PARTE II
ANÁLISE INICIAL DE DADOS

Introdução

Neste capítulo dedicado à Análise Inicial de Dados, começaremos por fazer uma revisão de forma muito breve de conceitos de que os alunos já conhecem do Ensino Secundário: técnicas de representação tabular e gráfica de amostras de dados (tabelas de frequências, diagramas de linhas, diagramas circulares, histogramas, diagramas de extremos-e-quartis) e algumas características amostrais (média, moda, mediana, quartis, variância, desvio padrão, amplitude interquartis).

Conceitos básicos

Como já referimos, a Estatística lida com dados, ou seja, registos de observações relativos a variáveis, obtidos, em geral, a partir de uma amostra de uma dada população.
As variáveis em estudo, podem ser de natureza qualitativa ou quantitativa.
As variáveis qualitativas – também ditas categóricas ou fatores – variam em tipo ou qualidade, mas não em quantidade; os seus valores são intrinsecamente não numéricos. Exemplos de variáveis qualitativas: cor do cabelo, grupo sanguíneo, dureza dos minerais (expressos na escala de Mohs).

Note-se que, por exemplo, no caso da dureza dos minerais, o que está em causa é uma característica (a resistência ao risco) a qual é intrinsecamente não numérica, sendo a escala numérica de Mohs, de 1 a 10, uma simples convenção.

As variáveis quantitativas variam em quantidade, sendo os seus valores intrinsecamente numéricos. São exemplos de variáveis quantitativas: altura, peso, idade, número de filhos, número de dentes sãos, temperatura (medida em graus Celsius), temperatura absoluta (medida em graus Kelvin).

Escalas de medida

As variáveis quantitativas podem ainda distinguir-se entre variáveis discretas – quando assumem apenas um número finito ou infinito numerável de valores e variáveis contínuas – quando podem tomar valores num conjunto infinito não numerável (por exemplo um certo intervalo de números reais).

Existem quatro tipos de escalas de medida que podemos usar quando lidamos com dados estatísticos:

1. Nominal (ou categórica)
2. Ordinal
3. Intervalar
4. Absoluta (ou de razões)

Escala nominal e escala ordinal

As escalas de tipo nominal e ordinal são usadas para variáveis qualitativas, para distinguir diferentes categorias ou classes de indivíduos.
Usam-se escalas de tipo nominal quando a ordem das categorias ou classes não tem significado; aos indivíduos da mesma categoria ou classe é atribuído o mesmo nome (ou código). Um exemplo de uma escala nominal, será, para a variável sexo, a que atribui F para o sexo feminino e M para o sexo masculino (ou 1 para o sexo feminino e 2 para o sexo masculino, se preferirmos).

Neste caso não faz sentido efetuar operações aritméticas sobre os nomes das classes ou categorias, mesmo que estes estejam representados por números.

Os métodos estatísticos apropriados para a análise de dados em escala nominal são os que se baseiam em contagens de efetivos de cada classe ou análise de proporções.

As escalas de tipo ordinal são usadas quando a ordem das categorias ou classes já tem significado. Por exemplo, a escala de Mohs para a dureza dos minerais.

Para os dados em escala ordinal, para além de usarmos métodos baseados em contagens e proporções, também podemos usar métodos baseados nas ordens (ou *ranks*) das observações.

Escala intervalar e escala absoluta

As escalas de tipo intervalar e absoluto são escalas numéricas, usadas para variáveis quantitativas.
Nas escalas de tipo intervalar pode haver um zero, mas esse zero é apenas convencional, não significando ausência da característica medida. Por exemplo, a escala Celsius ou a escala Fahrenheit para medir temperaturas: 0^o Celsius não significa ausência de calor (note-se que 0^oC corresponde a 32^o F).

Neste tipo de escalas faz sentido ordenar as medições (32^oC é uma temperatura superior a 16^oC), fazer operações aritméticas envolvendo somas e diferenças (por exemplo achar médias de temperaturas), mas não faz sentido fazer divisões ou razões.Por exemplo, não faz sentido dizer que 32^oC de temperatura é o dobro de 16^oC (basta converter na escala Fahrenheit para perceber porquê...)

