Estatística I

Estatística Descritiva

Maria João Cortinhal 08/09

Estatística

- **Objectivos**: recolha, compilação, análise e interpretação de dados

Estatística Descritiva e Inferência Estatística

- **Estatística descritiva**: cálculo de medidas, construção de tabelas e de gráficos que sintetizem a informação contida nos dados
- **Inferência estatística**: com base na análise de um conjunto limitado de dados caracterizar o todo
- **População**: conjunto de elementos com uma ou mais características em comum
  - **Ex:** eleitores inscritos na freguesia de Benfica
- **Amostra:** subconjunto representativo da população
  - **Ex:** grupo de eleitores seleccionados ao acaso e contactados pessoalmente

Variáveis

- Observamos características (sexo, cor dos olhos, idade, notas atribuídas, temperaturas,…) e registamo-las, ou seja atribuímos-lhe valores: construímos uma **variável**
- **Classificação das variáveis:**
  - **Qualitativas:** exprimem qualidade
    - **Nominais:** não é possível ordená-las. Ex: cor do cabelo, sexo, comida preferida,…
    - **Ordinais:** podem ser ordenadas. Ex: classificações, nível de escolaridade,..,
  - **Quantitativas:** exprimem quantidade
    - **Discretas:** podem assumir um número finito ou infinito numerável de valores. Ex: idade, número de livros lidos num mês,…
    - **Contínuas:** assumem um número infinito não numerável de valores. Ex:peso, altura, temperaturas, tempo gasto a ler um livro,…

Variáveis discretas

- Foi realizado um inquérito a 100 pessoas acerca do número de vezes que costumavam ir ao cinema, durante um mês . Obtiveram-se os seguintes resultados:
  3,4,3,3,2,5,4,3,3,2,4,3,3,5,4,6,4,1,4,3,7,6,2,4,3,4,7,1,3,2,5,6,1,4,5,6,
  7,2,3,6,4,4,5,2,7,5,2,1,2,6,5,7,5,2,5,3,4,6,3,7,6,3,2,4,6,4,1,2,2,5,4,7,
  3,4,6,4,3,6,1,7,2,2,4,1,5,7,3,5,2,4,5,3,2,4,3,6,5,3,4,4

- Neste caso, a variável em estudo, X, é o número de idas ao cinema durante um mês. A variável X é uma variável quantitativa discreta.
- Os valores registados constituem a amostra. Neste caso a dimensão da amostra é n=100

Tabelas de Frequências

- O primeiro passo é agrupar os dados, ou seja, construir uma tabela de frequências:
  - Ordenar a amostra:

1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7

Pouca utilidade !!!

  - Calcular as frequências absolutas F_i: contar o número de vezes que aparecem na amostra os diferentes valores que a constituem

X	F_i
1	7
2	16
3	20
4	22
5	14
6	12
7	9

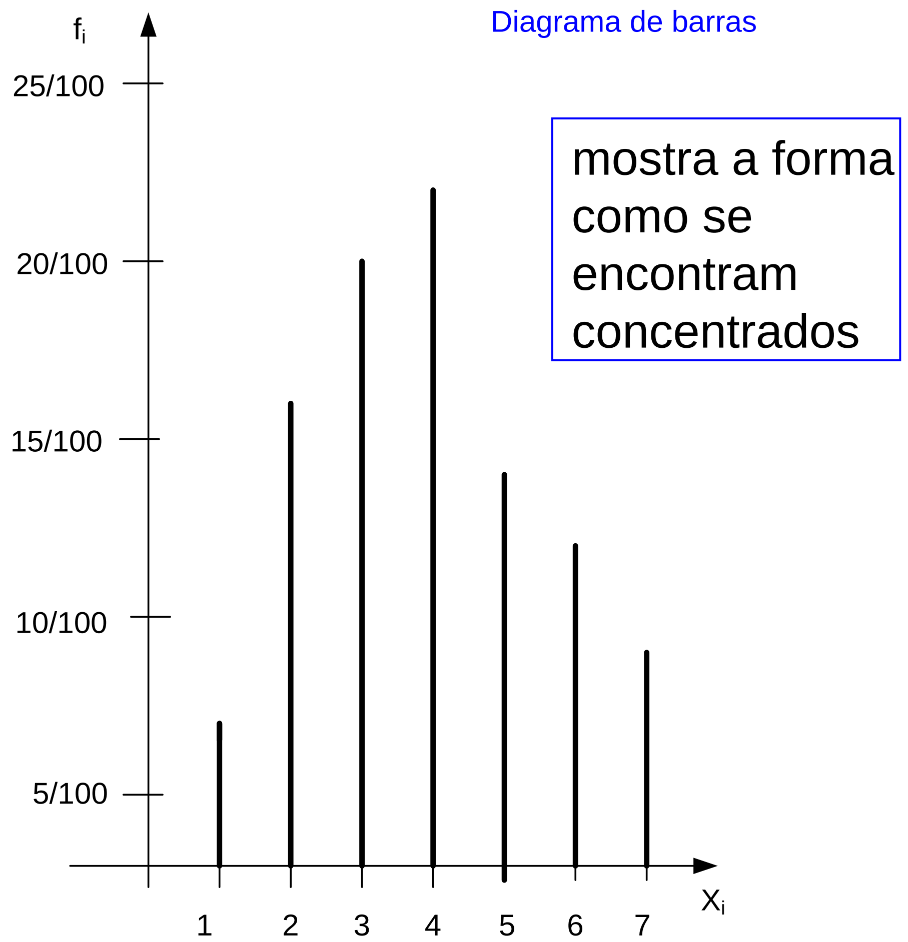
Tabelas de Frequências

- Nas tabelas de frequências é usual ter informação relativa às
  - Frequências relativas: $f_i = F_i / n$
  - Frequências absolutas acumuladas
  - Frequências relativas acumuladas

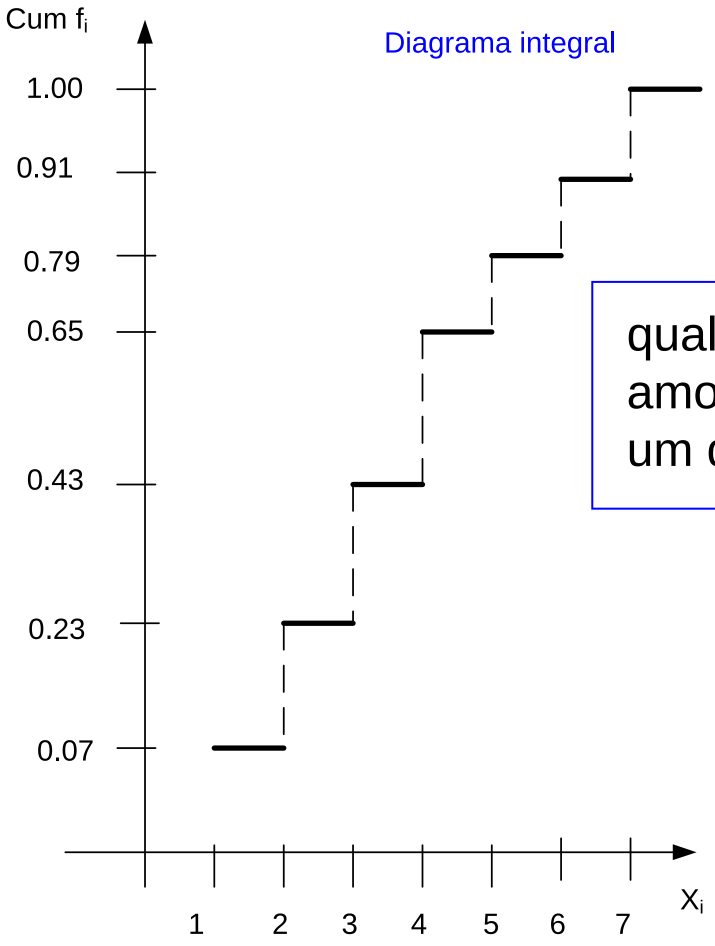
X	F_i	f_i	F_i Acum.	f_i Acum
1	7	0.07	7	0.07
2	16	0.16	23	0.23
3	20	0.20	43	0.43
4	22	0.22	65	0.65
5	14	0.14	79	0.79
6	12	0.12	91	0.91
7	9	0.09	100	1

