Organização e Tratamento de Dados

Departamento de Matemática - UA

Maria da Conceição Lopes Costa

Estatística (Oxford Dictionary of Statistical Terms): ciência que trata da recolha, organização, apresentação, análise e interpretação de dados.

Estatística Descritiva: conjunto de métodos estatísticos que visam sumariar e descrever os atributos mais proeminentes aos dados.

Estatística Inferencial: conjunto de métodos estatísticos que visam caracterizar ou inferir acerca de uma população a partir de uma parte dela (a amostra).

- A população é um conjunto de unidades individuais, que podem ser pessoas, animais, resultados experimentais, etc., com uma ou mais características em comum, que se pretendem analisar.
- A amostra é a parte da população que é observada com o objectivo de obter informação para estudar a característica pretendida.

A amostra deve ser representativa da população.

Muitas vezes a amostra é entendida como o conjunto das observações das variáveis de interesse.

exemplo

Estudando a característica **altura** na população portuguesa e selecionando aleatoriamente 100 indivíduos desta população podemos entender a amostra como sendo constituida pelas alturas dos 100 indivíduos.

A uma característica comum que assume valores diferentes de indivíduo para indivíduo chamamos variável.

- Uma variável é uma característica de um indivíduo à qual se pode atribuir um número ou uma categoria.
- Cada um dos indivíduos (ou objectos ou entidades) relativamente aos quais se recolhe a informação designa-se por unidade estatística.

Consoante os valores que uma variável pode assumir, as variáveis são classificadas em Qualitativas ou Quantitativas.

Variáveis Qualitativas ou Categóricas referem-se a características que não se podem contar nem medir mas apenas classificar, podendo assumir várias categorias.

- Variáveis Qualitativas Nominais: se apenas se podem atribuir categorias.

exemplos

nacionalidade de um indivíduo, região geográfica, área de estudo, cor do cabelo

- Variáveis Qualitativas Ordinais: se pode ser estabelecida uma ordem subjacente.

exemplos

grau de satisfação com atendimento (com as categorias *nada satisfeito*, *pouco satisfeito*, *satisfeito*, *muito satisfeito*)

Variáveis Quantitativas ou Numéricas referem-se a características que se podem quantificar e resultam em geral de contagens e medições.

- Variáveis Quantitatitvas Discretas: assumem valores em conjuntos numéricos finitos ou infinitos mas numeráveis

  Isto é, é possível distinguir os diferentes valores que a variável pode assumir. Estas variáveis resultam normalmente de contagens.

exemplos

- número de alunos numa sala de aula
- número de filhos por casal
- número de assoalhadas de um apartamento
- preço de um artigo no supermercado (em cêntimos)

Variáveis Quantitativas ou Numéricas referem-se a características que se podem quantificar e resultam em geral de contagens e medições.

- Variáveis Quantitativas Contínuas: podem assumir qualquer valor dentro de um qualquer intervalo de números reais

  Do ponto de vista estatístico, não se espera observar valores repetidos dentro de uma mesma amostra.

exemplos

- altura de um indivíduo
- comprimento de um terreno
- peso de um corpo
- temperatura
- tempo de espera num atendimento
- tempo de vida de um equipamento

Tarefa

Caracterize quanto ao tipo cada uma das seguintes variáveis

1. Número de pares de sapatos na montra de uma loja
2. Cor do cabelo da primeira pessoa que encontra ao sair da sala
3. Idade da pessoa da alínea anterior
4. Género da pessoa da alínea anterior
5. Grau de escolaridade da pessoa das alíneas anteriores
6. Número de livros que leu por completo no ano anterior
7. Tempo que demora a ler um livro
8. Número de mensagens que recebe por dia, no telemóvel
9. Tempo que demora a responder a uma mensagem, no telemóvel
10. Tempo de duração de uma chamada telefónica

Organização de Dados em Tabelas e Gráficos

Tabelas e Gráficos para Variáveis Qualitativas

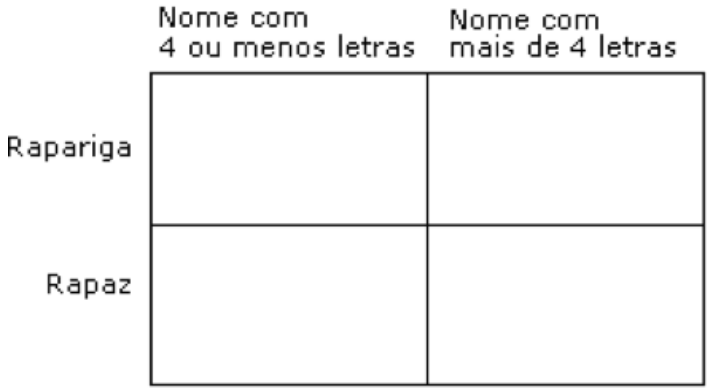
- Diagrama de Venn
- Diagrama de Carroll
- Esquemas de contagem (*tally charts*)
- Tabela de Frequências
- Gráfico de Pontos
- Pictograma
- Gráfico de Frequências (ou Gráfico de Barras)
- Gráfico Circular

Diagrama de Venn

Num diagrama de Venn (ou de Euler) os objectos são classificados de forma rápida. A representação baseia-se nas características que têm em comum.

