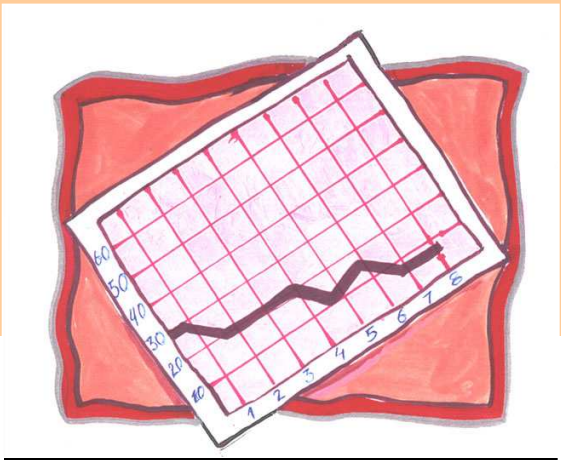
Educational Dossiers



Graphical Representations

Notes on the creation and presentation of several types of charts

Ana Alexandrino da Silva

Foreword

The ALEA project – Local Applied Statistics Initiative – contributes toward the creation of new statistics teaching support media for primary and secondary students and teachers.

The project arose from an idea jointly fostered by Tomaz Pelayo secondary school and Instituto Nacional de Estatística [National Statistics Institute of Portugal], founded on the requirements and structures that the intervening parties possess. The improvement of statistical literacy is thus a significant proviso in guaranteeing the provision of a service of public value. The teaching of statistics in lower and upper secondary education constitutes one of the most important instruments aimed at achieving this objective. ALEA's site on the internet is at the address: www.alea.pt.

The Educational Dossiers area was designed to support the creation of educational material on a range of topics. The dossier **Graphical Representations** is now presented, in PDF format.

Dossiers available in the English version of ALEA:

– Notes on the History of Statistics
– The Cinema in Numbers
– Graphical Representations
– Statistical Surveys

Contents

1.1. Introduction

Charts are present in nearly all information-provision channels, particularly in newspapers and magazines, school text books, public presentations and not even our individual reports can get by without them.

Nevertheless, producing a chart or a map that really provides information and is, at the same time, appealing, legible and data-coherent, is not an easy task.

The great advantage of charts is their capacity to tell a story in an interesting and attractive manner, allowing the viewer to rapidly understand phenomena that would be difficult to understand any other way. This virtue, nonetheless, does not mean that this process is performed in a simple manner. A large amount of work and care is required.

There are countless ways of displaying statistical information figuratively, and the huge range of graphical possibilities is such that we felt it more appropriate to restrict the scope of this dossier to the most common charts, rather than carry out an exhaustive analysis.

The history of graphics

The history of statistical graphics is relatively recent. The greatest advance occurred only about 200 years ago, in 1786, as a result of the work of William Playfair, who invented most of the charts that we use today: the bar graph, the line graph based on economic data and the pie chart.

While the 19th century was noted for the creation and broad dissemination of statistical charts amongst the scientific community, the exponential increase in their use in extensively published documents accessible to the general public occurred in the 20th century.

The dissemination of statistical graphics has taken huge leaps forward since the time of Playfair. Statistical charts are now used in some measure in all sectors - in schools, in the media etc., but the majority of charts currently used date back to Playfair's time (18th/19th century).

Studies in the field of graphs recommenced with the advent of personal computers. A compulsory reference in this field is Edward TUKEY (1977), who invented box-and-whiskers plots and stem-and-leaf diagrams, amongst others. These charts are essential in the exploratory analysis of data.

Some notes on the construction of charts

The production of charts, with the technology available today, is in everybody's grasp. But a certain degree of care must be taken.

This dossier sets out a set of criteria underlying the creation of a chart. This process begins as soon as a graphical representation of the data is decided on and it

finishes when the end result is deemed to be satisfactory.

Readers have become demanding in relation to graphs, given that they are constantly bombarded with them. A chart that is overloaded with data may cause viewers to retreat from it or, even if it is viewed with attention, viewers will not be capable of recalling the data afterwards. This distancing may also be caused by an excess of non-informative graphical components, giving rise to what TUFTE (1983) called “chart junk”.

One must, before anything else, ask whether graphical representation of the data is necessary. In certain cases the use of a chart may not be most appropriate when the objective is not to provide an image but to provide hard data. This is just as true in situations of small quantities of data as it is when the intention is to disclose large quantities of data.

Another problem one may be faced with when producing charts is space restriction, causing the data to be accumulated into one single chart or into a number of charts of reduced size, thereby hampering their interpretation.

WALLGREN (1996) summarises this preparatory phase into eight questions that cannot be separately answered to:

- Is a chart really the best option?
- What is the target public?
- What is the objective of the chart?
- What kind of chart should be used?
- How should the chart be presented?
- How big should the chart be?
- Should only one chart be used?
- What technical means should be used?

The construction of the chart only starts after the selection of the type of chart most appropriate to the particular context.

When one thinks that finally has the required chart, then it is essential that a critical analysis is carried out, to see if this is the most effective way of transmitting the initial message. A misunderstood chart can lead to the wrong interpretation. On the other hand, a visually unpleasing chart can cause readers to disconnect, instead of informing them, “*A poor chart is worse than no chart at all!*” (WALLGREN, 1996, p. 89).

An iterative process is commenced in an attempt to find the best image that satisfies all the initial requirements. This process only ends when a high level of legibility and pertinence is achieved. The use of a chart can, therefore, only be decided after the following questions have been suitably answered:

- Is the chart easy to read?
- Can the chart be misinterpreted?
- Does the chart have the right size and shape?
- Is the chart in the right place?
- Does the chart benefit from being in colour?
- Was the understanding of the chart tested with anybody?

Formatting a chart

Graphical representation is a complex subject that encompasses areas as diverse as statistics, drawing and psychology. A chart can correctly display the variables, contain all of the necessary elements, and yet not be attractive nor easy to read.

A chart can be redrawn by modifying or hiding some of the graphical components, without losing information (TUFTE, 1983). Nevertheless, many charts that are published require a certain degree of sophistication in this regard. It is common to find images that are visually alike, produced by the chart assistant of Microsoft Excel software, and which, due to the fact that they are very frequently seen, are tiring and unattractive to the reader.

MS Excel software allows a certain degree of visual manipulation of the range of charts it produces. There follows an example of how to improve the visualization of a chart by modifying its parameters.

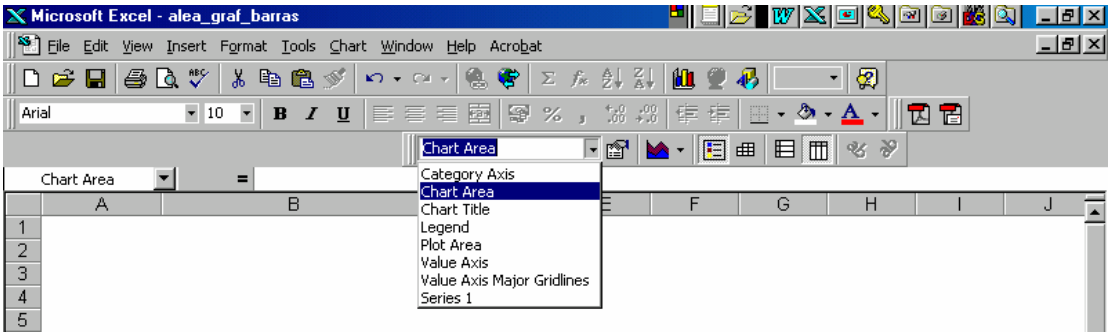
The first thing to take into account when drawing a chart is data organisation. The type of chart selected is influenced by the manner in which the data are organised. The best way is to place the data in a table with its labels, so that these may be used as the chart's title and labels.

Data table:

Academic qualifications of the 15-64 year age group

Gender / Academic qualifications	Male	Female
None	7,5%	11,3%
Compulsory education	69,3%	61,5%
Secondary education	15,7%	16,7%
Higher education	7,5%	10,5%

Chart formatting parameters in MS Excel



1. Chart area
2. Legend
3. Category axis
4. Plot area
5. Value axis
6. Gridlines
7. Series

Description of the formatting process

Using the criterion that at least two-thirds of the chart's area must be allocated to bars or, more generically, to the plot area (SCHMID, 1992), the space occupied by these was augmented.

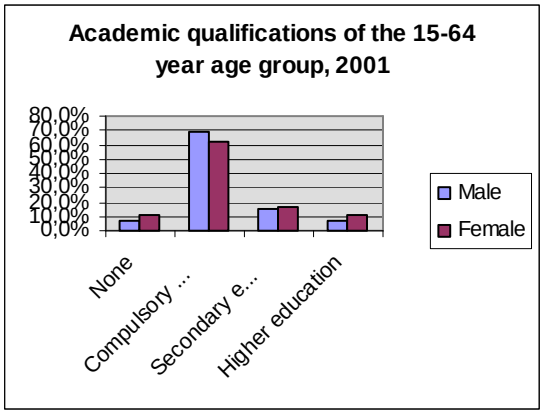
The decimals were removed from the value axis and some values were hidden. The respective gridlines were maintained. The % sign could have been removed from the 0 and 40 values and just kept for the last value. The value axis line was removed as well as the tick marks on the same. The amplitude of the interval of values was reduced, given that the largest of the bars did not exceed 80%.

The category axis line is provided with greater visual weight than any other auxiliary line and the labels are arranged horizontally to aid reading.

The borders of the chart, bar and legend were removed since there was no particular advantage in keeping them, as they unnecessarily overloaded the chart. The legend was placed within the chart's boundaries, to reduce the distance travelled by the eyes between the components and their identification. The colours of the bars were changed. The thickness of the bars was increased and the space between groups of bars was reduced.

The figure on the right is the result of the chart on the left, amended using the specific features of Excel software.

Before



After

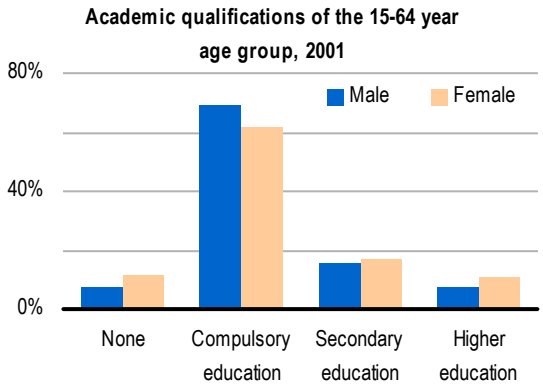
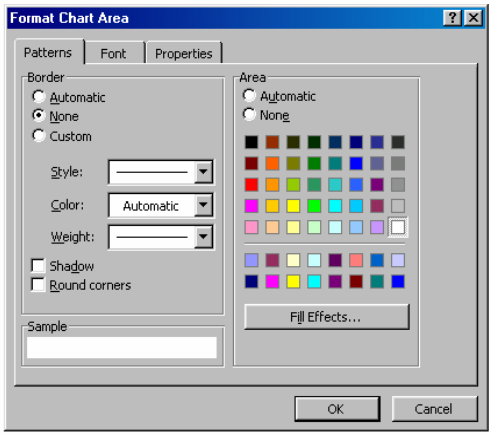


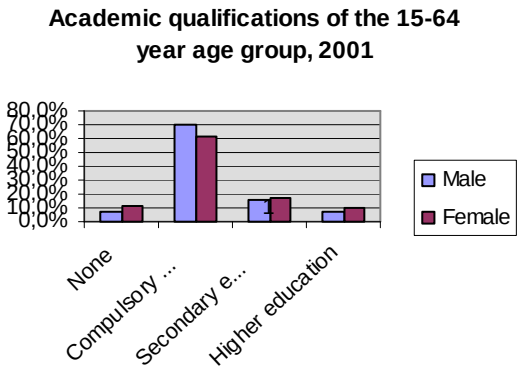
Figure 1 – Bar graph before and after being modified using Microsoft Excel software

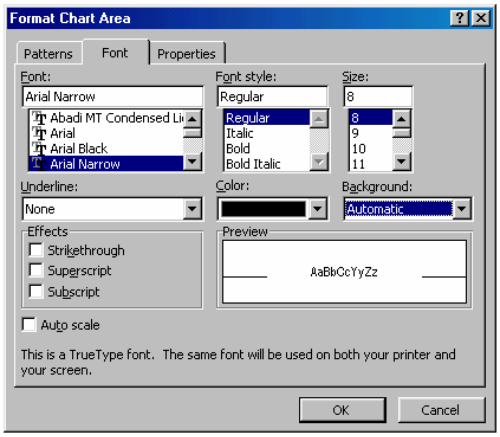
(Re)drawing a chart using Excel

1 – Chart area

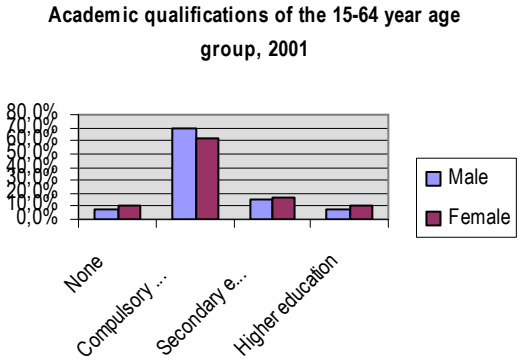


The chart without borders and white chart area…

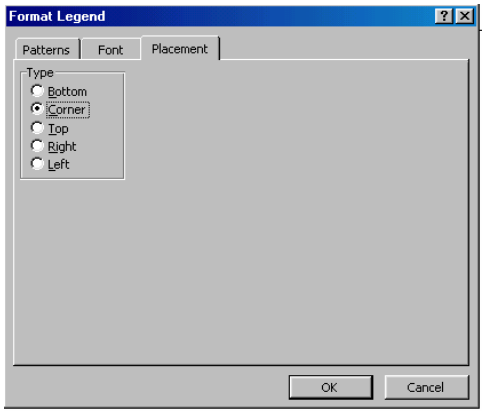




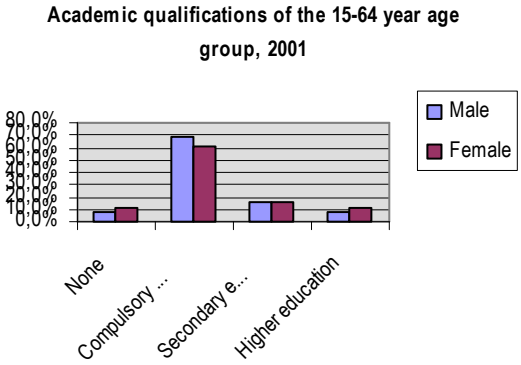
The chart with Arial Narrow size 8 font…