Representação gráfica da distribuição de frequências

- Após se obter a distribuição de frequências deve proceder-se à representação gráfica da mesma. Para o efeito utiliza-se:
  - Diagrama de barras ou diferencial
  - Diagrama integral ou função distribuição empírica da amostra
- Para o exemplo anterior temos:



Representação gráfica da distribuição de frequências



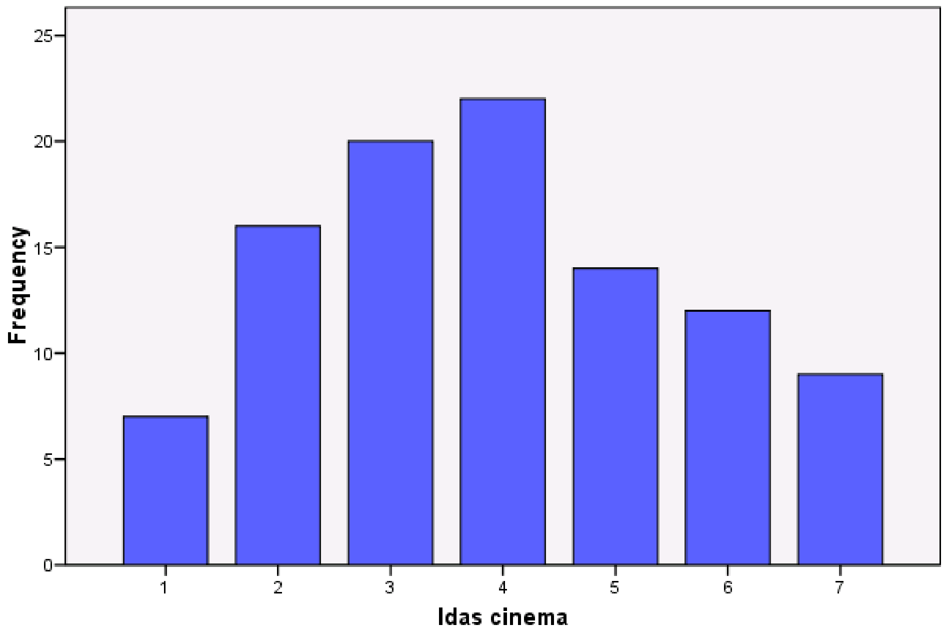
qual a percentagem de valores da amostra inferiores ou superiores a um determinado valor?

SPSS

Idas cinema

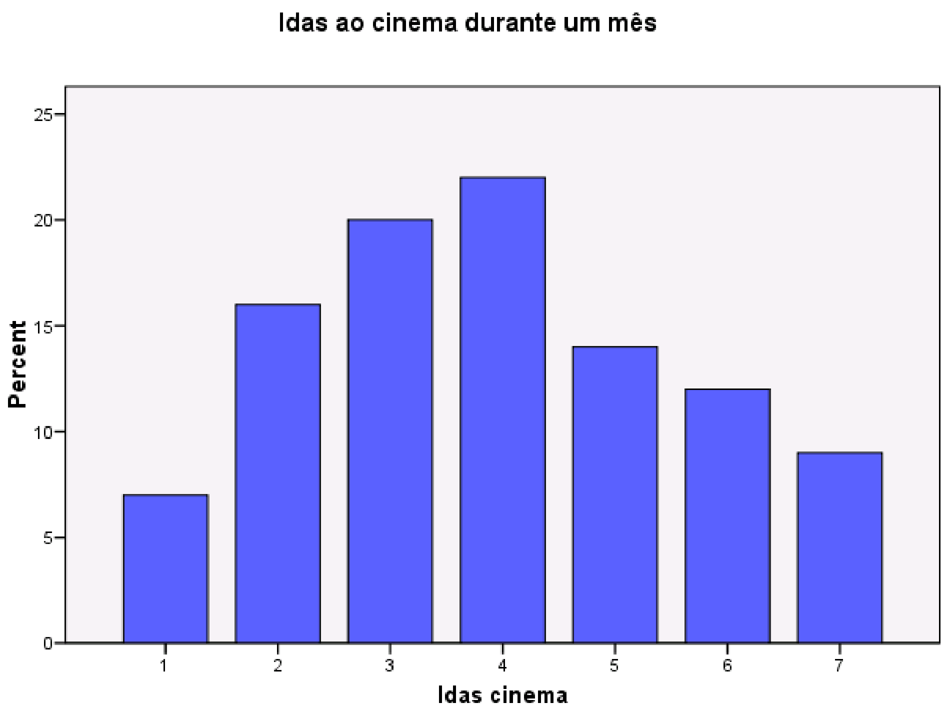
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	7,0	7,0	7,0
	2	16	16,0	16,0	23,0
	3	20	20,0	20,0	43,0
	4	22	22,0	22,0	65,0
	5	14	14,0	14,0	79,0
	6	12	12,0	12,0	91,0
	7	9	9,0	9,0	100,0
	Total	100	100,0	100,0	

