## 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)

## Exemplos:

- 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.

***População*** *- todas as peças produzidas pela fábrica durante aquele período de tempo.*

***Característica estatística*** *tem apenas duas modalidades: peça defeituosa e peça não defeituosa.*

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

***População*** *- famílias daquela cidade*

***Característica estatística*** *- rendimento familiar mensal.*

***Modalidade****s não se podem enumerar; são todos os valores desde, por exemplo, 10 0Euros até 10000 Euros.*

- 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.

***População*** *- trabalhadores da fábrica*

*C**aracterística estatística** 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** (e.g., o número diário de nascimentos no hospital de Viseu ou a altura dos alunas da ESTV):

  – **Discretos** (número finito ou infinito numerável de modalidades; e.g., o número diário de nascimentos no hospital de Viseu)

  –

  – **Contínuos** (pode assumir qualquer valor num intervalo de números reais; a distinção entre os dois é por vezes arbitrária; e.g., altura de um aluno da ESTV)

- **Qualitativos** (e.g., 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;
- Profissão de um indivíduo;
- Sexo dos indivíduos de uma população (**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, mas cuidado! não faz qualquer sentido usar operações aritméticas e calcular médias, desvios padrões, etc..

**Escala Ordinal** (dados qualitativos): as diversas categorias possuem uma ordem intrínseca (os códigos numéricos devem ter em conta essa ordem).

**Exemplos**:

– Habilitações literárias

– Num inquérito de opinião pede-se às pessoas que classifiquem um determinado produto como sendo, muito fraco, fraco, razoável, bom ou muito bom.

– 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).

**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$^{o}$ C está à mesma distância de 25$^{o}$ C do que 25$^{o}$ C de 30$^{o}$ C).

Não podemos atribuir um significado à razão entre dois valores (se na Guarda se registar uma temperatura de 40$^{o}$ C e em Viseu 20$^{o}$C, isto não significaria que na Guarda está 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$^{o}$ C não tem qualquer temperatura.

## 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.

**Exemplos:** peso; altura

A temperatura do ar não está definida numa escala de rácios. Note que

$10^{o}C = 50^{o}F$ e $30^{o}C = 86^{o}F$

mas,

$$\frac{10°C}{30°C} \neq \frac{50°F}{86°F}.$$

## Representação de Dados

População ou amostra de **n** indivíduos.

Atributo **A** com **p** modalidades: $\mathbf{A_1}$, $\mathbf{A_2}$,...,$\mathbf{A_p}$.

**Frequência absoluta** ou **efectivo** da modalidade $A_i \rightarrow$ $\mathbf{n_i}$, é o nº de indivíduos que apresentam a modalidade $A_i$.

**Frequência relativa** da modalidade $A_i \rightarrow$ $\mathbf{f_i}$, é a proporção de indivíduos que apresentam a modalidade $A_i$, $f_i = \frac{n_i}{n}$.

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

## *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 - Vendas

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:

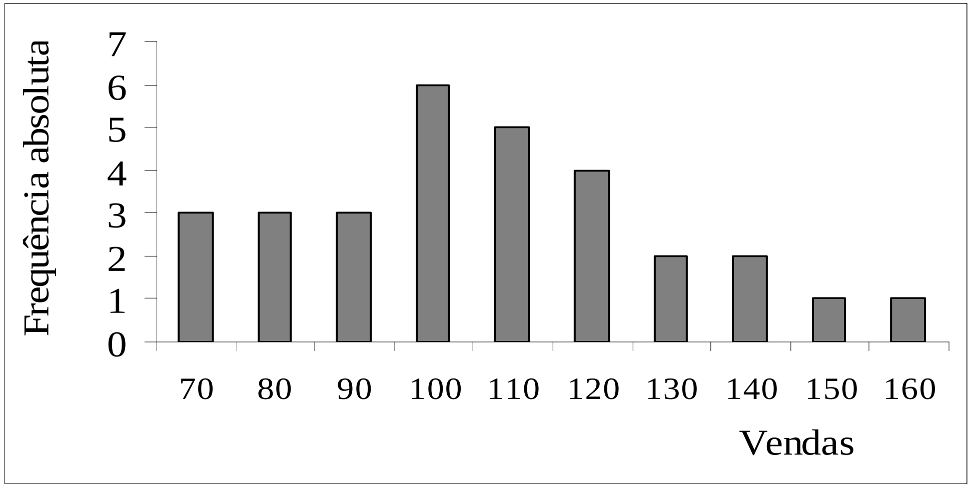
**Regra de Sturges** – nº de classes $\cong 1+\log_{10}(n)/\log_{10}(2)$

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

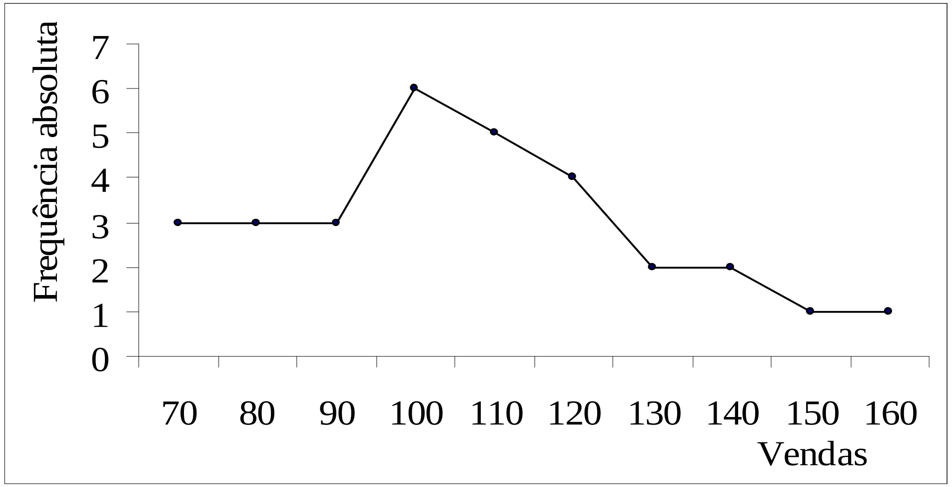
## Representação gráfica

### Dados Não Agrupados

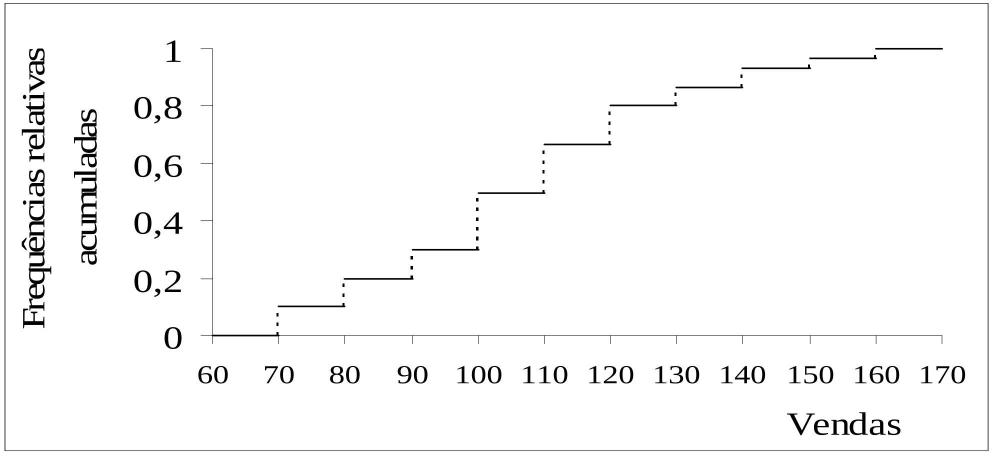
**Diagrama de barras**




**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:

$$h_i = \begin{cases} \dfrac{n_i}{a_{i+1} - a_i} & \text{(frequências absolutas)} \\[2ex] \dfrac{f_i}{a_{i+1} - a_i} & \text{(frequências relativas)} \end{cases}$$

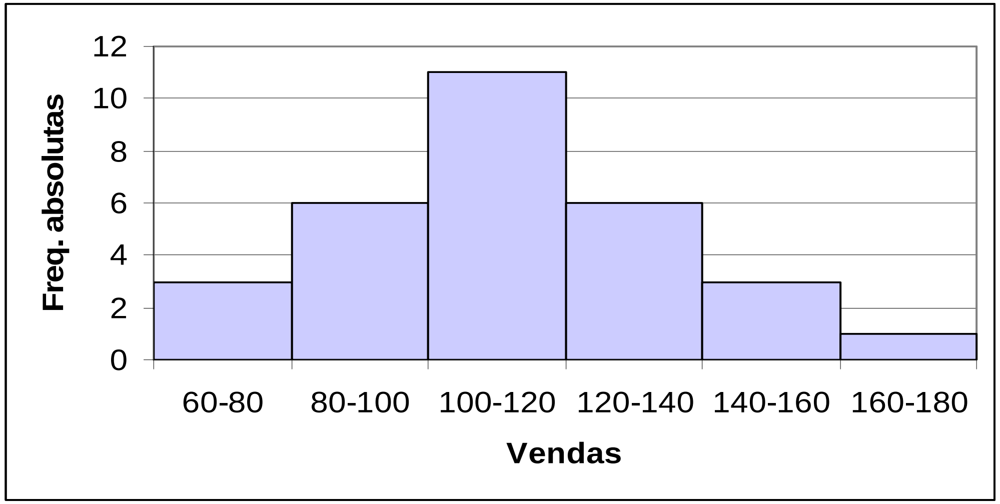
A área de cada rectângulo é então proporcional à frequência da classe respectiva:

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

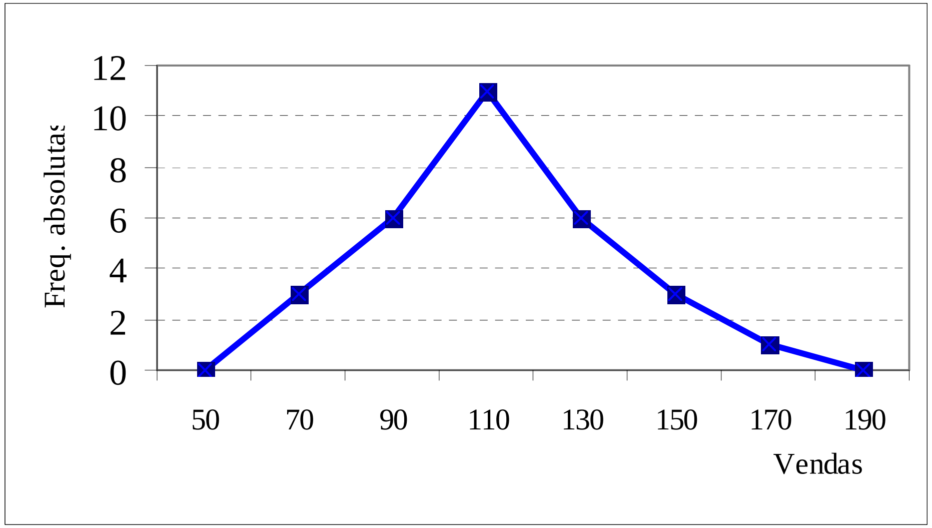
A área total do histograma é igual a ***n*** se foram usadas frequências absolutas e igual a **1** se foram usadas frequências relativas.