Diagrama de Carroll

Um diagrama de Carroll é uma tabela rectangular de duas entradas. Os dados são classificados de acordo com dois critérios, em simultâneo.



Esquemas de Contagem

Este tipo de esquemas permite registar e contar os dados à medida que são recolhidos.

Figura: Cor dos olhos dos alunos de uma turma

Tabela de Frequências

Para dados qualitativos, uma tabela de frequências tem normalmente apenas 3 colunas:

1ª : As diferentes categorias da amostra

2ª : A frequência absoluta, n_i de cada categoria ou classe

A frequência absoluta de uma categoria é o número de observações ou elementos da amostra pertencentes a essa categoria.

3ª : A frequência relativa, f_i de cada categoria

A frequência relativa de uma categoria é a frequência absoluta a dividir pela dimensão da amostra. A dimensão da amostra é o número total de observações que a constituem, n.

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

É comum incluir uma última linha com os totais ao longo das colunas, o que permite verificar que

- a soma das frequências abolutas é igual à dimensão da amostra, n;
- a soma das frequências relativas é igual a 1.

Categorias	Freq. Abs.	Freq. Rel.
Castanhos	15	0,625
Pretos	3	0,125
Verdes	2	0,083
Azuis	4	0,167
Total	24	1

Figura: Tabela de frequências da cor dos olhos dos alunos de uma turma

Gráfico de Pontos

As categorias são assinaladas sob um eixo horizontal, de forma equidistante. Sobre cada categoria representa-se um ponto por cada dado, na vertical.

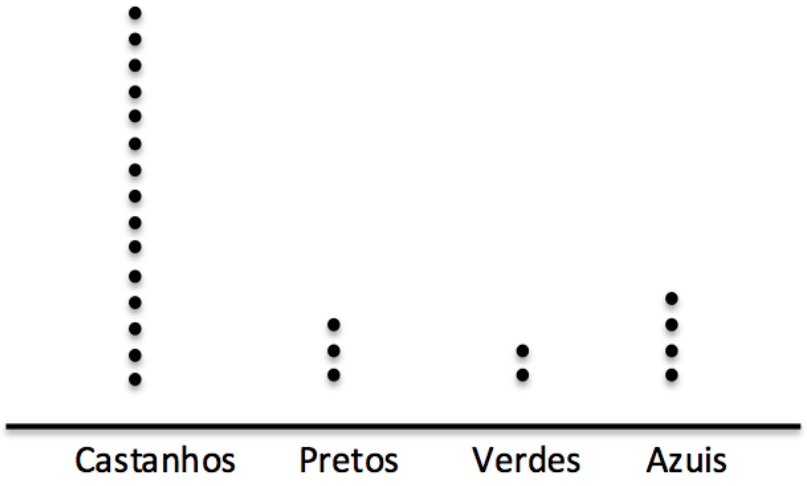


Figura: Gráfico de pontos da cor dos olhos dos alunos de uma turma

Pictograma

É semelhante a um grafico de pontos mas em vez de pontos são utilizados símbolos alegóricos às variáveis em estudo. Os símbolos devem ser todos do mesmo tamanho e cada símbolo representar um dado (ou, representando mais do que um dado, essa informação ser mencionada junto ao gráfico).

Figura: Pictograma da cor dos olhos dos alunos de uma turma

Gráfico de Frequências ou Gráfico de Barras

A altura de cada barra é igual à frequência absoluta (ou relativa) da categoria correspondente. Estes gráficos têm sempre dois eixos, o eixo das categorias e o eixo das frequências. O eixo das frequências deve conter uma escala onde se podem ler as frequências e a indicação do tipo de frequência representada (n_i ou f_i).

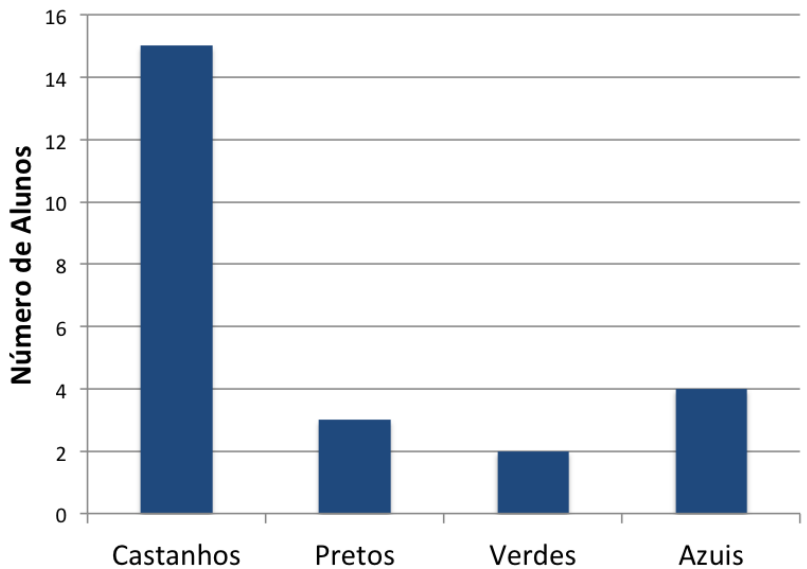


Figura: Gráfico de frequências da cor dos olhos dos alunos de uma turma

Gráfico Circular

Um círculo é dividido em setores circulares cuja amplitude é proporcional à frequência (absoluta ou relativa) das diferentes categorias. Cada setor representa uma fração do total dos dados e é usual utilizar percentagens para indicar a fração correspondente a cada setor ou categoria. As categorias devem estar identificadas no gráfico, através da correspondente designação ou de uma legenda de cores.