Idas ao cinema durante um mês



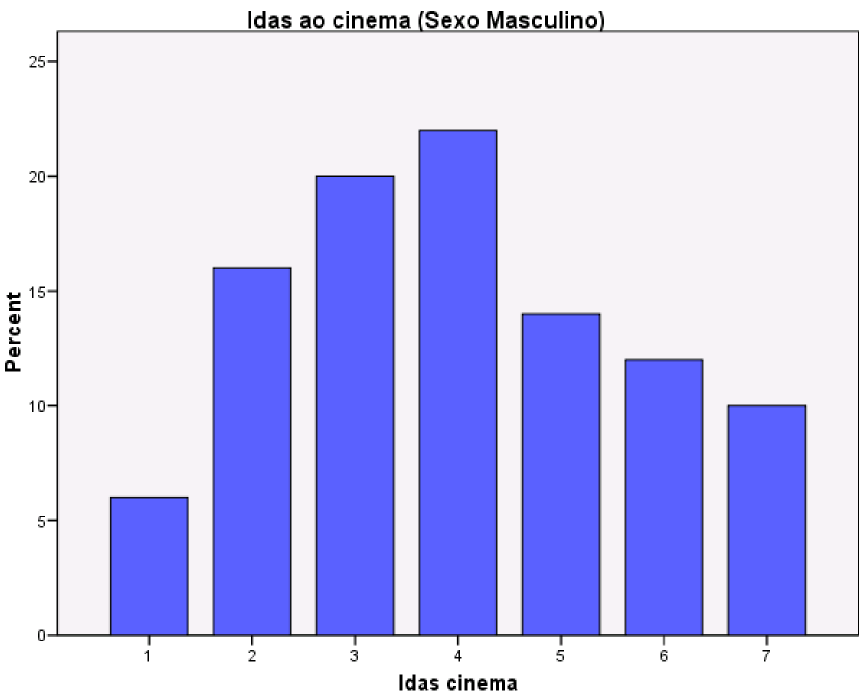
SPSS

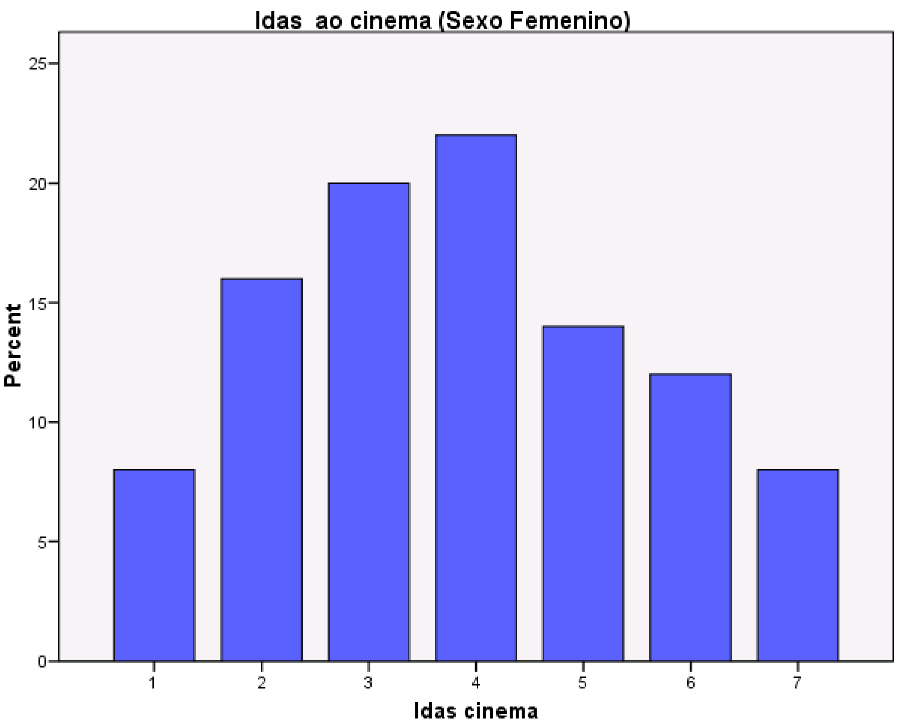
- Nota: Também é possível representar o diagrama de barras com as frequências relativas. Se quisermos comparar amostras, esse tipo de representação facilita



SPSS

- Consideremos o exemplo anterior, e suponhamos que para além do número de idas mensais ao cinema também foi identificado o sexo do inquirido. Como se repartem os inquiridos pelo sexo?





Stem-and-Leaf Plot

- Idas ao cinema Stem-and-Leaf Plot

```
Frequency    Stem &  Leaf

  7,00        1 .  0000000
 16,00        2 .  0000000000000000
 20,00        3 .  00000000000000000000
 22,00        4 .  0000000000000000000000
 14,00        5 .  00000000000000
 12,00        6 .  000000000000
  9,00        7 .  000000000

Stem width:        1
Each leaf:      1 case(s)
```

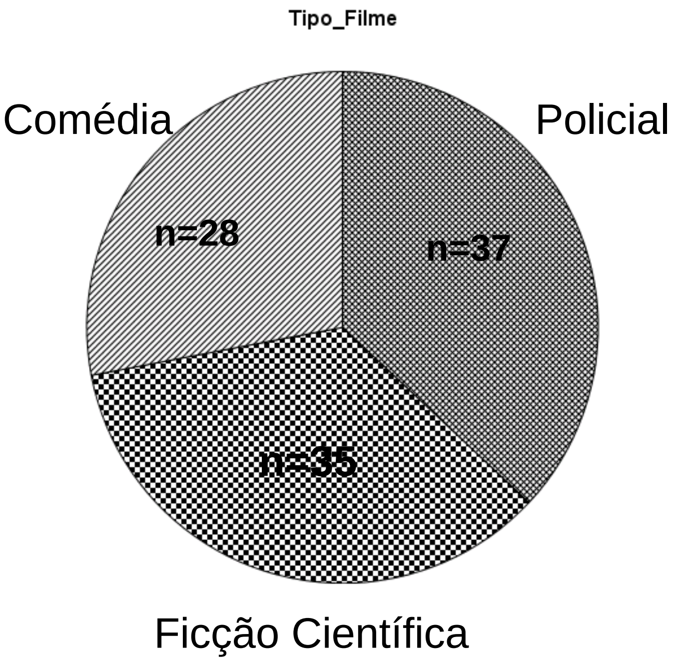
Os valores da caule representam as unidades e cada folha representa um caso

Diagramas sectoriais

- Suponhamos que os inquiridos também tiveram que identificar o género de filme preferido: policial (1), ficção científica(2) ou comédia (3).
- Neste caso, a variável em questão é qualitativa nominal

Tipo_Filme

		Frequency	Percent	Valid Percent
Valid	1	37	37,0	37,0
	2	35	35,0	35,0
	3	28	28,0	28,0
	Total	100	100,0	100,0



Medidas descritivas

- Já vimos alguns processos de resumir informação contida na amostra, utilizando os processos gráficos.
- Vamos agora ver um outro processo de resumir essa informação, utilizando determinadas medidas descritivas.
- Na população estas medidas designam-se por **parâmetros** enquanto que na amostra designam-se por **estatísticas.**
- Propriedades a que devem obedecer:
  - Definidas de forma objectiva;
  - Dependerem de todas as observações;
  - Pouco sensíveis a erros de amostragem.

Tipo de medidas descritivas

- Medidas de localização
  - Tendência central (Média, Mediana, Moda)
  - Tendência não central (Quartis, percentis)

- Medidas de dispersão
  - Medidas absolutas (Variância, Desvio Padrão, amplitude do intervalo inter-quartis)
  - Medidas relativas (Coeficiente de variação)

Medidas de tendência central- Média

- A média amostral ou simplesmente média, que se representa por $\overline{X}$, corresponde ao valor que a variável assumiria se todas as observações fossem iguais . Obtém-se a partir da seguinte expressão:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \boxed{\mu = \frac{\sum_{i=1}^{N} x_i}{n}} \longrightarrow$$

População (parâmetro)

onde $x_1, x_2, \ldots, x_n$ representam os elementos da amostra e **n** a sua dimensão.

- Para dados agregados (em distribuição de frequências):

$$\overline{X} = \frac{\sum_{i=1}^{n_1} x_i F_i}{n_1} = \sum_{i=1}^{n_1} x_i f_i$$

- Só pode ser calculada para variáveis quantitativas e pode ter um valor não assumido pela variável

Exemplo

- Consideremos o exemplo anterior:

  3,4,3,3,2,5,4,3,3,2,4,3,3,5,4,6,4,1,4, 3,7,6,2,4,3,4,7,1,3,2,5,6,1,4,5,6,7,2, 3,6,4,4,5,2,7,5,2,1,2,6,5,7,5,2,5,3,4, 6,3,7,6,3,2,4,6,4,1,2,2,5,4,7,3,4,6,4, 3,6,1,7,2,2,4,1,5,7,3,5,2,4,5,3,2,4,3, 6,5,3,4,4
- Para calcular a média basta somar os valores apresentados e dividir pela dimensão da amostra, ou seja, por 100. Neste caso o valor obtido é 3.92
- Se os dados forem apresentados sob a forma de uma tabela de frequências, então basta fazer

  (1x7+2x16+3x20+4x22+5x14+6x12 +7x8)/100=3.92

X	F_i	F_i Acum.
1	7	7
2	16	23
3	20	43
4	22	65
5	14	79
6	12	91
7	9	100

A média será sempre representativa dos dados ?