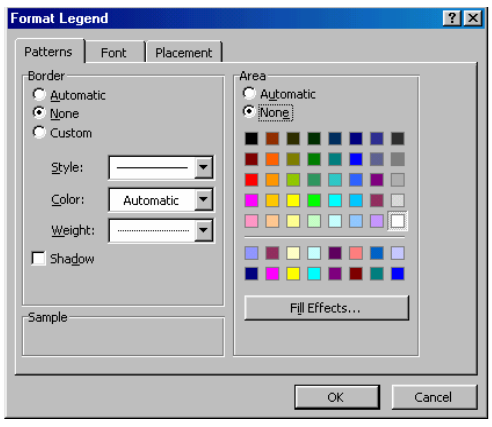


2 – Legend

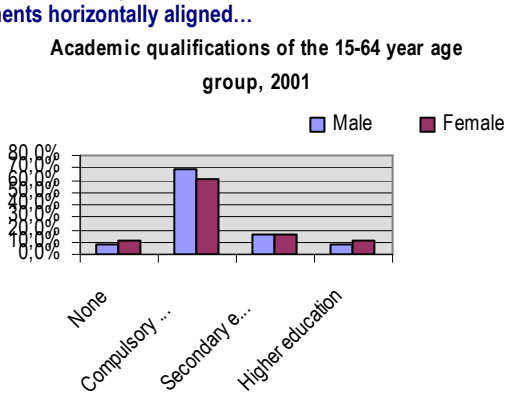


The chart with the legend in the top right-hand corner…

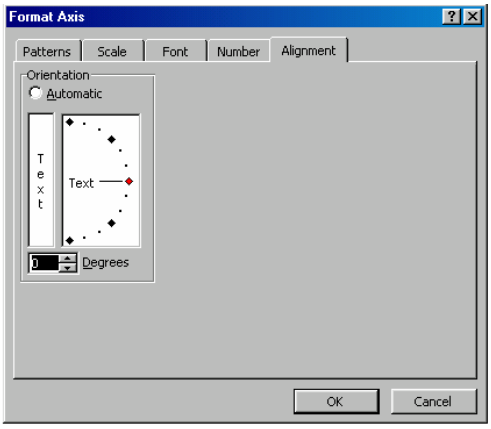




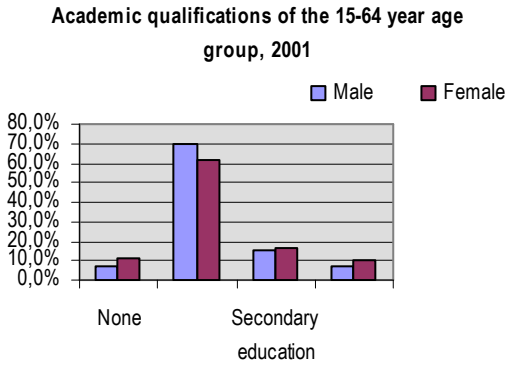
The chart with the legend without borders and fill and with components horizontally aligned…



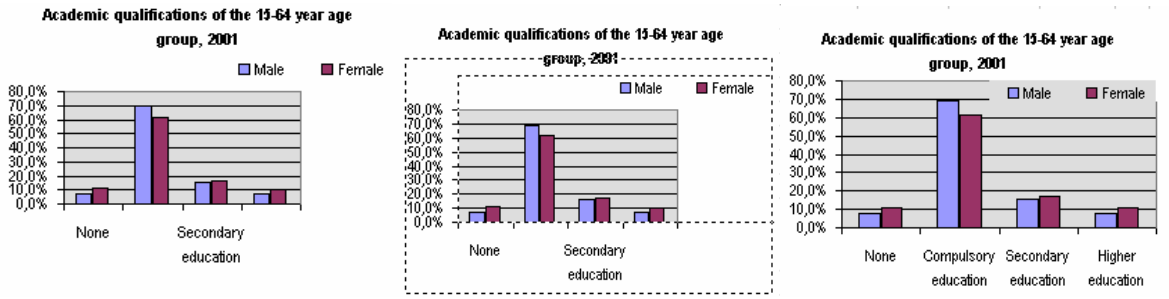
3 - Category axis



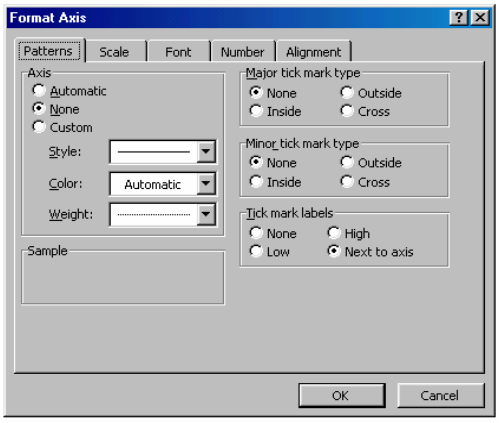
The chart with category labels horizontally aligned…



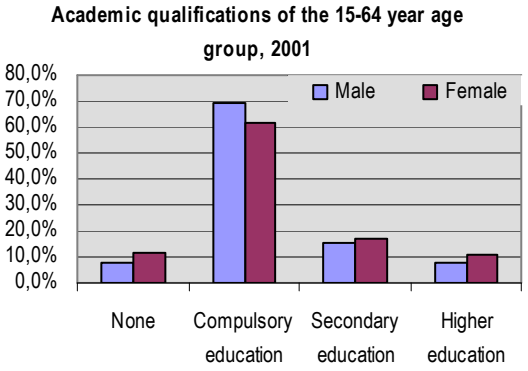
Modification of elements inside the chart - The chart area can be increased so that all the labels appear in the viewing field.

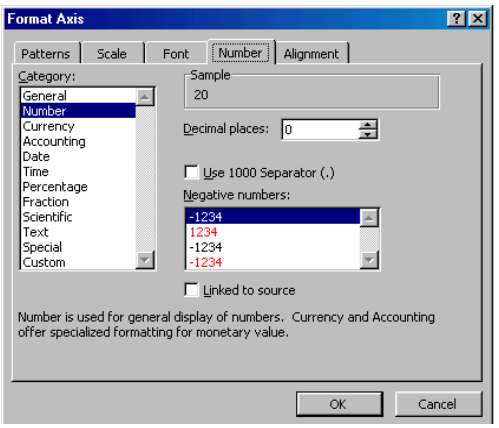


4 - Value axis

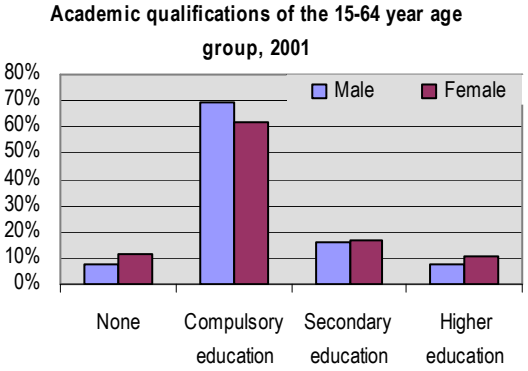


The chart without a value axis line and tick marks…

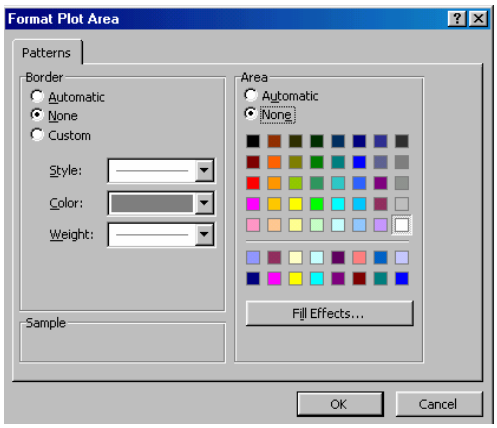




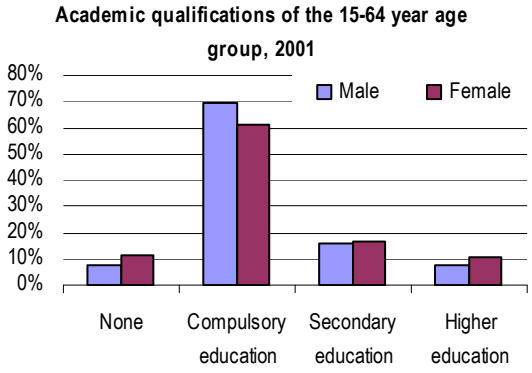
The chart with a value axis without decimals…



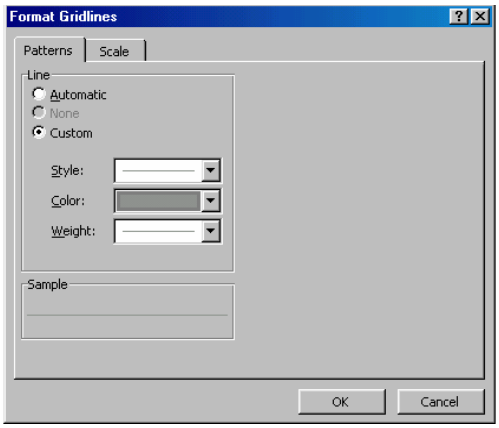
5 - Plot area



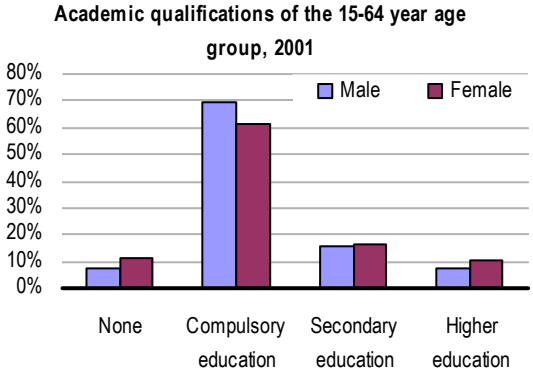
The chart with a white plot area without borders…

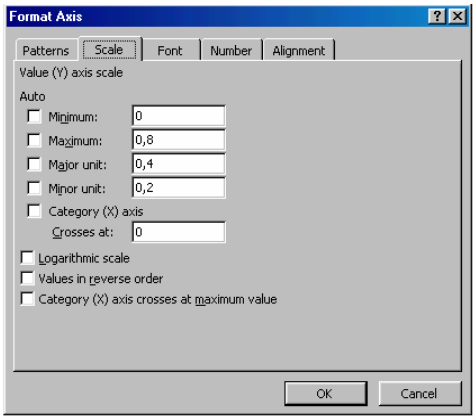


6 - Gridlines

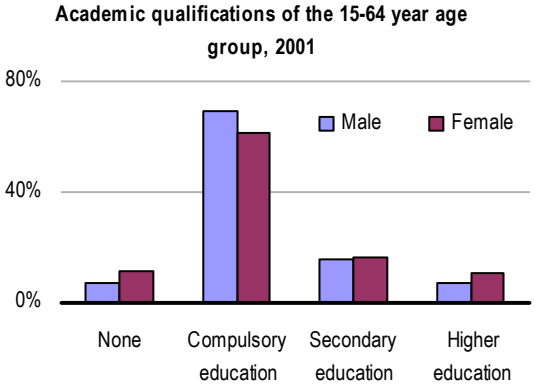


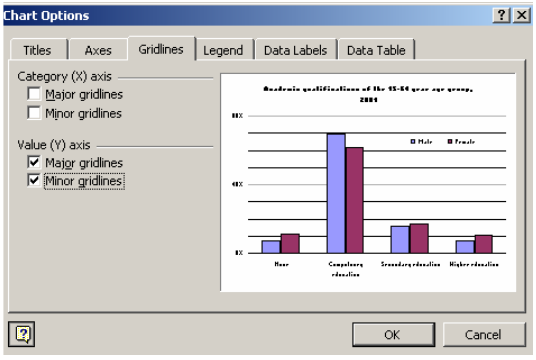
The chart with grey gridlines…

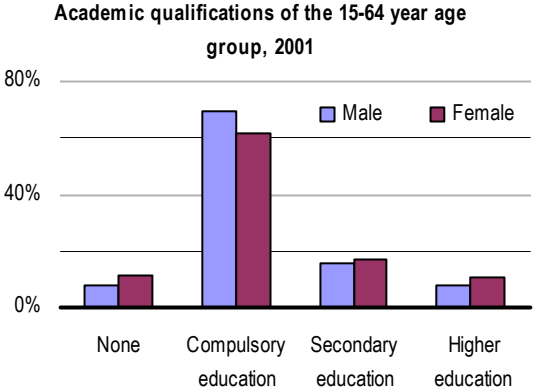




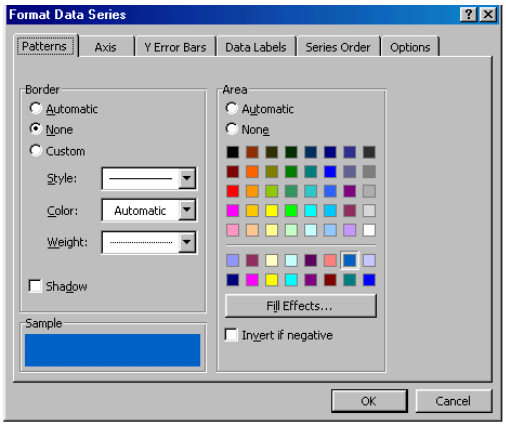
The chart with a scale for the two types of gridlines…