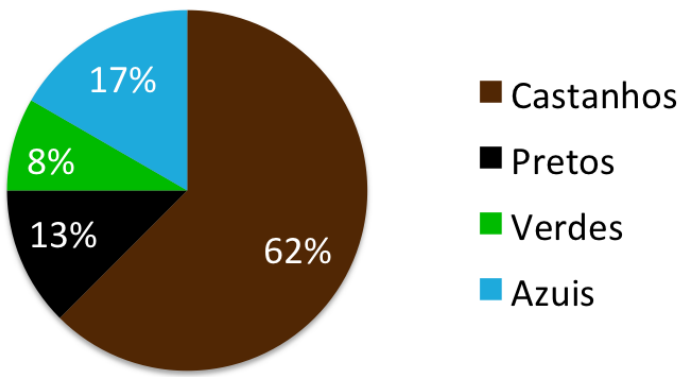


Figura: Gráfico circular da cor dos olhos dos alunos de uma turma

Tabelas e Gráficos para Variáveis Quantitativas Discretas

Todas as representações mencionadas atrás continuam a ser adequadas para variáveis quantitativas discretas. Apenas o gráfico circular é desaconselhado uma vez que uma variável quantitativa tem uma ordem natural e faz mais sentido fazer a sua representação através de um gráfico que contenha um eixo linear. Considerando estas características de uma variável quantitativa discreta as representações mais adequadas são:

- Tabela de Frequências
- Gráfico de Frequências (ou Gráfico de Barras)
- Diagrama de Caule e Folhas
- Diagrama de Extremos e Quartis (Caixa de Bigodes)
- Gráfico de Linhas

Tabela de Frequências

Para dados quantitativos discretos, uma tabela de frequências tem normalmente 5 colunas:

1ª : As diferentes observações distintas da amostra, ordenadas

2ª : A frequência absoluta, n_i, de cada observação distinta

3ª : A frequência relativa, f_i, de cada observação distinta

4ª : A frequência absoluta acumulada, N_i, de cada observação distinta

Soma das frequências absolutas até à observação de ordem i.

5ª : A frequência relativa acumulada, F_i, de cada observação distinta

Soma das frequências relativas até à observação de ordem i.

Perguntou-se a um grupo de 30 pessoas quantas vezes tinham ido ao cinema no último mês. Com os resultados foi construída a seguinte tabela de frequências:

Nº de idas ao cinema	Freq. Abs. (n_i)	Freq. Rel. (f_i)	Freq. Abs. Acum. (N_i)	Freq. Rel. Acum. (F_i)
0	8	0,267	8	0,267
1	12	0,400	20	0,667
2	5	0,167	25	0,833
3	3	0,100	28	0,933
4	2	0,067	30	1
Total	30	1		

Figura: Tabela de frequências da variável número de idas ao cinema no último mês.

Com base na tabela de frequências anterior é possível dar resposta de forma quase imediata a questões como:

1. Quantas pessoas foram apenas uma vez ao cinema? E não foram vez nenhuma?
2. Qual a percentagem de pessoas que foi três vezes ao cinema no último mês?
3. Qual a percentagem de pessoas que foi, no máximo, duas vezes ao cinema?
4. Qual a percentagem de pessoas que foi pelo menos uma vez ao cinema? E pelo menos duas?

Gráfico de Frequências

À semelhança do que acontecia para variáveis qualitativas, o gráfico de frequências é a imagem gráfica da tabela de frequências. Para variáveis quantitativas discretas são representadas as frequências absolutas ou frequências relativas em função das observações distintas da amostra, ordenadas.

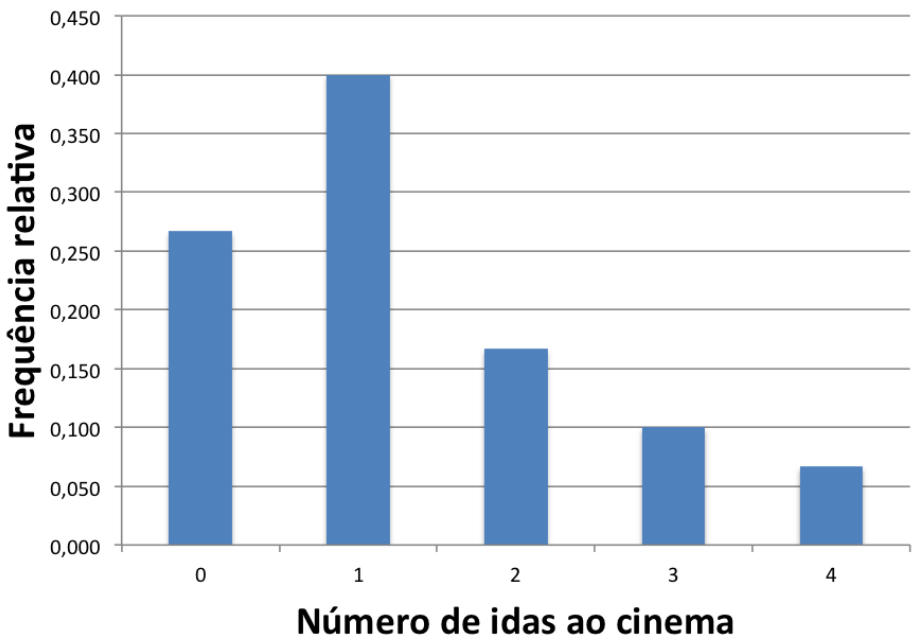


Figura: Gráfico de frequências do número de idas ao cinema.

