Graphical representation of data.

Learning Objectives

- Understand sources of information
- Construct tables
- Illustrate data using a graph
- Illustrate data using a pie chart
- Illustrate data using a bar chart
- Illustrate data using a pictogram
- Condense raw data using a frequency distribution
- Condense raw data using a grouped frequency distribution
- Construct a histogram
- Understand how statistics are misused

Introduction

A problem in statistics is that of condensing large sets of data. One of three processes may be involved:

1. to present a simple ‘neutral’ description of the facts of a certain situation

2. to draw conclusions from a set of facts (i.e. make an inference)

3. to refute an inference drawn by someone else, which may be done either by presenting new information or by demonstrating that existing information has been incorrectly interpreted

Introduction

Some examples of the sources of available information are:

- Official sites
- Semi-official sites
- Survey reports
- Results
- Journals

There are basic rules that should be followed once a set of data has been collected

- The data must be factual and relevant

Before presentation, always check:

- the source of the data
- that the data has been accurately transcribed
- that the figures are relevant to the problem

Tables

When constructing a table, it is important to determine which relationships should be emphasized

Example of a table

Table 1 The number of adults aged between 15 and 34 years of age in Australia who participate in swimming, soccer or cricket

Age (years)	Swimming	Soccer	Cricket	*Total*
15 – 17	78 700	127 500	54 000	260 200
18 – 24	176 600	121 300	87 300	385 200
25 – 34	318 300	91 600	115 200	525 100
Total	573 600	340 400	256 500	1 170 500

Source: Australian Bureau of Statistics, *Participation in Sport and Physical Education*, Cat. No. 4177.0

Tables

Some of the rules that should be followed when constructing a table are:

- Every table should have a clear and unambiguous number
- Each table should have a title that describes the types of information given
- Row and column labels should be precise and unambiguous
- Categories should not overlap
- The units of measurement must be clearly stated
- Any unimportant figures should be combined or omitted
- Any subheadings should be labeled clearly
- Any relevant totals, subtotals, percentages, and so on should be shown
- The correctness of any calculations should be verified

Graphs

When illustrating data by means of a graph, follow these basic rules:

1. Give the graph a clear title

2. Label the axes clearly

3. Do not put too many curves on the same graph

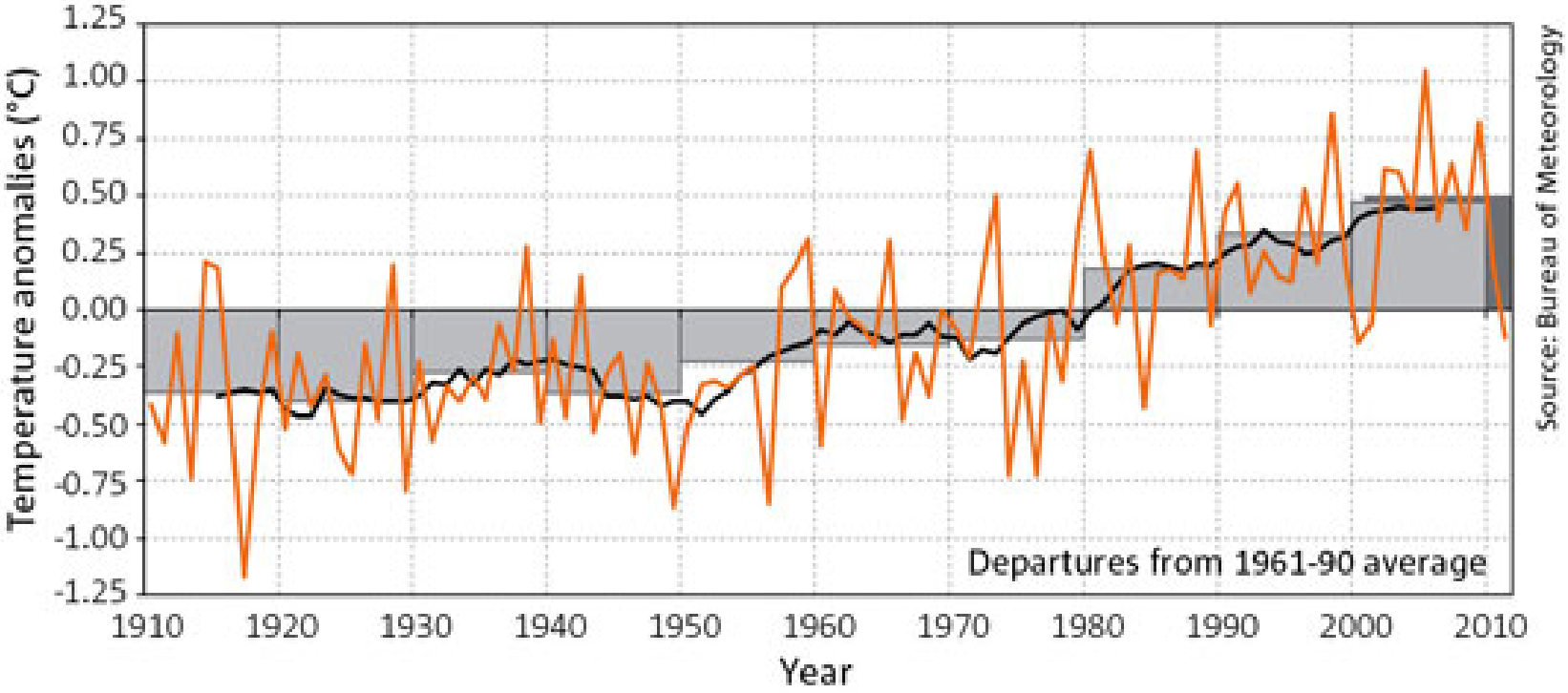
4. Quote all sources of data

5. If possible, accompany the graph with the table of the data that are represented by the graph

6. Use a zero on the scale of the vertical axis if possible

Graphs

Figure 1: Australian average temperatures over land



Changes in average temperature for Australia for each year (orange line) and each decade (grey boxes), and 11-year average (black line – an 11-year period is the standard used by the Intergovernmental Panel on Climate Change). Anomalies are the departure from the 1961-1990 average climatological period. The average value for the most recent 10-year period (2002–2011) is shown in darker grey.

Bar graph

Definition 1. *A graph of bars whose heights represent the frequencies (or relative frequencies) of respective categories is called a* **bar graph**.

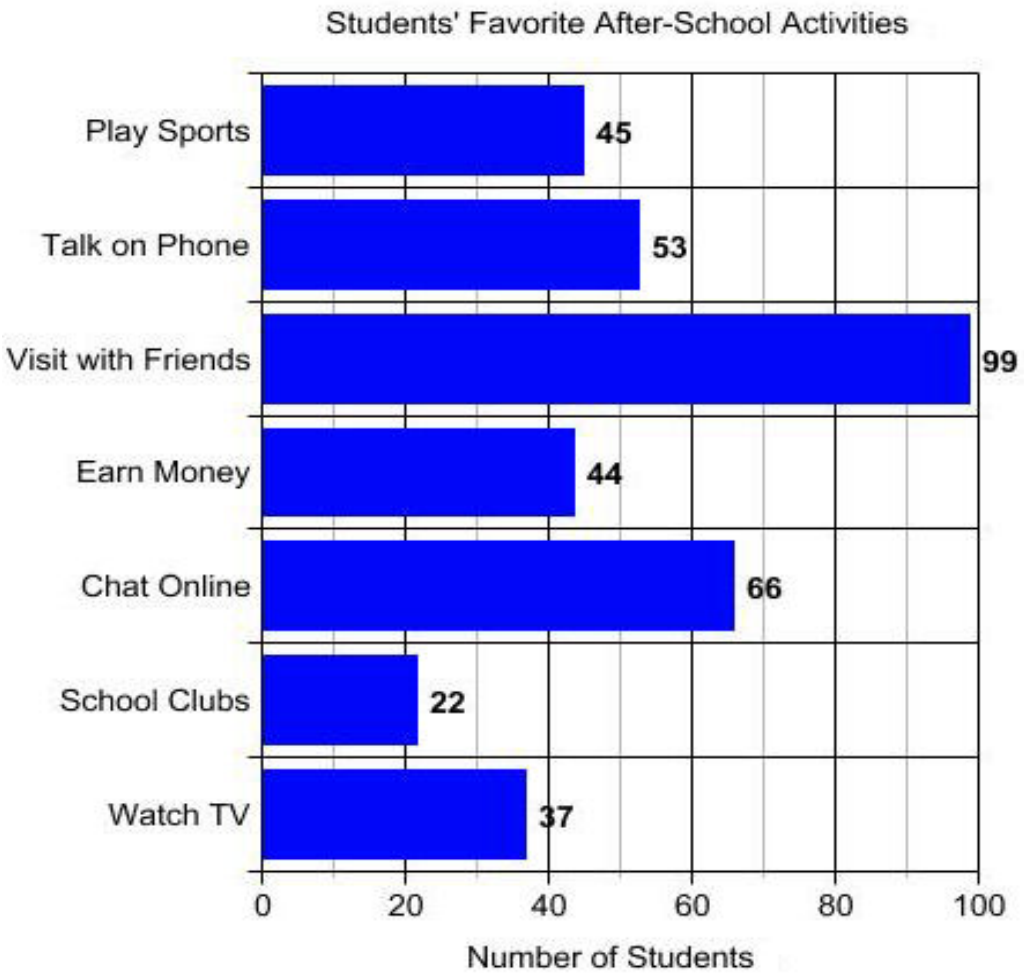
Use these guidelines when preparing a bar chart:

1. Make all bars the same *width*
2. If the bar graph is such that an ordering is feasible or desirable, use one of the most common orders
3. Clearly label the axes
4. Include keys to assist in the interpretation if necessary
5. Include footnotes, tables and sources of data

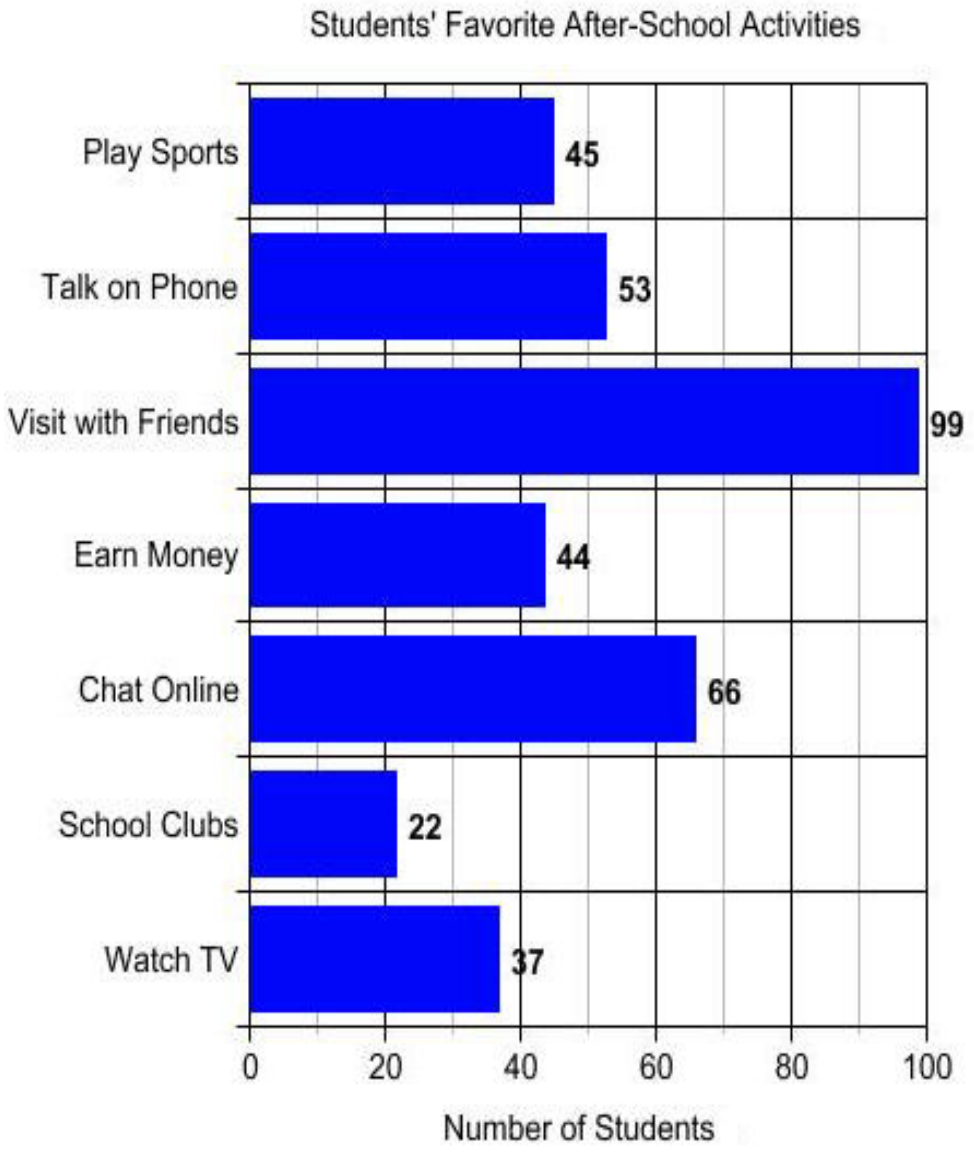
Example 1: A survey of students' favorite after-school activities was conducted at a school. The table below shows the results of this survey.

Students' Favorite After-School Activities	
Activity	Number of Students
Play Sports	45
Talk on Phone	53
Visit With Friends	99
Earn Money	44
Chat Online	66
School Clubs	22
Watch TV	37

Note that since the data in this table is not changing over time, a line graph would not be a good way to visually display this data. Each quantity listed in the table corresponds to a particular category. Accordingly, the data from the table above has been displayed in the bar graph below.



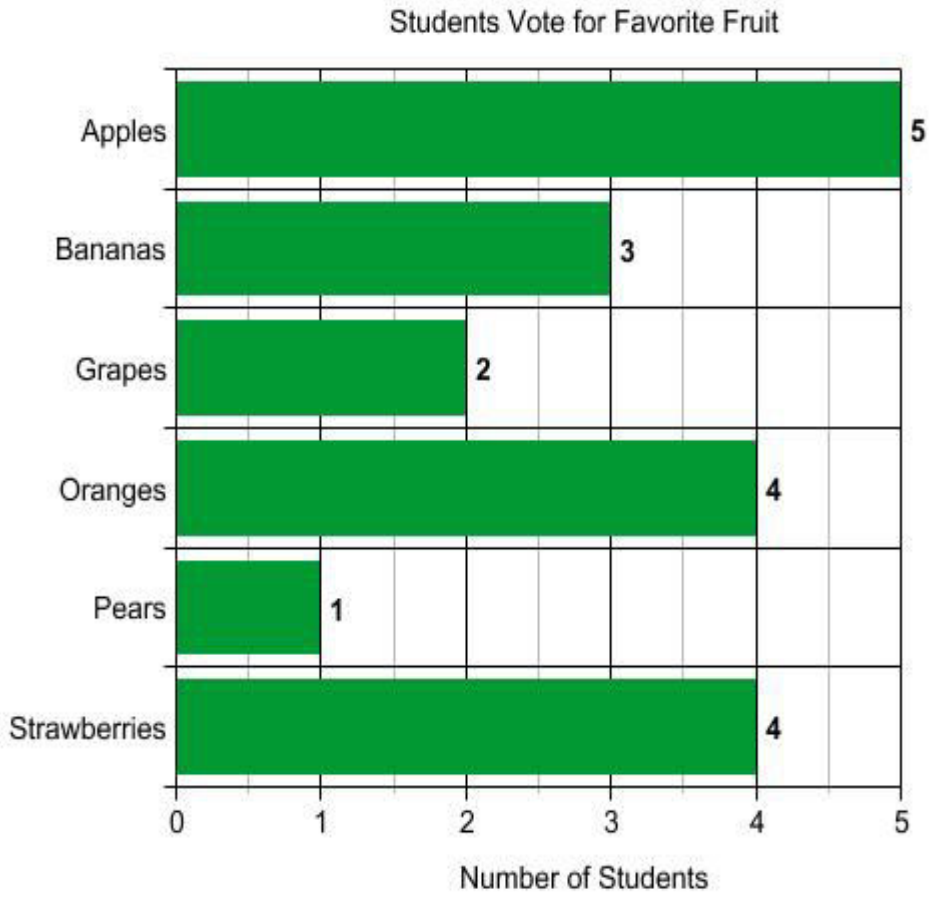
A **bar graph** is useful for comparing facts. The bars provide a visual display for comparing quantities in different categories. Bar graphs help us to see relationships quickly. Another name for a bar graph is a bar chart. Each part of a bar graph has a purpose.