Nas escalas absolutas, o zero não é uma simples convenção, mas corresponde, na realidade, a ausência da característica medida. Por exemplo, a escala Kelvin para temperatura ou a escala em Kg para o peso.

Neste caso, faz sentido, para além das operações aritméticas já referidas para a escala intervalar, fazer também divisões (comparações por quocientes). Um indivíduo com $100Kg$ pesa, efetivamente, o dobro de um com $50Kg$ (se converter o peso noutras unidades, a relação será

Organização/Representação de dados

Dados Qualitativos

Os dados qualitativos são geralmente descritos usando tabelas de frequências (absolutas ou relativas) e gráficos, tais como gráficos de barras ou diagramas circulares.

Exemplo

Foi recolhida uma amostra de 400 estudantes a viver em residência universitária e estes foram inquiridos relativamente ao seu grau de satisfação com as condições de alojamento, na seguinte escala: Muito Insatisfeito; Insatisfeito; Razoavelmente Satisfeito; Satisfeito; Muito Satisfeito. Os resultados foram os seguintes:
28 escolheram a categoria "Muito Insatisfeito",
60 escolheram a categoria "Insatisfeito",
160 escolheram a categoria "Razoavelmente Satisfeito",
120 escolheram a categoria "Satisfeito",
32 escolheram a categoria "Muito Satisfeito".

Exemplo

Uma tabela de frequências (absolutas e relativas) para os dados recolhidos seria:

Opinião de estudantes sobre o alojamento na residência

Grau de satisfação	Frequência absoluta	Frequência relativa	Frequência relativa (%)
Muito Insatisfeito	28	0.07	7
Insatisfeito	60	0.15	15
Razoavelemente Satisfeito	160	0.40	40
Satisfeito	120	0.30	30
Muito Satisfeito	32	0.08	8
Total	400	1	100

Tabelas de frequências acumuladas

Se os dados estiverem em escala ordinal, podemos também apresentar na tabela de frequências, as chamadas frequências acumuladas (absolutas ou relativas), as quais se obtêm adicionando as sucessivas frequências.

Exemplo

Considerando os dados do exemplo anterior, ter-se-ia a seguinte tabela.

Opinião de estudantes (alojamento na residência)

Grau satisfação	Freq. abs.	Freq. abs. acumulada	Freq. rel.	Freq. rel. acumulada
Mto Insatisfeito	28	28	0.07	0.07
Insatisfeito	60	88	0.15	0.22
Raz. Satisfeito	160	248	0.40	0.62
Satisfeito	120	368	0.30	0.92
Mto Satisfeito	32	400	0.08	1
Total	400		1	

Gráficos de barras e diagramas circulares

Na construção de um gráfico de barras ou diagrama circular, deve ter-se em atenção que:

1. as alturas das barras (num gráfico de barras) ou as amplitudes dos diversos sectores (num diagrama circular) são proporcionais às respectivas frequências;
2. num gráfico de barras, estas devem estar igualmente distanciadas umas das outras, e podem ser representadas na horizontal ou na vertical;
3. os diagramas circulares utilizam-se apenas quando o número de categorias em que a variável é classificada é pequeno.

Os comandos do Mathematica apropriados para esboçar gráficos de barras são BarChart ou BarChart3D.
Para diagramas circulares podemos usar os comandos PieChart ou PieChart3D.

Gráficos de barras

Exemplo

Gráficos de barras para os dados do exemplo que estamos a considerar poderiam ser os seguintes:

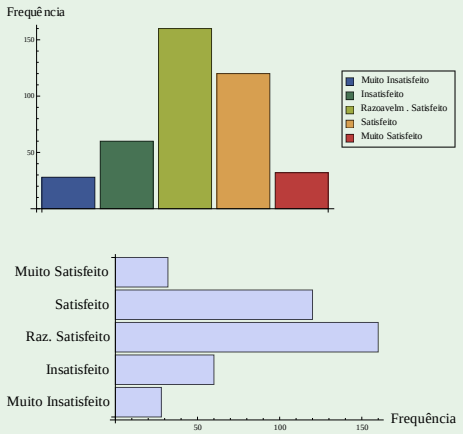
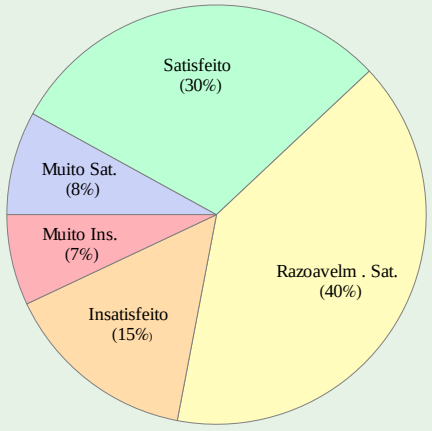


