico a cada categoria da característica em estudo

Não realizar operações aritméticas e não calcular médias, desvios padrões,

Escalas de Medida de Dados Estatísticos

Escala Ordinal (dados qualitativos)

As diversas categorias possuem uma ordem intrínseca
Os códigos numéricos devem ter em conta essa ordem

Exemplos

- O sistema de graduação militar: ***Soldado***, ***Cabo***, ***Sargento***, ...
- Num inquérito de opinião pede-se às pessoas que classifiquem um determinado produto como sendo: ***muito fraco***, ***fraco***, ***razoável***, ***bom*** ou ***muito bom*** (*escala de Likert*)
- Classificação dos clientes de um banco, segundo o volume de capital que movimentam mensalmente: ***pouco importantes***, ***importantes*** ou ***muito importantes***
- Classificação dos alunos de uma escola segundo a sua altura: ***baixos*** (menos de 155 cm), ***médios*** (entre 155 e 170 cm) ou ***altos*** (mais de 170 cm)

Escalas de Medida de Dados Estatísticos

Escala de Intervalo (dados quantitativos)

Os dados podem ser ordenados e a diferença entre dois valores desta escala pode ser calculada e interpretada

Exemplo

Temperatura do ar em graus *Fahrenheit* ou em graus centígrados

$$F = 9/5C + 32$$

Distâncias numericamente iguais implicam as mesmas alterações na característica que está a ser medida

20^oC está à mesma distância de 25^oC, do que 25^oC de 30^oC

Não podemos atribuir um significado à razão entre dois valores

Se na Guarda se registasse uma temperatura de 40^oC isto não significaria que na Guarda estava duas vezes mais calor do que em Viseu

O valor zero não tem o significado de "nada"

Não se pode dizer que uma cidade onde se registe uma temperatura de 0^oC não tem qualquer temperatura

Escalas de Medida de Dados Estatísticos

Escala de Razões ou de Rácios (dados quantitativos)

Tem todas as características de uma escala de intervalo e, além disso, o valor zero representa a ausência total da característica que está a ser medida

Com dados medidos nesta escala, não só é possível atribuir um significado à diferença (distância) entre dois valores como também à razão entre eles

Alterações nas unidades de medida não afectam os rácios entre dois valores

por exemplo: peso, altura

A temperatura do ar não está definida numa escala de rácios

Note que $10^oC = 50^oF$ e $30^oC = 86^oF$ mas, $\frac{10^oC}{30^oC} \neq \frac{50^oF}{86^oF}$

Representação de Dados

População ou amostra de n individuos

Atributo A com p modadilades: $A_1, A_2, ..., A_p$

$n_i \leftarrow$ frequência absoluta ou efectivo da modadilade A_i: nº de indivíduos que apresentam a modalidade A_i

$f_i \leftarrow$ frequência relativa da modadilade A_i: proporção de indivíduos que apresentam a modadidade A_i,

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

$$\sum_{i=1}^{p} n_i = n \qquad e \qquad \sum_{i=1}^{p} f_i = 1$$

Representação Tabular – Quadros de Frequências

Modalidades	Frequências absolutas	Frequências relativas	Frequências absolutas acumuladas	Frequências relativas acumuladas
A_1	n_1	$f_1=n_1/n$	n_1	f_1
A_2	n_2	$f_2=n_2/n$	n_1+n_2	f_1+f_2
⋮	⋮	⋮	⋮	⋮
A_P	n_p	$f_p=n_p/n$	$n_1+n_2+...+n_p=n$	$f_1+f_2+...+f_p=1$
Total	n	1	-	-

Exemplo 1:

Os dados que se seguem são relativos às vendas (em contos) de 30 vendedores da ElectroNoLar durante o mês de Outubro passado.