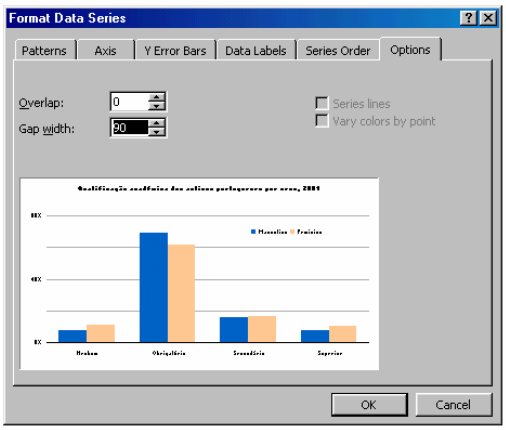




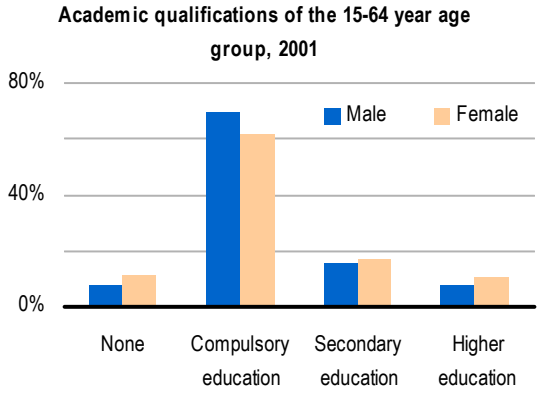


7 - Data series

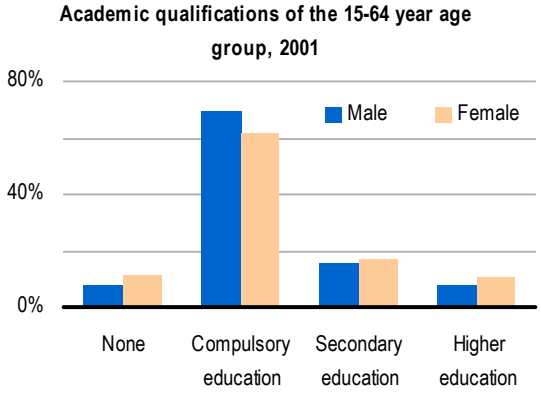




The chart with different coloured bars and without a border…



The chart with an amended gap width…



Perception studies

Graphical perception is one of the most important components to be taken into consideration when drawing a chart, since it provides a scientific foundation for the construction of the chart and provides endorsement for the selection of one kind of chart instead of another. The readability of images can be limited by the difficulty in correctly estimating the displayed data.

During the construction phase, the information is encoded in the chart using symbols, lengths, slopes of segments of lines, areas, texture or colour. When a chart is analysed, the encoded information is visually decoded and this decoding process is known as graphical perception, which is a factor that controls a chart's capacity to transmit information (CLEVELAND, MCGILL, 1987).

The extraction of information from charts involves perceptive tasks performed by the eye-brain visual system. These tasks are ranked in the following table according to their accuracy in the extraction of quantitative information. Less accurate perceptive tasks produce greater reading errors. In other words, there is greater variation between the perceived and correct values.

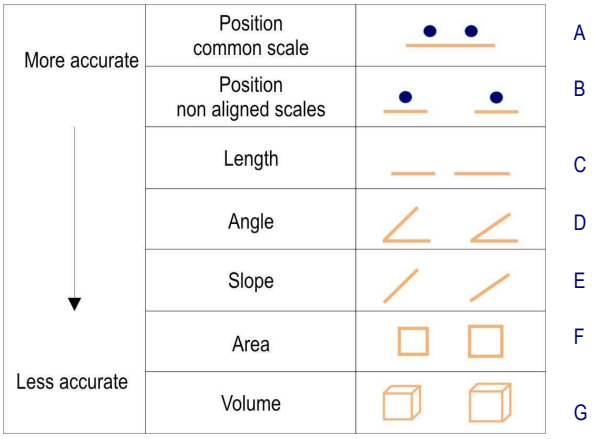


Figure 2 – Assessment of perceptual tasks ranked by accuracy
(adapted from CLEVELAND, MCGILL, 1984, 1987)

In grouped bar graph, for example, the viewer estimates the values through the position of the bars on the same scale or on separate scales, depending on how the data is presented.

The comparison of bars close to each other (Figure 3 - A) is better than the comparison of bars further apart (Figure 3 - B), in other words, it is harder for a viewer to estimate values in the latter case.

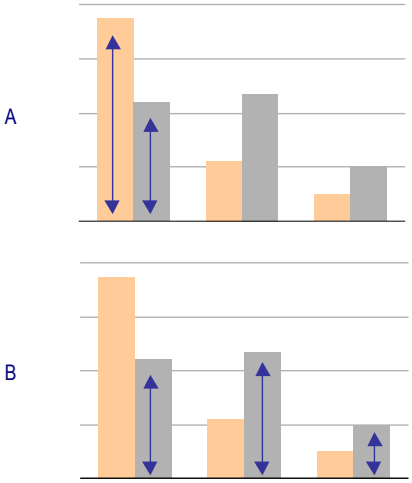


Figure 3 – Examples of tasks A and B

Bar graphs are more perceptively suitable than pie charts, given that estimating lengths is known to be two times more accurate than estimating angles. Let's analyse the case of Norte and the Lisboa e Vale do Tejo region, for example. In the pie chart we cannot ascertain which one is the largest. But the bar graph, on the other hand, clearly displays the difference.

It is common to come across 3-D charts in which the depth does not describe any variable at all. Since volume is the measure that causes the greatest problems in terms of perception, it should not be used.

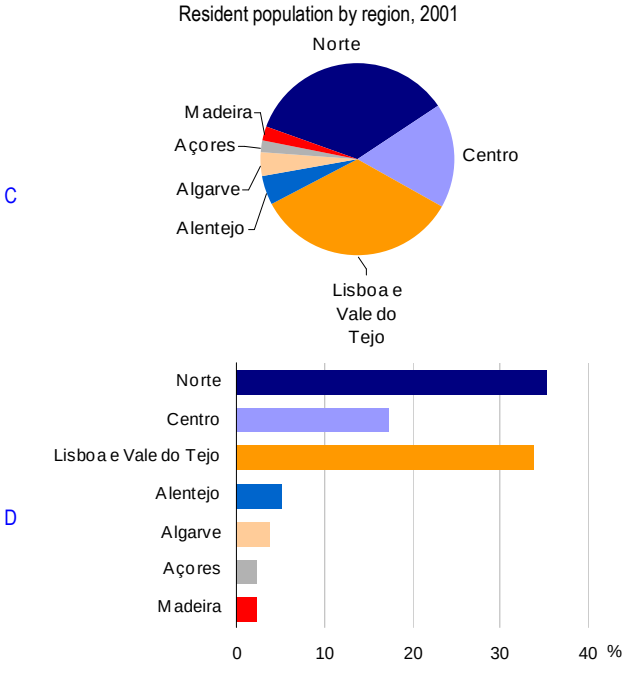


Figure 4 – Examples of tasks C and D

Graphical components

Charts contain the following set of components: the title, the value and category axes (based on the coordinate system), the legend, the labels for data and gridlines (Figure 5)

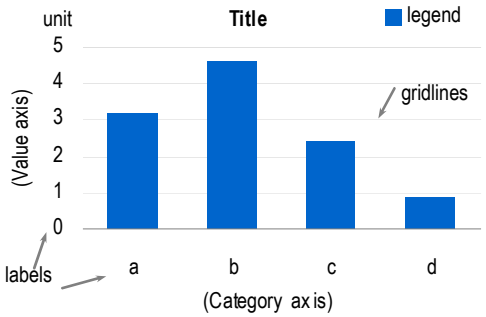


Figure 5 – Components of a chart

These components are composed of graphical symbols (points, lines, numbers, letters etc.) and their variation (colour, value, etc.).

The area occupied by the graph can contain all or just some of these components, arranged into two complementary areas: the plot area, which contains the actual graphical plot, and the chart area, in which the auxiliary interpretation components are normally placed (the title, legend and labels).

Chart area:

Title

Any type of graphical plot must possess a title, which must orientate the viewer with regard to the interpretation of the chart. As such, it must be written in a manner that answers the following questions: What, Where and When. It must also be concise, relevant and clear. In other words, it must only contain information that is essential to the correct interpretation of the chart. For example, a chart forming part of an area-specific publication, relative to a given region or a certain timeframe, does not necessarily need to systematically include the same regional or time reference. It is likewise suggested that the title of the chart be placed immediately above the chart, horizontally centred (SCHMID, 1992) or aligned to the left (WALLGREN, 1996), thereby functioning as a header.

Labels

This general term encompasses all written information placed in the chart area e.g. the category and value axes labels, the indication of the respective units and any notes (sources, clarifications, etc.).

All words must preferably be aligned horizontally and in the language's reading direction, left to right in the case of English or Portuguese.

Great precision in the data presented in the majority of charts or tables is not required.

An excessive number of decimal places (separated from the whole number by a comma, in Portugal or a point, in England or USA), or even one decimal place in the case of high values, is a level of accuracy that is unnecessary and hampers interpretation. Values in thousands can be formatted using a

space instead of a point, in Portugal or a comma, in England or USA, to make them more legible.

The values of a scale must be specified in rounded-up whole numbers that are multiples of 1, 2 and 5 (e.g. 5, 10, 25, 50, 100 etc.). The display of numbers with more than 5 digits is not recommended. The unit may be adapted, if necessary, to thousands or millions.

Legend

A good legend must do more than simply label the chart's components. It should tell us what is important and what is the chart's objective: to inform the viewer and compel the person creating the chart to structure the information (CLEVELAND, MCGILL, 1984a).

The legend is composed of symbols and their respective labels. The symbols (colours or any other) should be used in a way that ensures that there is no visual mix-up between them and, consequently, that there is a clear connection between the symbols and the component represented. The labels, in turn, must be clear and concise and any clarifications should be consigned to the corresponding notes.

The symbols must be ranked in the same order as the respective components: horizontally when side by side (Figure 6) and vertically when one over the other (WALLGREN, 1996).

It is recommended to use the same legend with charts in which the components occur more than once (Figure 6).

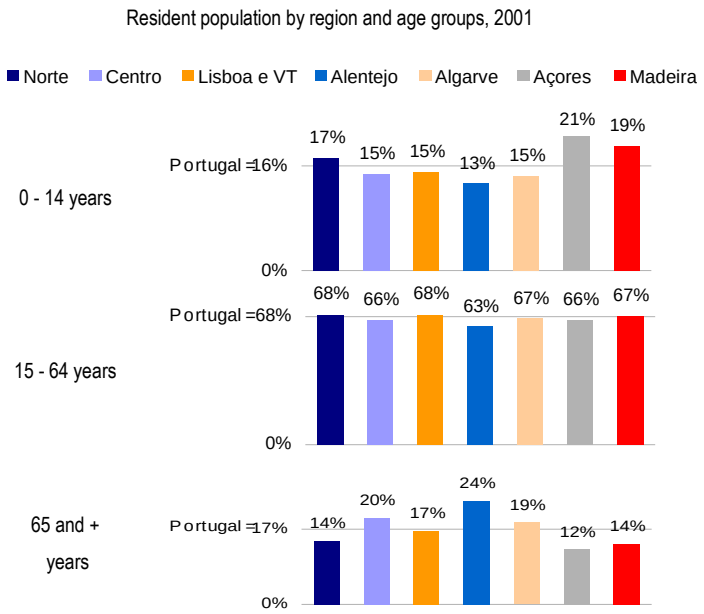


Figure 6 – Series of charts with a common legend

Placing the legend in the chart area compels the visual system (eye-brain) to alternate between the legend and the plot, thereby hampering any immediate interpretation. It is therefore recommended that the legend be omitted whenever possible and the labels be placed next to the components they describe, particularly in line graphs (see Figure 8) and pie charts.

The legend labels can be moved from the chart area to the plot area, allowing not only the chart to occupy less space as a whole but also reducing the visualization distance between the legend and plot (see Figure 10, where the labels are located beside the data lines).