Cuidados a ter na representação dos dados:

- O tipo de frequência representado num gráfico de frequências deve ser indicado com clareza no eixo das frequências.
- A escala utilizada no eixo vertical pode por vezes distorcer a leitura do gráfico:

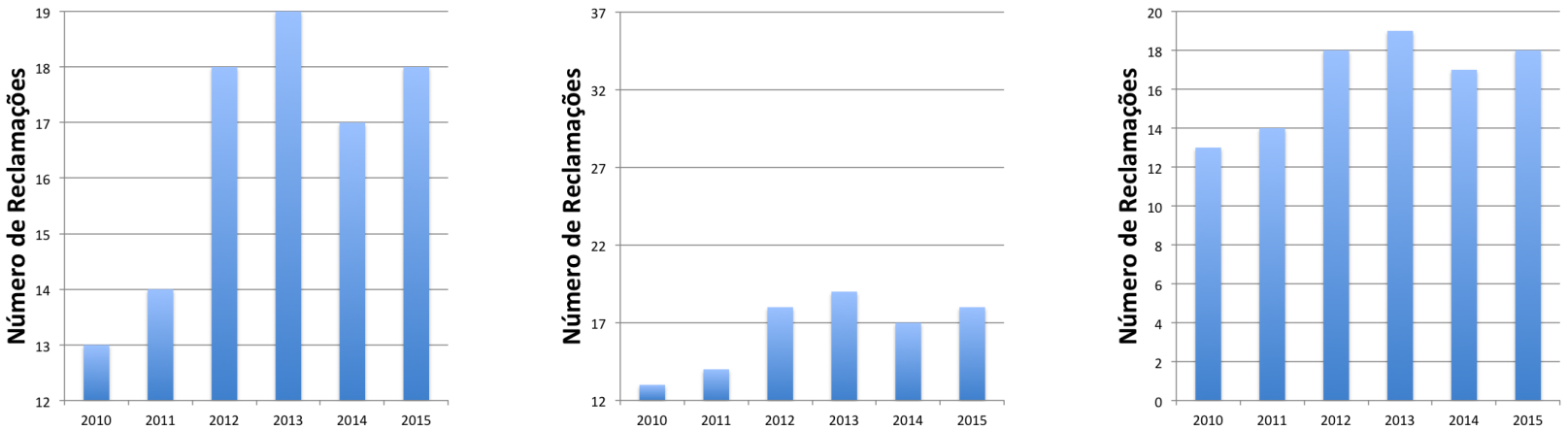


Figura: Número de reclamações recebidas por ano num serviço de atendimento ao público.

- Nem todos os gráficos com barras são gráficos de frequências! Um gráfico de frequências tem que representar frequências.

  No exemplo anterior, as colunas representam os dados, constítuidos pelo número de reclamações por ano, desde 2010 a 2015:

  $$13, 14, 18, 19, 17, 18$$

  Estes dados formam a amostra que tem 6 observações e apenas uma observação repetida. Como tal, o gráfico representado não é um gráfico de frequências: o eixo vertical não representa o número de vezes que uma determinada observação se repete na amostra! Este é um gráfico que apenas representa os dados e não se pode chamar de gráfico de frequências.

Diagrama de Caule e Folhas

Cada observação da amostra é dividida em duas partes: a parte da direita designada por folha e que deve conter apenas um algarismo e a parte da esquerda, designada por caule e que pode conter qualquer número de algarismos.

O diagrama de caule e folhas é uma representação entre a tabela e o gráfico. Tem a grande vantagem de se poder reconstruir a amostra a partir da sua observação e, em simultâneo, ter ideia da distribuição de frequências da amostra (imaginando uma rotação de 90º no sentido anti-horário).

2	3 7
3	5 7 7 8 8 8 9
4	0 1 4 6 6 7 8 8
5	2 3 4 5 7 9
6	0 2 3 5 8

2|3 significa 23

Figura: Tempo sem respirar (em s) num grupo de 28 alunos de natação.

Se considerar como folha o algarismo das unidades resultar em poucas classes podemos subdividir cada linha em duas ou em cinco linhas.

Divisão em duas linhas: a linha 1 tem as folhas 0 a 4 (cinco valores possíveis) e a linha 2 tem as folhas 5 a 9 (os restantes cinco valores). Para distinguir a primeira da segunda linha, normalmente o caule da segunda linha é identificado com um $*$.