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.

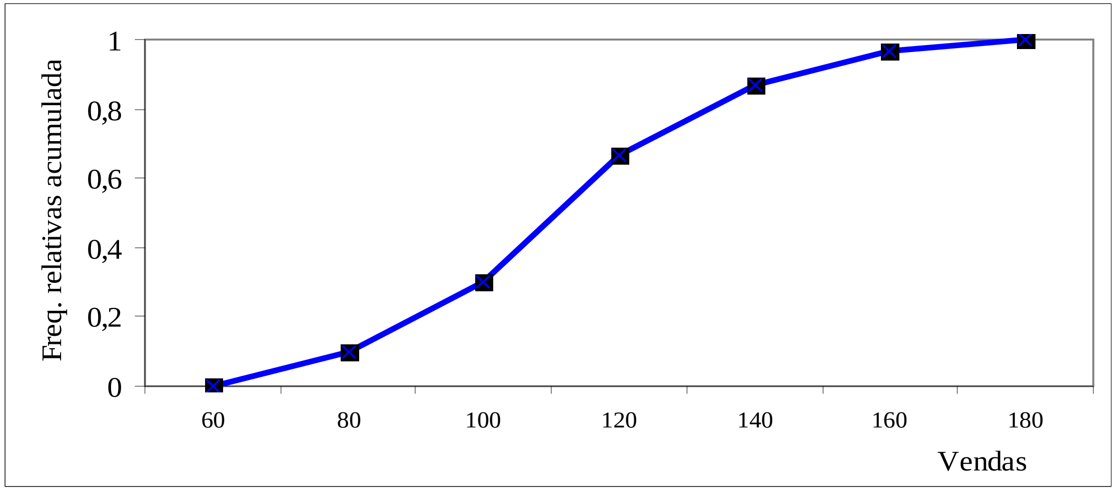
## Histograma




## Polígono de frequências



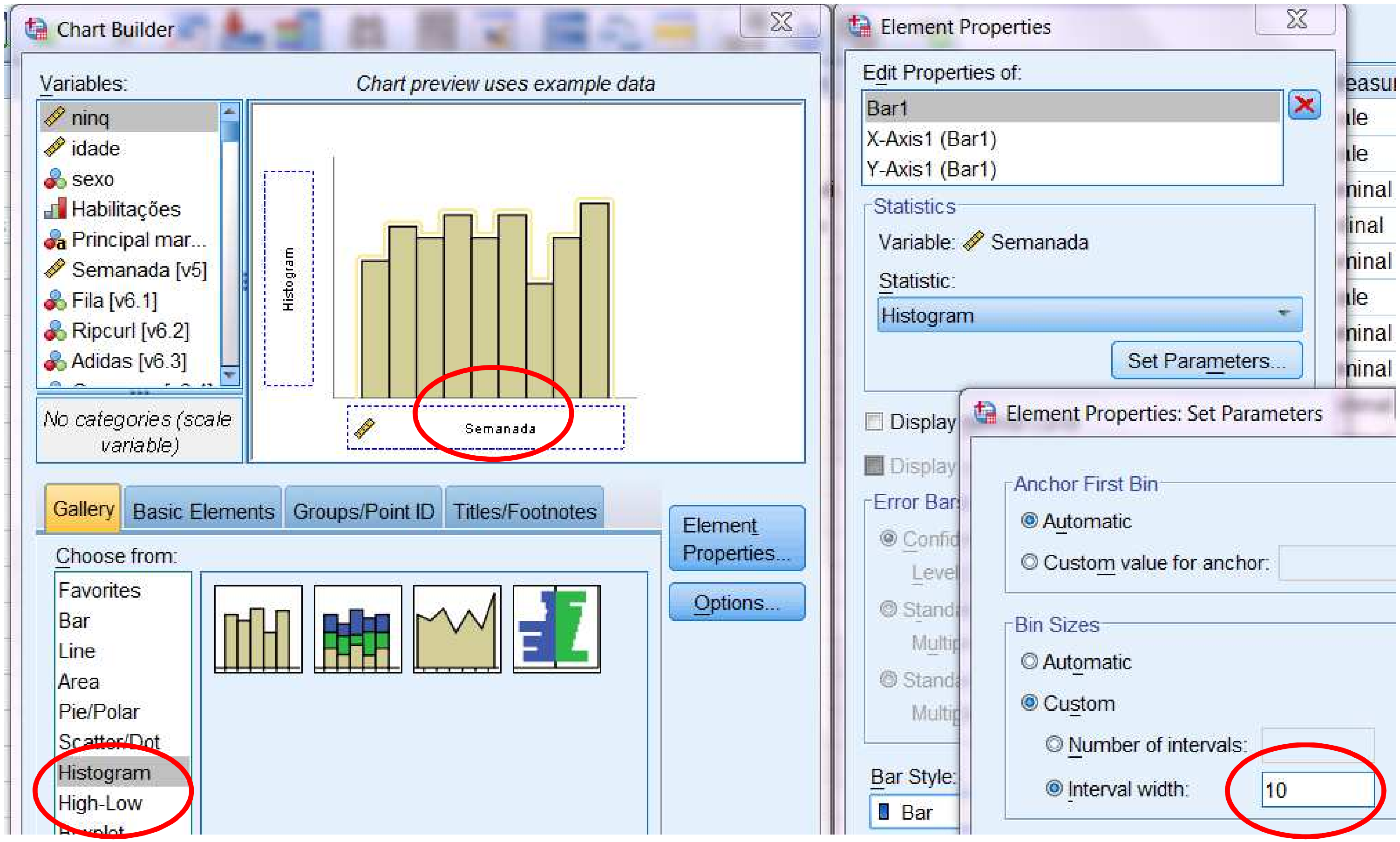

## Polígono de frequências acumuladas

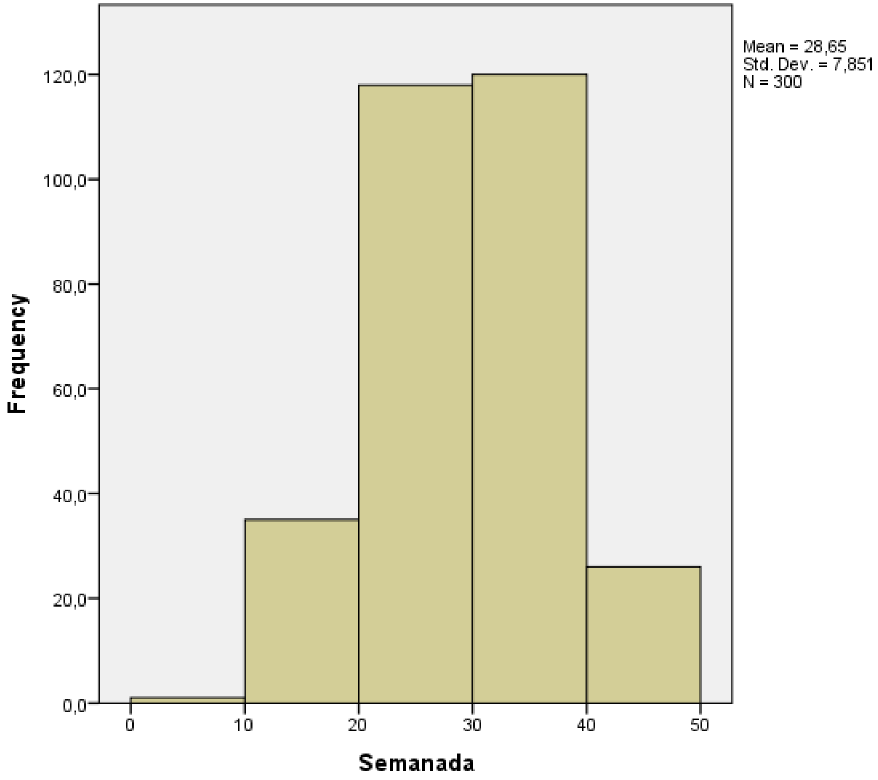


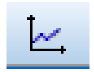
## Exemplo – Calçado Desportivo

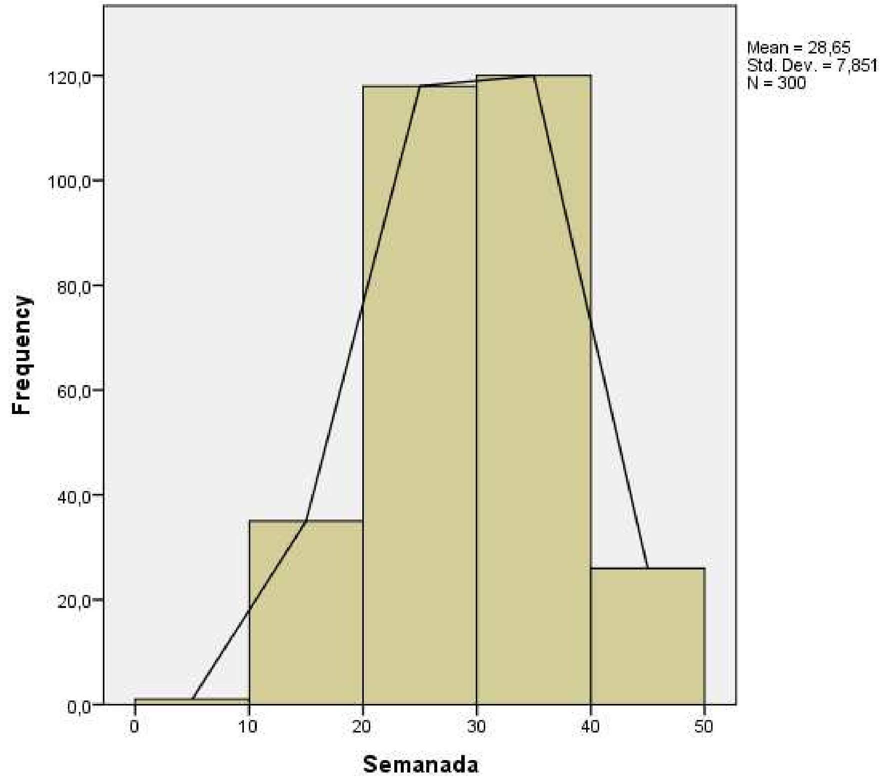
Os dados estão registados no ficheiro *Compra Calçado Desportivo.sav*.
Construa-se o histograma para a variável Semanada (v5).