120	130	80	100	110	100	90	70	140	120
140	110	100	100	110	70	90	90	130	150
160	80	70	120	100	110	110	80	100	120

Tabela de frequências - dados não agrupados

x_i	Freq. absolutas n_i	Freq. relativas f_i	Freq. absolutas acumuladas	Freq. relativas acumuladas
70	3	3/30	3	3/30
80	3	3/30	6	6/30
90	3	3/30	9	9/30
100	6	6/30	15	15/30
110	5	5/30	20	20/30
120	4	4/30	24	24/30
130	2	2/30	26	26/30
140	2	2/30	28	28/30
150	1	1/30	29	29/30
160	1	1/30	30	1
Total	30	1	-	-

Tabela de frequências com dados agrupados.

Classes de valores	Freq. absolutas n_i	Freq. relativas f_i	Freq. absolutas acum.	Freq. relativas acum.
[60, 80[	3	3/30	3	3/30
[80, 100[	6	6/30	9	9/30
[100, 120[	11	11/30	20	20/30
[120, 140[	6	6/30	26	26/30
[140, 160[	3	3/30	29	29/30
[160, 180[	1	1/30	30	30/30
Total	30	1	-	-

- Os intervalos de classe podem ter a mesma amplitude ou amplitudes diferentes dependendo da natureza dos fenómenos a estudar.
- Agrupar os dados implica perda de informação.
- Regras práticas para a determinação do nº de classes:

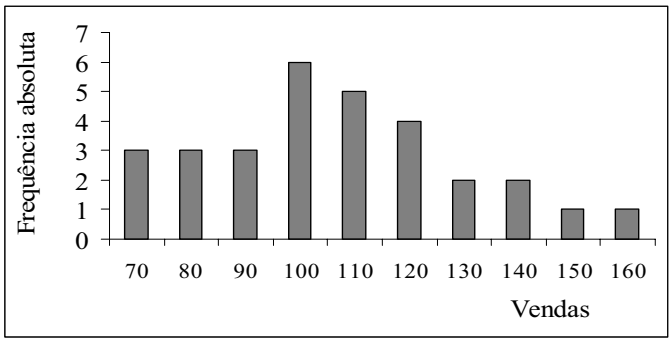
  Regra de Sturges – nº de classes $\cong 1+\log_{10}(n)/\log_{10}(2)$

  Outra – nº de classes $\cong \sqrt{n}$ (usualmente empregue quando n>25).

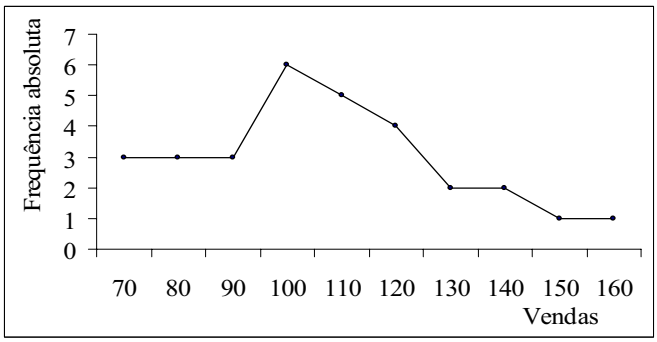
Representação gráfica

Dados Não Agrupados

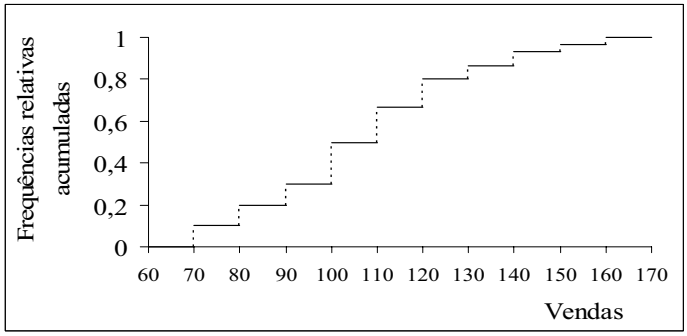
Diagrama de barras



Polígono de frequências



Representação gráfica das frequências acumuladas



Dados Agrupados

Histograma

No histograma tomamos rectângulos justapostos, cada um com base proporcional à amplitude da classe respectiva e altura h_i dada por:

$$h_i = \begin{cases} \dfrac{n_i}{a_{i+1} - a_i} & \text{(frequências absolutas)} \\ \dfrac{f_i}{a_{i+1} - a_i} & \text{(frequências relativas)} \end{cases}$$