Plot area:

Category or variables' axis

The variables or categories to be displayed are placed along this axis. In relation to charts displaying time series, the time periods are associated to this axis, where each month, quarter, year or other measure of time will correspond to one point or one bar of the chart. This relationship is obviously univocal, in other words it makes no sense at all to include annual and half-yearly values, for example, in the same bar, or years and decades along the same axis, or annual and quarterly values in the same space (TUFTE, 1983).

The category axis must be visually bolder than all other gridlines (Figure 5) (SCHMID, 1992).

Gridlines

Gridlines are amongst the most visually monotonous graphical components. They should, therefore, be hidden or masked in such a manner that their presence becomes implied. The majority of dark gridlines, even if they aid the interpretation of the chart, have great visual impact and they sometimes even cover up the most important part of the chart: the information. When they are really necessary, a neutral colour must be used for them - grey being the best colour in the particular case of a white background (Figure 7).

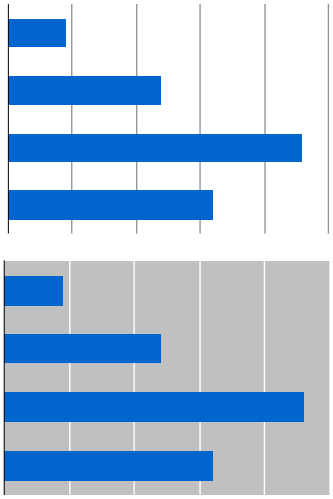


Figure 7 – Gridlines on a white and coloured background

The use of vertical gridlines in certain cases, particularly with time series, may be considered an important aid to the reading of charts, in order to complement the reading of the series progression with specific values obtained from the chart (Figure 8).

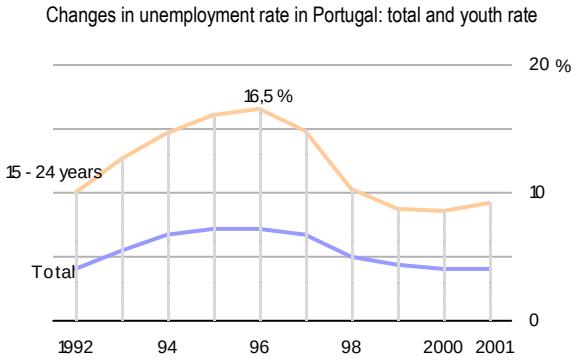


Figure 8 - Vertical gridlines on a line graph

Value axis

The most recent data in the majority of time series charts is located on the extreme right, distant from the markers on the value axis, which are normally located to the left (Figure 9). This fact means that the human eye has to alternately switch between the data and the values along the chart.

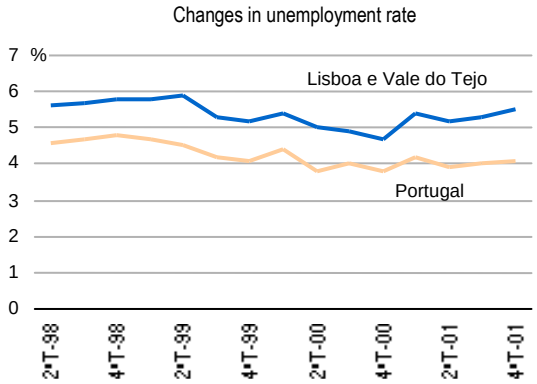


Figure 9 – Value axis with values on the left

This imprecision in visualization may be reduced somewhat by placing the value axis on the right-hand side, close to the most recent data (see Figure 8), by duplicating the axis on both sides of the chart (Figure 10), or by positioning the values next to the respective coordinates (TUFTE, 1983).

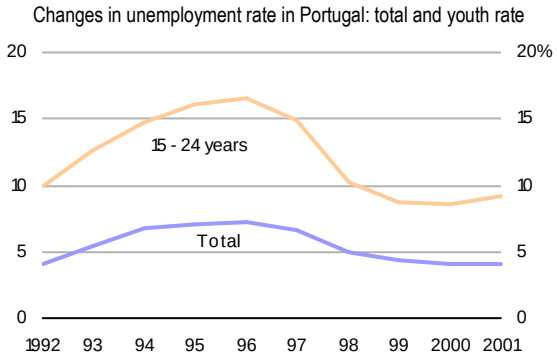


Figure 10 – Chart with duplicated value axis

Charts with two distinct axes are normally used when the data includes different units of measure (Figure 11) or there are considerable differences in the values of the category of a variable. These types of charts must be avoided since they are usually quite difficult to read and are often quite confusing (SCHMID, 1992).

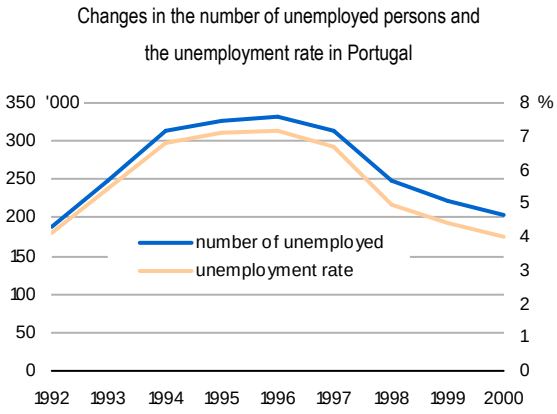


Figure 11 – Chart with two distinct axes

Scale break

A complete scale should be preferentially used (i.e. one that starts at zero or any other reference value), thereby presenting the data unmanipulated (Figure 12 - A). A break is, however, considered reasonable in situations where the information consists of small variations, as long as the break is accompanied by symbols easily understandable to the viewer (Figure 12 - B).

In order to better understand the data in the exploratory analysis phase, scales may be manipulated and any variations extrapolated, but caution must be used in the publication phase in order not to graphically demonstrate alterations to the data that did not really occur.

A scale break is an example of how the message transmitted can be distorted. When the variation on the data is significantly different to the graphical variation, the values appear to be visually under or over-evaluated (TUFTE, 1983).

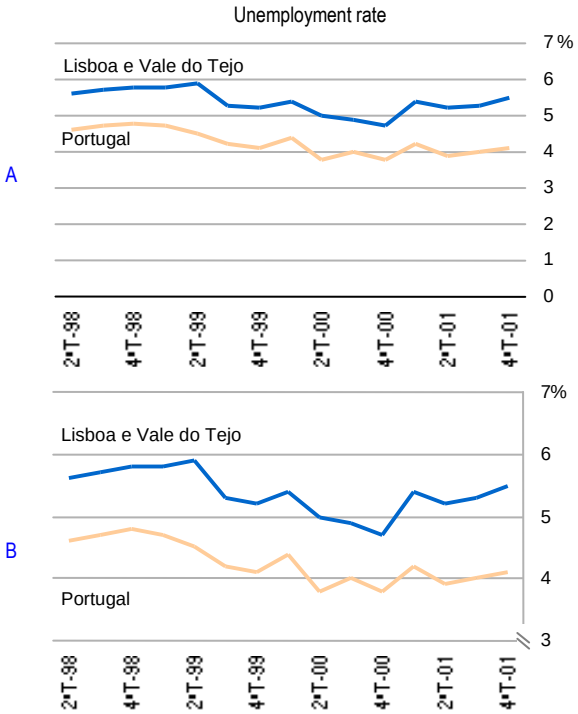


Figure 12 – Charts without and with scale breaks

A graph with a more than one time series can be read in two possible ways: a vertical comparison, in which the relative dimension of one series is evaluated against another (e.g. Portugal has an unemployment rate that is equivalent to around three-quarters of that of Lisboa e Vale do Tejo) and the comparison of slopes, which involves an analysis of the development of both series.

In the case of two series that are apparently constant, any comparison between them can only be performed vertically, given that it is very difficult to visually detect any variations in their development.

In such a case, the use of a scale break provides for better detection of differences in slopes, but the vertical comparison of the lines in such a situation is no longer valid (WALLGREN, 1996). This is the reason why scale breaks must not be used in vertical bar charts, since no vertical comparison of bars is possible if the chart contains a scale break.

Visual variables

Jacques BERTIN, in his book *Sémiologie graphique* (1973, 2nd edition), was the first person to organize knowledge regarding the visual appearance of graphical symbols. He classified graphical symbols using the following visual variables.

Location - provided by the two planar dimensions, x and y;

Size - the variation in length, width or area, evidently connected to the numerical value of the data;

Value - refers to the (perceived) light-dark variation of the colour or black-white variation.

Texture - the size and spacing of the graphical elements that constitute the symbol (points, lines or others), expressed by the number of these elements that are repeated per length unit;

Colour – the sense used to differentiate between specific segments of the electromagnetic spectrum, in other words between blue, green, red etc.;

Orientation – also known as direction, corresponds to the angle to the reading line;

Shape – can be geometric (such as squares or circles) or it can be irregular;

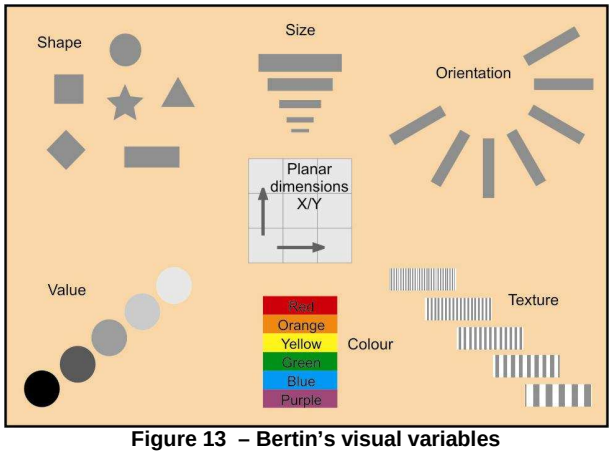


Figure 13 – Bertin's visual variables

1.2. Bar graphs

Bar graphs are one of the most popular ways of graphically presenting information, partly due to the ease with which they can be constructed as well as their straightforward interpretation.

They are used to compare discrete or qualitative variables, in absolute or relative terms, or to compare categories of quantitative variables. Moreover, they can represent the change of a variable over time.

In graphs of this type, the viewer extracts data values by visualizing the position of the bars relative to a common scale (CLEVELAND, MCGILL, 1984).

The bars normally begin at the category axis, which facilitates the comparison of relative positions.

Simple bar charts (vertical or horizontal)

With a bar graph, the frequencies can be arbitrarily plotted on the x-axis or the y-axis. In other words, the bars can be horizontal or vertical (Figure 14).

Despite the fact that vertical bar graphs are more popular, there are situations in which the other type of layout is more desirable. An horizontal bar graph is reckoned to be easier to read when the difference between the maximum and minimum value of the variable is significant.

In a situation where the space to insert the graph is limited, the horizontal bar graph is also the preferred option given that it allows the inclusion of different categories without significantly increasing the space occupied by the chart.

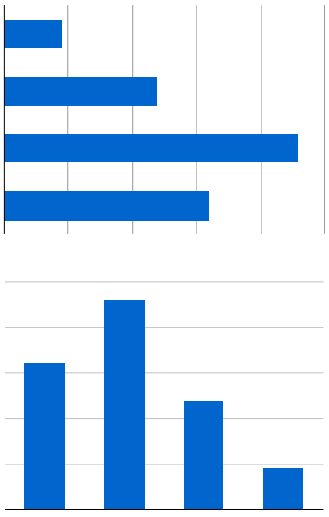


Figure 14 – Horizontal and vertical bar graphs

The horizontal bar graph is recommended for variables with extensive labels, since the space for inserting the name of the category in a vertical bar graph is more limited (Figure 15). Note that the category labels should not be abbreviated nor positioned in a manner that makes them difficult to read (vertically or diagonally). This means that sometimes they take up more space than the plot itself.

Also of note is the fact that horizontal bar graphs display the differences between data in a clearer way, given that they have a more extensive value axis. Figure 15 is an example of this: despite the fact that both charts occupy the same area, the visualization of the categories with the greatest frequencies produces different visual impacts.

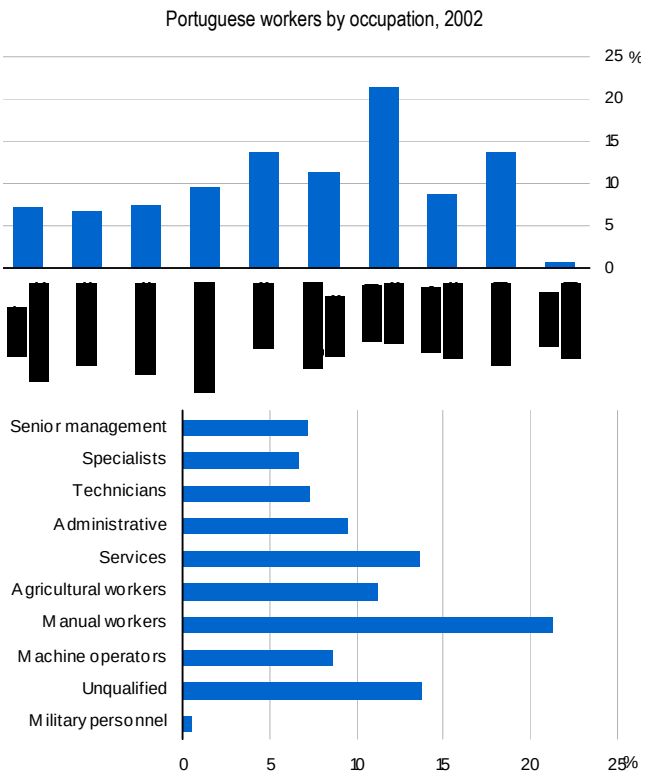


Figure 15 – Category labels in a vertical and horizontal bar graph

The representation of negative values

The plotting of negative values is not recommended in horizontal bar graphs, given that a descending bar is conventionally associated to negative values.