2	0 2 2 3 3 4
2*	5 5 6 6 6 6 7 7 7 7 9
3	0 0 0 1 1 1 2 2 4
3*	7 7 8 8 8 9
4	0 1 1 1 2 3 3 4
4*	5

2 | 0 significa 20

Figura: Idades de 42 participantes de uma turma de Pilates.

Divisão em cinco linhas: a linha 1 tem as folhas 0 e 1; a linha 2 tem as folhas 2 e 3 e o caule é identificado com um *t* (de *two and three*); a linha 3 tem as folhas 4 e 5 e o caule é identificado com um *f* (de *four and five*); a linha 4 tem as folhas 6 e 7 e o caule é identificado com um *s* (de *six and seven*); a linha 5 tem as folhas 8 e 9 e o caule é identificado com um $*$.

0^f	5 5 5 5 5 5
0^s	6 6 6 6 6 7 7 7 7 7 7 7 7
0^*	8 8 8 8 9 9 9 9
1.	0 0 0 1 1 1
1^t	2 2 3 3
1^f	4

0 | 5 significa 5

Figura: Idades de 40 crianças de uma escola do 1º e 2º ciclos.

Diagrama de Extremos e Quartis ou Caixa de Bigodes

Para a construção da caixa de bigodes são necessárias cinco medidas amostrais: mínimo, máximo, 1º e 3º quartis e mediana. A sua construção e vantagens serão abordadas após o estudo destas medidas amostrais.

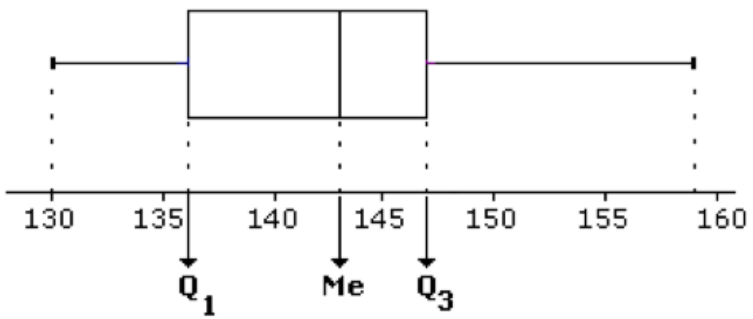


Figura: Altura de um aluno de uma escola do 1º ciclo (em cm).

Gráfico de Linhas

Um gráfico de linhas é um caso especial de um gráfico de dispersão que representa a forma como uma variável evolui relativamente a outra. É particularmente adequado quando no eixo horizontal se representa o tempo e no eixo vertical a evolução da característica em estudo ao longo do tempo.

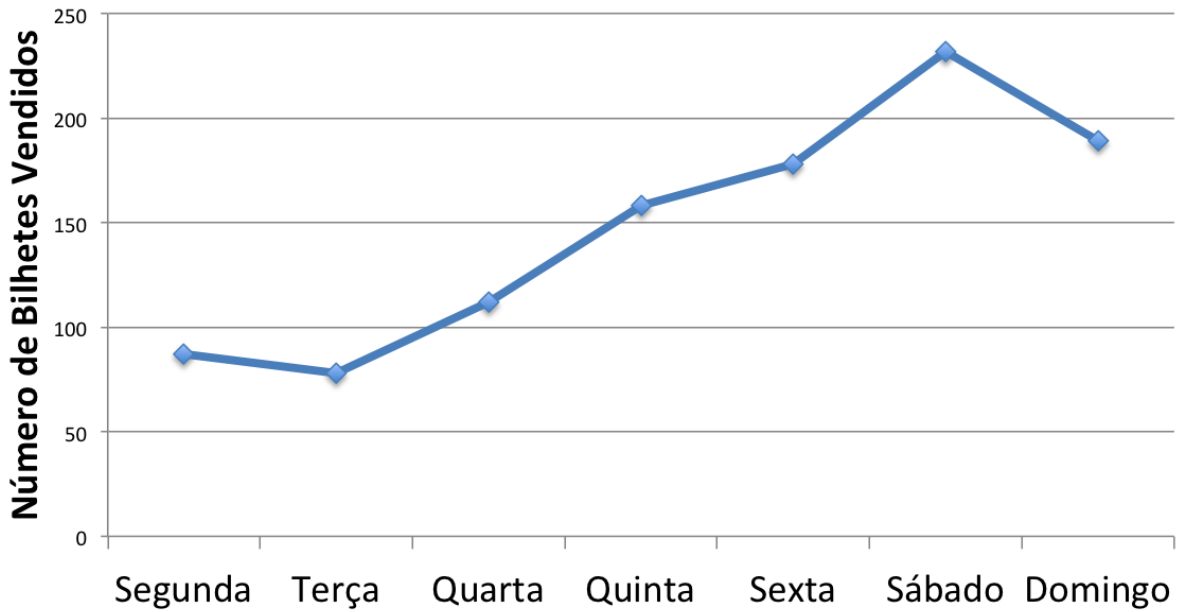


Figura: Número de bilhetes de cinema vendidos numa semana.

Tabelas e Gráficos para Variáveis Quantitativas Contínuas

- Tabela de Frequências
- Histograma
- Diagrama de Caule e Folhas
- Diagrama de Extremos e Quartis (Caixa de Bigodes)
- Gráfico de Linhas

Apenas a tabela de frequências e o histograma são representações diferentes das que foram mencionadas para variáveis quantitativas discretas e serão apresentadas de seguida.