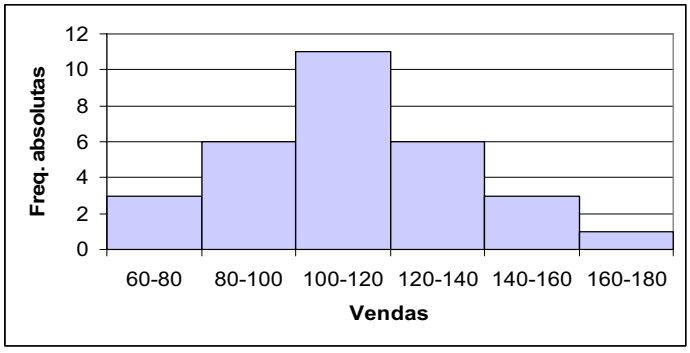
A área de cada rectângulo é então proporcional à frequência da classe respectiva:

$$\text{área do } i\text{ - ésimo rectângulo} = \begin{cases} n_i & \text{(frequências absolutas)} \\ f_i & \text{(frequências relativas)} \end{cases}$$

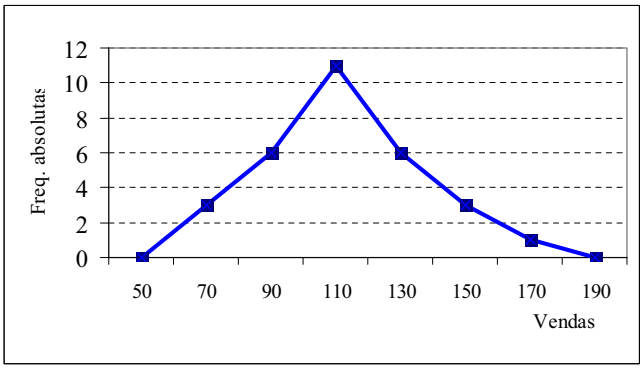
A área total do histograma é igual a ***n*** se foram usadas frequências absolutas e igual a **1** se foram usadas frequências relativas.

Note-se porém que, quando as classes têm todas a mesma amplitude é costume, para facilitar a representação, tomar para altura de cada rectângulo a frequência absoluta ou relativa da classe a que respeita.

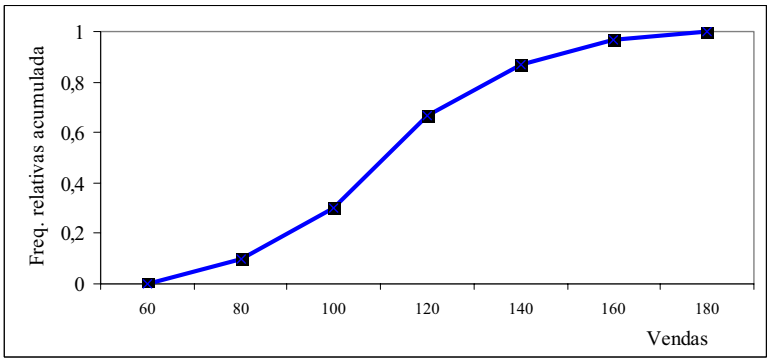
Histograma



Polígono de frequências



Polígono de frequências acumuladas



Medidas Descritivas

Medidas de Localização ou de Tendência Central

Estas medidas dão-nos uma ideia do "centro" ou "localização" da distribuição dos dados.

Média aritmética

Sejam $x_1, x_2, \ldots, x_p$ os valores distintos de um conjunto de **n** dados, cada um deles com frequência absoluta n_i e frequência relativa f_i. Então a média aritmética representa-se por $\bar{x}$ e é dada por:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \sum_{i=1}^{n} f_i x_i .$$

Para **dados agrupados em classes** toma-se para x_i o ponto médio da *i-ésima* classe; n_i e f_i serão, naturalmente, a frequência absoluta e relativa da *i-ésima* classe, respectivamente.