- Ao determinar a média dos seguintes valores

16	13	13	11	15	13	12	100	14

  obteve-se o valor = **23**

- Embora todos os dados, menos um, estejam no intervalo [10, 16], o valor obtido para a média está "bem afastado" daquele intervalo! O que aconteceu é que a média é muito sensível a valores muito grandes ou muito pequenos. No caso do exemplo foi o valor 100 que inflacionou a média.
- A razão pela qual a média é uma medida muito popular deve-se a dois factos:
  - é uma medida **muito simples de calcular e fácil de interpretar**
  - Mostra-se que quando a distribuição dos dados é "normal", então a melhor medida de localização do centro, é a média. Sendo a Distribuição Normal uma das distribuições mais importantes e que surge com mais frequência nas aplicações, esse facto justifica a grande utilização da média.

Propriedades da média

- A média dos desvios dos valores das variáveis em relação à média é nulo, ou seja

$$\sum_{i=1}^{N}(x_i - \mu) = 0$$

- A média dos quadrados dos desvios dos valores das variáveis em relação à média é mínima, ou seja

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

é mínima, o que significa que a média é o valor mais próximo do conjunto das variáveis

Medidas de tendência central- Mediana

- A mediana, **Me**, é definida do seguinte modo:
  - *Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, pelo menos 50% dos elementos da amostra são menores ou iguais à mediana.*
- Cálculo da mediana para variáveis discretas:
  - ordenar a amostra: $\mathbf{x_{1:n}}$ **,** $\mathbf{x_{2:n}}$ **, ... ,** $\mathbf{x_{n:n}}$
  - Calcular

$$Me = \begin{cases} x_{\frac{n+1}{2}:n} & \text{se n impar} \\ \dfrac{x_{\frac{n}{2}:n} + x_{\frac{n+1}{2}:n}}{2} & \text{se n par} \end{cases}$$

- Não pode ser calculada em dados qualitativos que se apresentem numa escala nominal

Exemplo – n par

X	F_i	F_i Acum.
1	7	7
2	16	23
3	20	43
4	22	65
5	14	79
6	12	91
7	9	100

n=100 (par) ⇨ (n+1)/2=50.5 e n/2=50

$$Me = \frac{x_{51:100} + x_{50:100}}{2} = \frac{4+4}{2} = 4$$

Exemplo – n impar

X	F_i	F_i Acum.
1	7	7
2	16	23
3	20	43
4	25	68
5	14	82
6	12	94
7	9	103

n=103 (impar) ⇨ (n+1)/2=52

$$Me = x_{52:100} = 4$$

Média ou Mediana?

- Consideremos o seguinte exemplo:
  - um aluno obteve as seguintes notas: 10, 10, 10, 11, 11, 11, 11,12
  - A média e a mediana da amostra anterior são respectivamente 10.75 e 11
  - Admitamos que uma das notas de 10 foi substituída por uma de 18. Neste caso a mediana continuaria a ser igual a 11, enquanto que a média subiria para 11.75!

- A mediana não é tão sensível, como a média, às observações que são muito maiores ou muito menores do que as restantes (outliers). Por outro lado, a média reflecte o valor de todas as observações.
- Assim, não se pode dizer, em termos absolutos, qual destas medidas de localização é preferível, dependendo do contexto em que estão a ser utilizadas.

Medidas de tendência central - Moda

- A moda, **Mo**, é definida do seguinte modo:
  - *é o valor mais frequente da distribuição, ou seja, é o valor que mais observações apresenta*
- Cálculo da moda para variáveis discretas:
  - A moda é o elemento com maior frequência absoluta
- Não é influenciada por valores extremos!!!
- Esta medida é especialmente útil para reduzir a informação de um conjunto de dados qualitativos, apresentados sob a forma de nomes ou categorias, para os quais não se pode calcular a média e por vezes a mediana (se não forem susceptíveis de ordenação, ou seja, estejam numa escala nominal).

X	F_i	F_i Acum.
1	7	7
2	16	23
3	20	43
4	22	65
5	14	79
6	12	91
7	9	100

Mo=4

Medidas de tendência não central - Quantis

- Generalizam a noção de mediana **Me**
- **A mediana é a medida de** localização, tal que, pelo menos, 50% dos elementos da amostra são menores ou iguais a **Me**
- **Quartil: Divide a amostra em 4 partes**
  - Q_1: é a medida de localização, tal que pelo menos 25% dos elementos da amostra são menores ou iguais a $\mathbf{Q_1}$
  - Q_2: é a mediana
  - Q_3: é a medida de localização, tal que, pelo menos, 75% dos elementos da amostra são menores ou iguais a $\mathbf{Q_1}$
  - **Percentil: Divide a amostra em 100 partes**
- **Quantil de ordem p, com $0<p<1$**: valor **Qp** tal que, pelo menos, 100p% dos elementos da amostra são menores ou iguais a **Qp**

Cálculo dos quantis

- Generalizando a expressão para o cálculo da mediana, vem

$$Qp = \begin{cases} x_{\lceil np \rceil + 1:n} & \text{se np não é inteiro} \\ \dfrac{\left\lceil x_{np+1:n} + x_{np:n} \right\rceil}{2} & \text{se np é inteiro} \end{cases}$$

em que $\lceil a \rceil$ representa o maior inteiro contido em a

- Para o exemplo respeitante ao número de idas ao cinema durante um mês, tem-se:
  - Primeiro quartil Q_1: como p=1/4 então np=100x1/4=25 que é inteiro. Logo tem $(x_{26:100}+x_{25:100})/2=(3+3)/2=3$

X	F_i	F_i Acum.
1	7	7
2	16	23
3	20	43
4	22	65
5	14	79
6	12	91
7	9	100

Medidas de dispersão absolutas - Variância

- As medidas de dispersão dão-nos uma ideia da concentração dos valores da variável e por isso servem para verificar a representatividade das medidas de localização
- A **variância** mede a concentração dos valores em torno da média e é calculada a partir da seguinte expressão:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

- Nota: se se pretender que o estimador seja centrado então calcula-se

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\overline{x}^2$$

- Só pode ser calculada para variáveis numéricas.
- É expressa no quadrado das unidades da variável : calcula-se o desvio padrão (o que são idas ao cinema ao quadrado???)