Diagrama circular

Exemplo

Um diagrama circular para este exemplo, poderia ser:



Organização/representação de dados quantitativos

Os dados quantitativos correspondentes a variáveis que não tomem um grande número de valores podem, de modo análogo ao que é feito para variáveis qualitativas, ser apresentados através de tabelas de frequências. Quanto à representação gráfica destes dados, é usual usar gráficos de linhas (mas, por vezes, também gráficos de barras) para os representar, podendo também ser usados diagramas circulares.

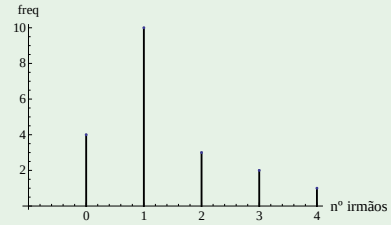
Gráfico de linhas

Exemplo

Registou-se o número de irmãos de cada um dos 20 alunos de uma dada turma de um liceu, tendo-se obtido os seguintes dados:

$$1, 1, 3, 2, 0, 1, 0, 3, 1, 1, 4, 0, 2, 2, 1, 1, 0, 1, 1, 1$$

Um gráfico de linhas correspondente a estes dados (obtido após a determinação das frequências dos diferentes dados) é o seguinte:



Dados provenientes de variáveis contínuas

No caso de dados provenientes de uma variável contínua, ou no caso de dados de uma variável discreta que tome um elevado número de valores, a sua apresentação numa tabela de frequências e respectiva representação gráfica pressupõe um agrupamento prévio dos dados em classes, as quais, salvo casos excepcionais, devem ser intervalos de igual amplitude, designada por intervalo de classe.
Isto significa que os dados são categorizados.
As classes devem ser:

- exaustivas (i.e. a união das classe deve conter todos os dados)
- mutuamente exclusivas (i.e. cada um dos dados apenas pode pertencer a uma classe).

A escolha das classes é bastante importante. Um regra prática frequentemente utilizada para determinar o número de classes a adoptar é a chamada regra de Sturges que estipula que, para uma amostra de dimensão n, se usem k classes, onde k é o menor inteiro para o qual se tem $2^{k-1} \geq n$.

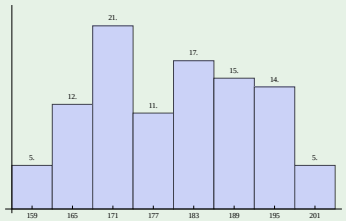
Histogramas

Os histogramas são gráficos de barras contíguas, cujas alturas são proporcionais às frequências das classes.

O Mathematica tem um comando apropriado para desenhar histogramas, Histogram, com diversos parâmetros para que possamos controlar a sua forma - variar o número de classes, escolher a área total do histograma como sendo igual a 1, etc.; por defeito, o Mathematica aproxima o número de classes pela regra de Sturges.

Exemplo

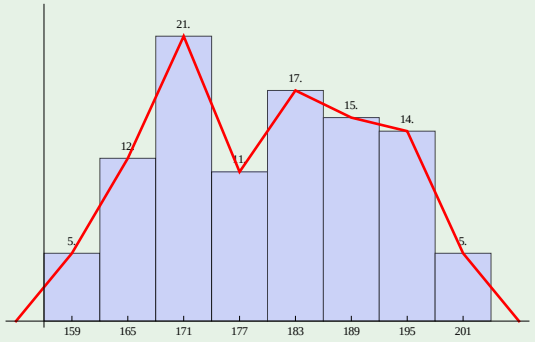
Exemplo de um histograma obtido com o comando Histogram:



Polígono de frequências

Por vezes, desenha-se o chamado polígono de frequências, unindo por segmentos de recta os pontos médios do topo dos retângulos; os pontos médios dos retângulos inicial e final são unidos até aos pontos, situados no eixo dos xx, cujas abcissas são os pontos médios das classes que seriam adjacentes às classes inicial e final.