Exemplo 2:

A tabela de frequências que se segue é relativa ao número de pneus produzidos por dia na fábrica MAVOR, para uma amostra de 30 dias.

$\mathbf{x_i}$	Freq. absoluta $\mathbf{n_i}$	Freq. relativa $\mathbf{f_i}$	Freq. abso. acum.	Freq. relat. acum.s	$\mathbf{n_i x_i}$
18	2	0.06667	2	0.06667	36
20	3	0.1	5	0.16667	60
21	5	0.16667	10	0.33334	105
24	7	0.23333	17	0.56667	168
25	6	0.2	23	0.76667	150
28	4	0.13333	27	0.9	112
29	3	0.1	30	1	87
Total	30	1	-	-	718

A média de pneus produzidos diariamente, para os 30 dias considerados é:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i = \frac{718}{30} = 23.9333.$$

Mediana

Trata-se do valor que divide o conjunto de dados, ordenados por ordem crescente, em duas partes iguais. Isto é, a mediana, como o próprio nome indica, é o ponto mediano de um conjunto de dados ordenados em ordem crescente.

Sejam $x_1, x_2, \ldots, x_n$, **n** observações ordenadas por ordem crescente dos seus valores, e que constituem o conjunto de dados em análise.

$$Me = \begin{cases} x_{(n+1)/2} & se \quad n \quad é \quad impar \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & se \quad n \quad é \quad par \end{cases}$$

Exemplo 2, como n é par: $Me = \dfrac{x_{30/2} + x_{30/2+1}}{2} = \dfrac{x_{15} + x_{16}}{2} = \dfrac{24+24}{2} = 24.$

Para **dados agrupados em classes**, procuramos a classe mediana, sendo esta tal que a sua frequência absoluta (resp. relativa) acumulada é $\geq n/2$ (resp. 1/2) e a frequência absoluta (resp. relativa) acumulada da classe anterior é $< n/2$ (resp. 1/2).

Depois de encontrada a classe mediana, $[a_j, a_{j+1}[$, encontra-se a mediana por interpolação linear:

$$Me = a_j + \frac{n/2 - \sum_{i=1}^{j-1} n_i}{n_j}\left(a_{j+1} - a_j\right)$$

Moda

É o valor mais frequente num conjunto de dados.

- {2, 3, 4, 4, 5} → Mo=4 (distribuição unimodal);
- {2, 2, 3, 4, 4, 5} → Mo=2 e 4 (distribuição bimodal);
- Exemplo 2 → Mo=24.

Havendo mais de 2 valores modais, a distribuição diz-se multimodal.

Quando os **dados estão agrupados em classes**, a classe modal é aquela que tem maior frequência por unidade de amplitude. Nestes casos não podemos determinar o valor exacto da moda pois não sabemos como estão distribuídas as observações dentro de cada classe. Podemos, no entanto, obter uma aproximação da Moda usando uma das seguintes fórmulas:

Fórmula de King: $Mo = a_j + \dfrac{n_{j+1}}{n_{j-1} + n_{j+1}}\left(a_{j+1} - a_j\right)$

Fórmula de Czuber: $Mo = a_j + \dfrac{n_j - n_{j-1}}{\left(n_j - n_{j-1}\right) + \left(n_j - n_{j+1}\right)}\left(a_{j+1} - a_j\right)$

onde, $[a_j, a_{j+1}[$ é a classe modal; n_j é a freq. abso. desta classe; n_{j+1} e n_{j-1} são, resp., a freq. abso. da classe anterior e posterior à modal.

Medidas de Localização não Central – Quantis: Q_p

A mediana divide o conjunto de dados em duas partes iguais. Quando o conjunto de dados ordenados é dividido em 4 partes iguais, os pontos de divisão são chamados os ***quartis***:

- $Q_{1/4}$, 1º quartil – valor que tem cerca de 25% dos dados abaixo dele;
- $Q_{2/4}$, 2º quartil – valor que tem cerca de 50% dos dados abaixo dele – trata-se da Mediana;
- $Q_{3/4}$, 3º quartil – valor que tem cerca de 75% dos dados abaixo dele.