Medidas de dispersão absolutas- Desvio padrão

- O desvio padrão é a raiz quadrada positiva da variância

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

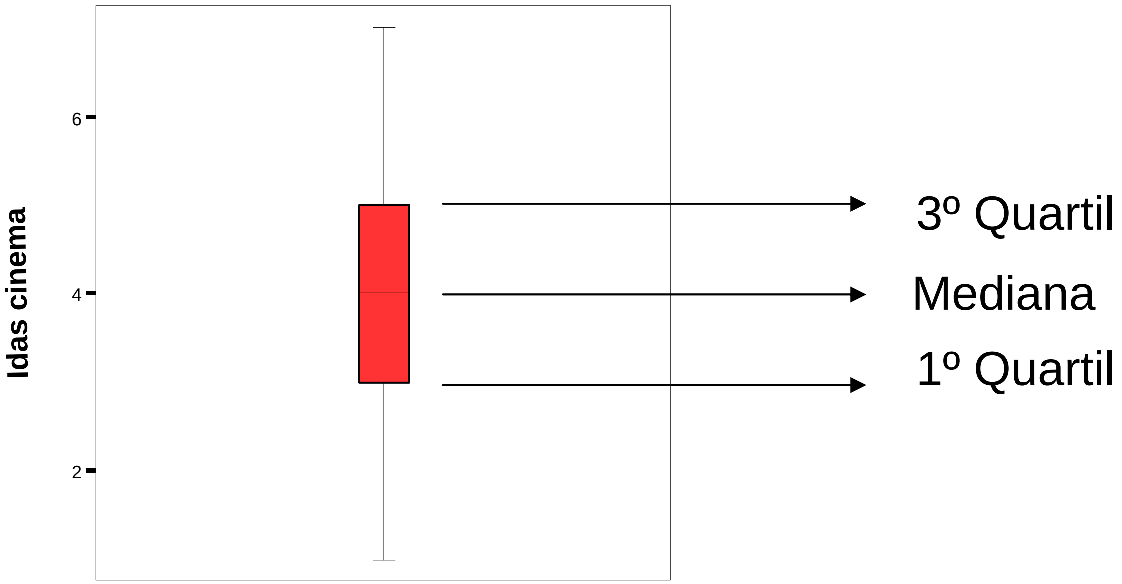
- O desvio padrão é uma medida que **só pode assumir valores não negativos,** e quanto maior for maior será a dispersão dos dados.

- A variância e, consequentemente, o desvio padrão são afectados por todos os valores observados

Medidas de dispersão absolutas- amplitudes do Intervalo de variação e intervalo inter-quartis

- A amplitude do intervalo de variação é definido por
  - $R = X_{max} - X_{min}$
- Pode dar uma ideia errada acerca da dispersão uma vez que só considera os valores extremos
- A amplitude do intervalo inter-quartis, é o intervalo que engloba 50% das observações centrais. é definido por:
  - $I_Q = Q_3 - Q_1$
- Ignora os valores pertencentes aos extremos

Gráfico de caixa de bigodes



Medidas de dispersão relativas

- O coeficiente de variação, tal como a variância e o desvio padrão, indica a concentração em torno da média, e é calculada pela seguinte expressão:

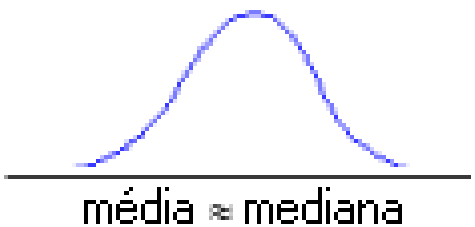
$$CV = 100 \times \frac{S}{\bar{x}}$$

- Esta medida não depende das unidades em que a variável está expressa sendo por isso útil para efeitos de comparação

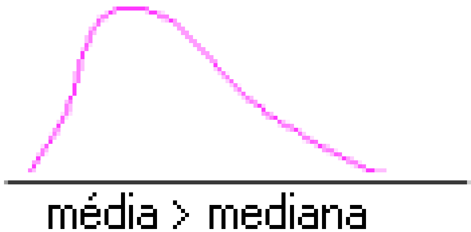
- Note-se que quanto maior for o seu valor menor será a representatividade da média . Por exemplo, se CV>50% podemos de imediato concluir que o grau de dispersão é elevado

Enviesamento

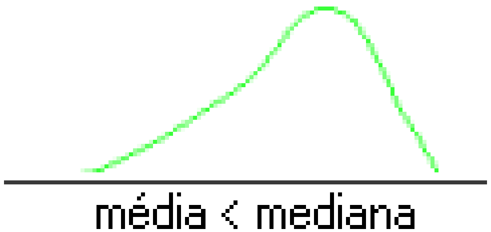
- A comparação das medidas de tendência central permite-nos avaliar o enviesamento dos dados



Distribuição simétrica



Distribuição assimétrica positiva



Distribuição assimétrica negativa

Medidas - SPSS

Descriptives

			Statistic	Std. Error
Idas cinema	Mean		3,92	,171
	95% Confidence Interval for Mean	Lower Bound	3,58	
		Upper Bound	4,26	
	5% Trimmed Mean		3,91	
	Median		4,00	
	Variance		2,923	
	Std. Deviation		1,710	
	Minimum		1	
	Maximum		7	
	Range		6	
	Interquartile Range		2	
	Skewness		,176	,241
	Kurtosis		-,839	,478

Statistics

Idas cinema

N	Valid	100
	Missing	0
Mean		3,92
Std. Error of Mean		,171
Median		4,00
Mode		4
Std. Deviation		1,710
Variance		2,923
Skewness		,176
Std. Error of Skewness		,241
Kurtosis		-,839
Std. Error of Kurtosis		,478
Range		6
Minimum		1
Maximum		7
Sum		392
Percentiles	10	2,00
	25	3,00
	50	4,00
	75	5,00

Coeficiente de assimetria ou enviesamento:
Se >0 então distribuição assimétrica positiva
Se < 0 então distribuição assimétrica negativa
Se = 0 então distribuição simétrica positiva

Coeficiente de achatamento:
Se >0 então distribuição leptocártica (pontiaguada)
Se < 0 então distribuição platicúrtica (achatamento)
Se = 0 então distribuição mesocúrtica

Variáveis contínuas

- O valor diário das vendas (em u.m.) dum supermercado, ao longo dos últimos 40 dias foi:
138,146,168,146,161,164,161,164,158,126,173, 145,150,140,138, 142,135,132,147,176,147,142,144,136,163,135,150,125,148,119,153,156,149,152,154,140,145,157,144,165,135,128

- Neste caso, a variável em estudo, X, é o valor diário das vendas durante um 40 dias. Esta variável pode ser considerada uma variável contínua.

- Os valores registados constituem a amostra. Neste caso a dimensão da amostra é n=40

Tabelas de Frequências

- O objectivo é agrupar os dados, em classes, que se forem bem construídas dão uma melhor perspectiva da amostra
- Construção das classes:
  - Devem ter amplitudes iguais
  - Nenhuma classe deve ter frequência nula
  - Cada elemento da amostra só pode pertencer a uma classe
  - Em geral, vão de 4 a 14
- Número de classes (k) :
  - Regra 1: $\text{Se } n < 25 \text{ então } k = 5 \text{ senão } k \cong \sqrt{n}$
  - Regra 2 (Sturges): $k \cong 1 + 3.22 \times \ln n$
  - Outras regras: por exemplo, quando se utilizam dados oficiais
- Com amplitude (a):

$$a = \frac{\text{valor máximo da amostra - valor mínimo da amostra}}{k}$$

Tabelas de frequências

- Como se constroem as classes:
  - Se existir um limite inferior finito $[\underline{L}, \underline{L}+a[, [\underline{L}+a, \underline{L}+2a[, \ldots$
  - Se existe um limite superior finito $\ldots,]-2a+\overline{L}, -a+\overline{L}],]-a+\overline{L}, \overline{L}]$
  - Centradas na média $\ldots, \left[\bar{x}-\frac{a}{2}, \bar{x}+\frac{a}{2}\right], \ldots$
  - Pode ainda aparecer menos 10, 10-20, 20-30,… o que representa $]-\infty, 10[, [10, 20[, [20, 30[, \ldots$
- Centro da classe: só está definido para classes com amplitude finita e não é mais que o seu ponto médio
- Quando se agrupa os dados em classes, passa-se a trabalhar com as frequências das classes e os valores individuais deixam de ser identificados