SPSS: `Graphs – Chart Builder- Ok`



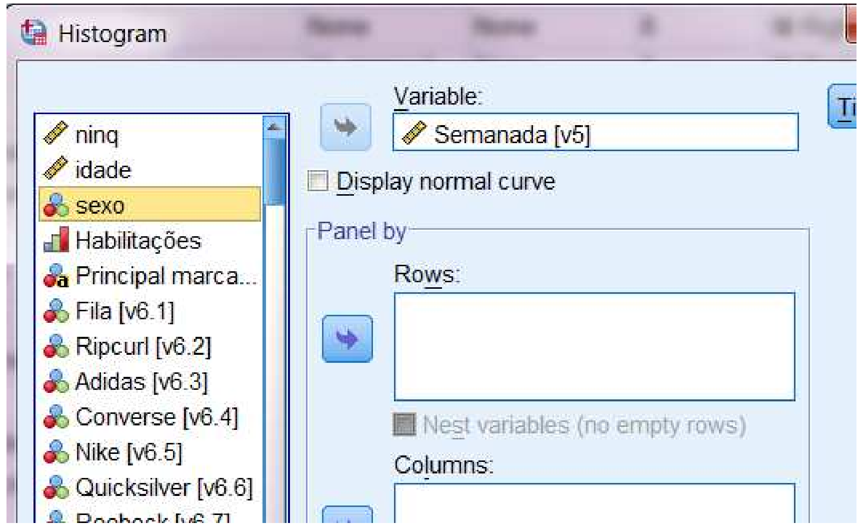


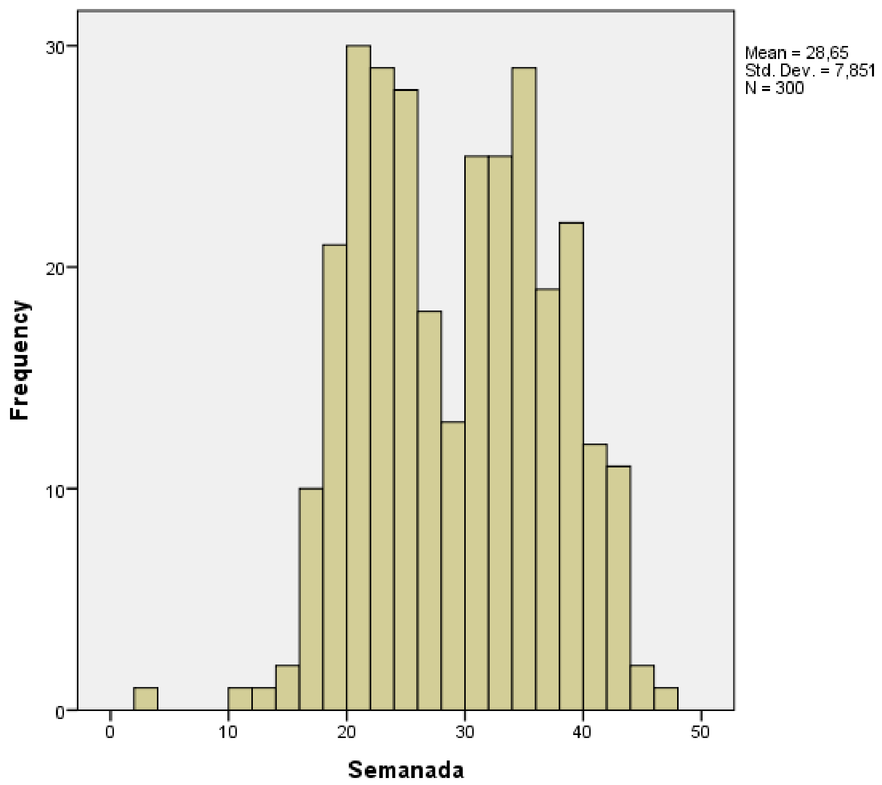

Clicando 2x no gráfico podemos editá-lo e, por exemplo, clicar em  ( na barra superior) para obter:



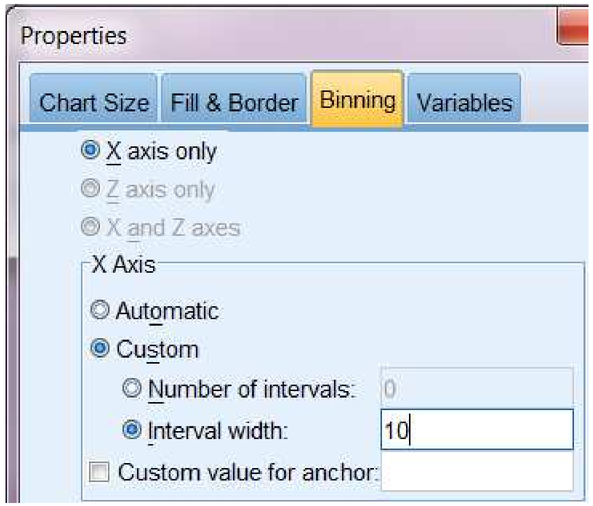
Alternativamente,

SPSS: `Graphs – Legacy Dialogs - Histogram`



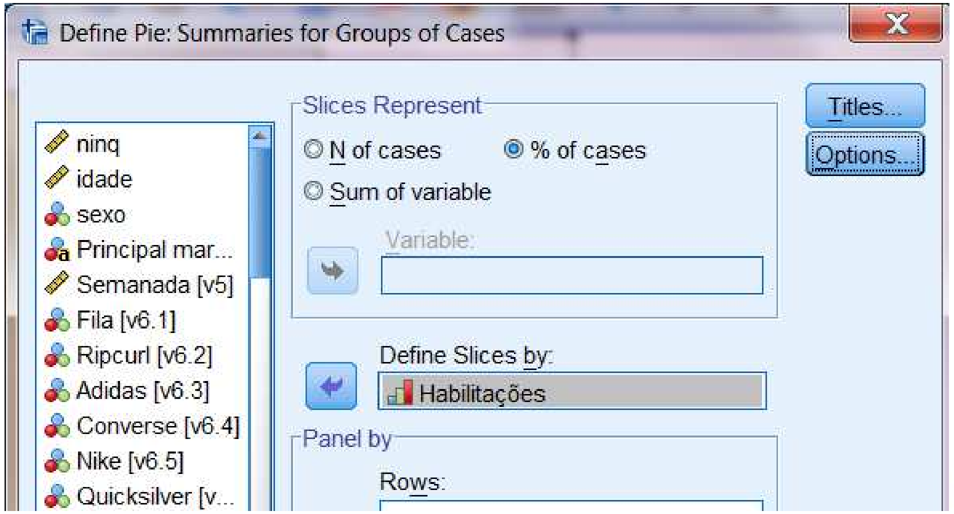

Clicando 2x no gráfico podemos editá-lo e, por exemplo, alterar a amplitude das classes para obter um histograma igual ao anterior.

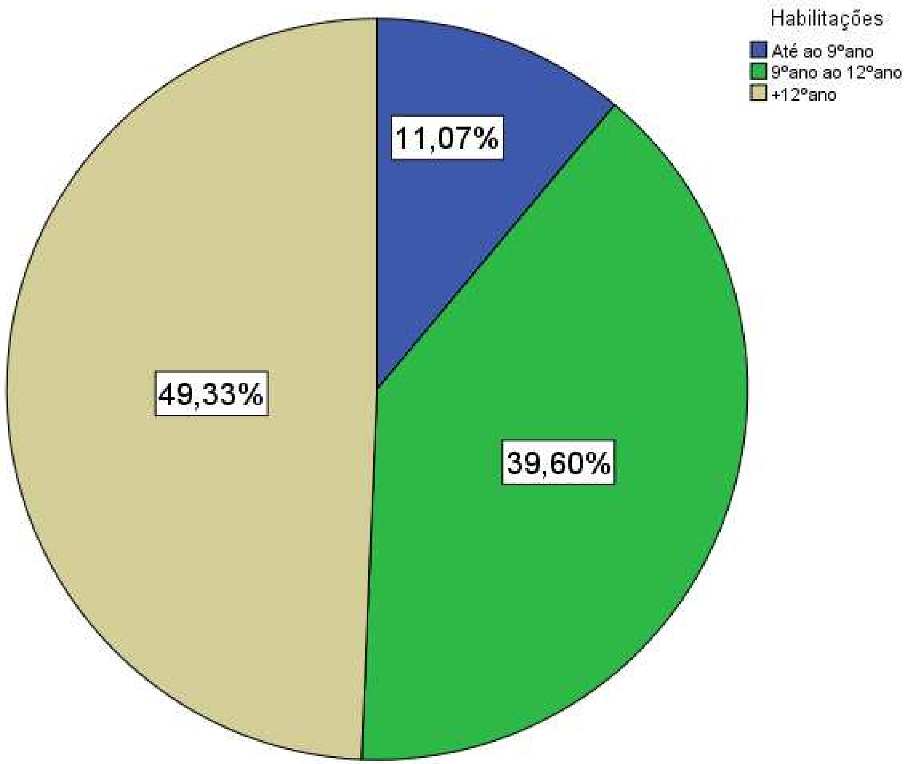


Para as habilitações podemos construir vários tipos de gráficos:

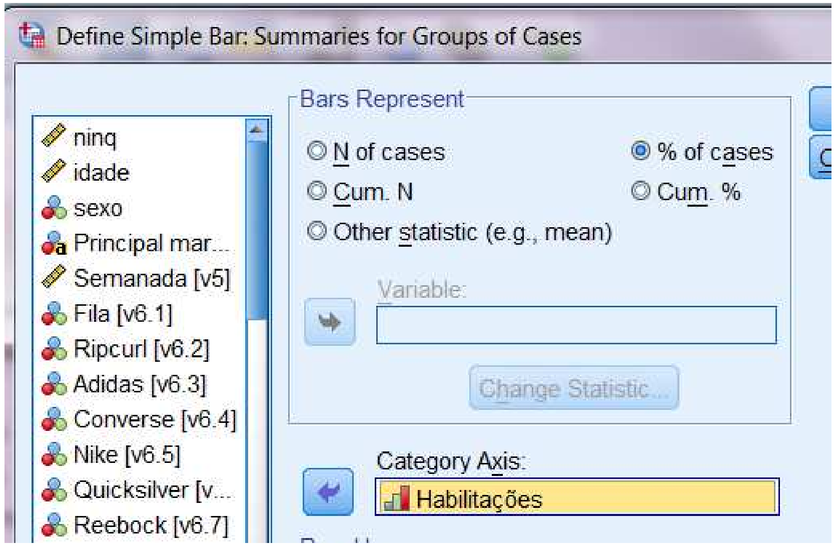
Eis alguns exemplos:

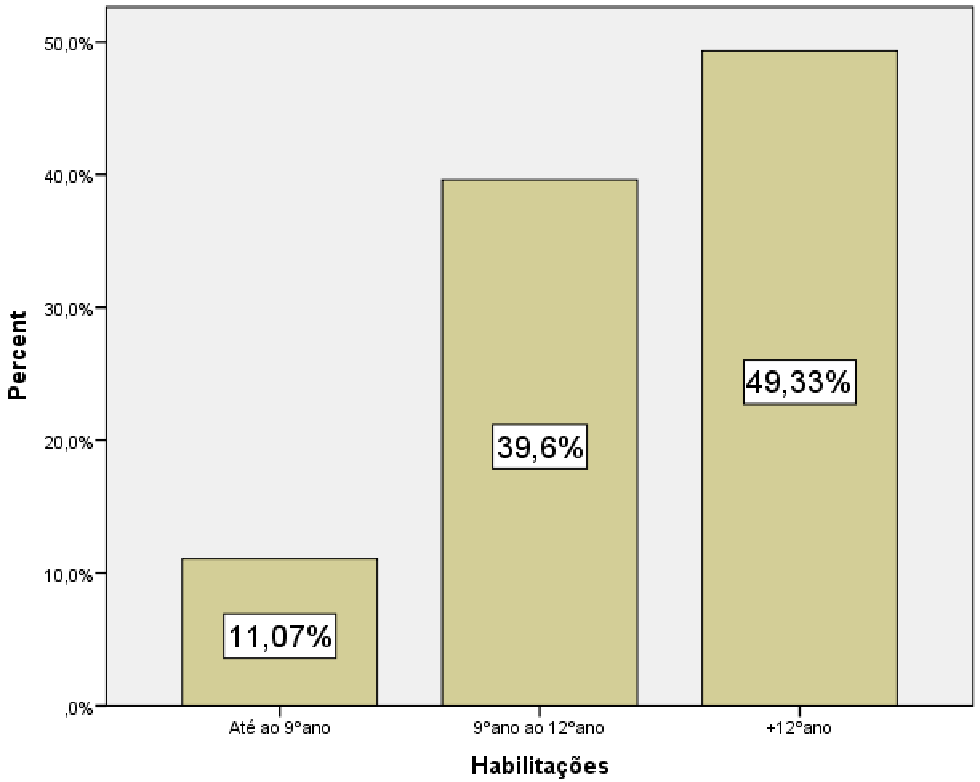
`Graphs – Legacy Dialogs – Pie – Summaries for groups of cases`



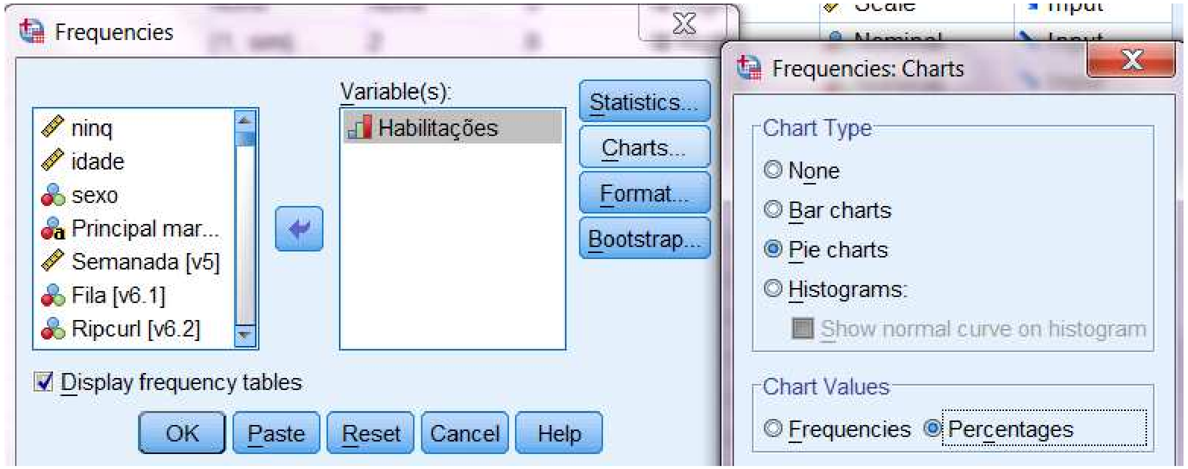
# Graphs – Legacy Dialogs – Bar – Simple



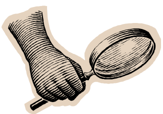
A tabela de frequências e os gráficos anteriores podem também ser obtidos em:

`Analyse – Descriptive Statistics – Frequencies`




**Habilitações**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Até ao 9ºano | 33 | 11,0 | 11,1 | 11,1 |
| | 9ºano ao 12ºano | 118 | 39,3 | 39,6 | 50,7 |
| | +12ºano | 147 | 49,0 | 49,3 | 100,0 |
| | Total | 298 | 99,3 | 100,0 | |
| Missing | System | 2 | ,7 | | |
| Total | | 300 | 100,0 | | |



Há muito para explorar no SPSS…

## Exemplo - Vendas (atividade proposta para trabalho individual)

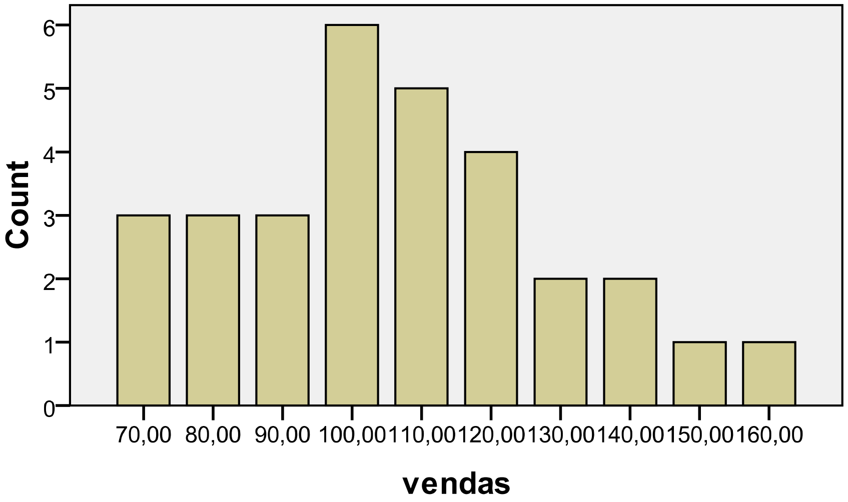
Construa um ficheiro em SPSS com os dados deste exemplo.

Obtenha a tabela de frequências e os gráficos abaixo.

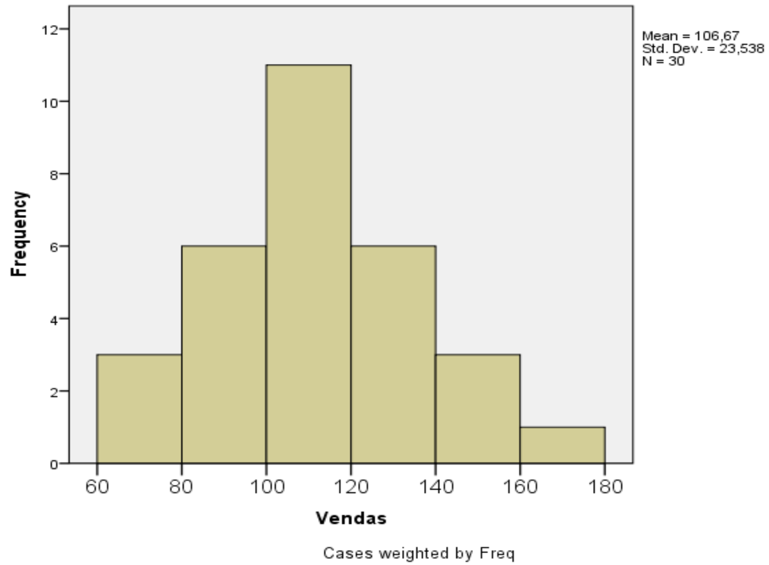
**vendas**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 70,00 | 3 | 10,0 | 10,0 | 10,0 |
| | 80,00 | 3 | 10,0 | 10,0 | 20,0 |
| | 90,00 | 3 | 10,0 | 10,0 | 30,0 |
| | 100,00 | 6 | 20,0 | 20,0 | 50,0 |
| | 110,00 | 5 | 16,7 | 16,7 | 66,7 |
| | 120,00 | 4 | 13,3 | 13,3 | 80,0 |
| | 130,00 | 2 | 6,7 | 6,7 | 86,7 |
| | 140,00 | 2 | 6,7 | 6,7 | 93,3 |
| | 150,00 | 1 | 3,3 | 3,3 | 96,7 |
| | 160,00 | 1 | 3,3 | 3,3 | 100,0 |
| | Total | 30 | 100,0 | 100,0 | |


Graphs – Legacy Dialogs – Line - Simple

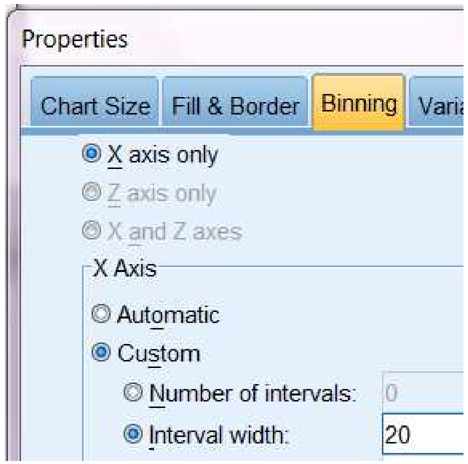



Graphs – Legacy Dialogs – Bar - Simple




Graphs – Legacy Dialogs – Histogram

(para obter um histograma igual ao anterior, é preciso clicar 2x no gráfico e alterar a amplitude das classes para 20)



# Medidas Descritivas

## Medidas de Localização e de Tendência Central

As medidas que descrevemos seguidamente dão-nos uma ideia do "centro" ou "localização" da distribuição dos dados.