Podemos ainda calcular os ***quintis***, ***decis***, ***percentis***,…

Cálculo do quantil de ordem p, Q_p: *Dados não agrupados em classes*

Sejam $x_1, x_2, \ldots, x_n$, n observações ordenadas por ordem crescente dos seus valores.

Se np não é um inteiro, então $Q_p = x_k$, onde k é o inteiro imediatamente seguinte a np. Caso contrário, sendo np um inteiro, então $Q_p = (x_{np} + x_{np+1})/2$.

Cálculo do quantil de ordem p, Q_p: *Dados agrupados em classes*

Seja $[a_j, a_{j+1}[$ a classe que contém Q_p, i.e., que contém o valor ao qual corresponde a frequência absoluta (resp. relativa) acumulada de np (resp. p). Por interpolação linear obtém-se Q_p:

$$Q_p = a_j + \frac{np - \sum_{i=1}^{j-1} n_i}{n_j}\left(a_{j+1} - a_j\right)$$

Posição relativa da média, mediana e moda

As distribuições de frequências podem ser simétricas ou não.

Considerando apenas distribuições unimodais, temos:

Distribuições simétricas $\rightarrow \ \bar{x} = Me = Mo$

Distribuições assimétricas positivas $\rightarrow \ Mo < Me < \bar{x}$

A cauda direita é mais longa e menos abrupta do que a esquerda.

Distribuições assimétricas negativas $\rightarrow \ \bar{x} < Me < Mo$

A cauda esquerda é mais longa e menos abrupta do que a direita

Nas distribuições assimétricas os valores extremos da cauda mais longa puxam a média para o lado direito. A mediana, como divide a área em duas partes iguais, para compensar a redução de área no lado abrupto, afasta-se também da moda, mas menos do que a média.

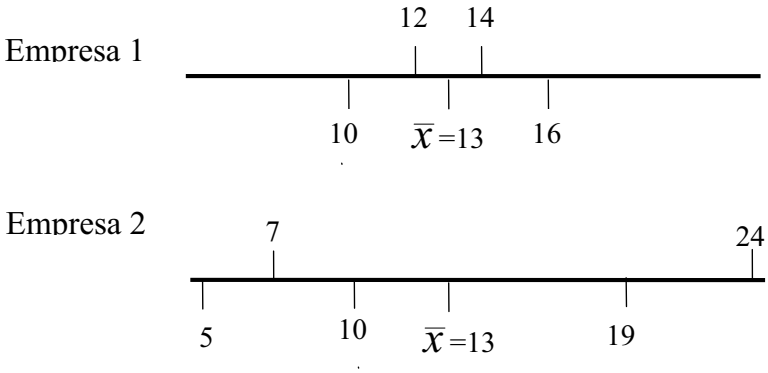
Medidas de Dispersão

Exemplo:

Duas empresas concorrentes com sede em Viseu, obtiveram os seguintes lucros nos 5 últimos anos:

	Lucros em unidades monetárias (u. m.)				
Empresa 1	10	13	12	14	16
Empresa 2	7	5	10	19	24

O lucro médio das duas empresa nos últimos 5 anos é o mesmo, 13 u.m., no entanto a Empresa 2 apresenta uma maior variabilidade nos lucros do que a Empresa 1.



O **intervalo interquartis**, $[Q_{1/4}, Q_{3/4}]$ contém 50% das observações. A amplitude deste intervalo, **amplitude interquartis**, é uma medida de dispersão.

As medidas de dispersão mais utilizadas são o **desvio padrão** e a **variância** que definimos a seguir.

Sejam $x_1, x_2, \ldots, x_p$ os valores distintos de um conjunto de n dados, cada um deles com frequência absoluta n_i e frequência relativa f_i.