Tabela de Frequências

Numa amostra de dados contínuos espera-se que o número de observações distintas seja muito elevado. Assim, a metodologia apresentada para dados discretos (considerar como classes todas as observações distintas da amostra) não é adequada.

Neste caso devem-se dividir os valores da amostra em classes na forma de intervalos e usar o procedimento já conhecido tendo por base esses intervalos.

Tabela de Frequências

Para dados contínuos, uma tabela de frequências terá então 6 colunas:

- **1ª** : As diferentes classes da amostra, na forma de intervalos
- **2ª** : O representante ou marca da classe: o ponto médio de cada intervalo
- **3ª** : A frequência absoluta, n_i, de cada classe
- **4ª** : A frequência relativa, f_i, de cada classe
- **5ª** : A frequência absoluta acumulada, N_i, de cada classe
- **6ª** : A frequência relativa acumulada, F_i, de cada classe

Analisou-se a altura dos alunos de uma escola do 1º ciclo e com base nos dados foi construída a seguinte tabela de frequências:

Classes	Representante da Classe	Freq. Abs. (n_i)	Freq. Rel. (f_i)	Freq. Abs. Acum. (N_i)	Freq. Rel. Acum. (F_i)
[130, 135[	132,5	7	0,14	7	14%
[135, 140[	137,5	9	0,18	16	32%
[140, 145[	142,5	11	0,22	27	54%
[145, 150[	147,5	14	0,28	41	82%
[150, 155[	152,5	5	0,10	46	92%
[155, 160[	157,5	4	0,08	50	100%
Total		50	1		

Figura: Tabela de frequências da variável altura de um aluno de uma escola do 1º ciclo (em cm).

Histograma

A partir da tabela de frequências pode ser construído um gráfico semelhante ao gráfico de frequências em que a altura de cada coluna corresponde à frequência de cada classe. Este gráfico designa-se por histograma. Num histograma não há espaçamento entre colunas.

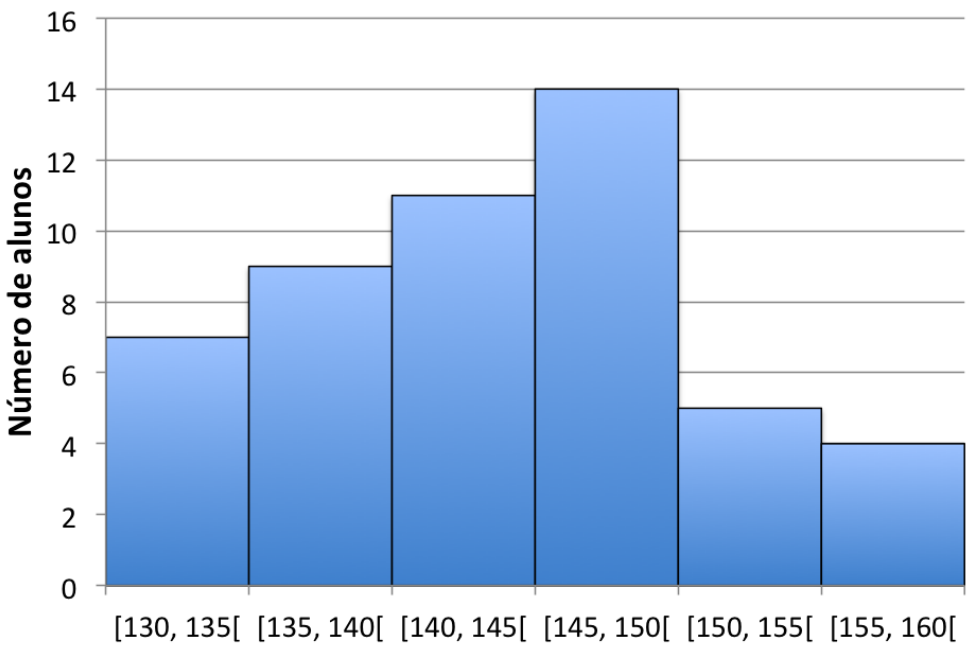


Figura: Altura de um aluno de uma escola do 1º ciclo (em cm).

- Pode ser construído um histograma cumulativo se forem usadas as frequências relativas acumuladas (absolutas ou relativas). Esta representação é útil para o cálculo aproximado de certas medidas amostrais tais como a mediana ou os quartis.
- Para a construção do histograma é necessário começar pela formação das classes. Para a escolha do número de classes convém saber que:

1. quanto maior o número de classes, maior é a variabilidade que se observa; quanto menor, maior quantidade de informação se perde;
2. a amplitude das classes e o início de cada classe devem ser um valores inteiros, preferencialmente;
3. é habitual utilizar-se a Regra de Sturges[1] mas a maior parte dos programas de Estatística decide o número de classes de forma automática.

[1] Segundo a Regra de Sturges, o número de classes, L, é calculado através de $L = 1 + 3,322 \log n$, em que n é a dimensão da amostra.