In fact, the visual association between left and right and positive and negative values respectively, may not be direct for a less experienced viewer. Vertical bar graphs should therefore be used to show negative values.

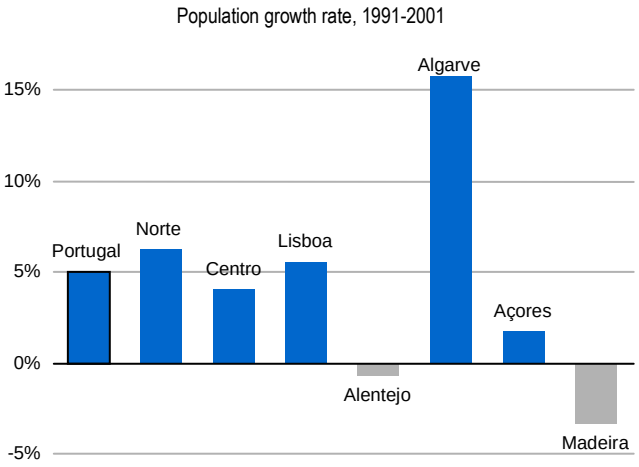


Figure 16 – Representation of negative values

Some rules related to the construction of bar graphs

The scale on the value axis

Scale breaks in bar graphs are not acceptable since they prevent the vertical comparison of categories.

A scale break is misleading, since it visually indicates the existence of great data variation where this does not in fact exist (Figure 17 A and B).

Analysing Figure 17 B, for example, a less attentive reader might conclude that in 1991 Portugal had about one-third of the people it had in 2001, which is false.

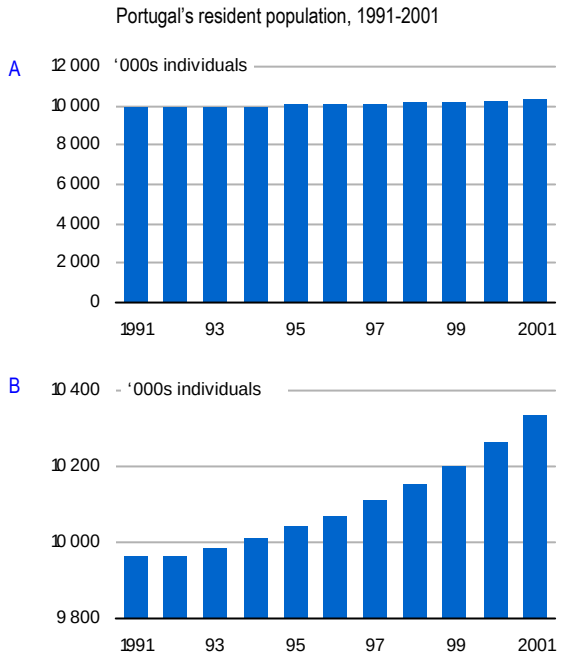


Figure 17 – A chart without scale break and a chart erroneously with a scale break

Nevertheless, when one of the bars has an abnormal value that occupies a lot of image space, its shortening is permitted. This must be done in a clear manner that is comprehensible to the reader. Its value, for example, can be displayed along with a symbol(s) clearly indicating that the bar was cut short (Figure 18).

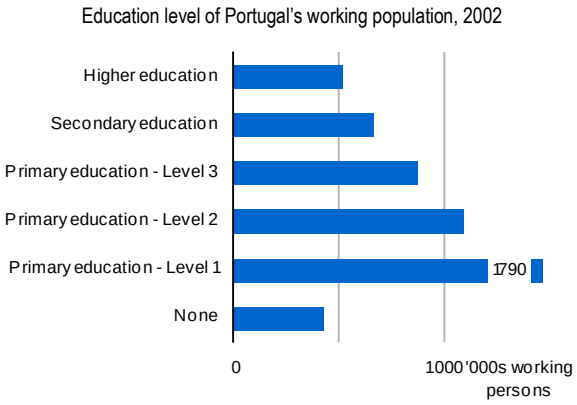


Figure 18 – Chart with curtailed bar

In certain cases, a scale of 0 to 100% can be used (Figure 19) so that the reader can perceive how much each bar is short of attaining 100%.

It is recommended, whenever possible, that the categories are compared with the total - Portugal, in this case - thereby enriching the information to be obtained from the chart (Figure 19)

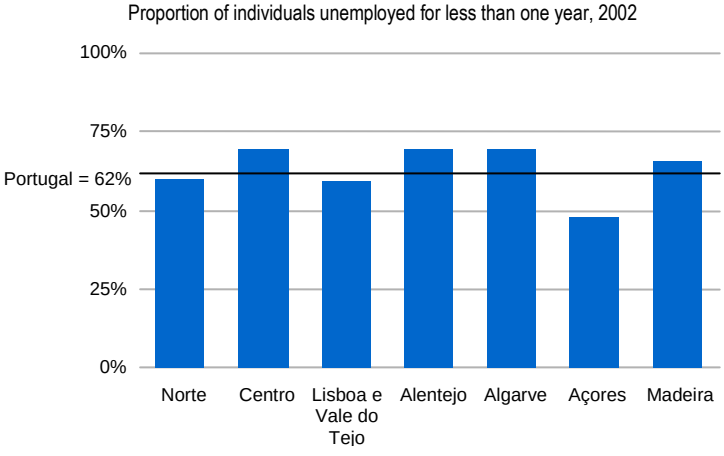


Figure 19 - Chart with a scale from 0 to 100%

Visual equilibrium: gaps between bars and gridlines

The gaps between bars must be constructed in a manner that does not make their comparison difficult (Figure 20 - B) nor gives the chart the appearance of a histogram (C), suggesting continuity when the displayed variable is, in fact, discrete. The recommended gap between the bars should be approximately equal to the bars' size (A).

The gridlines are there to aid visualization of comparisons and to read approximate values. A graph with too many gridlines (B) gives too much visual significance to these secondary elements without providing any resulting significant advantages in the reading of approximate values. A graph with too few gridlines on the other hand, does not provide any value added to the chart's interpretation (C) (WALLGREN, 1996).

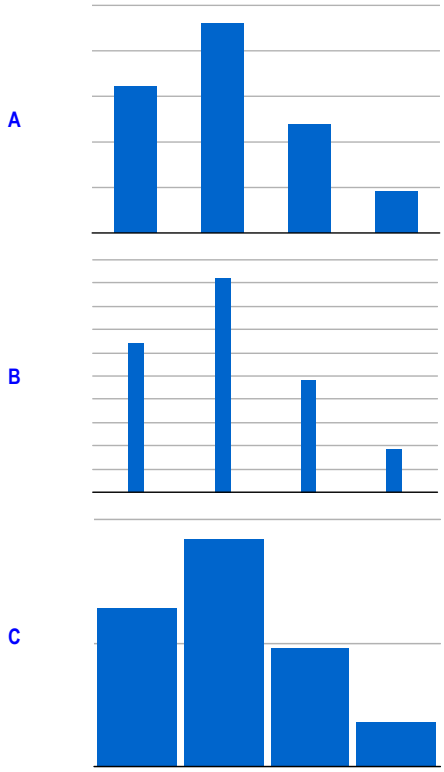


Figure 20 – Gaps between bars, and gridlines

Ranking

It is often necessary, when displaying information, to organise the categories in ascending or descending order (Figure 21) so that certain phenomena inherent to them can be better understood.

Likewise, it is common to alphabetically (or geographically) rank category labels, especially when these consist of countries or any other type of administrative unit. This, however, is not always the best option.

If the same set of categories is used in more than one chart, then the relative position of each category should remain unchanged. In other words, the categories should appear in the same order in all the charts. Likewise, the size and scale of the charts should be the same if the idea is to compare them with each other.

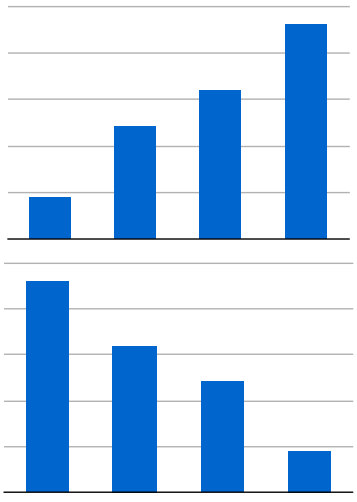


Figure 21 – Bar graph in ascending or descending order

When not all of the categories are defined, and there is, for example, one category, called 'Others', under which all the remaining categories are grouped, it is recommended that this category is not included in any sort and should be left to occupy the last place (WALLGREN, 1996; SCHMID, 1992) (Figure 20). If colours are used to differentiate categories, then the 'Others' category should have a colour that does not make it stand out, since it is the least important category (e.g. grey).

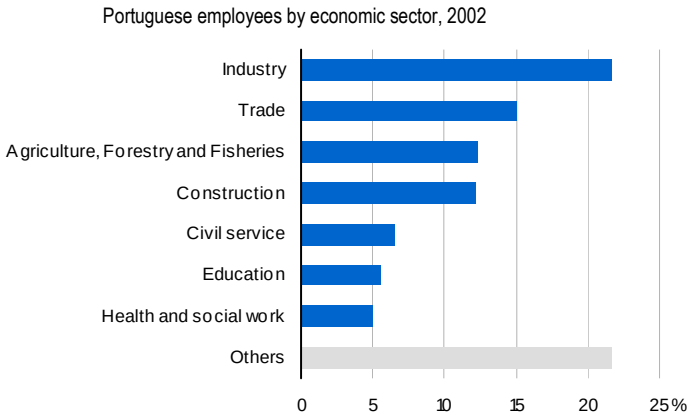


Figure 22 – Ranked categories

Grouped bar graphs

Grouped bar graphs are used to simultaneously describe two or more categories for a given discrete variable, or when the objective is to highlight the value of the categories instead of the total value of the variables (WALLGREN, 1996).

The different categories are represented by bars and are distinguished using visual variables (colour or value). The groups of entities should be separated by a blank space, but no gap should exist between the categories in each group.

Given that, in perceptive terms, the comparison between the estimated values of adjacent bars is more efficient than that made between bars further apart, the type of grouping used must be in agreement with the categories to be spotlighted. Thus, in visual terms, first the categories contained in the legend are compared and only after this is the breakdown of the different variables correlated (Figure 23 - A and B).

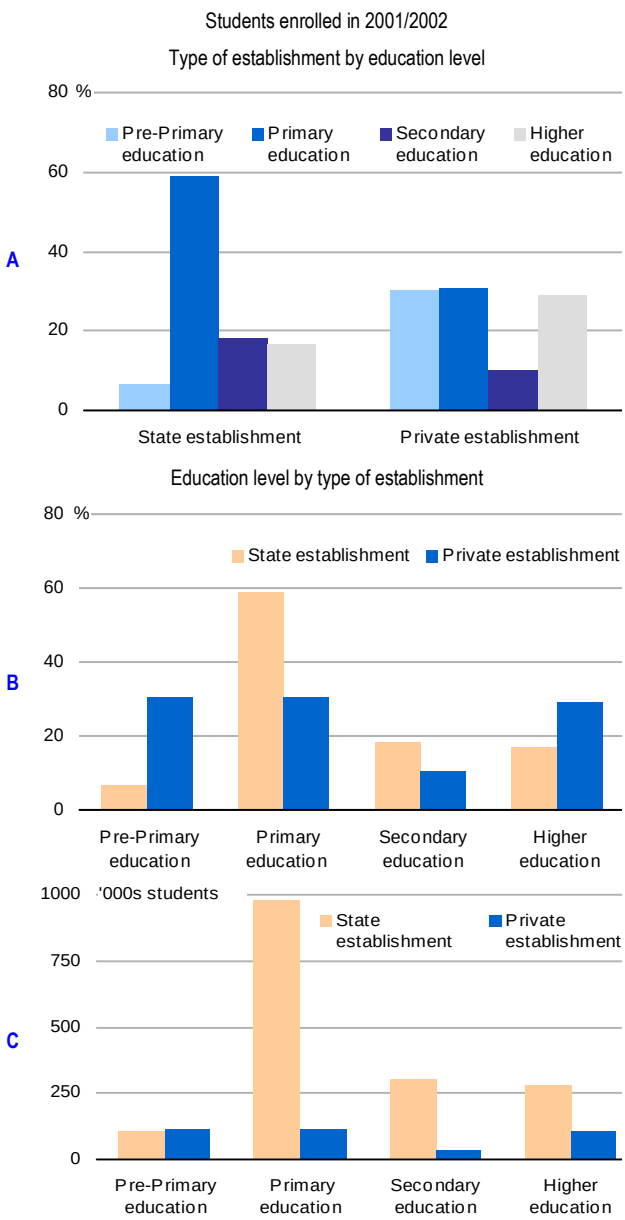


Figure 23 – Chart with bars grouped into four and two categories, and into relative and absolute values

The bars can indifferently display relative or absolute values, according to the type of analysis. Sometimes the plot of both types of value is of great interest when there are significant differences (Figure 23 – B and C).

The greater the number of displayed categories, the less legible this process becomes. No more than three or four categories per variable are recommended in a chart.

The construction of different charts is preferable to the accumulation of all the information in one single chart when there are various groups composed of different categories.