### Média aritmética

Sejam $x_1$, $x_2$, ..., $x_p$ 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 $\overline{x}$ e é dada por:

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

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 – Pneus

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.

| $x_i$ | Freq. absoluta $n_i$ | Freq. relativa $f_i$ | Freq. abso. acum. | Freq. relat. acum.s | $n_i x_i$ |
|---|---|---|---|---|---|
| 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:

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

## 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 – Pneus:** 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 ≥ *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} $\rightarrow$ Mo=4 (distribuição unimodal);
- {2, 2, 3, 4, 4, 5} $\rightarrow$ Mo=2 e 4 (distribuição bimodal);
- **Exemplo – Pneus** $\rightarrow$ 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; $\boldsymbol{n_j}$ é a freq. abso. desta classe; $\boldsymbol{n_{j+1}}$ e $\boldsymbol{n_{j-1}}$ 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 (Mediana)– valor que tem cerca de 50% dos dados abaixo;

– $Q_{3/4}$, 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 $x_1, x_2, ..., x_n$, $n$ 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*

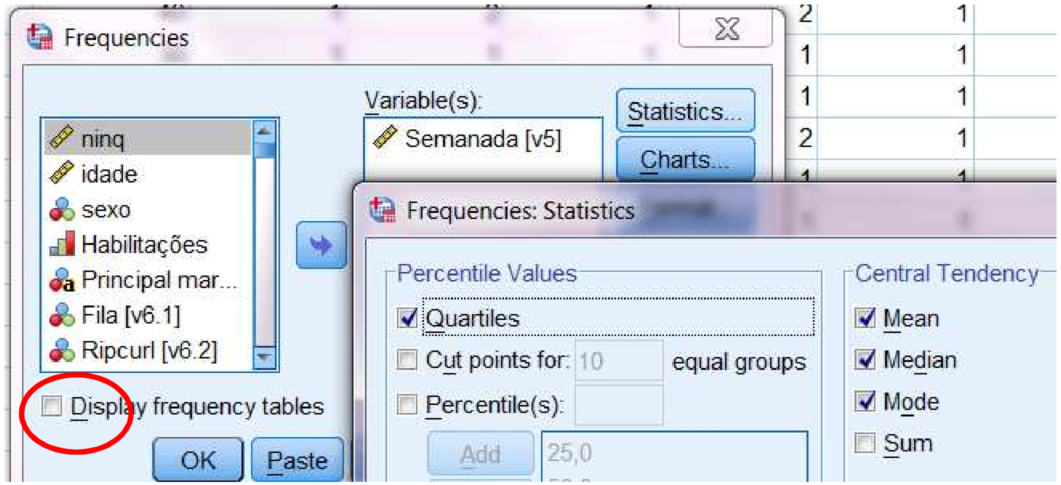
Seja $[a_j, a_{j+1}[$ a classe que contém $Q_p$, i.e., que contém o valor ao qual corresponde a frequência absoluta (resp. relativa) acumulada de $np$ (resp. $p$). Por interpolação linear obtém-se $Q_p$:

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

## Exemplo – Calçado Desportivo

Recorrendo ao SPSS, para a variável Semanada (v5) ...

Analyse – Descriptive Statistics – Frequencies




**Statistics**

Semanada

| | | |
|---|---|---|
| N | Valid | 300 |
| | Missing | 0 |
| Mean | | 28,65 |
| Median | | 28,50 |
| Mode | | 22 |
| Percentiles | 25 | 22,00 |
| | 50 | 28,50 |
| | 75 | 35,00 |

Interprete os valores fornecidos no quadro.

## Exemplo - Vendas (atividade proposta para trabalho individual)

Recorrendo ao SPSS obtenha a tabela seguinte.

**Statistics**

Vendas

| | | |
|---|---|---|
| N | Valid | 30 |
| | Missing | 0 |
| Mean | | 106,6667 |
| Median | | 105,0000 |
| Mode | | 100,00 |
| Percentiles | 25 | 90,0000 |
| | 50 | 105,0000 |
| | 75 | 120,0000 |

Complete as seguintes afirmações:

No mês de Outubro o volume médio de vendas por vendedor foi de ______________ .

Cerca de metade dos vendedores venderam menos de _____________ .

Cerca de 25% dos vendedores venderam menos de  ________________ .

Cerca de 25% dos vendedores venderam mais de _________________ .

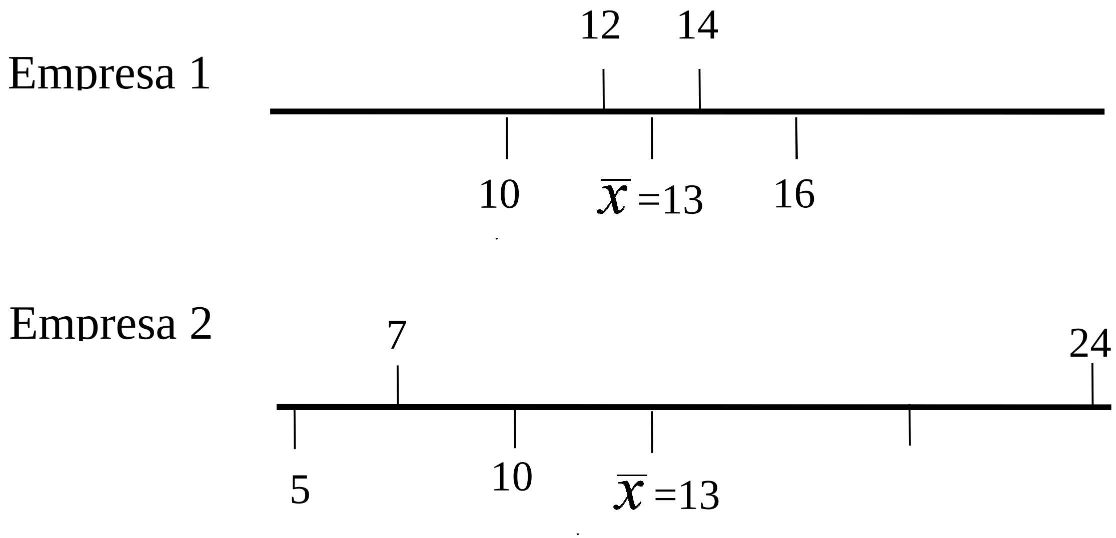
## Medidas de Dispersão

### 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 $\boldsymbol{n}$ 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 $\sigma^2$ e é calculada da seguinte maneira:

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

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

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

O **desvio padrão** é a raiz quadrada da variância e denota-se por $\boldsymbol{\sigma}$ ou por $\boldsymbol{s}$.

## Exemplo – Pneus

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

| $\mathbf{x_i}$ | Frequências absolutas $\mathbf{n_i}$ | Frequências relativas $\mathbf{f_i}$ | $\mathbf{n_i x_i}$ | $\mathbf{n_i x_i^2}$ |
|---|---|---|---|---|
| 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,

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

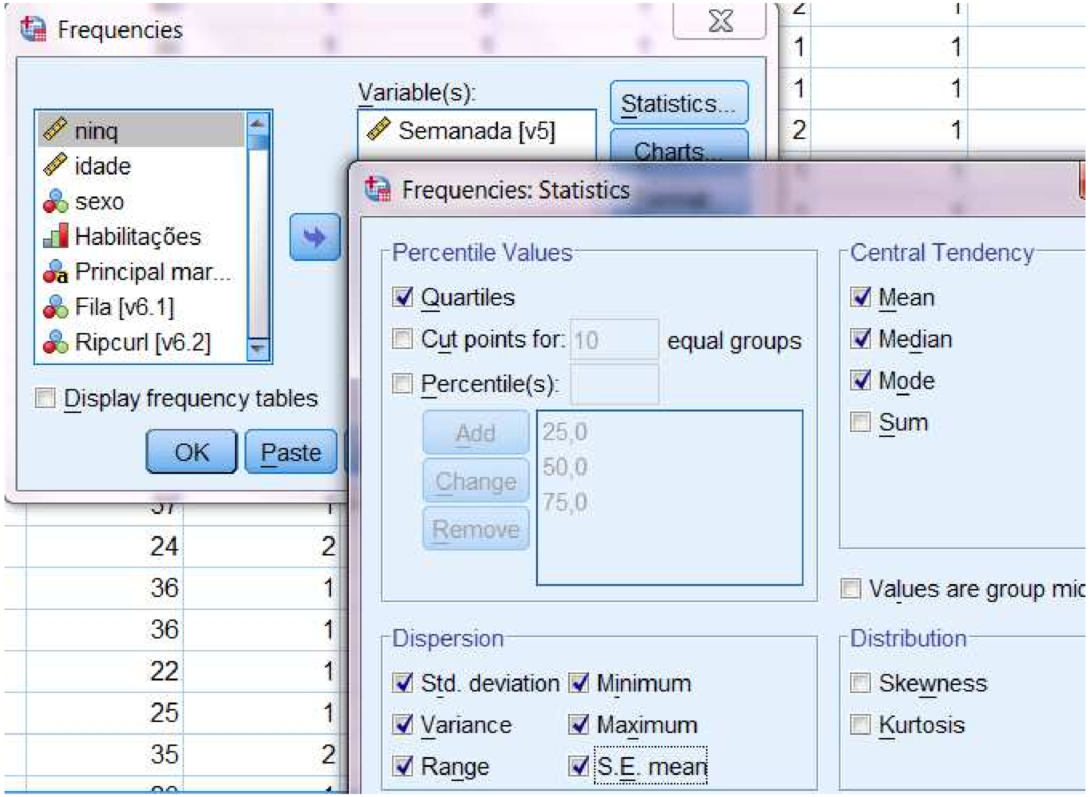
As medidas descritivas usadas com maior frequência em cada tipo de dados (escala de medida) são dadas no quadro seguinte:

<table>
<tr><th rowspan="2">Escala de medida</th><th colspan="2">Estatísticas descritivas</th></tr>
<tr><th>Medidas de localização</th><th>Medidas de dispersão</th></tr>
<tr><td>Nominal (sem qualquer relação de ordem).<br>● espécie ou género</td><td>Moda: valor mais frequente da amostra</td><td>Não há....</td></tr>
<tr><td>Ordinal (ordenável, mas sem quantificar).<br>● nível hierárquico<br>● ordem de chegada</td><td>Moda;<br>Quartis: Valor abaixo do qual estão 25%($Q_{1/4}$), 50%($Q_{2/4}$) ou mediana) ou 75%($Q_{3/4}$) dos valores da amostra</td><td>Amplitude interquartilica<br>AIQ= $Q_{3/4}$- $Q_{1/4}$<br>Intervalo interquartis=[ $Q_{1/4}$, $Q_{3/4}$]</td></tr>
<tr><td>Quantitativa (ordenável sendo possível quantificar as diferenças).<br>● nº de células<br>● altura<br>● temperatura</td><td>Moda<br>Quartis<br>Média</td><td>AIQ<br>Desvio padrão<br>$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{p} n_i\,(x_i - \overline{x})^2}$$<br>Erro padrão da média $s/\sqrt{n}$</td></tr>
</table>

## Exemplo – Calçado Desportivo

Recorrendo ao SPSS, podemos agora obter, para a variável Semanada (v5), as medidas de dispersão abordadas.

`Analyse – Descriptive Statistics – Frequencies`




**Statistics**

Semanada

| | | |
|---|---|---|
| N | Valid | 300 |
| | Missing | 0 |
| Mean | | 28,65 |
| Std. Error of Mean | | ,453 |
| Median | | 28,50 |
| Mode | | 22 |
| Std. Deviation | | 7,851 |
| Variance | | 61,645 |
| Range | | 44 |
| Minimum | | 3 |
| Maximum | | 47 |
| Percentiles | 25 | 22,00 |
| | 50 | 28,50 |
| | 75 | 35,00 |

## Exemplo - Vendas (atividade proposta para trabalho individual)

Recorrendo ao SPSS obtenha a tabela seguinte.