Exemplo



Caraterísticas Amostrais

As tabelas de frequências e os gráficos constituem processos de redução de dados. No caso de dados quantitativos, é possível resumir de uma forma mais drástica esses dados, calculando certas características amostrais ou (empíricas).
Chamamos estatísticas às características numéricas da amostra e parâmetros às características numéricas da população; as características amostrais (empíricas, obtidas a partir dos dados) são estimativas das correspondentes características populacionais.

De um modo geral, no Mathematica, todos os comandos referidos para cálculo de caraterísticas populacionais (por exemplo, `Mean` para o valor médio, `Quantile` para os quantis teóricos, etc) podem também ser usados para calcular as respetivas caraterísticas amostrais.

Medidas de localização
Média amostral

No que se segue, consideramos que dispomos de uma amostra $x_1, x_2, \ldots, x_n$ e denotaremos por $x_{(1)}, \ldots, x_{(n)}$ a sequência desses dados ordenados por ordem crescente, i.e.

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Definição

A média amostral, denotada por $\bar{x}$, é definida por

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

A média amostral $\overline{x}$ é uma estimativa do valor médio $\mu = E(X)$ teórico, que, recorde-se é dado por $\mu(X) = \sum_k x_k p_k$ (no caso discreto) e por $\int_{-\infty}^{\infty} x f(x) dx$ (no caso contínuo).

Medidas de localização
Quantis amostrais

Definição

Para $0 < p < 1$, o quantil-p amostral, que denotaremos por por q_p é, *grosso modo*, o valor que separa os $p \times 100\%$ primeiros valores da amostra ordenada dos $(1-p) \times 100\%$ valores restantes.

A "definição" anterior é um pouco ambígua, porque existem diversas formas de definir os quantis amostrais, nem sempre coincidentes. Por exemplo, por vezes exige-se que o quantil seja um dos valores da amostra, outras vezes o quantil pode ser um valor situado entre dois valores da amostra, determinado de acordo com um certo critério.

No Mathematica, ao usarmos o comando `Quantile` para determinar os quantis amostrais, por defeito, estes serão sempre valores da amostra. O comamdo `Quantile` pode, no entanto, ser usada com outros parâmetros para permitir calcular os quantis de acordo com outras definições.

Medidas de localização

Mediana e quartis

Alguns quantis especialmente importantes são: o quantil $q_{1/2}$, que é designado por mediana amostral e os quantis $q_{1/4}, q_{2/4}, q_{3/4}$ que são os chamados quartis amostrais (respectivamente primeiro quartil, segundo quartil – mediana – e terceiro quartil).

O Mathematica tem comandos próprios para calcular a mediana e os três quartis: `Median` e `Quartiles`, respetivamente.

Nota: A mediana calculada com a função `Median` do Mathematica corresponde ao uso da fórmula

$$q_{1/2} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{se } n \text{ for par} \end{cases}.$$

Note-se que, quando n é par, o valor da mediana é uma média dos dois valores "centrais" da amostra, podendo não ser nenhum dos valores da amostra. Isto significa que o valor da mediana obtido com o comando `Median` pode não coincidir com o valor obtido usando o comando `Quantile` com os parâmetros por defeito. Uma situação idêntica se passa com os quartis quando calculados com os comandos `Quantile` e

Medidas de localização

Moda amostral

Definição

A moda amostral é o valor (ou valores) da amostra que ocorre com maior frequência. Pode haver uma ou mais modas amostrais.[a]

[a] No caso de dados de variáveis contínuas, agrupados em classes, fala-se na classe modal (como sendo aquela de maior frequência) e, e geral, considera-se como moda o centro dessa classe.

As medidas de localização referidas (média, moda, mediana e quantis) acompanham transformações lineares dos dados; por exemplo, tem-se

$$y_k = a + bx_k \Rightarrow \bar{y} = a + b\bar{x}.$$

Medidas de dispersão

Variância amostral e desvio padrão amostral

Definição

A variância amostral, que representamos por $\mathrm{var}(\mathbf{x})$ ou s^2 é a média (corrigida) dos quadrados dos desvios dos dados em relação à média amostral, sendo dada pela fórmula

$$\mathrm{var}(\mathbf{x}) = s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2.$$

Nota: A razão de considerarmos a média corrigida – isto é, de dividirmos por $n-1$ e não por n – será explicada posteriormente; note-se, no entanto que, para valores de n grandes, o uso de n ou $n-1$ no denominador fará pouca diferença.