QUESTION

1. What is the title of this bar graph?
2. What is the range of values on the (horizontal) scale?
3. How many categories are in the graph?
4. Which after-school activity do students like most?
5. Which after-school activity do students like least?
6. How many students like to talk on the phone?
7. How many students like to earn money?
8. Which two activities are liked almost equally?
9. List the categories in the graph from greatest to least.

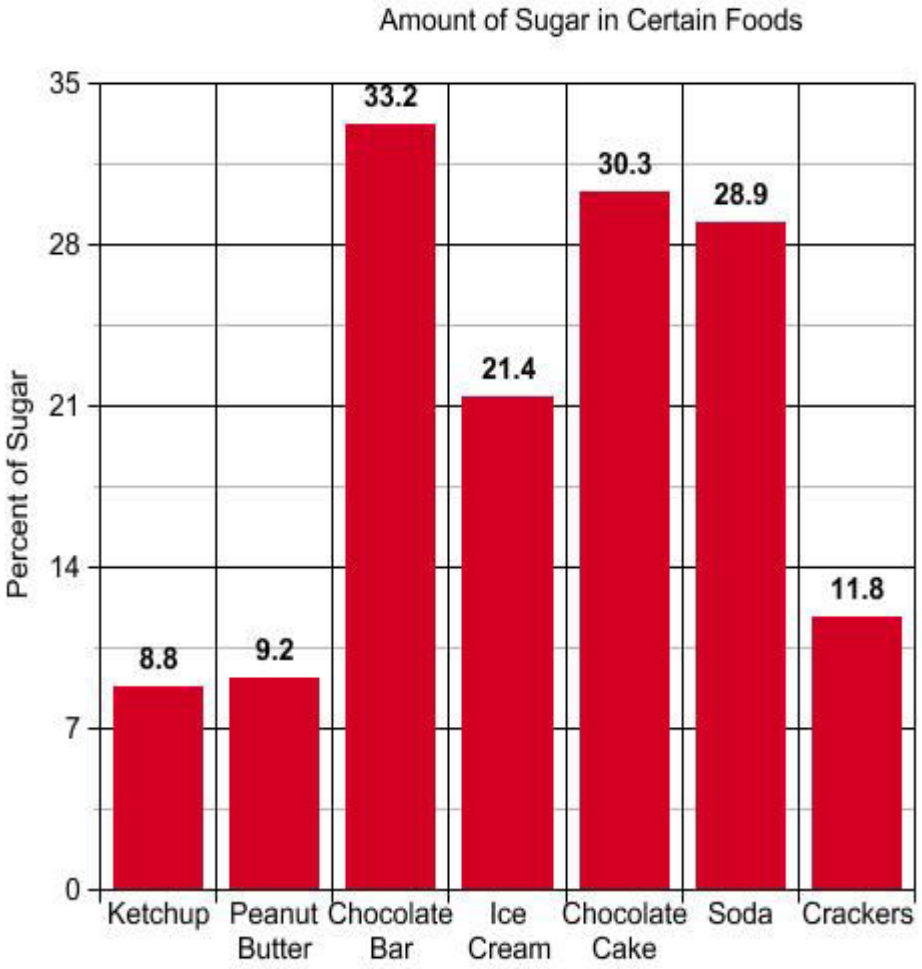
Example 2: Students in a class voted on their favorite fruit. Each student voted once. The bar graph below summarizes the data collected from the class vote.



QUESTION

1. What is the range of values on the (horizontal) scale?
2. How many categories are in the graph?
3. Which fruit had the most votes?
4. Which fruit had the least votes?
6. How many students voted for bananas?
7. How many students voted for grapes?
8. Which two fruits had the same number of votes?
9. List the categories in the graph from least to greatest

Example 3: The amount of sugar in 7 different foods was measured as a percent The data is summarized in the bar graph below**.**



<u>QUESTION</u>

1. What is the title of this bar graph?
2. What is the range of values on the (vertical) scale?
3. How many categories are in the graph?
4. Which food had the highest percentage of sugar?
5. Which food had the lowest percentage of sugar?
6. What percentage of sugar is in soda?
7. What is the difference in percentage of sugar between ice cream and crackers?

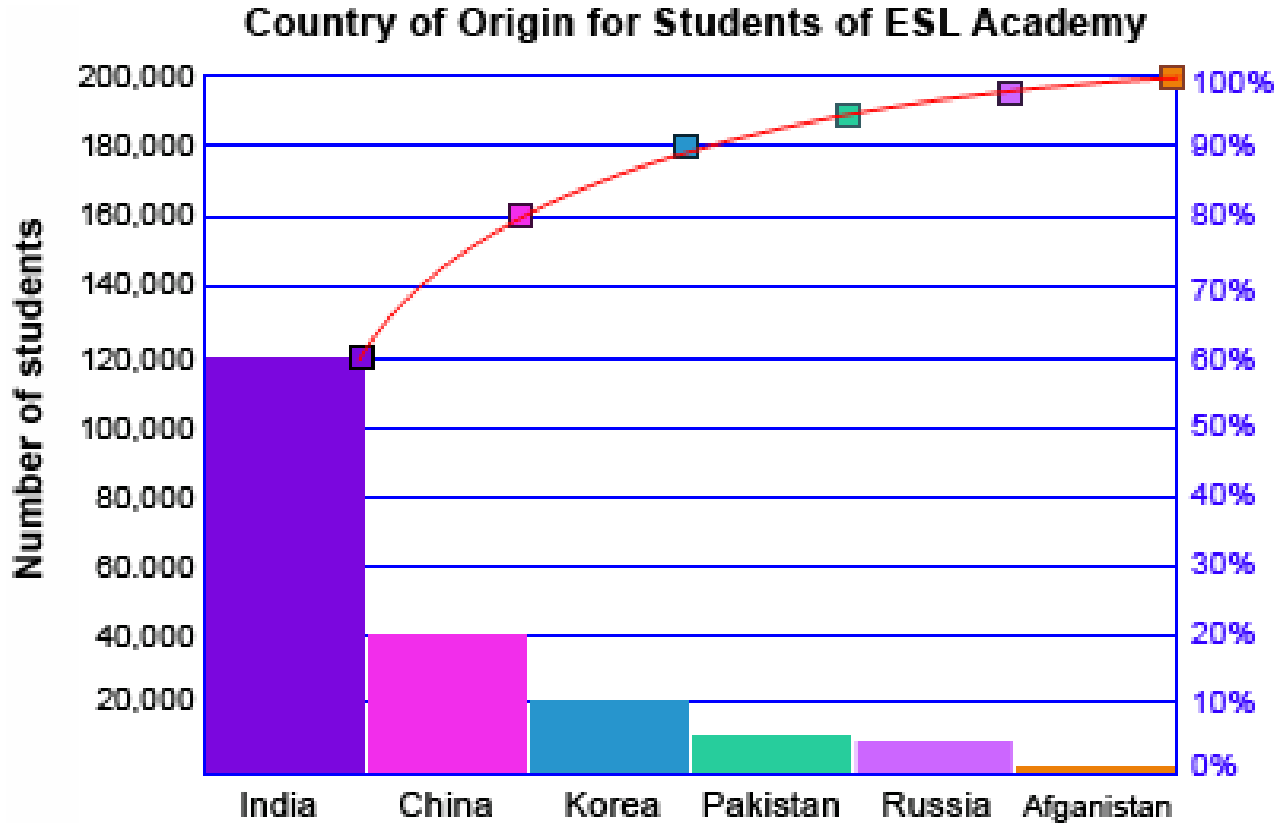
Pareto graph

Definition 2. *A **Pareto Graph** (also known as a Pareto Chart) is a graph that is used to show the differences between groups of data – and the relative importance of each data group.*