**Statistics**

Vendas

| | | |
|---|---|---|
| N | Valid | 30 |
| | Missing | 0 |
| Mean | | 106,6667 |
| Std. Error of Mean | | 4,29738 |
| Median | | 105,0000 |
| Mode | | 100,00 |
| Std. Deviation | | 23,53769 |
| Variance | | 554,023 |
| Range | | 90,00 |
| Minimum | | 70,00 |
| Maximum | | 160,00 |
| Percentiles | 25 | 90,0000 |
| | 50 | 105,0000 |
| | 75 | 120,0000 |

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

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

## Exemplos (atividades propostas para trabalho individual)

**1.** Uma transportadora pretende comparar a variabilidade do peso com a variabilidade do volume das caixas que transporta habitualmente. Sabe-se que as caixas têm um peso médio de 10 quilos com um desvio padrão de 3 kg. O volume médio é de 30 $dm^3$ com um desvio padrão de 6 $dm^3$. Ajude a transportadora neste estudo.

**2.** Compare a variabilidade do peso dos patos adultos e dos elefantes asiáticos adultos sabendo que os patos de uma determinada espécie de têm um peso medio de 4 kg com desvio padrão de 1,2 e que os elefantes asiáticos pesam em média 4200 kg com um desvio padrão de 400 kg.

**3.** Qual dos dois conjuntos de dados tem maior dispersão, A (em unidades monetárias) ou B (em metros)?

A={1, 13, 2, 6, 9} B={10.3, 126.52, 29.76, 5.931, 7.463 }

## Coeficientes de assimetria

As distribuições de frequências podem ser simétricas ou não.

Considerando apenas distribuições unimodais, temos, em geral:

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

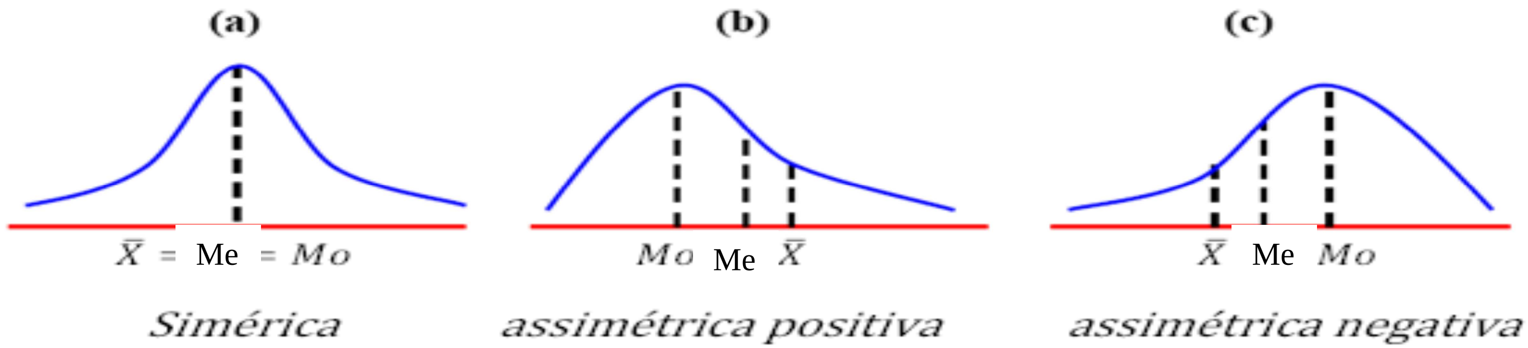
Distribuições assimétricas positivas $\rightarrow \; Mo < Me < \bar{x}$

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.



Há vários coeficientes para medir a assimetria; estes, de um modo geral, interpretam-se da seguinte maneira:

Distribuição **simétrica** $\rightarrow$ ***coeficiente*** $\cong$**0**

Distribuição **assimétrica positiva** $\rightarrow$ ***coeficiente >0***

Distribuição **assimétrica negativa** $\rightarrow$ ***coeficiente <0***

## Coeficientes de achatamento

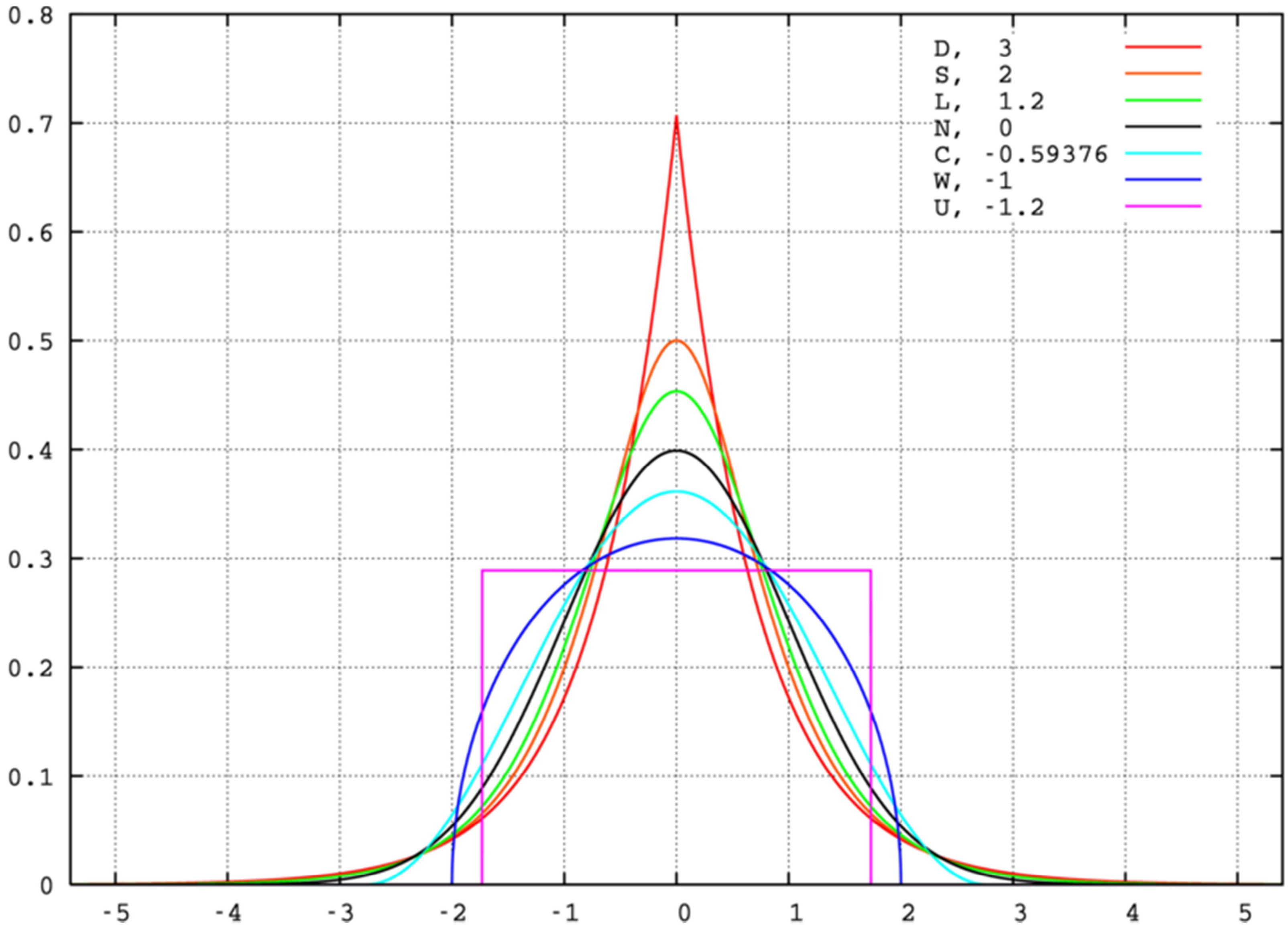
Estes medem o grau de achatamento de uma distribuição, considerado em relação ao da distribuição normal. Para o coeficiente calculado pelo SPSS a interpretação é a seguinte.

Distribuição **mesocúrtica** $\rightarrow$ ***coeficiente*** $\cong$**0** (achatamento igual ao da normal)

Distribuição **leptocúrtica** $\rightarrow$ ***coeficiente >0*** (achatamento inferior ao da normal)

Distribuição **platicúrtica** $\rightarrow$ ***coeficiente <0*** (achatamento superior ao da normal)

(http://pt.wikipedia.org)

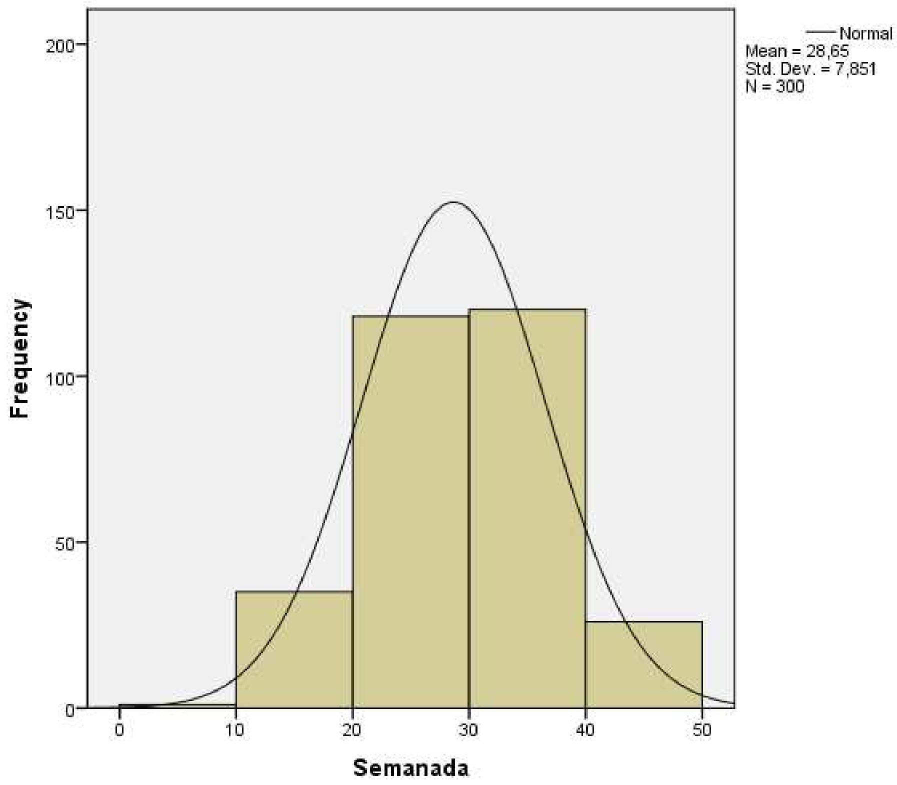



A curva normal está representada pela linha preta.