Definição

A raiz quadrada da variância amostral, s, é chamada desvio padrão amostral.

Variância de dados transformados

Quando os dados sofrem uma transformação linear

$$y_k = a + bx_k$$

tem-se a seguinte expressão para a variância dos dados transformados:

$$\mathrm{var}(\mathbf{y}) = b^2\,\mathrm{var}(\mathbf{x}).$$

Note-se que a constante aditiva, que está associada a uma mudança de localização dos dados, não afeta a variância; já a constante multiplicativa, associada a uma mudança de escala nos dados, afeta a variância.

Medidas de dispersão

Amplitude da amostra e amplitude interquartis

Definição

A amplitude amostral, denotada por R, é a diferença entre o valor máximo e o valor mínimo da amostra, isto é, é dada por

$$R = x_{(n)} - x_{(1)}.$$

Definição

A amplitude interquartis da amostra, AIQ, é a diferença entre o terceiro e o primeiro quartis amostrais, isto é, é dada por

$$AIQ = q_{3/4} - q_{1/4}.$$

Outliers

Os *outliers* são observações extremas, demasiado afastadas dos 50% valores centrais da amostra (ordenada).

Definição

Um dado x_k da amostra é dito um *outlier*, se verificar

$$x_k < q_{1/4} - 1.5AIQ \quad \text{ou} \quad x_k > q_{3/4} + 1.5AIQ.$$

Esse *outlier* será considerado um *outlier* severo, se verificar

$$x_k < q_{1/4} - 3AIQ \quad \text{ou} \quad x_k > q_{3/4} + 3AIQ.$$

Um *outlier* que não seja um *outlier* severo, diz-se um *outlier* moderado.

Nota Os *outliers* são elementos a que devemos dar especial atenção, porque podem desvirtuar totalmente uma análise estatística. A primeira coisa a fazer é ver se não houve erros de registo. Se tal não aconteceu, será aconselhável fazer uma análise com e sem esses dados, de forma a avaliar o efeito que eles têm na análise e na interpretação dos resultados.

Exemplo

Considere-se a seguinte amostra, já ordenada

$$1, 700, 800, 900, 1000, 1000, 1100, 1200, 2500$$

Temos

- Os quartis (calculados usando a função `Quartiles`) são $q_{1/4} = 775$, $q_{1/2} = 1000$ e $q_{3/4} = 1125$, pelo que a amplitude inter-quartis é $AIQ = 1125 - 775 = 350$.
- $775 - 1.5 \times 350 = 775 - 525 = 250$
- $775 - 3 \times 350 = 775 - 1050 = -275$
- $1125 + 1.5 \times 350 = 1125 + 525 = 1650$
- $1125 + 3 \times 350 = 1125 + 1050 = 2175$

Logo,há dois *outliers*: 1, que é um *outlier* moderado e 2500, que é um *outlier* severo.

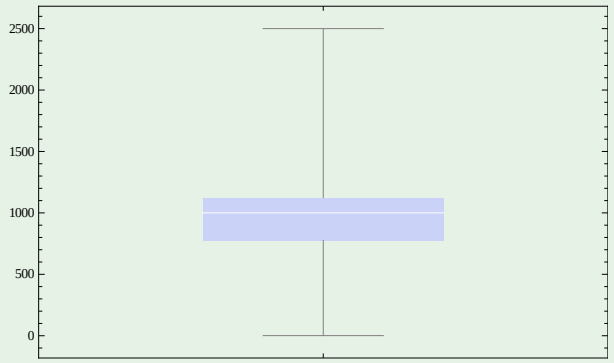
Diagramas de caixa-com-bigodes

O chamado diagrama de caixa-com-bigodes ou diagrama de extremos-e-quartis é um gráfico com duas caixas centrais limitadas pelos três quartis e umas linhas que se prolongam para fora das caixas até ao valor mínimo e máximo da amostra ou, se quisermos evidenciar os *outliers*, até ao menor e maior valores que não sejam *outliers*, sendo os *outliers* identificados por símbolos próprios.