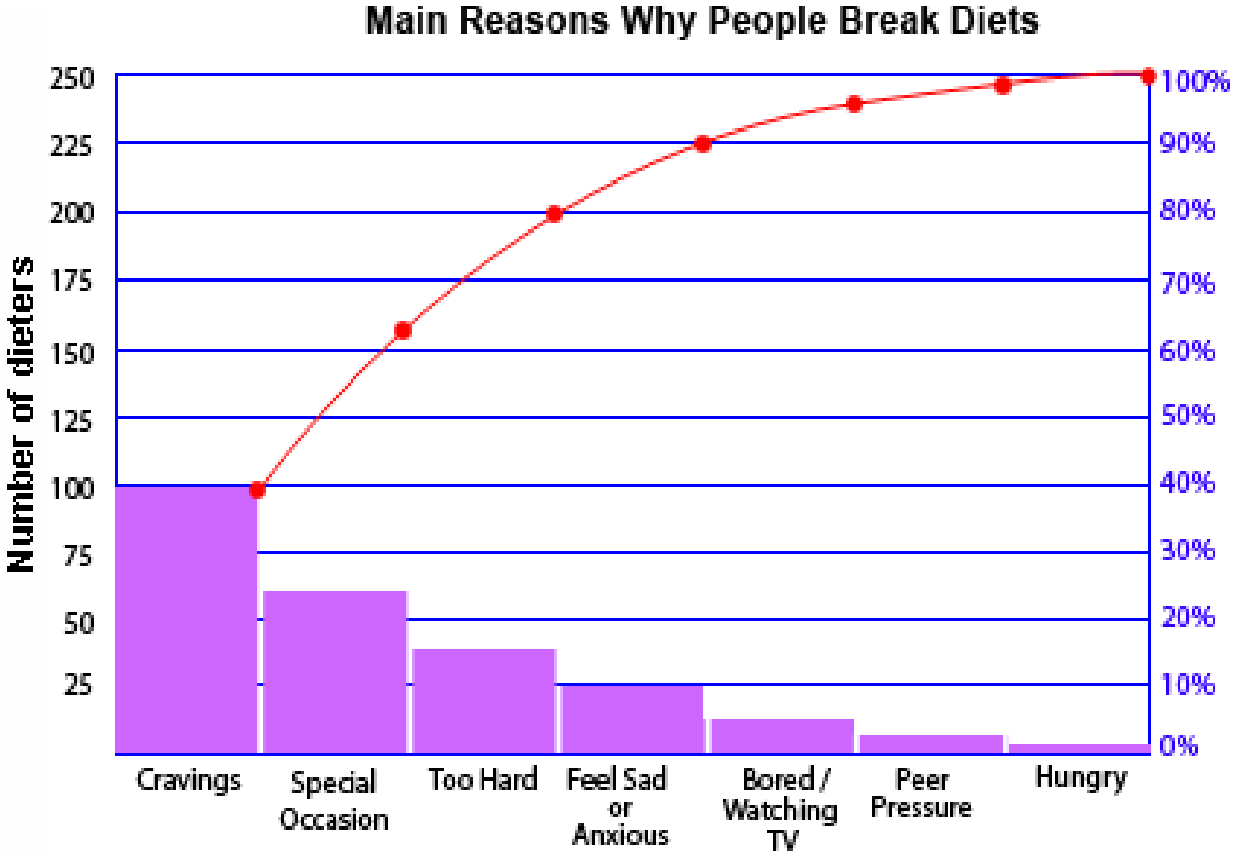
A Pareto Graph includes both lines and bars**.**

This style of graph was named after **Vilfredo Pareto** (1848 – 1923), an Italian industrialist, economist and sociologist, who also introduced the **80/20** rule (80% of a company's income comes from 20% of its customers, 20% of students do 80% of the volunteering in school, etc.).
A Pareto Graph will often depict the **80/20** rule for many situations – even though the percentages may change a bit.

For instance, if this rule were applied in the brewing industry to identify the best beer customers, you would find that 4% of customers are responsible for 65% of the gallonage sold!
A *Pareto Graph* is known as a *Juran Diagram* in some schools



The *left vertical axis* usually shows the frequency with which an event occurs, the cost of a unit, percentages, the number of people represented and other importance measurements.
The *right vertical axis* usually shows the cumulative percentage of the occurrence, the total costs, etc.

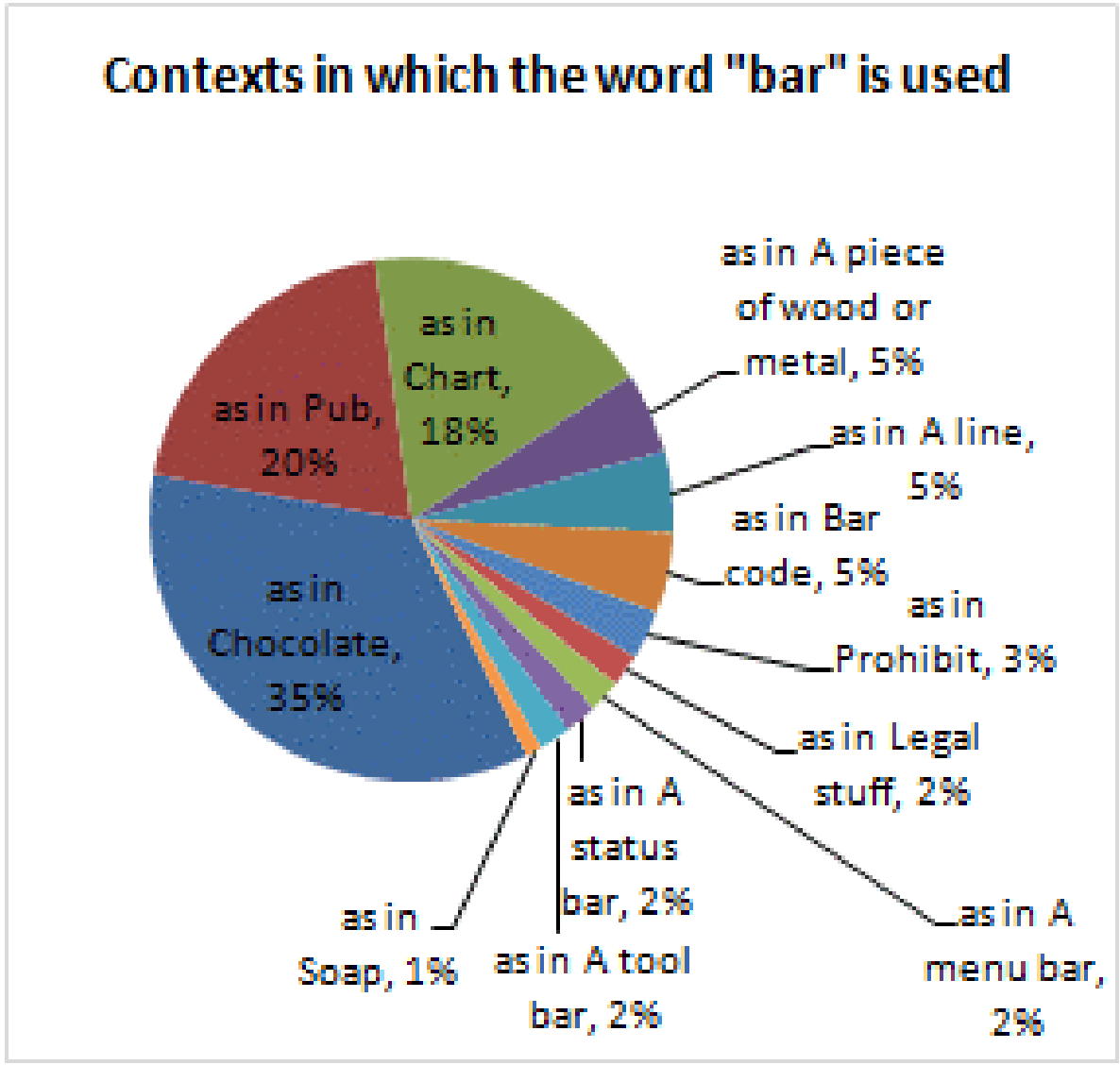


The bars show the relative importance of each group of data and are arranged in descending order (i.e. from the highest to the lowest). The *line graph* is used to show the cumulative totals of each category (represented by the bars), going from left to right.