Superimposition in grouped bar graph

In grouped graphs, the bars representing the categories of each group can touch or even be superimposed on each other (SCHMID, 1992). Superimposition allows the categories to be ranked and also saves space and allows more information to be included. It must be noted that bars that are placed on a plane perceived to be more distant (and possessing a paler colour than other bars) are perceived as being less important (Figure 24).

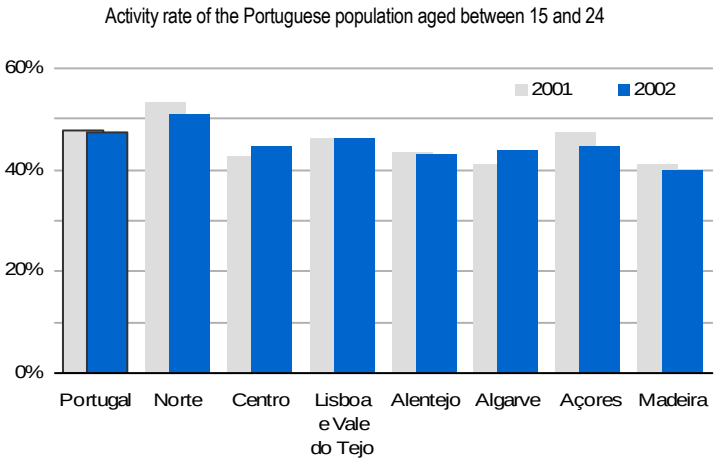


Figure 24 - Partial superimposition in a grouped bar graph

The superimposition of bars is also permitted in those situations where the values of one category are consistently less than the values in another (Figure 25). Highlighting values or occurrences is also a way of analysing data. It is often important to visually emphasize a specific value or category.

In this example the auxiliary line relative to 50% of the employed population was made thicker - this is the only line with an associated numerical value. All the other values are read using the unnumbered gridlines (Figure 25).

To highlight the category referring to Portugal, a frame can be used (A) or a darker colour (B). Only the values of the categories deserving analysis are inputted (e.g. A - the Açores has the greatest gender difference), so as not to overload the graph with information.

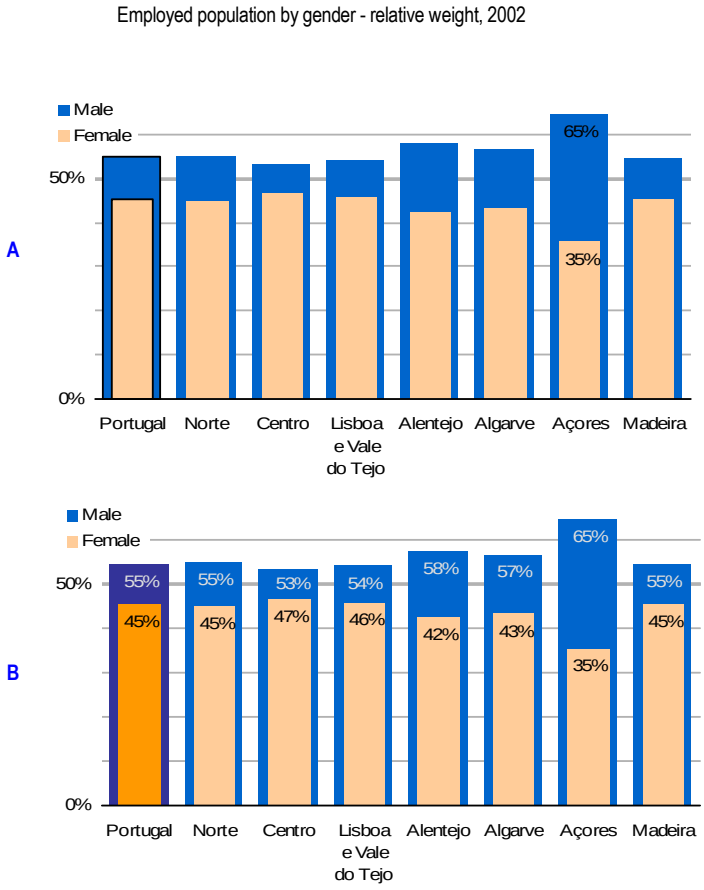


Figure 25 – Total superimposition in a grouped bar graph

Stacked bar graphs

Stacked bar graphs (Figure 26) are used in the same context as grouped bar graphs: when the data set contains two or more categories.

In these type of charts, each bar is sub-divided into at least two categories, with distinct colours or fills, thereby allowing the display of the relationship between each category (Men/Women) and the respective sub-total (e.g. Commerce and Administration). The categories are, thus, positioned one on top of the other in the case of a vertical bar graph (or side by side in a horizontal bar graph). The height (or width) of each component corresponds to the absolute or relative value of the category.

The absolute value graph (A) is suitable for those cases where the emphasis focuses more on the total value of the variables than on the respective categories (WALLGREN, 1996), given that the total value is more accurately perceived than the component parts. This accuracy relative to the total is the result of the comparison of the relative position on the same scale, while the estimate of the values of the categories involves the comparison and ranking of the respective sizes.

If the main objective of these charts is to graphically indicate the total sum, ahead of visually estimating the respective categories, then a valid question to ask would be why not just plot the totals or replace this type of chart with a different type.

In the relative value chart (B), the value of the categories can only be estimated examining the size of the associated bars.

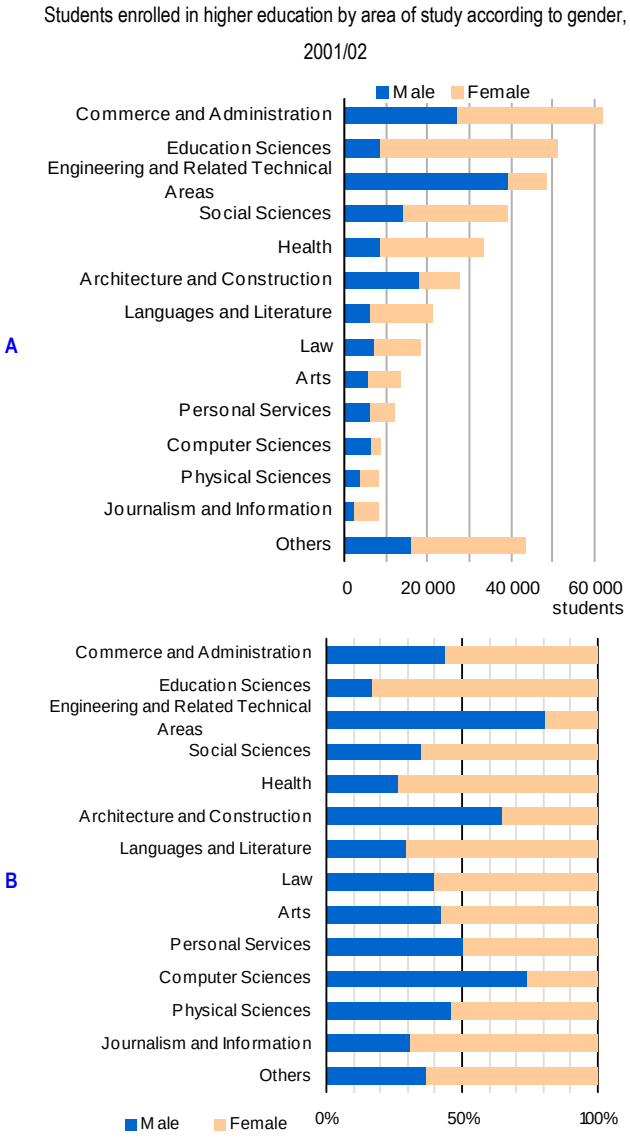


Figure 26 – Horizontally stacked bar graph of absolute and relative values

It is easier to estimate values with only two categories, given that the base and top of the scale serve as a reference mark. With more than two categories, however, the reading of values is significantly more difficult.

Disadvantages of stacked bar graphs

The first components of the chart are easily comparable since they start close to the axis. With subsequent components, though, it is only possible to infer the values approximately. The greater the variation existing to the first category the more difficult this task becomes (Figure 27). As a result, the fluctuation and excessive weight of the first category can compromise the reading of the other variables plotted in the chart.

If the size-based comparison between categories can encompass non-negligible errors between the real value and visually estimated values, then the inter-category ranking of the same bar may even be incorrectly performed, seriously hampering the validity of this means of representing information (CLEVELAND, MCGILL, 1984a).

This is the reason why stacked bar graphs must be limited to a restricted set of variables and categories. In certain cases the use of a grouped bar graph is preferable since it improves the estimation of individual values, despite the fact that it does not facilitate inter-category comparison.

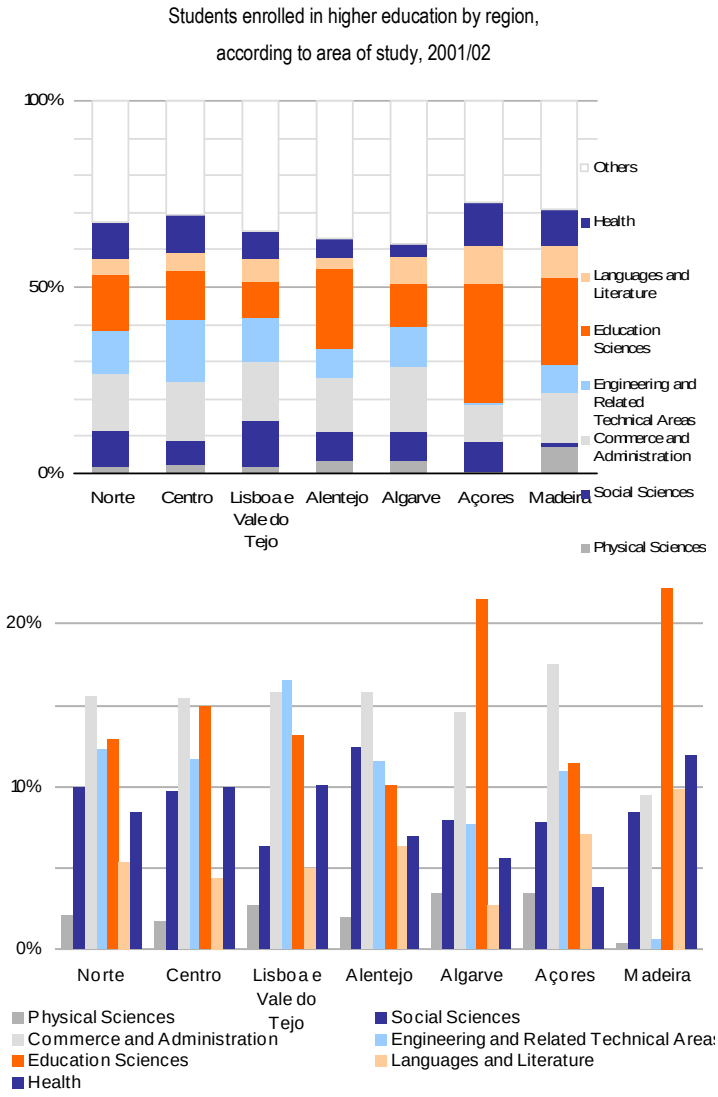


Figure 27 – Vertically stacked bar graph and grouped bar graph

Histogram

A histogram shows the distribution of a continuous variable via a graph of joined bars. Histograms are normally represented by bars of equal width, where the height (or length) varies according to the relative or absolute frequency. When the intervals have the same width, the area depends only on the height. But, when the classes are of different dimensions, the area of each bar is no longer proportional to its height. The height should be calculated so that the area of each rectangle is proportional to the relative frequency of each class. In the first case, the value axis provides the information regarding the relative

frequency of each class. Whereas with classes of different dimensions this axis has no meaning at all and the viewer is obliged to compare areas in order to extract information, which is quite a difficult task to perform.

This kind of chart allows the display of extreme values and skewed data, visually indicating whether the variable follows a normal distribution. The plotting of percentages also allows data sets of different dimensions to be compared.

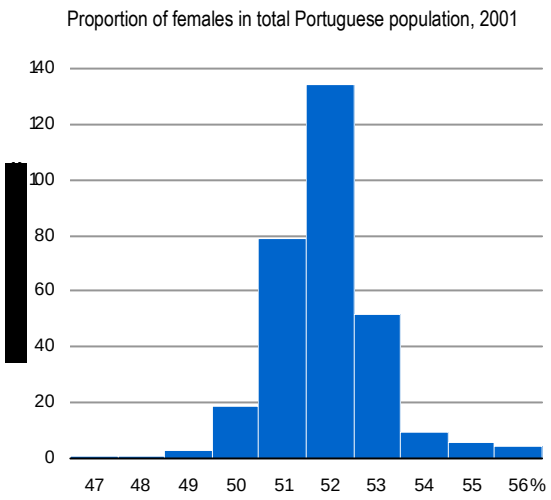


Figure 28 – Histogram

Population pyramid

The population pyramid is also a histogram and it is widely used in demographic analyses since it allows the distribution of the population by age and the concurrent comparison between genders to be visualized in one single chart. The plot consists of two horizontal axes (one for males and the other for females), which can be in absolute or relative values.

Age is plotted on the vertical axis, covering both plots. Age is normally represented in five-year age bands, but a scale of individual years can also be used.

A plot of absolute values provides information on the dimension of the data but prevents any kind of space or time comparison. Such a comparison is only possible if the relative data values are plotted (NAZARETH, 1996; INE, DRLVT, 2001).