Medidas de Localização

Dizem respeito à localização dos dados, isto é, onde se situa a amostra. Apenas vamos considerar as seguintes medidas:

- Medidas de localização central
  - Média
  - Mediana
  - Moda
- Outras medidas de localização
  - Mínimo
  - Máximo
  - Quartis

Convém referir que todas estas medidas dizem respeito a variáveis quantitativas. Para variáveis qualitativas apenas a moda pode ser usada.

Para definir as medidas amostrais referidas convém introduzir a seguinte notação para representar os dados.

Considere-se uma amostra de dimensão n. Os seus elementos ou observações são representados por:

$$x_1, x_2, x_3, \cdots, x_n$$

Esta primeira notação não pressupõe a ordenação dos dados. Se as observações da amostra forem ordenadas, passamos a identificá-las por:

$$x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n)}$$

Normalmente numa amostra há observações repetidas (em particular tratando-se de variáveis quantitativas discretas). Os valores

$$x_1^*, x_2^*, x_3^*, \cdots, x_k^*$$

com $k \leq n$, representam as observações distintas da amostra, ordenadas.

Média, $\overline{x}$

A média amostral ou simplesmente média representa-se por $\overline{x}$ e é a soma de todas as observações a dividir pela dimensão da amostra, n.

- Quando se tem acesso a **todas as observações da amostra**:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Quando temos **dados discretos agrupados numa tabela de frequências** a média pode ser calculada a partir de uma das seguintes expressões:

$$\overline{x} = \frac{x_1^* n_1 + x_2^* n_2 + \cdots + x_k^* n_k}{n} \quad \text{ou} \quad \overline{x} = x_1^* f_1 + x_2^* f_2 + \cdots + x_k^* f_k$$

- Quando temos **dados contínuos agrupados em classes**, não temos acesso a cada uma das observações da amostra. Neste caso, para obter um valor aproximado para a média podemos usar qualquer uma das duas últimas expressões acima, substituindo $x_1^*, x_2^*, x_3^*, \cdots, x_k^*$ pelos representantes das classes.

Alguns comentários acerca da interpretação da média:

- A média é uma medida que utiliza a soma de todas as observações e é útil em problemas que envolvem essa quantidade, mas não permite concluir acerca de observações individuais.

exemplo

Sabendo que numa empresa há 10 trabalhadores que em média ganham 1200 euros sabemos que são necessários 12 000 euros para pagar os salários mensais. Mas não sabemos como se distribuem os salários: podem todos ganhar 1200 euros ou podem 3 receber 2400 euros e os restantes 600 que a média dá o mesmo valor.

- Uma forma de interpretar a média consiste em imaginar uma régua graduada onde se colocam pesos unitários nos pontos correspondentes aos valores das observações.

- Se houver observações repetidas são colocados tantos pesos quanto o valor da frequência absoluta de cada observação. A média será o ponto de equilíbrio ou o *centro de massa* da régua, ou seja, o ponto sob o qual pode ser colocado um eixo que equilibra a régua. Sendo a média um ponto de equilíbrio da amostra um (ou mais) valor(es) distante(s) do centro das observações pode provocar grandes desvios no valor da média. Diz-se por isso que a média é sensível a variações extremas e pode por vezes não representar o grosso das observações.

exemplo

Voltando ao exemplo atrás, se o gerente receber 4800 euros mensais e os restantes 9 trabalhadores apenas 800 euros cada um, a média dá na mesma 1200 euros e este valor não representa o grosso das observações nem a tendência central da amostra.

- A média apenas representa a tendência central da amostra quando a distribuição de frequências da amostra é simétrica e não existem observações extremas.

Mediana, Me

A mediana é o valor que divide a amostra ao meio: metade das observações são menores ou iguais à mediana e a outra metade é superior ou igual à mediana. Para determinar a mediana é necessário ordenar os dados.

- Quando se tem a **amostra ordenada** e a dimensão da amostra é ímpar, a mediana é a observação central. Se a dimensão da amostra é par, a mediana é dada pela média das duas observações centrais da amostra ordenada.
- Quando temos **dados discretos agrupados numa tabela de frequências** a mediana é o primeiro valor da amostra em que se atinge ou ultrapassa os 50% na coluna da frequência relativa acumulada, F_i.
- Quando temos **dados contínuos agrupados em classes**, não podemos calcular a mediana de forma exacta mas podemos identificar a classe mediana como sendo a classe que pela primeira vez ultrapassa os 50% na coluna da frequência relativa acumulada, F_i.

Moda, Mo

A moda de uma amostra é a observação mais frequente dessa amostra, ou seja é o valor x_i para o qual n_i ou f_i são máximos.

- Uma amostra pode ter mais do que uma moda: pode haver mais do que uma observação com o mesmo valor máximo de frequência absoluta.
- Quando se trata de uma variável quantitativa contínua a moda pode não ter grande significado e apenas existir por arredondamento ou insuficiência na precisão da leitura dos dados.

Mínimo e Máximo

O mínimo é o primeiro valor da amostra ordenada, $x_{(1)}$.

O máximo é o último valor da amostra ordenada, $x_{(n)}$.

Quartis

Os quartis dividem a distribuição dos dados em quatro partes, cada uma com igual percentagem de observações. Para o cálculo dos quartis o procedimento a seguir é:

1. Ordenar os dados e calcular a mediana, Me;
2. O 1^o quartil ou Q_1 é a mediana dos dados que ficam à esquerda de Me;
3. O 2^o quartil ou Q_2 coincide com a Me;
4. O 3^o quartil ou Q_3 é a mediana dos dados que ficam à direita de Me.