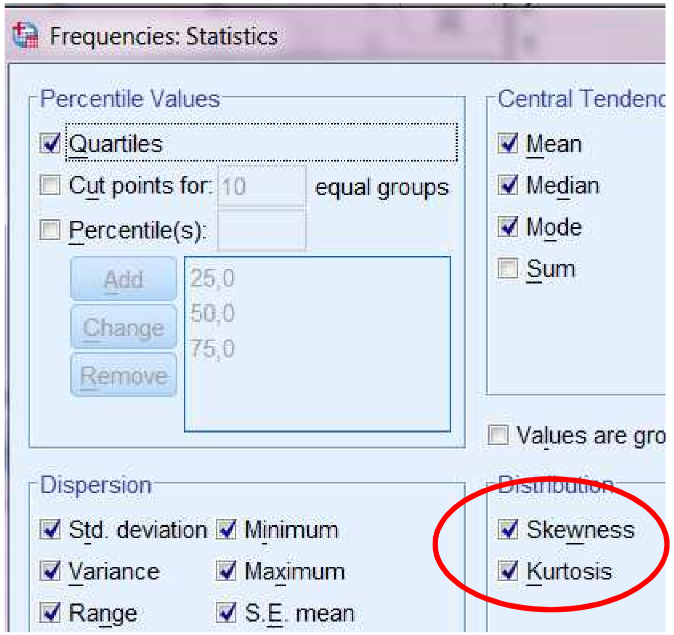
## Exemplo – Calçado Desportivo

Recorde o histograma da varável semanada:

A distribuição aparenta ser aproximadamente simétrica…




Analyse – Descriptive Statistics – Frequencies



**Statistics**

Semanada

| | | |
|---|---|---|
| N | Valid | 300 |
| | Missing | 0 |
| Mean | | 28,65 |
| Std. Error of Mean | | ,453 |
| Median | | 28,50 |
| Mode | | 22 |
| Std. Deviation | | 7,851 |
| Variance | | 61,645 |
| Skewness | | -,007 |
| Std. Error of Skewness | | ,141 |
| Kurtosis | | -,778 |
| Std. Error of Kurtosis | | ,281 |
| Range | | 44 |
| Minimum | | 3 |
| Maximum | | 47 |
| Percentiles | 25 | 22,00 |
| | 50 | 28,50 |
| | 75 | 35,00 |

Coeficiente de assimetria $\cong 0$ → distribuição simétrica

Coeficiente de achatamento<0 → distribuição platicúrtica

# Outras Representações Gráficas

## Caixas de bigodes ou diagramas de extremos e quartis (boxplots)

Neste gráfico representam-se a mediana, os quartis, os valores máximo e mínimo e eventuais valores extremos ou outliers.

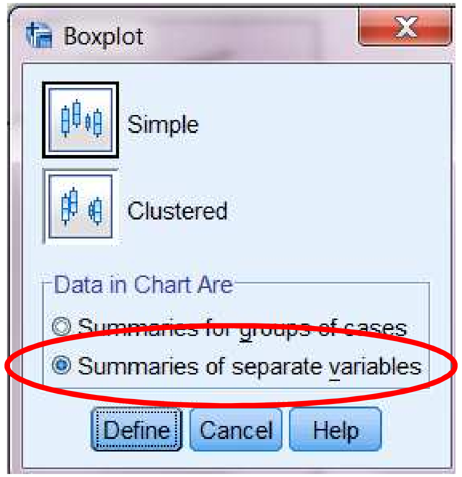
Os valores **extremos** são assinalados com um asterisco e correspondem a valores acima do 3º quartil, ou abaixo do 1º quartil, pelo menos 3 vezes a amplitude interquartis.

Os **outliers** são assinalados com uma bola e correspondem a valores acima do 3º quartil, ou abaixo do 1º quartil, de 1.5 a 3 vezes a amplitude interquartis.

### Exemplo – Calçado Desportivo

Para a variável Semanada (v5)…

`Graphs – Legacy Dialogs – Boxplot – Simple`

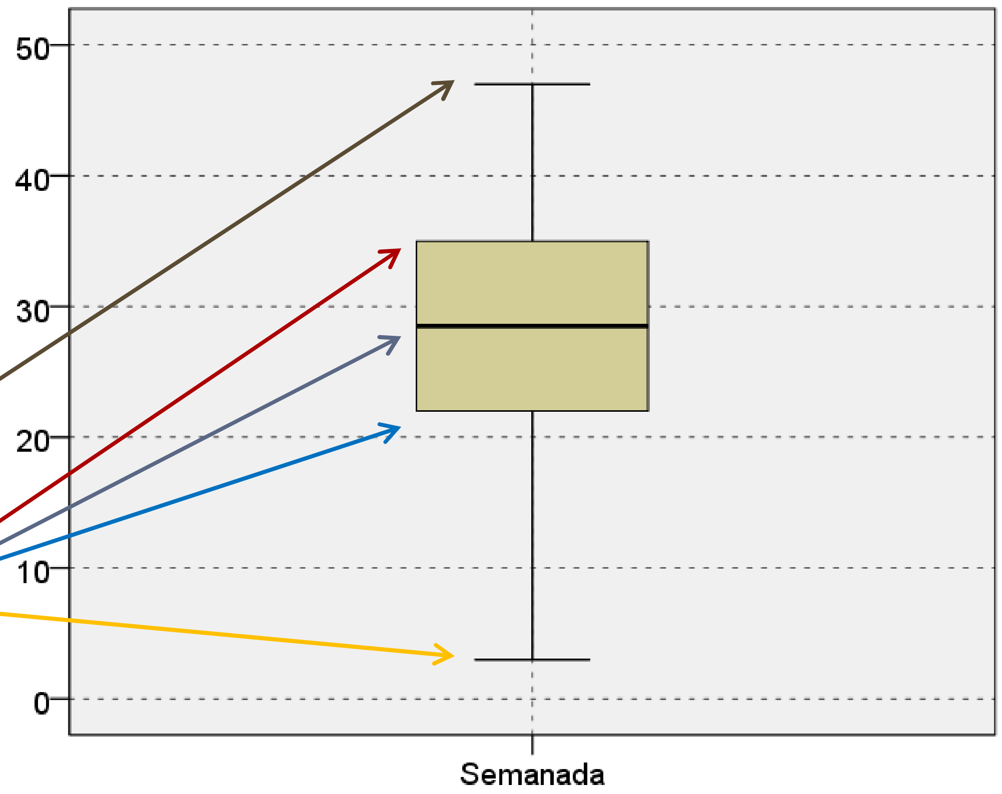



Recorde a tabela seguinte

**Statistics**

Semanada

| | | |
|---|---|---|
| N | Valid | 300 |
| | Missing | 0 |
| Mean | | 28,65 |
| Std. Error of Mean | | ,453 |
| Median | | 28,50 |
| Mode | | 22 |
| Std. Deviation | | 7,851 |
| Variance | | 61,645 |
| Range | | 44 |
| Minimum | | 3 |
| Maximum | | 47 |
| Percentiles | 25 | 22,00 |
| | 50 | 28,50 |
| | 75 | 35,00 |

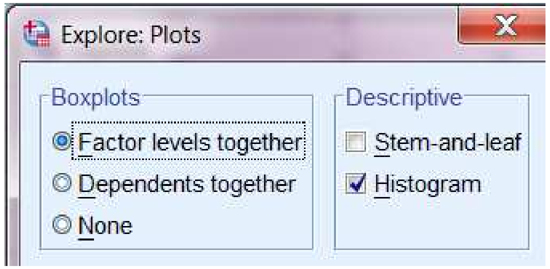



Este gráfico, bem como uma tabela com algumas estatística descritivas, pode também ser obtida através de

`Analyse – Descriptive Statistics – Explore…`

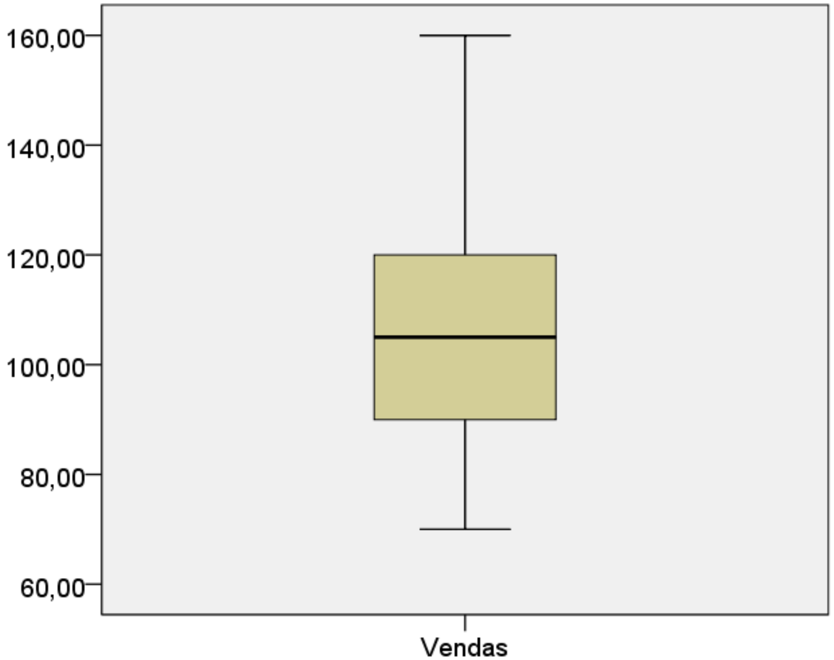
… colocando a variável `Semanada` na caixa `Dependent List;`

...no botão `plots` podemos também pedir o histograma




**Exemplo - Vendas** (atividade proposta para trabalho individual)

Recorrendo ao SPSS obtenha o diagrama de extremos e quartis.

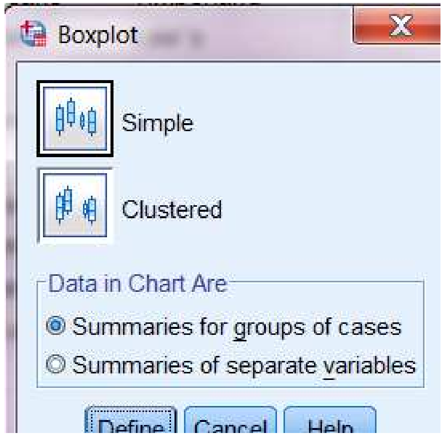


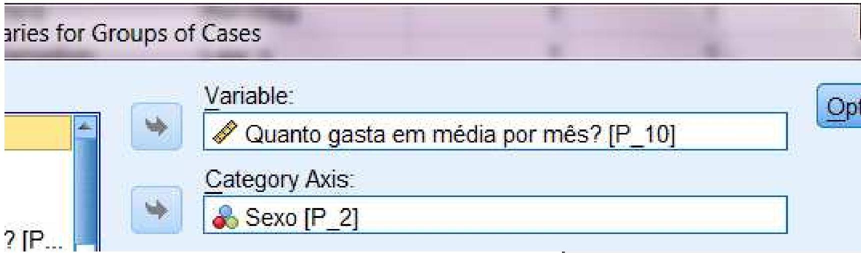
# Comparando distribuições…

## Exemplo – Vestuário

Para a variável P10 – Quanto gasta em média por mês, comparemos os géneros…

`Graphs – Legacy Dialogs – Boxplot – Simple`





Também podíamos ter feito:

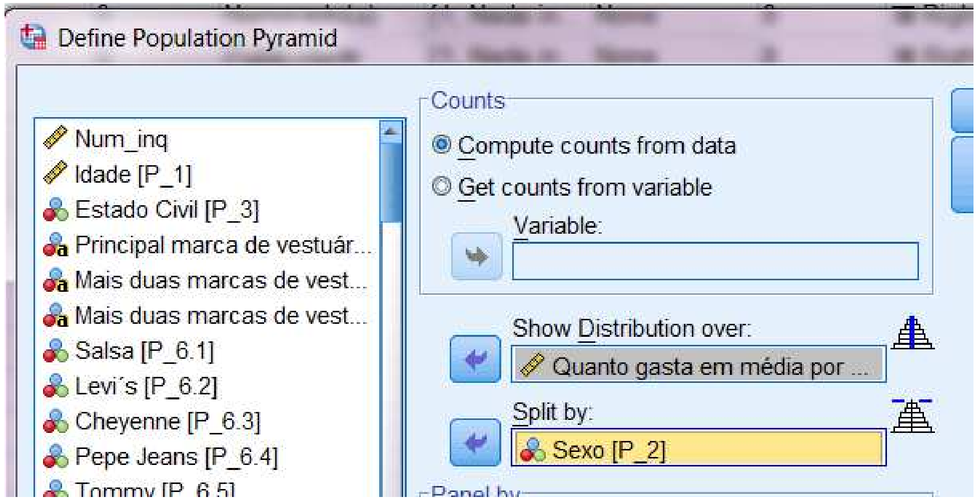
`Analyse – Descriptive Statistics – Explore…`

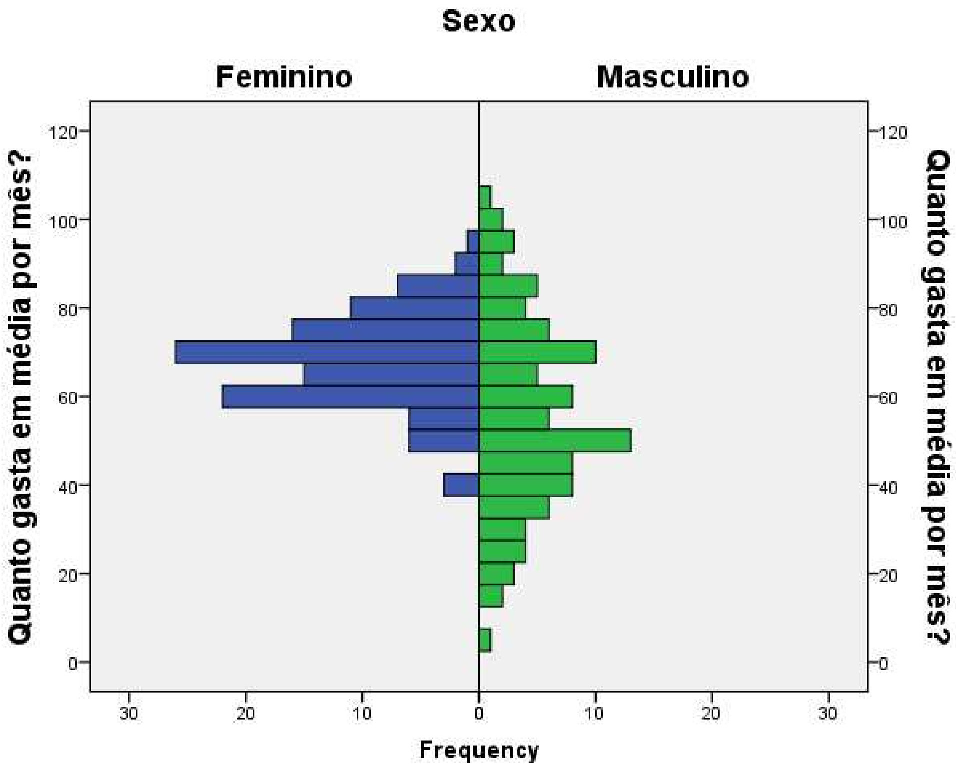
Colocando a variável sexo na caixa `Factor List`

Comente… (em termos de localização e de dispersão)

Outra forma gráfica para fazer a comparação é:

`Graphs – Legacy Dialogs – Population Pyramid`




Este gráfico evidencia bastante claramente uma maior dispersão dos valores dos gastos no género masculino do que no feminino.

Quanto à tendência central, o género feminino apresenta valores superiores, apoiando a ideia estereotipada de que as Senhoras, em relação às compras, …

## Exemplo - Vestuário (atividade proposta para trabalho individual)

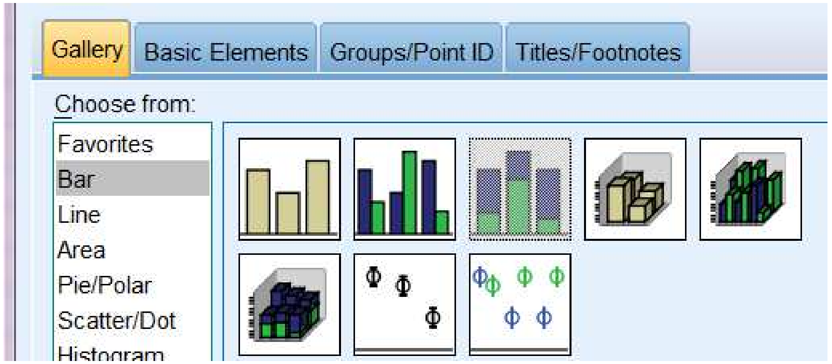
Compare os dois géneros quanto ao tempo gasto numa sessão de compras (P_13) e comente.

## Ainda, comparando distribuições…

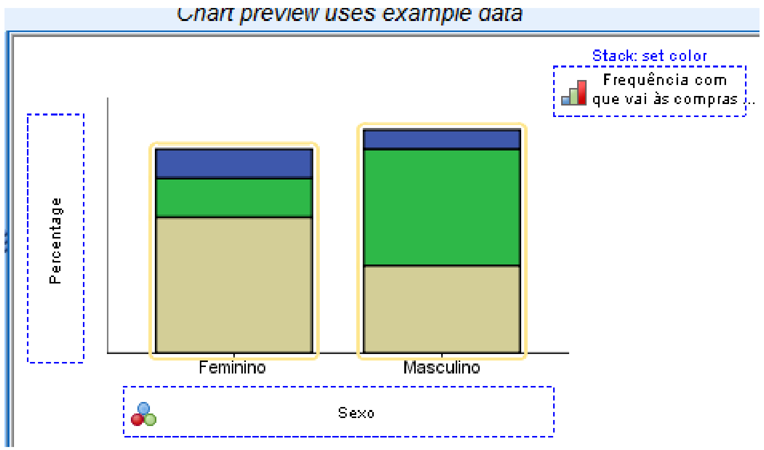
## Exemplo – Vestuário

Comparando os géneros quanto à frequência de compras (P11) …

SPSS: `Graphs – Chart Builder- Ok`
`Bar – 3º gráfico da 1ª linha`

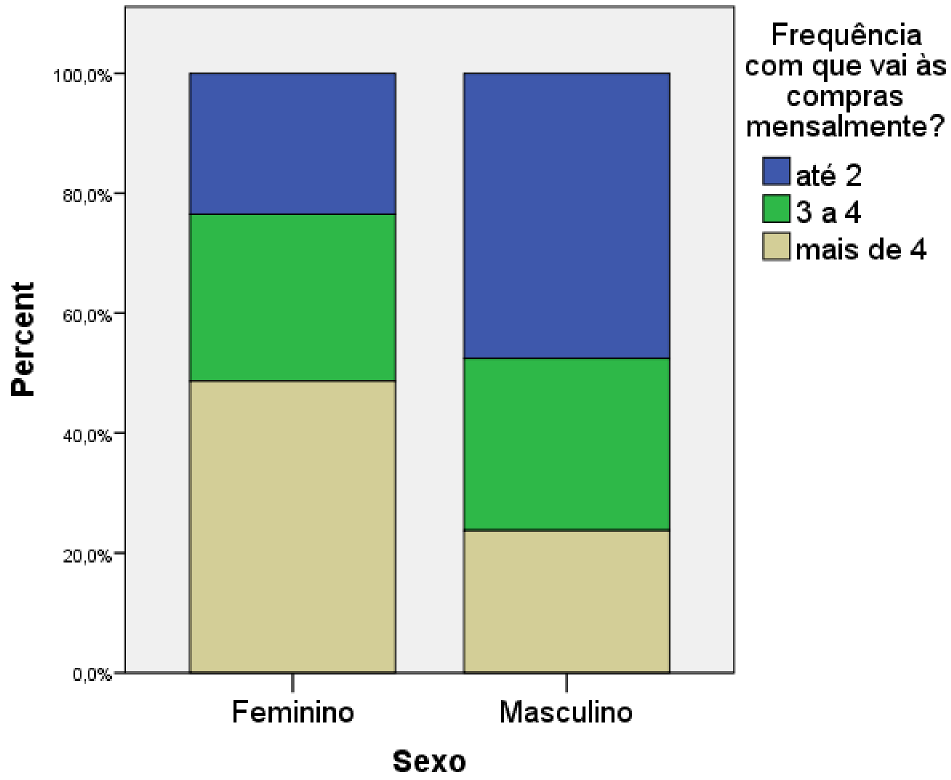


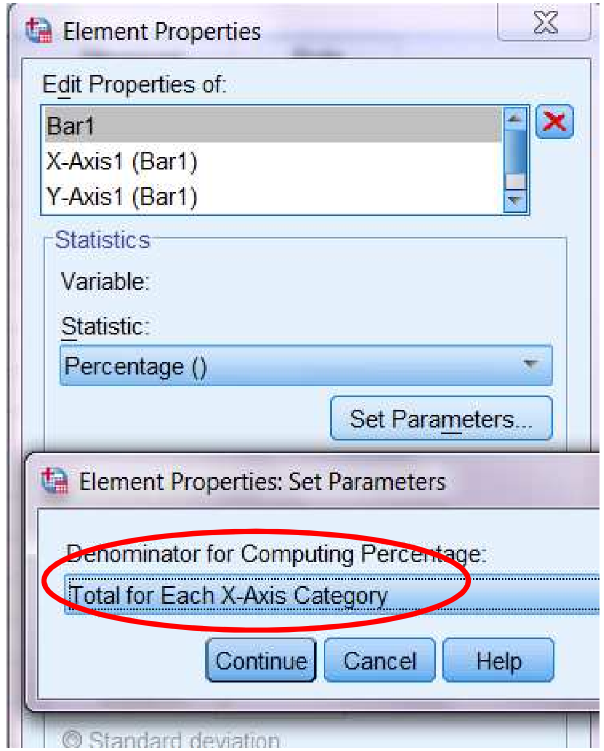

Arrastar a variável `Sexo` para o retângulo `X-axis`
Arrastar a variável `Frequência com que vai às compras` (P11) para o retângulo `Stack: set color`



Clicar no botão `Element Properties`

Clicar no botão `Apply`




No género feminino é maior a percentagem de pessoas que vai às compras mais de 4 vezes por mês…

Há muito para explorar no SPSS…