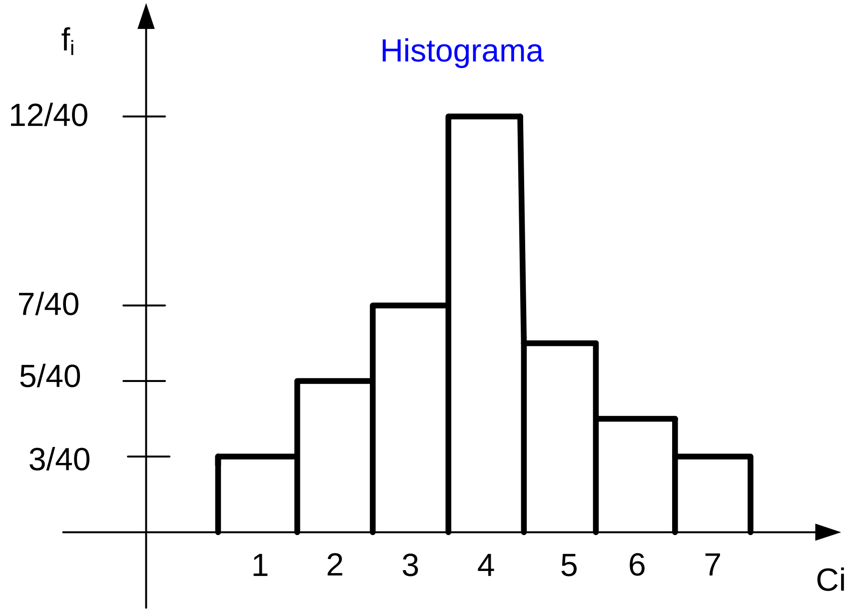
Tabelas de frequências

- Para o exemplo anterior:
  - Pela regra 1, vem $k \cong \sqrt{40} = 6.325$
  - Pela regra de Sturges: $k \cong 1 + 3.22 \ln 40 = 6.159$
  - Ambas as regras conduzem a **k=7**
  - O valor mínimo e máximo da amostra é 119 e 176 respectivamente, donde a amplitude das classes será **a_i=(176-119)/7=8.15**

C_i	F_i	f_i	Cum F_i	Cum f_i
[119,127.15[	3	3/40	3	3/40
[127.15,135.3[	5	5/40	8	8/40
[135.3,143.45[	7	7/40	15	15/40
[143.45,151.6[	12	12/40	27	27/40
[151.6,159.75[	6	6/40	33	33/40
[159.75,167.9[	4	4/40	37	37/40
[169.7,176.5[	3	3/40	40	1

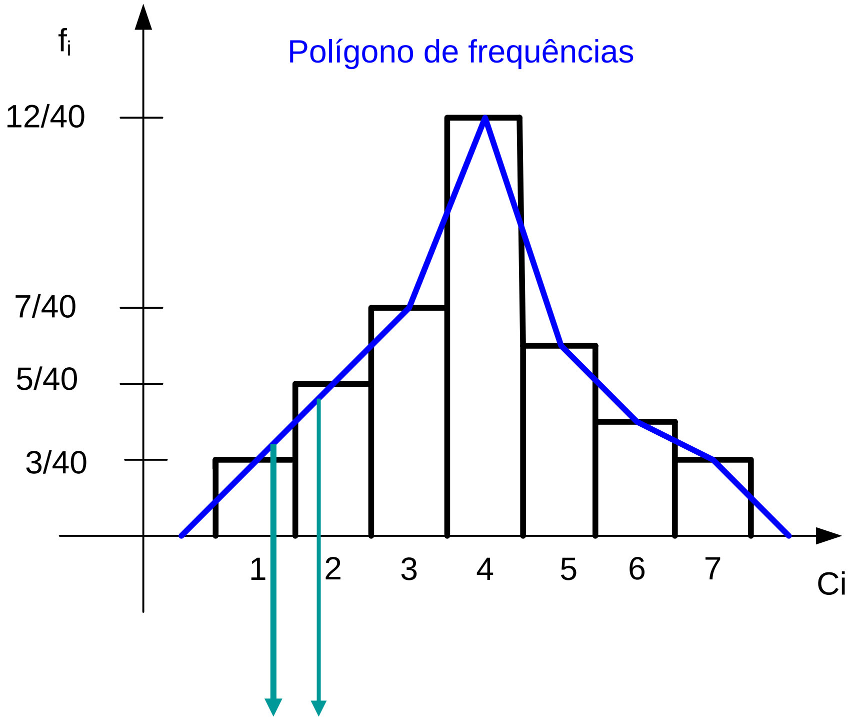
Representação gráfica da distribuição de frequências

- Após se obter a distribuição de frequências deve proceder-se à representação gráfica da mesma. Para o efeito utiliza-se:
  - Histograma
  - Polígono de frequência
- Histograma: é um gráfico de barras, em que se representam as classes versus as frequências absolutas (ou frequências relativas)
- Para o exemplo anterior, temos:
- Também podemos construir histograma de frequências acumuladas



Representação gráfica da distribuição de frequências

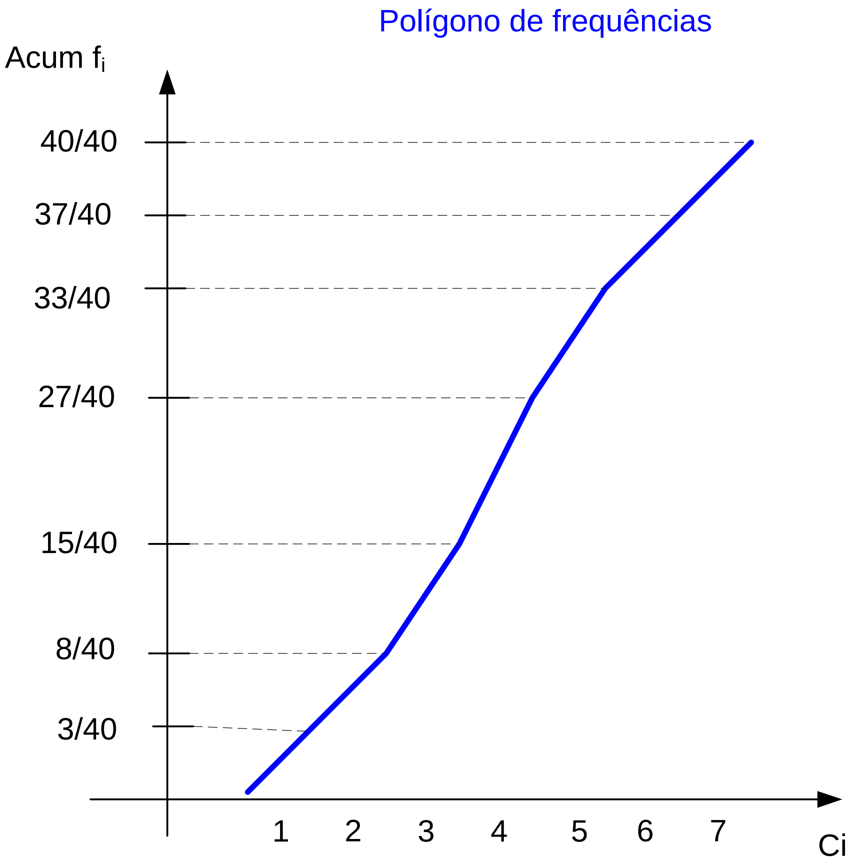
- Polígono de frequências: é uma linha que une os pontos médios dos topos das barras representados no histograma
- Para fechar essa linha, o polígono, cria-se uma classe adicional em cada um dos extremos, com frequência nula e amplitude igual às restantes.
- Para o exemplo anterior, temos