Quartis

- Quando se tem a **amostra ordenada** e a dimensão da amostra é par

exemplo

$2, 2, 3, 4, 5, 5, 5, 8$: $Q_1 = \frac{2+3}{2} = 2,5$ $Q_2 = Me = \frac{4+5}{2} = 4,5$ $Q_3 = \frac{5+5}{2} = 5$

- Quando se tem a **amostra ordenada** e a dimensão da amostra é ímpar, o elemento central considera-se como pertencente às duas metades da amostra para o cálculo dos quartis

exemplo

$2, 2, 3, 4, 5, 5, 5, 8, 8$: $Q_1 = 3$ $Q_2 = Me = 5$ $Q_3 = 5$

- Quando temos **dados discretos agrupados numa tabela de frequências** o 1^o quartil é o primeiro valor da amostra em que se atinge ou ultrapassa os 25% na coluna da frequência relativa acumulada, F_i. O 2^o quartil e o 3^o quartil são os mais pequenos valores da amostra que apresentam F_i de pelo menos 50% e 75%, respectivamente.

- Quando temos **dados contínuos agrupados em classes**, não podemos calcular os quartis de forma exacta mas apenas identificar as classes dos quartis, de forma semelhante ao caso anterior.

Medidas de Dispersão

Dizem respeito à variabilidade dos dados, isto é, se variam muito ou pouco. Apenas vamos considerar as seguintes medidas:

- Amplitude
- Distância inter-quartil
- Variância
- Desvio padrão

Amplitude

A amplitude da amostra é a diferença entre o máximo e o mínimo. É uma medida simples que fornece a amplitude de 100% das observações da amostra.

Distância inter-quartil

A distância inter-quartil é dada por $Q_3 - Q_1$ e fornece a amplitude de 50% das observações da amostra, a metade mais central da amostra.

Não é influenciada pela existência de observações de valor extraordinariamente elevado ou reduzido.

É a medida que fornece o comprimento da caixa de bigodes.

Variância e Desvio padrão

A medida de dispersão mais utilizada é o desvio padrão, s, que se obtém a partir da variância, s^2:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

A variância quantifica a variabilidade dos dados em torno da média e calcula-se como a média dos desvios de cada observação relativamente à média amostral (elevados ao quadrado para não haver cancelamento de termos positivos com negativos). Como para o cálculo da variância são considerados os quadrados dos desvios, a variância é uma medida que não tem a mesma unidade de medida que a média. Para criar uma medida de variabilidade na mesma unidade de medida que as observações, define-se o desvio padrão.

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}}$$

Variância corrigida e Desvio padrão corrigido

Em amostras de pequena dimensão é comum usarem-se as versões corrigidas da variância,

$$s_c^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

e desvio padrão,

$$s_c = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}.$$

Diagrama de Extremos e Quartis ou Caixa de Bigodes

Para a construção da caixa de bigodes são necessárias cinco medidas amostrais: mínimo, máximo, 1º e 3º quartis e mediana. A sua construção segue os seguintes passos:

1. Representar um eixo com uma escala graduada, que contenha todas as observações da amostra.
2. Nessa escala identificar a posição das medidas mínimo, máximo, 1º e 3º quartis e mediana.
3. Desenhar um retângulo com um lado coincidente com Q_1 e o outro coincidente com Q_3. Este retângulo terá de comprimento $Q_3 - Q_1$, isto é, a distância inter-quartil.

Diagrama de Extremos e Quartis ou Caixa de Bigodes

4. Na posição do mínimo representar um pequeno traço, correspondente ao bigode inferior. Na posição do máximo representar um pequeno traço, correspondente ao bigode superior.
5. Unir os bigodes à caixa.
6. No interior da caixa, fazer um traço correspondente à posição da mediana.

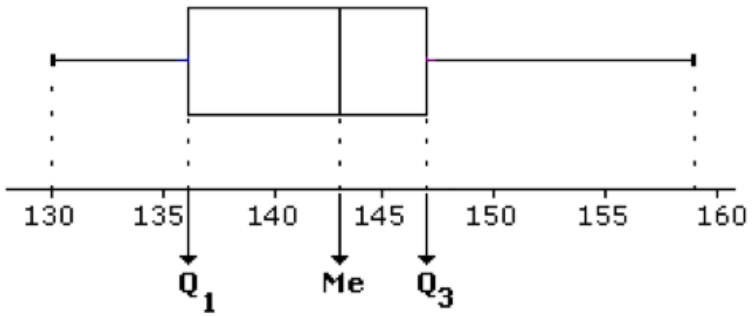


Figura: Altura de um aluno de uma escola do 1º ciclo (em cm).

Alguns comentários acerca das caixas de bigodes:

- As caixas de bigodes podem aparecer na horizontal ou na vertical.
- De um só diagrama é possível recolher informação quanto à localização (mínimo, máximo, mediana e quartis), quanto à dispersão (amplitude e distância inter-quartil) e quanto à forma da distribuição de frequências.
- São também muito úteis para representações comparativas de várias amostras ou de vários grupos dentro de uma amostra.