This form of plot can, nevertheless, be used to display other types of demographic data (such as the level of education, for example) or even to plot continuous variables with a common axis (WALLGREN, 1996).

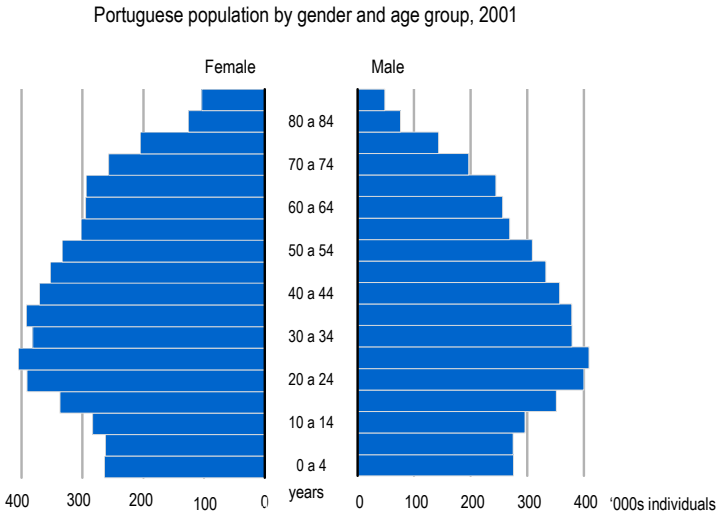


Figure 29 – Population pyramid

Time series in bar graphs

The category axis of a vertical bar graph can consist of dates, thereby making it possible to plot changes over time.

Bar graphs can replace line graphs when the data series is quite short. They are also to be preferred when vertical comparisons between certain variables over a specific period of time are required, that is, when the value of the variable in each time period is the focus and the foremost objective is to relate individual quantities.

Both possibilities (bar and line graphs) are appropriate for displaying trends with a single data series. However, line graphs are clearly preferable for displaying more than one data series (JACOBS, 1997). Bar graphs are, therefore, not recommended for plotting several data series. When one set of variables has continually lower values then it is still possible to follow the development of the values (Figure 30). However, when the sets of variables cross over, the chart becomes illegible (Figure 31).

In those cases where the chart contains so much information that its correct visualization is not possible, then the replacement of the chart with a data table or the division of the data amongst various charts should be considered.

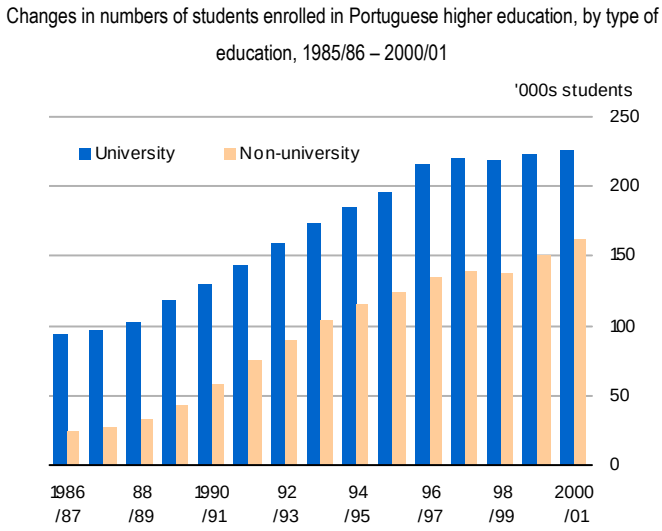


Figure 30 – Bar graph with two time series

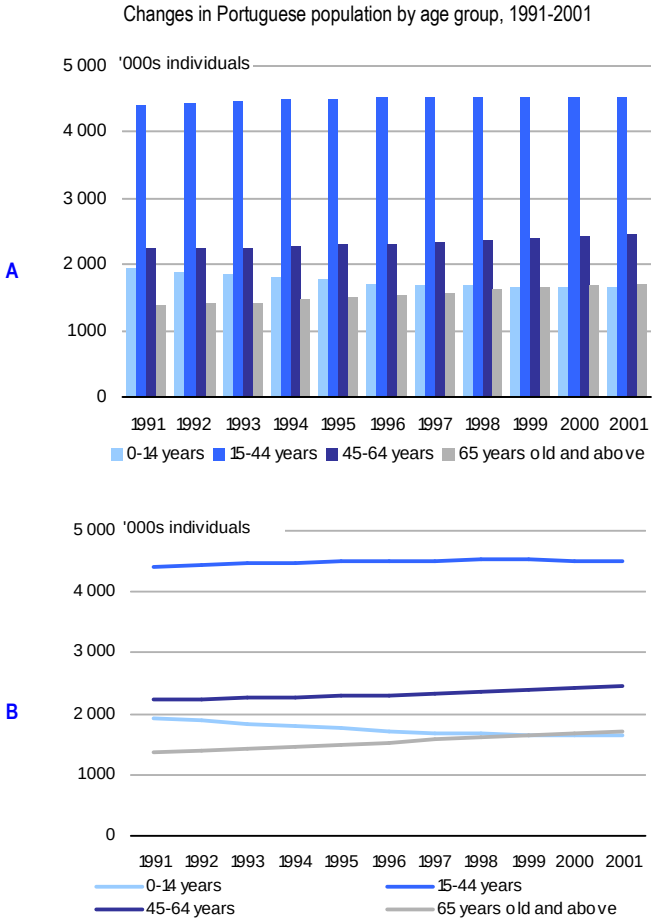


Figure 31 – Time series graphs: bar and line graphs

1.3. Line graphs

A line graph is recommended for showing trends and the development of a continuous variable against another continuous variable.

The most common line graph displays time (or chronological) series, where a specific continuous variable is analysed over time. The y-axis measures the variable(s) under analysis and the x-axis contains the chronologically ranked time units in equal time intervals, starting on the left with the oldest data element (Figure 32).

The series in a line graph, contrary to those of a bar graph, can be extensive. The objective in these graphs is to compare the slopes of the curves, so that questions such as the following can be answered: In which periods was the change significant? When were the inflection points? (WALLGREN, 1996).

For a specified data set, the points (x and y coordinate pair) are joined by a line that visually suggests continuity.

A graph must not contain more than three lines, otherwise it becomes too difficult to read (SCHMID, 1992; TUFTE, 1983). In the event of many criss-crossing lines (Figure 33), the line graph should be replaced by various graphs.

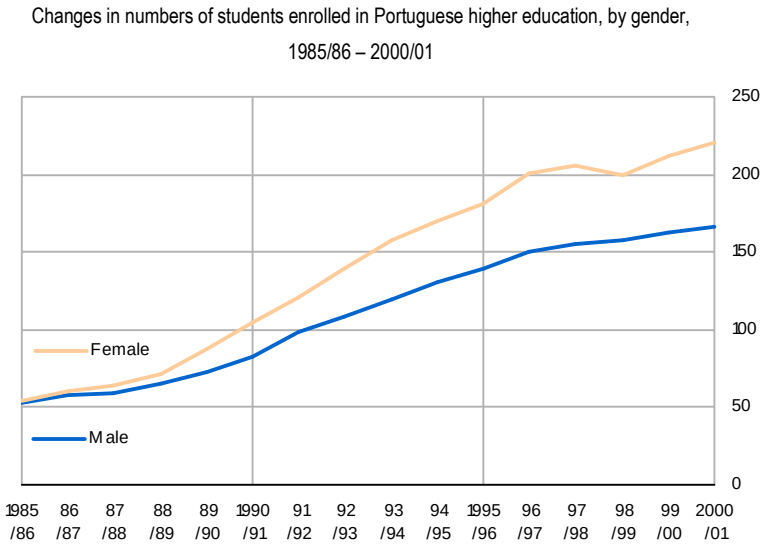


Figure 32 – Time series graph

A different style must be used for the plot of each line, using colour, shape, form and value to this end. It may be necessary to use a technique besides colour to distinguish lines, in order to aid interpretation in the event of black and white printing or photocopying. However, this option may visually rank the lines in a way that does not coincide with their real significance, since a dashed line, for example, is visually less important than an unbroken line.

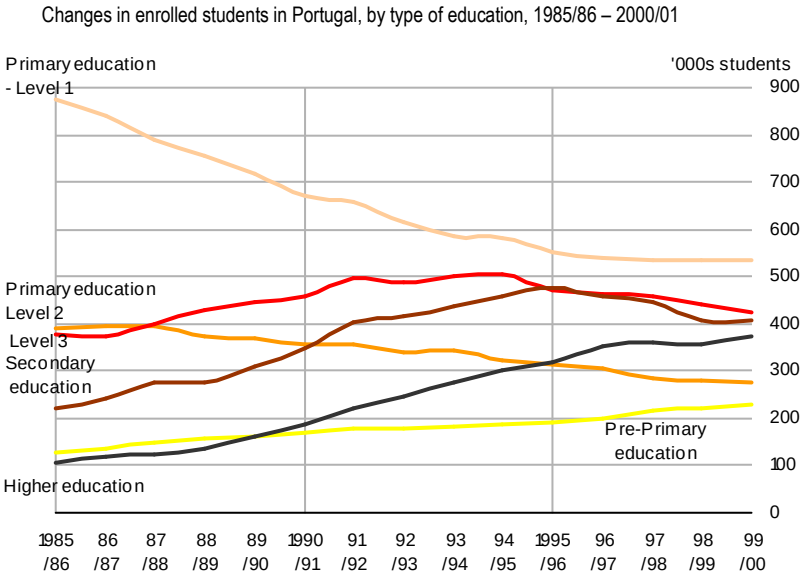


Figure 33 – Graph with too many lines

The variable measured on the category axis of line graphs cannot be qualitative (Figure 34). In such cases, the development of the series has no significance at all. Between the Algarve and Madeira, for example, one cannot say that there is a drop in the data series. All one can say is that Açores has a lower value. It is also not possible to estimate the intermediate values of the variable between the categories. In this case, we cannot say that there are x% unemployed persons in the Atlantic Ocean (correct graph: Figure 19, page 20)

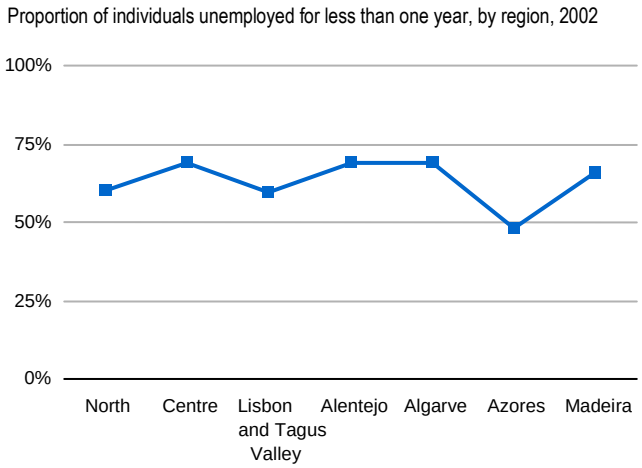


Figure 34 – Incorrect line graph

The periods should be equally spaced if they are consecutive and proportionally spaced if discontinuous. In other words, the spacing of the columns is adjusted to show any irregular time intervals. For example, the gap between data relative to 1998 and 2000 must be double the gap between 2000 and 2001 (Figure 35).

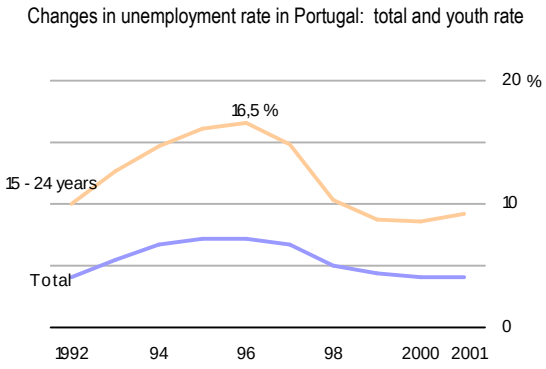


Figure 35 – Spacing of values on the category axis

If the objective is to compare two curves displaying very similar plots (Figure 36 - A), it is recommended that the difference between them is plotted, between men and women (Figure 36 - B) in this case, instead of just plotting the curves.

A sudden alteration in data can be hidden if the graph starts after this alteration, thereby falsely displaying stability (WAINER, 1984). On the other hand, a modification can become sudden if the graph just displays that period and does not place it in the context. This can be the case with data series with strong seasonal influence.

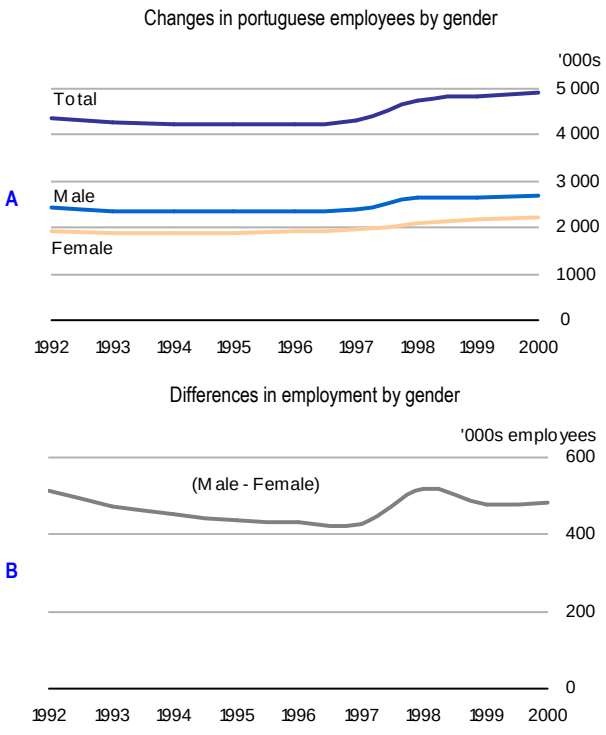


Figure 36 – Comparison of parallel series

Area graphs

Area graphs are used to simultaneously visualize the development of totals and of the respective component parts. This type of plot, however, just like the stacked bar graph, has few advantages given that it is not possible to instantly answer questions on growth or reduction over time, particularly when the first of the component parts oscillates significantly.