O que se retira é igual ao que se acrescenta

Representação gráfica da distribuição de frequências

- É também possível representar o polígono de frequências acumuladas.
- Neste caso, os pontos para marcação são colocados no início de cada classe.
- Para o exemplo anterior, temos



Representação gráfica da distribuição de frequências

- E quando as classes não têm a mesma amplitude?
- E quando existem classes sem um dos extremos (inicial ou final)?
- Exemplo:

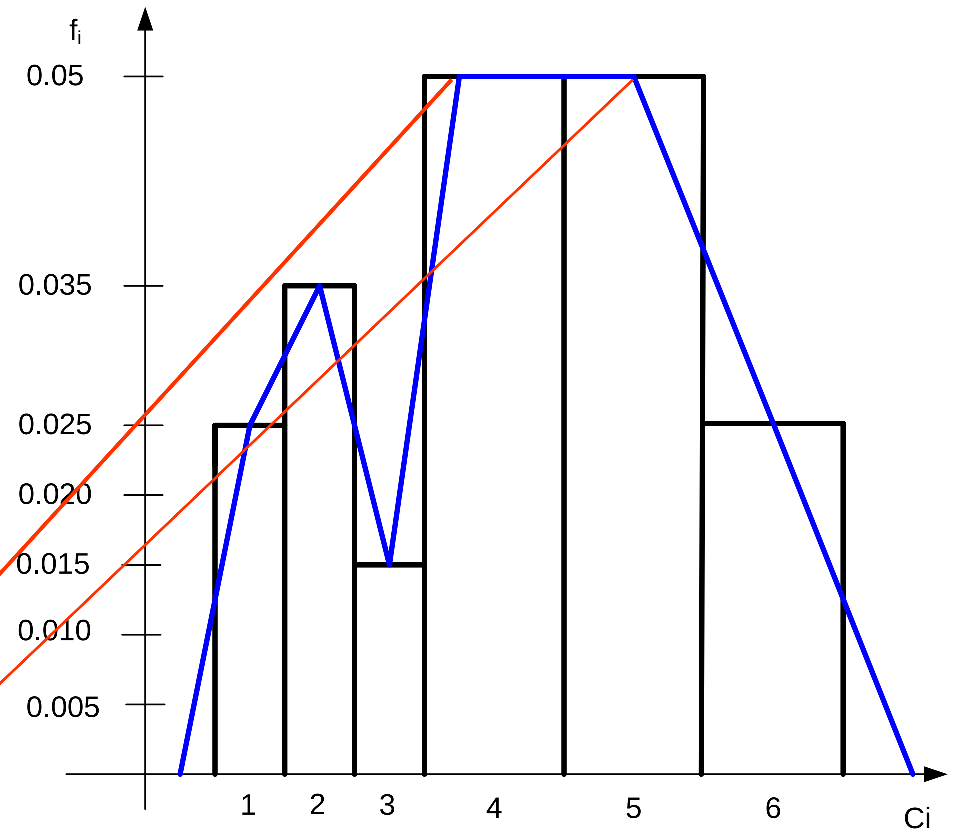
C_i	F_i	f_i
<20	25	0.25
20-30	35	0.35
30-40	15	0.15
40-60	10	0.10
60-80	10	0.10
>=80	5	0.05

- Consideram-se as classes sem um dos extremos como classes com amplitude igual à classe anterior ou posterior. Neste caso, ficaria 10-20 e 80-100

Representação gráfica da distribuição de frequências

- Dado que nem todas as classes têm a mesma amplitude para a construção do histograma e do polígono de frequências, é necessário normalizar as frequências relativas dividindo-as pela amplitude da respectiva classe.
- Neste caso, viria

C_i	F_i	f_i	f_i norm.
[10,20[	25	0.25	0.025
[20,30[	35	0.35	0.035
[30,40[	15	0.15	0.015
[40,60[	10	0.10	0.05
[60,80[	10	0.10	0.05
[80,100]	5	0.05	0.005



Para marcar o polígono considera-se que a classe tem a mesma amplitude que a anterior

Medidas de tendência central- Média

- Para dados agregados em classes:

$$\overline{X} = \frac{\sum_{i=1}^{n_1} C_i F_i}{n_1} = \sum_{i=1}^{n_1} C_i f_i$$

Centro da classe

Número de classes

- Ainda que, para a elaboração de tabelas de frequências, histogramas e polígonos de frequências seja necessário agrupar os dados em classes, sempre que seja possível deve calcular-se a média com os dados desagregados para evitar erros de cálculo

Medidas de tendência central - Mediana

- Cálculo da mediana para dados agrupados em classes:
  - Encontrar a classe mediana, ou seja a classe cujas frequências absolutas acumuladas contêm o valor n/2
  - Determinação numérica da mediana:

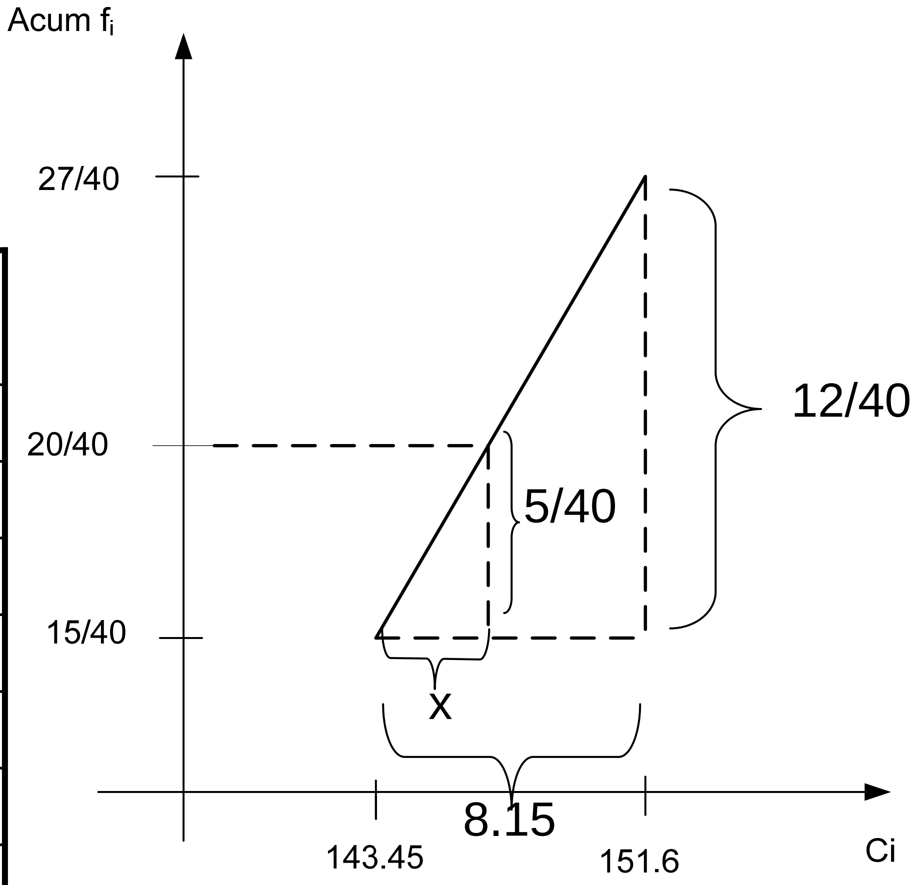
$$Me = li(Me) + \frac{0.5 - \text{cum } f(Me-1)}{f(Me)} \times a_{Me}$$

em que $li(Me)$ é o limite inferior da classe mediana, $\text{cum } f(Me-1)$ é a frequência relativa acumulada da classe que antecede a classe mediana, $f(Me)$ é a frequência relativa da classe mediana e a_{Me} é a amplitude da classe mediana