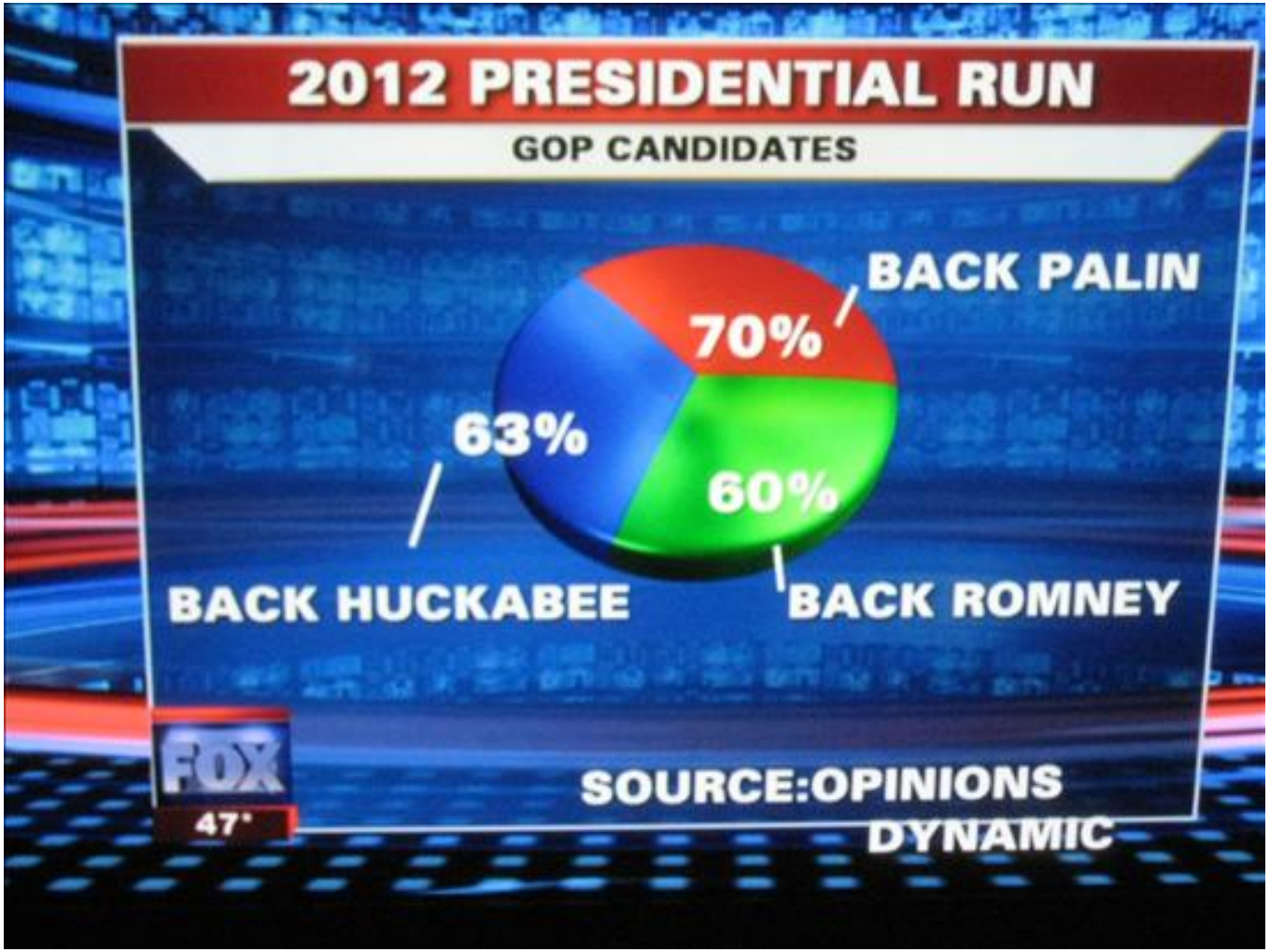
Pie chart

Definition 3. *A circle divided into sectors that represent the percentages of a population or a sample that belongs to different categories is called a* **pie chart**.

- A *pie chart* is often used to give a visual presentation of data to indicate the proportions that make up a given total
- The pie chart is a circle that is divided by radial lines into sectors in such a way that the area of each sector is proportional to the size of the quantity represented by that sector
- In a case where the sectors are large enough to do so, the percentage is often written inside the sector. Otherwise it is written outside the chart

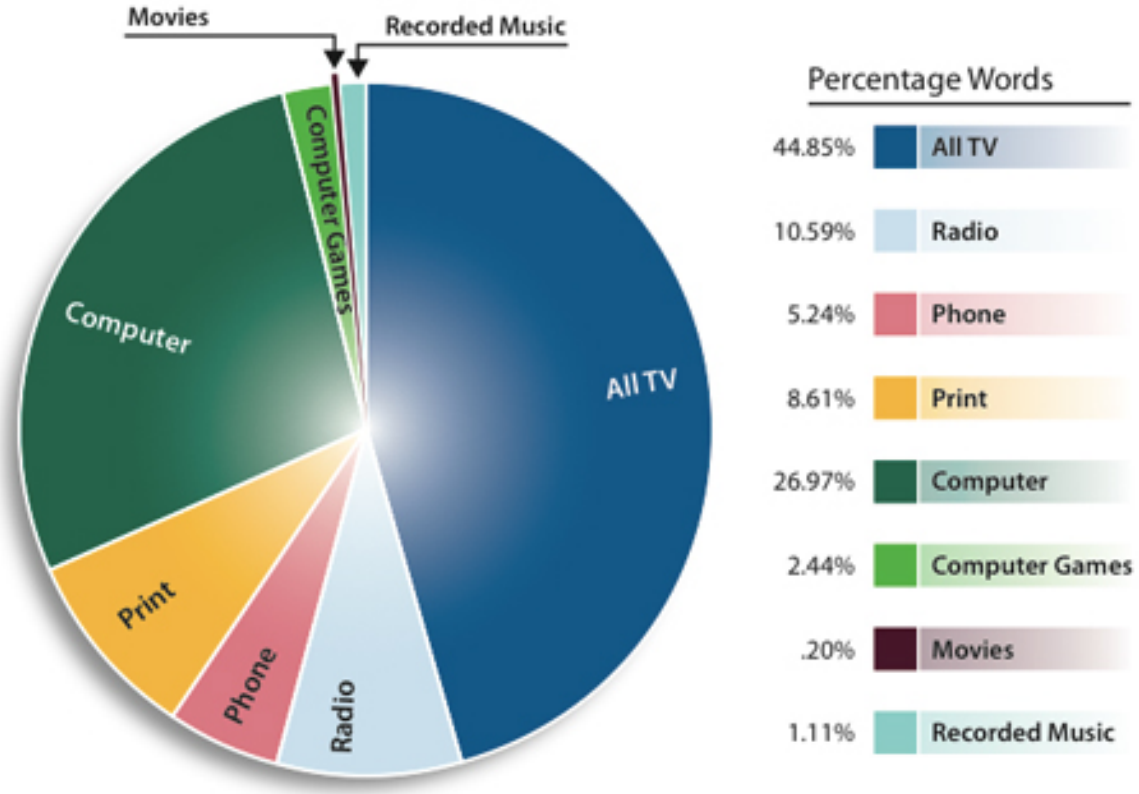

Contexts in which the word "bar" is used
as in Chart, 18%
as in A piece of wood or metal, 5%
as in A line, 5%
as in Bar code, 5%
as in Prohibit, 3%
as in Legal stuff, 2%
as in A menu bar, 2%
as in A status bar, 2%
as in A tool bar, 2%
as in Soap, 1%
as in Chocolate, 35%
as in Pub, 20%