Se estes dados constituem observações feitas sobre toda a população, a **variância** denota-se por σ^2 e é calculada da seguinte maneira:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i (x_i - \bar{x})^2 ,$$

ou equivalentemente,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2 = \sum_{i=1}^{p} f_i x_i^2 - \bar{x}^2 .$$

Se, pelo contrário, o conjunto de dados constitui uma amostra da população, então a **variância** denota-se por s^2 e é dada por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 \quad \Leftrightarrow \quad s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{p} n_i x_i^2 - n\bar{x}^2\right).$$

O **desvio padrão** é a raiz quadrada da variância e denota-se por σ ou por s.

Exemplo 2

Como dispomos de uma amostra, temos: $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{p} n_i\, x_i^2 - n\bar{x}^2\right)$.

x_i	Frequências absolutas n_i	Frequências relativas f_i	$n_i x_i$	$n_i x_i^2$
18	2	0.06667	36	648
20	3	0.1	60	1200
21	5	0.16667	105	2205
24	7	0.23333	168	4032
25	6	0.2	150	3750
28	4	0.13333	112	3136
29	3	0.1	87	2523
Total	30	1	718	17494

Então a variância e o desvio padrão são, respectivamente,

$s^2 = \frac{1}{29}\left(17494 - 30 \times 23.9333^2\right) = 10.6867$ *(u.m.)²* e $s = \sqrt{106867} = 3.269$ *u.m.*.

Coeficiente de dispersão e de variação

Medidas de dispersão absolutas: expressas na mesma unidade dos dados a que se referem

Medidas de dispersão relativas: independentes da unidade de medida dos dados a que se referem

A variância e o desvio padrão são medidas de dispersão absolutas.

Se pretendermos comparar a dispersão de dois conjuntos de dados que não estejam expressos na mesma unidade de medida, teremos de adoptar uma medida de dispersão relativa, por exemplo:

Coeficiente de dispersão: $cd = \frac{s}{\bar{x}}$ ou $\frac{\sigma}{\bar{x}}$ **Coeficiente de variação**: $cv = cd \times 100\%$

Estes coeficientes só se empregam quando a variável toma valores de um só sinal.

Momentos

Chama-se momento simples de ordem *r* ou momento ordinário de ordem *r* a

$$m'_k = \sum_{i=1}^{p} f_i x_i^k = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^k$$

Chama-se momento centrado de ordem *r* a

$$m_k = \sum_{i=1}^{p} f_i (x_i - \bar{x})^k = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^k$$

Se a distribuição for simétrica os momentos centrados de ordem ímpar são nulos, pois para cada desvio negativo há um desvio positivo com o mesmo valor absoluto.

Alguns momentos:

$$m'_o = \sum_{i=1}^{p} f_i = 1 \qquad m_o = 1$$

$$m'_1 = \sum_{i=1}^{p} f_i x_i = \bar{x} \qquad m_1 = \sum_{i=1}^{p} f_i (x_i - \bar{x}) = \bar{x} - \bar{x} = 0$$

$$m'_2 = \sum_{i=1}^{p} f_i x_i^2 = \sigma^2 - \bar{x}^2 \qquad m_2 = \sum_{i=1}^{p} f_i (x_i - \bar{x})^2 = \sigma^2$$

Coeficientes de assimetria e achatamento

Coeficiente de assimetria: $g_1 = \dfrac{m_3}{\sqrt{m_2^3}}$

Distribuição **simétrica** → $g_1 = 0$

Distribuição **assmétrica positiva** → $g_1 > 0$

Distribuição **assmétrica negativa** → $g_1 < 0$

Embora as proposições recíprocas não sejam sempre verdadeiras é costume tomar g_1 como medida de assimetria.

Coeficiente de achatamento ou curtose: $g_2 = \dfrac{m_4}{m_2^2}$

Este coeficiente mede o grau de achatamento de uma distribuição, considerado em relação ao da distribuição normal, para a qual se tem $g_2 = 3$.

Distribuição **mesocúrtica** → $g_2 = 3$ (achatamento igual ao da normal)

Distribuição **leptocúrtica** → $g_2 > 3$ (achatamento inferior ao da normal)

Distribuição **platicúrtica** → $g_2 < 3$ (achatamento superior ao da normal)