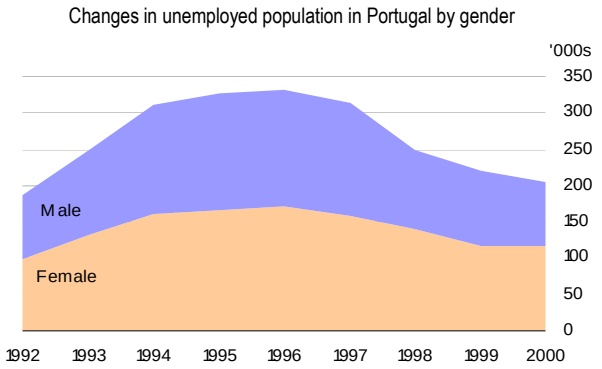


Figure 37 – Stacked area graph

Area graphs are used as an alternative to line graphs. However, they cause increased difficulty when the areas intersect because it becomes impossible to follow the development of the component parts.

1.4. Pie charts

Pie charts have become very common in publications aimed at a broad audience, but they have also been subject to widespread criticism due to their lack of capacity to provide information (WAINER, 1990; TUFTE, 1983; BERTIN, 1977, etc.).

Pie charts divide their whole into segments, similar to slices of cake, hence their name in English 'pie chart'. For a set period of time, the variable under analysis is plotted in a circle where each component part occupies an angle such that their total adds up to 360º (Figure 38).

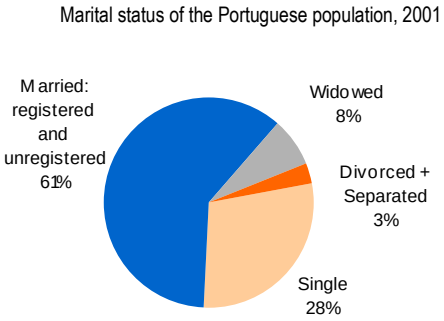


Figure 38 – Pie chart

The use of pie charts is not recommended when more than one time period is to be compared, nor is it recommended for variables with more than five component parts or when these component parts have approximately the same size. In this latter case, the use of a bar graph instead of a pie chart is preferred (SCHMID, 1992). In a pie chart with many slices or with slices that are too thin, making the chart difficult to interpret, it is necessary to add the respective values to complement the chart (Figure 38) or associate a subset of values to another pie chart with a size that is proportional to the quantity represented (Figure 39).

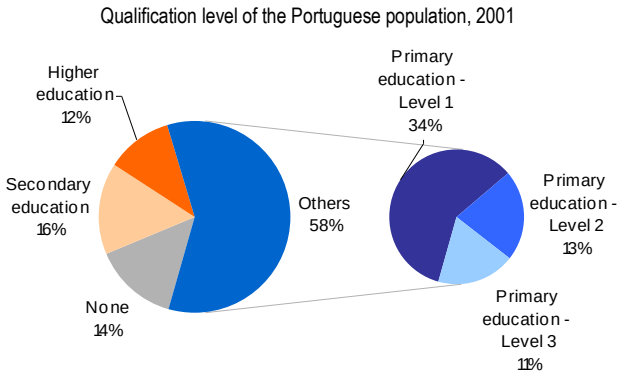


Figure 39 – Sub-divided pie chart

Thus, the use of pie charts is only positively recommended in cases where one or two components dominate the total, in order to give a general idea of the data. But, even so, the use of a pie chart in preference to a table is always debatable.

It is common to find distorted pie charts, that is, charts that are not circular, in order to save space or for other unknown reasons. To make a pie chart into an ellipse is highly misleading, particularly in relation to the thinnest segments, and this practice must be avoided since it completely misrepresents the original chart.

Another current practice is the separation of the slices, moving them radially outwards from the centre, causing different slices to be moved unequal distances from the centre. As it is necessary to position the slices in a non-circular fashion in order to maintain equal separation distances, none of the options are formally correct (BOUNFORD, 2000).

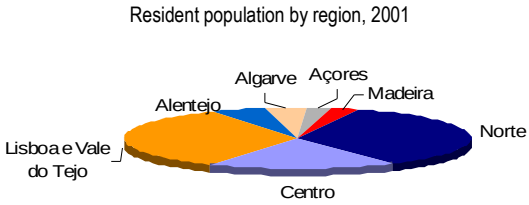


Figure 40 – Distorted pie chart

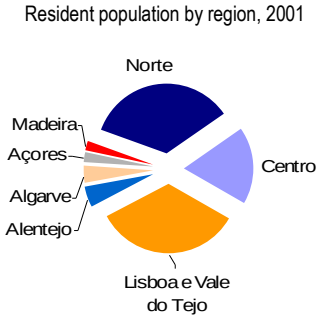


Figure 41 – Pie chart with separated slices

1.5. Pictorial charts

Pictorial charts are common graphs with decorative characteristics. They are suitable for a superficial presentation, where contact with the image is brief, for example in generalist newspapers or magazines or when the target public has a medium to low education level.

The most used pictorial charts are those based on size criteria: where the variation in area of the size of the shapes used is proportional to the change in the plotted variable (Figure 42 - A).

It is quite common, however, particularly in the media, to find pictorial charts in which the heights and widths have been simultaneously increased, but not the area, causing the pictogram to be out of proportion and, therefore, transmitting completely the wrong idea.

Analysing Figure 42 - B we see that Portugal has 3 times more students than Lisboa e Vale do Tejo, in both genders. Thus, the area of the figure relative to Portugal must be three times larger. This is why this type of plot is deemed to be one of the most misleading (SCHMID, 1992; TUFTE, 1983).

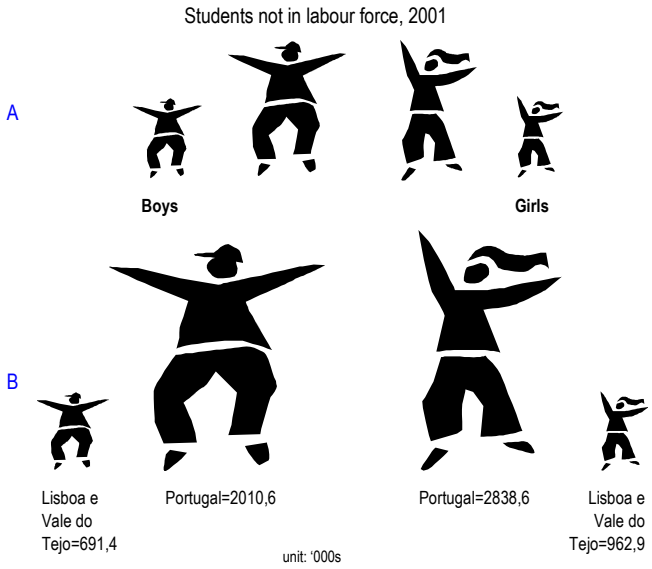


Figure 42 – Pictogram based on size criteria

Pictorial unit charts made up of symbols are also quite widely used. Each element is allotted a value, which means that the number of elements is equivalent to the extent of the variable.

One of the most well known examples of this is the population pyramid in which the bars are replaced by symbols representing people. One of the problems is how to handle decimals. Modley (1952, in SCHMID, 1992) says that fractions of symbols must be minimised and all values should preferably be rounded up or down. It is in fact common to find in these type of population pyramids bars in which the last symbol is incomplete, ending in an upper body, legs or a head (Figure 43).

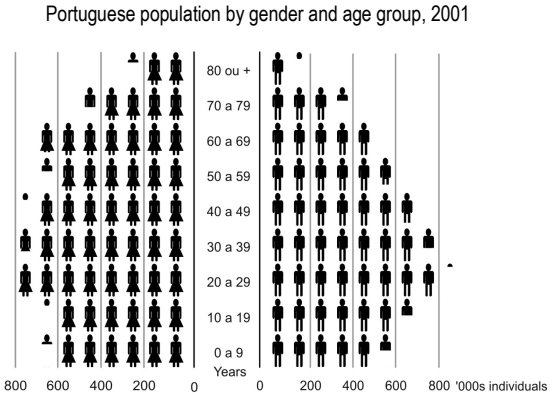


Figure 43 – Pictogram: population pyramid

1.6. See also...

This dossier provides a brief overview of some of the most important issues related to statistical graphs, in particular those related to the construction of the most well-known and widely used graphs.

The information used for the graphs included in this dossier is quite recent and can be found at the site www.ine.pt. All of the figures, except for the last one, were constructed using Microsoft Excel software.

This text is based on the Master's degree thesis bearing the title *The graphical and cartographical representation of statistical information*, which was submitted in June 2003, in ISEGI/Universidade Nova de Lisboa University.

Numerous books, articles and web sites discuss statistical graphics, graphs and statistics, the most important of which are as follows:

Publications: books (📖) and journal articles (🗎)

📖 ALEXANDRINO SILVA, Ana (2006) – *Gráficos e mapas : representação de informação estatística*. Lisboa: Lidel-edições técnicas.

🗎 BENIGER, James R.; ROBYN, Dorothy L. (1978), "Quantitative graphics in statistics: A brief history", *The American Statistician*, 32 (1), p. 1-11.

📖 BERTIN, Jacques (1973) 2nd ed. (1st ed. 1967) - *Sémiologie graphique*. Paris: Gauthier-Villars.

📖 CHAMBERS, John C.; CLEVELAND, William. S.; KLEINER, Beat; TUKEY, Paul A. (1998) 2nd ed. (1st ed. 1983) - *Graphical methods for data analysis*. USA: Chapman & Hall.

🗎 CLEVELAND, William S.; MCGILL, Robert (1987a), "Graphical perception: The visual decoding of quantitative information on graphical displays of data", *Journal of the Royal Statistical Society*, 150, p. 192-229.

🗎 CLEVELAND, William S.; MCGILL, Robert (1984a), "Graphical perception: Theory, Experimentation, and application to the development of graphical methods", *Journal of the American Statistical Association*, 82, p. 419-423.

📖 GRAPHICS GUIDELINES: *The theory and practice of presenting statistical data graphically, together with proposals for education of statisticians in appropriate use of graphics for presentation* (1994). COMISSION OF THE EUROPEAN COMMUNITIES - EUROSTAT. Kent: White Waghorn Limited.

🕮 HUFF, Darrell (1991) 3rd ed. (1st ed. 1954) - *How to lie with statistics*. England: Penguin Books.

🗎 INE, DRLVT (2001), *As pirâmides de idades* [Population Pyramids], *Revista de Estudos Regionais* nº 2 (Conceitos e metodologias) [Estudos Regionais magazine, no.2 (Concepts and methodologies), *Instituto Nacional de Estatística* [National Statistics Institute of Portugal], p. 75-78.

@ JACOBS, Bernhard (1997), "Experimental analysis of the graphical presentation of data in line graphs and bar charts in superposition and juxtaposition", http://www.uni-saarland.de/philfak/MZ/graph/gesamtue.html.

🕮 NAZARETH, J. Manuel (1996) - *Introdução à demografia - Teoria e prática*. [Introduction to demography - Theory and practice] Lisbon: Editorial Presença.

🕮 SCHMID, Calvin F. (1992) 2nd ed.; (1983, 1ª ed.) - *Statistical graphics - Design principles and practices*. Krieger.

🕮 TUFTE, Edward R. (1983) - *The visual display of quantitative information*. Cheshire-Connecticut: Graphic Press.

🕮 TUKEY John W. (1977) - *Exploratory data analysis*. USA: Addison-Wesley.

🗎 WAINER, Howard (1990), "Graphical Visions from William PLAYFAIR to John TUKEY", *Statistical Science*, 5 (3), p. 340-346.

🗎 WAINER, Howard (1984), "How to display data badly", *The American Statistician*, 38 (2), p. 137-147.

🕮 WALLGREN, Anders; WALLGREN, Britt; PERSSON, Rolf; JORNER, Ulf; HAALAND, Jan-Aage (1996) (English translation from Swedish "Statistikens Bilder - Att Skapa Diagram" Statistics Sweden 1995) - *Graphing statistics & data: Creating better charts*. California: SAGE Publications.

Web sites @

American statistical association - Section on Statistical Graphics: http://www.amstat-online.org/sections/graphics/

Journal of computational and graphical statistics: http://www.amstat.org/publications/jcgs/

Others:

http://www.edwardtufte.com/tufte/ (one of the best authors on this subject – see books)

http://www.mhhe.com/business/opsci/bstat/vistat.mhtml (visual statistics)

http://www.nas.nasa.gov/Groups/VisTech/visWeblets.html (links to scientific visualization)

http://www.bell-labs.com/topic/societies/asagraphics/resources.html (software, books, magazines, etc.)

I would like to thank Dr. Sandra Servinho and Dr. Carolina Macedo for their relevant comments.