Fox News pie chart

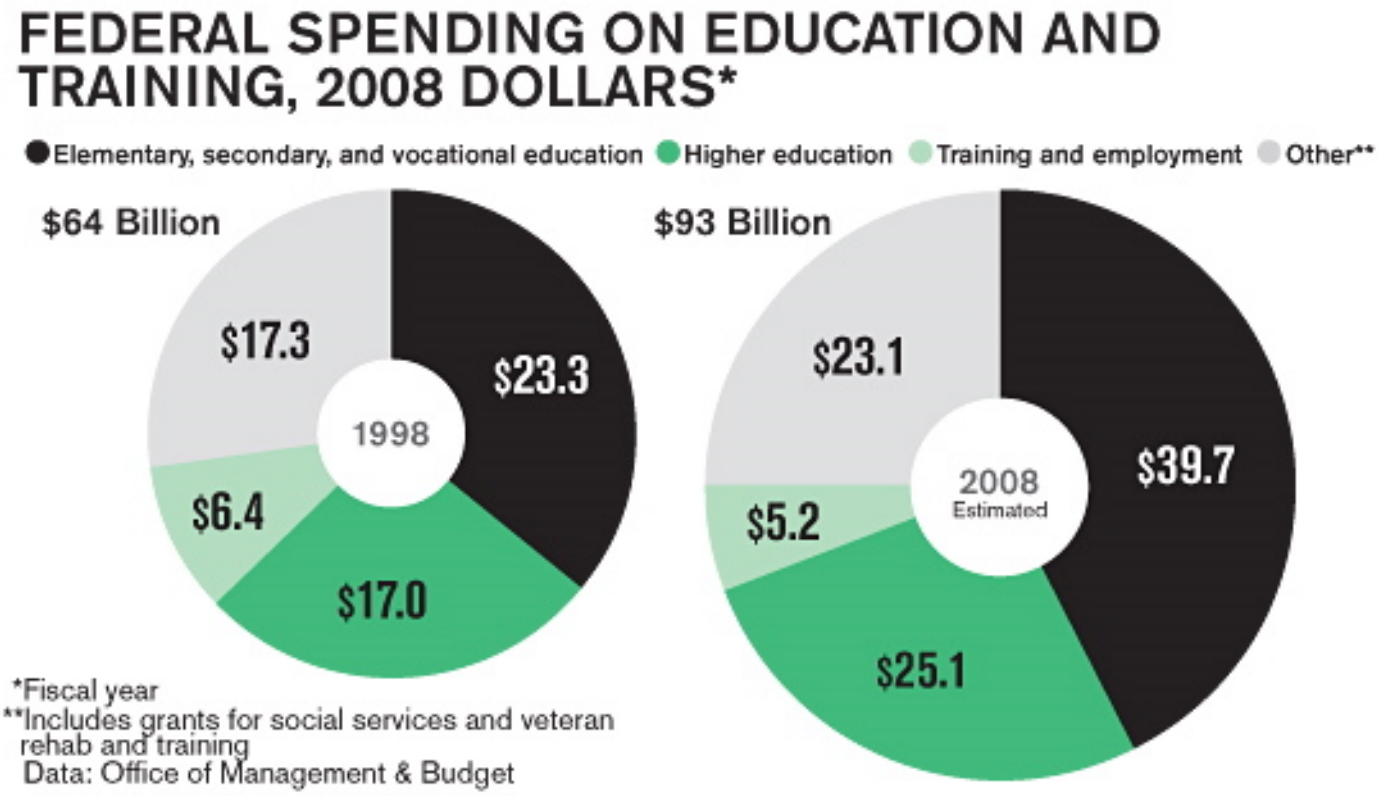


Pie Chart Mistakes

Care must be taken to retain the salient feature of the chart: the center. In a report on the number of words consumed each day, a pie chart was prettified with a highlight in the middle. That obscures the spot where the lines meet, and thus makes it impossible to judge angles, making the comparison more difficult.



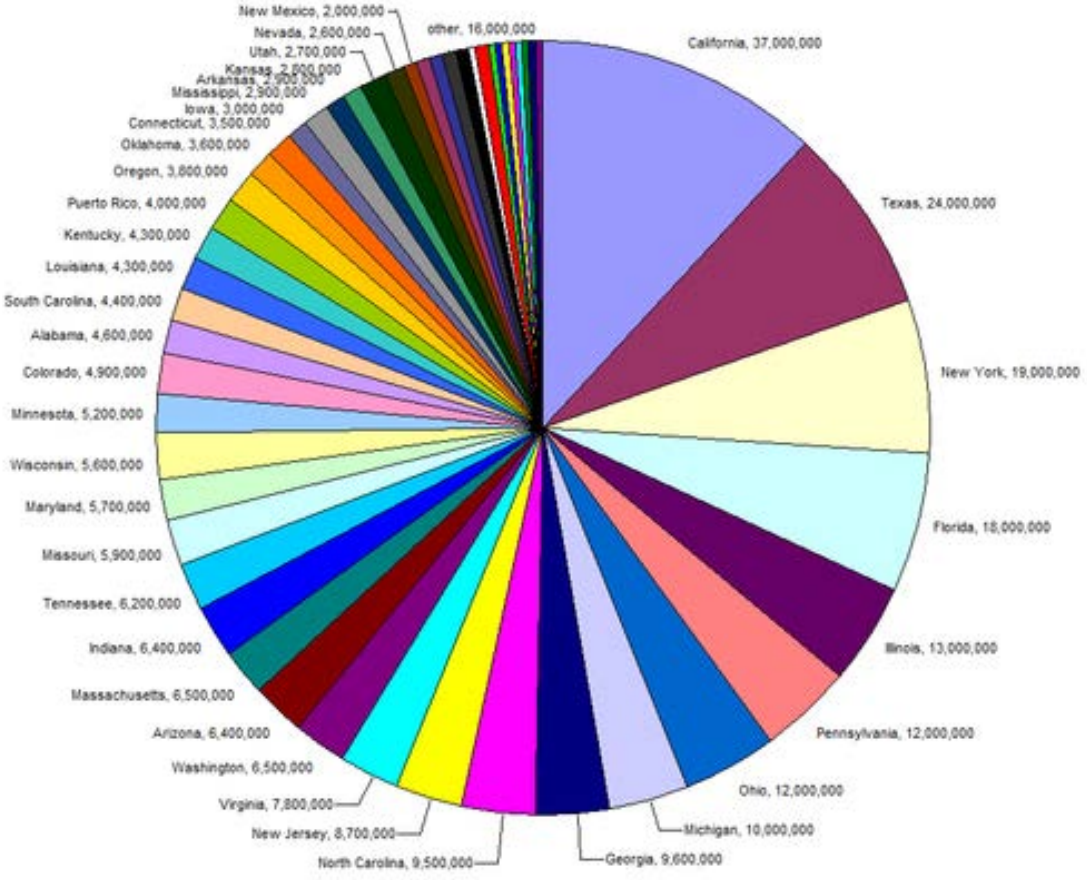
This is similar to the problem with a colleague of the pie chart, the donut chart. It is similar to the pie chart, but is missing a circular area in the center.



While the center may be a convenient spot for labeling, it degrades the chart's readability. The comparison between separate pie or donut charts is also largely meaningless, and should be avoided. To show progression over time, line and bar charts are much better suited. To compare two different kinds of data (absolute numbers and fractions), it makes more sense to split them up by data to compare than by year.

Number of Slices

The most common problem is trying to show too many categories in a single pie chart. Wikipedia has this beautiful specimen on the page on U.S. states by population. The first four states are clearly larger than any of the rest, and from there the chart turns from a meaningful visualization of numbers into a colorful pattern.



A bar chart would have been a much better idea here, because it would have allowed easier comparison between the states. Grouping together states of similar size into separate charts with different scales would have made it possible to clearly see the differences for all of them, not just the most populous ones.

When to Use Pie Charts

There are some simple criteria that you can use to determine whether a pie chart is the right choice for your data.

Do the parts make up a meaningful whole? If not, use a different chart. Only use a pie chart if you can define the entire set in a way that makes sense to the viewer.

Are the parts mutually exclusive? If there is overlap between the parts, use a different chart.

Do you want to compare the parts to each other or the parts to the whole? If the main purpose is to compare between the parts, use a different chart. The main purpose of the pie chart is to show part-whole relationships.