O Mathematica tem um comando específico para desenhar diagramas de caixas-com-bigodes: `BoxWhiskerChart`. Podemos desenhar diagramas com ou sem indicação de *outliers*. O Mathematica usa símbolos diferentes (pontos de 'cor" diferente) para os *outliers* severos e moderados.

Exemplo

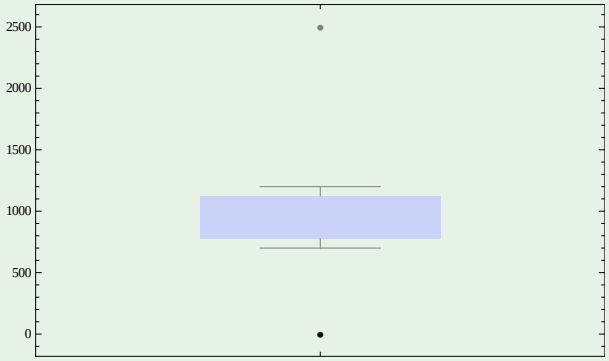
O diagrama de caixa-com-bigodes, sem indicação dos *outliers*, para o exemplo anterior, obtido com o comando `BoxWhiskerChart`, seria:



Nota Recorde que o valor mínimo da amostra é 1 e o valor máximo é 2500.

Exemplo

O diagrama de caixa-com-bigodes, com indicação dos *outliers*, para esse mesmo exemplo seria:



Nota Observe a forma diferente de representar os *outliers*: 1 é um *outlier* moderado e 2500 é um *outlier* severo.

Medidas de forma

Coeficiente de assimetria

A forma de uma distribuição dos dados é medida em duas perspectivas: assimetria e achatamento.

Definição

O momento central (empírico) de ordem r, denotado por m_r, é definido por:

$$m_r = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^r, \quad \text{para } r = 1, 2, \ldots.$$

Definição

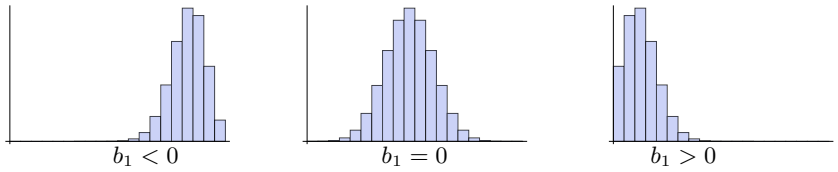
O coeficiente de assimetria (empírico), denotado por b_1, é dado por

$$b_1 = \frac{m_3}{m_2^{2/3}} = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^3}{\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right)^{2/3}}.$$

Coeficiente de assimetria

O coeficiente de assimetria dá indicação sobre o peso relativo das *caudas* (direita e esquerda).

Note-se que a assimetria pode ser negativa, positiva ou nula.



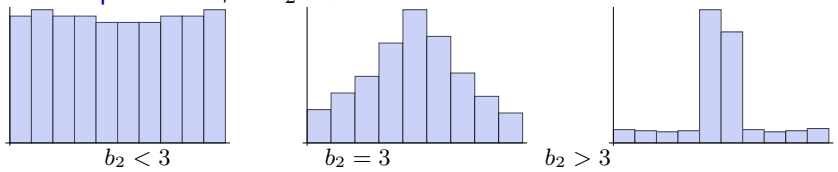
Definição

O coeficiente de achatamento (ou de curtose) empírico, denotado por b_2, é dado por

$$b_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^4}{\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right)^2}$$

Curtose

A curtose dá indicação sobre o maior ou menor achatamento da distribuição dos dados (ou, de outro modo, sobre a maior ou menor concentração destes à volta do valor médio, ou ainda, se quisermos, sobre a existência de caudas "leves" ou de caudas "pesadas").
Note-se que o achatamento é sempre positivo.
Dizemos que a distribuição é platicúrtica, se $b_2 << 3$, mesocúrtica, se $b_2 \approx 3$ e leptocúrtica, se $b_2 >> 3$.



Nota Só faz sentido falar de achatamento para distribuições que sejam (quase) simétricas.