Medidas de tendência central- Mediana

- Determinação gráfica da mediana:
  - Faz-se o gráfico das frequências relativas acumuladas no intervalo da classe mediana
  - Para o exemplo anterior vem que:

C_i	Cum F_i	Cum f_i
[119,127.15]	3	3/40
[127.15,135.3[	8	8/40
[135.3,143.45[	15	15/40
[143.45,151.6[	27	27/40
[151.6,159.75[	33	33/40
[159.75,167.9[	37	37/40
[169.7,176.5[	40	1



Classe mediana: acumula n/2=20 frequências

Medidas de tendência central- Mediana

- Desta forma vem que

$$\frac{12/40}{5/40} = \frac{8.15}{x} \Leftrightarrow x = 3.39$$

- Pelo que
  - Me=143.45+3.39=146.84

- Nota: Não existe problema se as classes tiverem diferente amplitude

Medidas de tendência central - Moda

- Cálculo da moda quando as classes têm igual amplitude:
  - Determinar a classe modal, ou seja a classe com maior frequência absoluta ou relativa
  - Determinação numérica da moda (Formula de King)

$$Mo = li(Mo) + \frac{f(Mo+1)}{f(Mo-1) + f(Mo+1)} \times a_{Mo}$$

Também se pode usar frequências absolutas

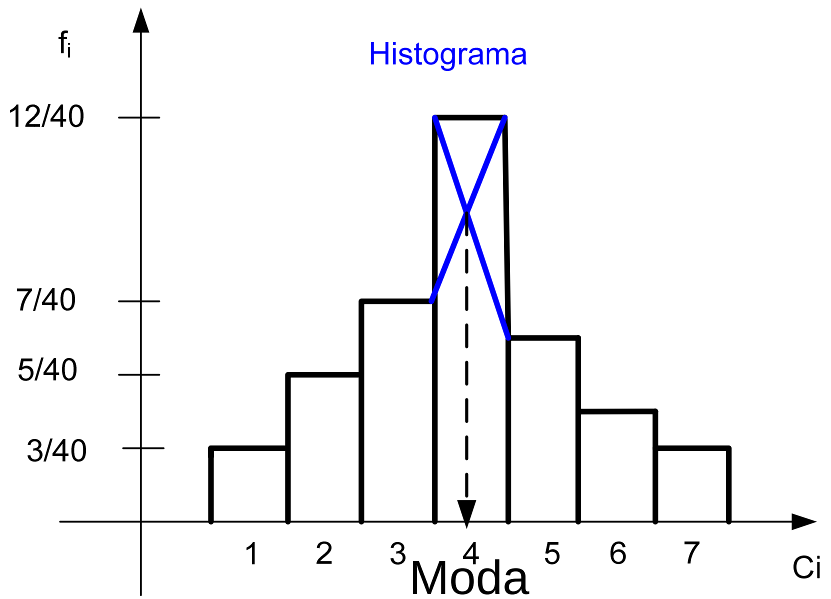
em que li(Mo) é o limite inferior da classe modal, f(Mo + 1) é a frequência relativa da classe que antecede a classe modal, f(Mo + 1) é a frequência relativa da classe que sucesse a classe modal a_{Mo} é a amplitude da classe modal

- Se as classes tiverem diferente amplitude, procede-se de forma análoga só que se trabalha com f_i/a_i em vez de f_i.

Medidas de tendência central- Moda

- Determinação gráfica da moda:
  - Faz-se o histograma das frequências relativas
  - Une-se o inicio da classe modal com o inicio da classe posterior e o fim da classe modal com o fim da classe anterior
  - O ponto de intersecção é a moda
  - Para o exemplo anterior vem que:

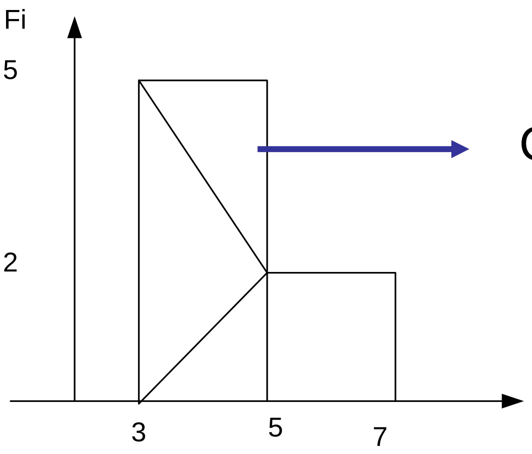
C_i	F_i
[119,127.15[	3
[127.15,135.3[	5
[135.3,143.45[	7
[143.45,151.6[	12
[151.6,159.75[	6
[159.75,167.9[	4
[169.7,176.5[	3



Classe modal: maiores frequências
Moda=143.45+(6/(7+6))x8.15=147.21

Medidas de tendência central- Moda

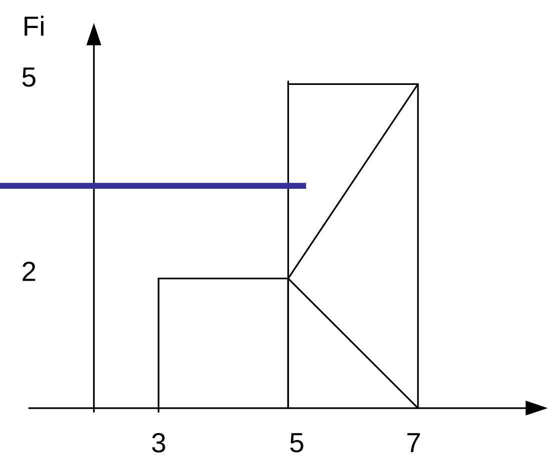
- E se a classe modal for a primeira classe?



Classe modal. Moda=3+(2/(0+2))*2=5

- E se a classe modal for a última?

Classe modal. Moda=5+(0/(2+0))*2=5



Medidas de Dispersão

- O calculo da variância para dados agrupados em classes é determinado por:

$$S^2 = \frac{1}{n} F_i \sum_{i=1}^{k} (C_i - \overline{x})^2 = \sum_{i=1}^{n} f_i (C_i - \overline{x})^2$$

- O desvio padrão será a raíz quadrada positiva da variância.

- O coeficiente de variação, bem como o intervalo de variação e o intervalo inter-quartis são calculados da mesma forma à descrita para variáveis discretas.