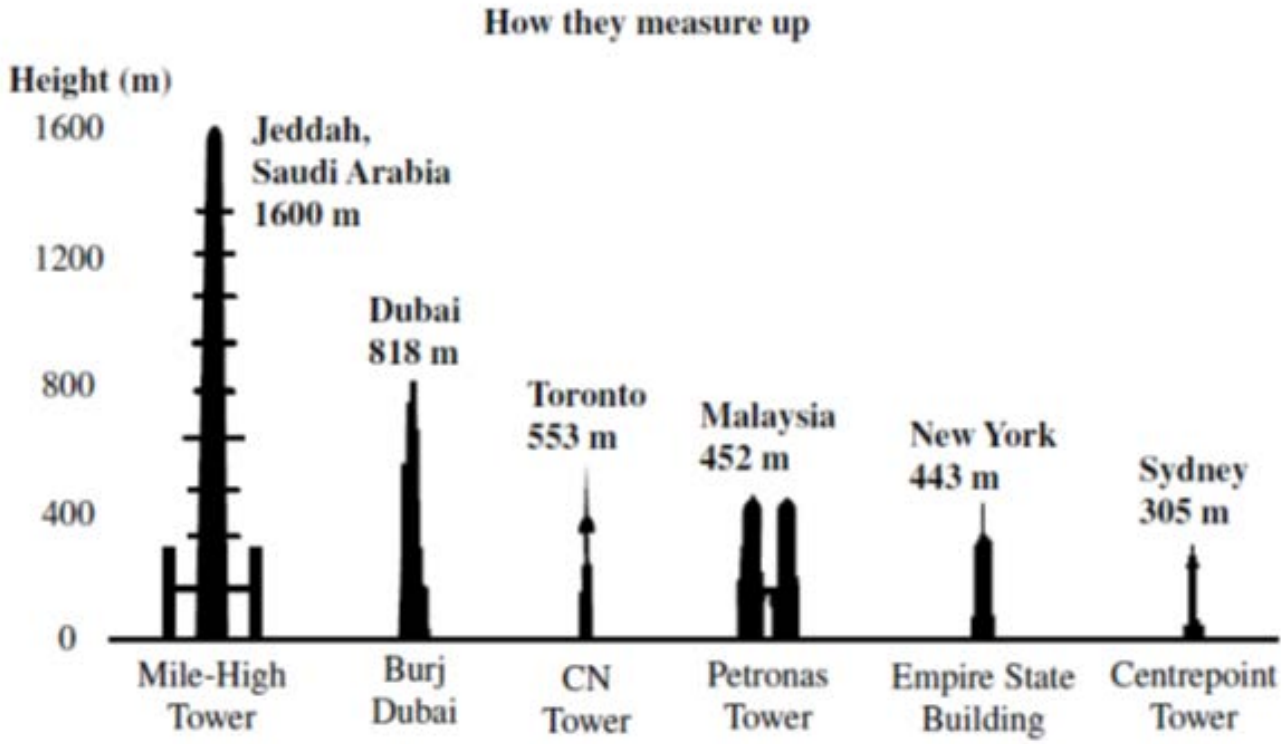
How many parts do you have? If there are more than five to seven, use a different chart. Pie charts with lots of slices (or slices of very different size) are hard to read.

In all other cases, do not use a pie chart. The pie chart is the wrong chart type to use as a default; the bar chart is a much better choice for that. Using a pie chart requires a lot more thought, care, and awareness of its limitations than most other charts.

Pictograms

A pictogram is a graph in which data are displayed using pictures rather than the traditional methods

There are no real rules for drawing a pictogram



Frequency distributions

- When data is collected it is initially presented with the *observations* (i.e. sample values) in some random order
- Information before it is analyzed is called ‘raw’ because it is unprocessed by statistical technique
- There are various techniques for condensing data into a comprehensible form

Arrays

- The first step in simplifying data is to arrange it in an array; that is, to arrange the data in some order
- One method is to arrange the data in order of magnitude from the smallest to the largest

Frequency distributions

- The frequency of an observation is the number of times that observation occurs
- Series of statistical observations are sometimes said to have a certain *distribution*
- A *frequency distribution* gives a listing of different observations (in order of magnitude), each with the corresponding frequency alongside it

Definition 4. *Let* f_i *denote the frequency of the class i and let n be sum of all frequencies. Then the* **relative frequency** *for the class i is defined as the ratio fi/n. The* **cumulative relative frequency** *for the class i is defined by* $\sum_{k=1}^{i} f_k/n.$

Example 4. The following data give the lifetime of 30 incandescent light bulbs (rounded to the nearest hour) of a particular type.

872	931	1146	1079	915	879	865	1112	979	1120
1150	987	958	1149	1057	1082	1053	1048	1118	1088
868	996	1102	1130	1002	990	1052	1116	1119	1028

Construct a frequency, relative frequency, and cumulative relative frequency table.

Solution. Note that there are n = 30 observations and that the largest observation is 1150 and the smallest one is 865 with a range of 285. We will choose six classes each with a length of 50.

Class	Frequency f_i	Relative frequency $\frac{f_i}{\sum f_i}$	Cumulative relative frequency $\sum_{k=1}^{i} \frac{f_k}{n}$
850–900	4	4/30	4/30
900–950	2	2/30	6/30
950–1000	5	5/30	11/30
1000–1050	3	3/30	14/30
1050–1100	6	6/30	20/30
1100–1150	10	10/30	30/30

Definition 5. *A* **histogram** *is a graph in which classes are marked on the horizontal axis and either the frequencies, relative frequencies, or percentages are represented by the heights on the vertical axis. In a histogram, the bars are drawn adjacent to each other without any gaps.*

GUIDELINE FOR THE CONSTRUCTION OF A FREQUENCY TABLE AND HISTOGRAM

1. Determine the maximum and minimum values of the observations. The range,

$$R = \text{maximum value} - \text{minimum value}.$$

2. Select from five to 20 classes that in general are nonoverlapping intervals of equal length, so as to cover the entire range of data. The goal is to use enough classes to show the variation in the data, but not so many that there are only a few data points in many of the classes. The class width should be slightly larger than the ratio

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

3. The first interval should begin a little below the minimum value, and the last interval should end a little above the maximum value. The intervals are called class intervals and the boundaries are called class boundaries. The class limits are the smallest and the largest data values in the class. The class mark is the midpoint of a class.

4. None of the data values should fall on the boundaries of the classes.

5. Construct a table (frequency table) that lists the class intervals, a tabulation of the number of measurements in each class (tally), the frequency $\boldsymbol{f_i}$ of each class, and, if needed, a column with relative frequency, $\boldsymbol{f_i}/n$, where n is the total number of observations.

6. Draw bars over each interval with heights being the frequencies (or relative frequencies).

Example 6. The following data refer to a certain type of chemical impurity measured in parts per million in 25 drinking water samples randomly collected from different areas of a county.

11	**19**	**24**	**30**	**12**	**20**	**25**	**29**	**15**	**21**
24	31	16	23	25	26	32	17	22	26
35	18	24	18	27					

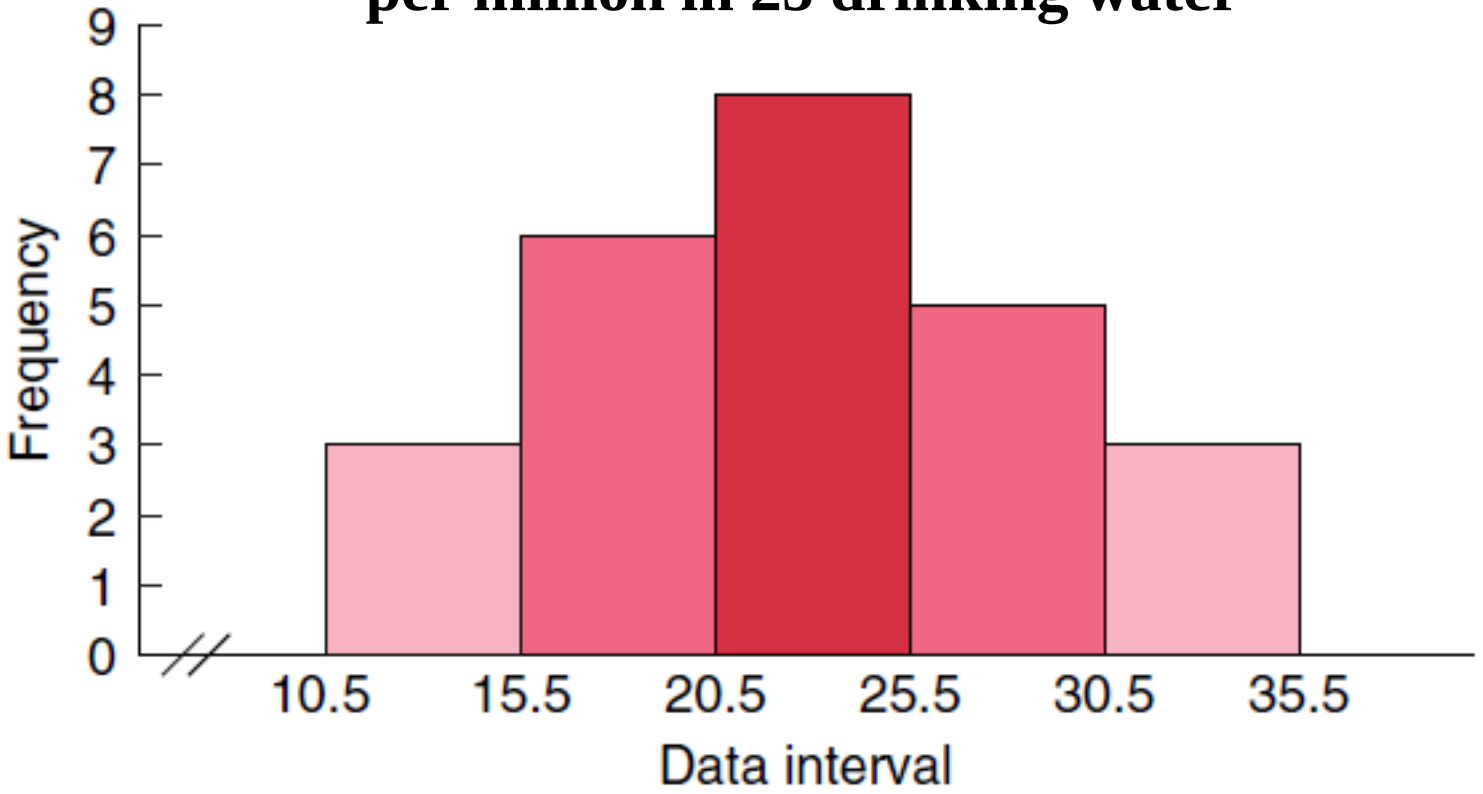
(a) Make a frequency table displaying class intervals, frequencies, relative frequencies, and percentages.
(b) Construct a frequency histogram.

Solution

(a) We will use five classes. The maximum and minimum values in the data set are 35 and 11. Hence the class width is $(35 - 11)/5 = 4.8 \approx 5$. Hence, we shall take the class width to be 5. The lower boundary of the first class interval will be chosen to be 10.5. With five classes, each of width 5, the upper boundary of the fifth class becomes 35.5. We can now construct the frequency table for the data.

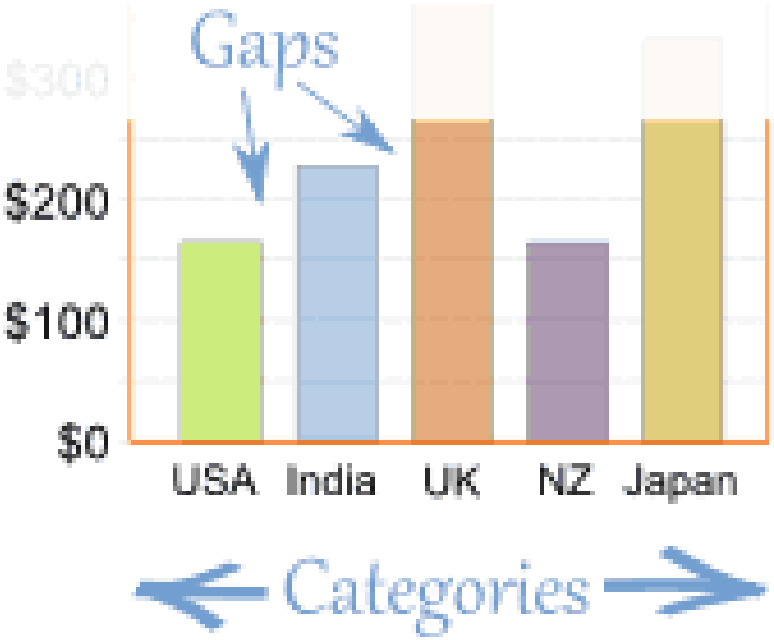
Class	Class interval	f_i = frequency	Relative frequency	Percentage
1	10.5 – 15.5	3	3/25 = 0.12	12
2	15.5 – 20.5	6	6/25 = 0.24	24
3	20.5 – 25.5	8	8/25 = 0.32	32
4	25.5 – 30.5	5	5/25 = 0.20	20
5	30.5 – 35.5	3	3/25 = 0.12	12

Certain type of chemical impurity measured in parts per million in 25 drinking water

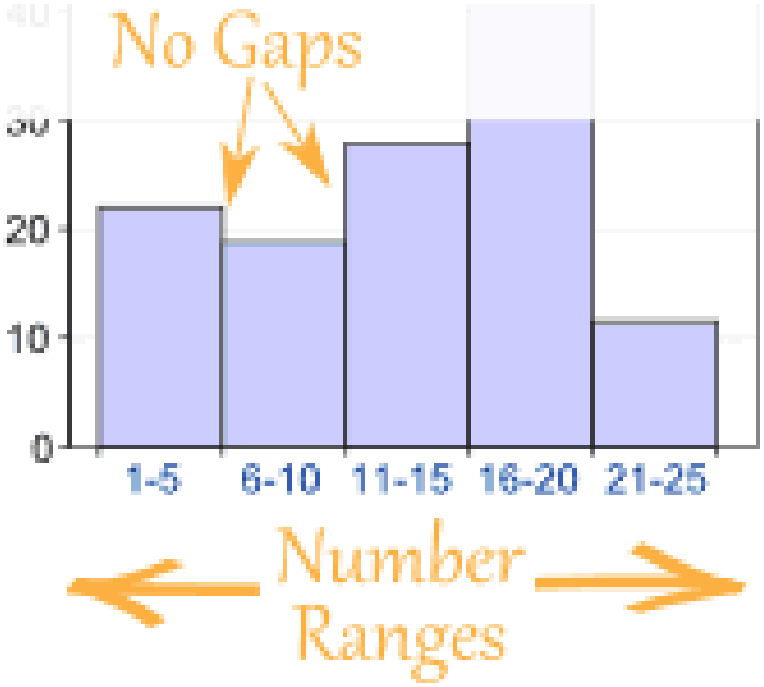


From the histogram we should be able to identify the center (i.e., the location) of the data, spread of the data, skewness of the data, presence of outliers, presence of multiple modes in the data, and whether the data can be capped with a bell-shaped curve.

Difference between Bar Graph and Histogram



Bar Graph



Histogram

Abuse of statistics

There are many instances in practice where statistical information has been presented in such a manner as to purposefully (or otherwise) mislead the reader

Meaningless statements

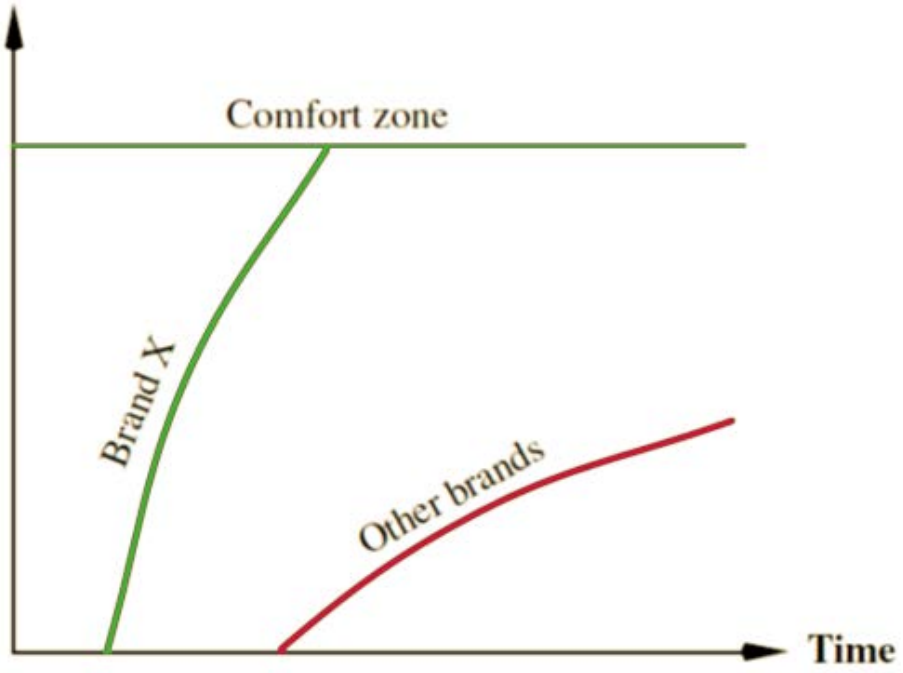
E.g. ‘Brand *X* washing detergent will wash your clothes 50% cleaner and brighter.’

Comment: What does ‘50%’ mean? Cleaner and brighter than what? Nothing at all, another product or...? What is the basis for comparison?

Abuse of statistics

Misleading graphs

- One of the most common ways to distort data is by means of a graph
- Example: the graph has no scale, includes meaningless quantities and is quite useless



Summary

- We have looked at some of the ways in which otherwise uninteresting data can be displayed in an eye-catching manner

- This is particularly important in the media, where graphic and other artists are assigned the task of creating a product that will attract the attention of the reader

- The task of condensing information to make it more comprehensible requires a degree of skill to decide just how much detailed information can afford to be lost in this process

- Determining the shape of the data is also a vital aspect for statisticians who wish to conduct further scientific analysis with a view to drawing appropriate conclusions