DESCRIPTIVE STATISTICS

This chapter is concerned with the nature of data, how they can be presented and how to summarise information. **This** is the domain of descriptive statistics. Data may be classified as being discrete if the variable can take only a finite number of values, for example, number of pregnancies, number of males in an animal population; or continuous if the variable can (at least within a certain range) take any value along the number line, for example, height, plasma cholesterol level, blood pressure.

Tables, Graphs, and Charts

Data analysis is an important component of scientific practice. **To** analyze data effectively, a researcher must first become familiar with the data before applying analytic techniques. The researcher may begin by examining individual records such as those contained in a list, but will quickly progress to summarizing the data with tables. **When** the amount of data is small and relationships are straightforward, the resulting tables are the only analysis that is needed. **When** the data are more complex, graphs and charts can help the researcher visualize broader patterns and trends and identify variations from those trends. Variations may represent important new findings or only errors in typing or coding which need to be corrected. **Thus**, tables, graphs, and charts are essential to the verification and analysis of the data. Once an analysis is complete, tables, graphs, and charts further serve as useful visual aids for describing the data to others.

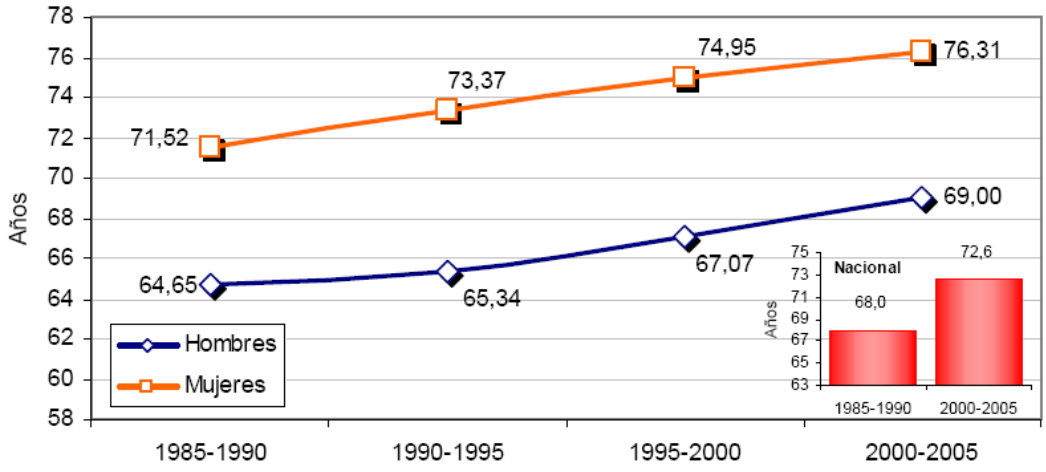
Tables

A table is a set of data arranged in rows and columns. Almost any quantitative information can be organized into a table. Tables are useful for demonstrating patterns, exceptions, differences, and other relationships. **In addition**, tables usually serve as the basis for preparing more visual displays of data, such as graphs and charts, where some of the detail may be lost. Tables designed to present data to others should be as simple as possible. Two or three small tables, each focusing on a different aspect of the data, are easier to understand than a single large table that contains many details or variables. A table should be self-explanatory. **If** a table is taken out of its original context, **it** should still convey all the information necessary for the reader to understand the data.

Table 1 Life expectancy for men and women in Colombia

Year	1985-1990	1990-1995	1995-2000	2000-2005
Men	71.52	73.37	74.95	76.31
Women	64.65	65.34	67.07	69.00

Line Graphs

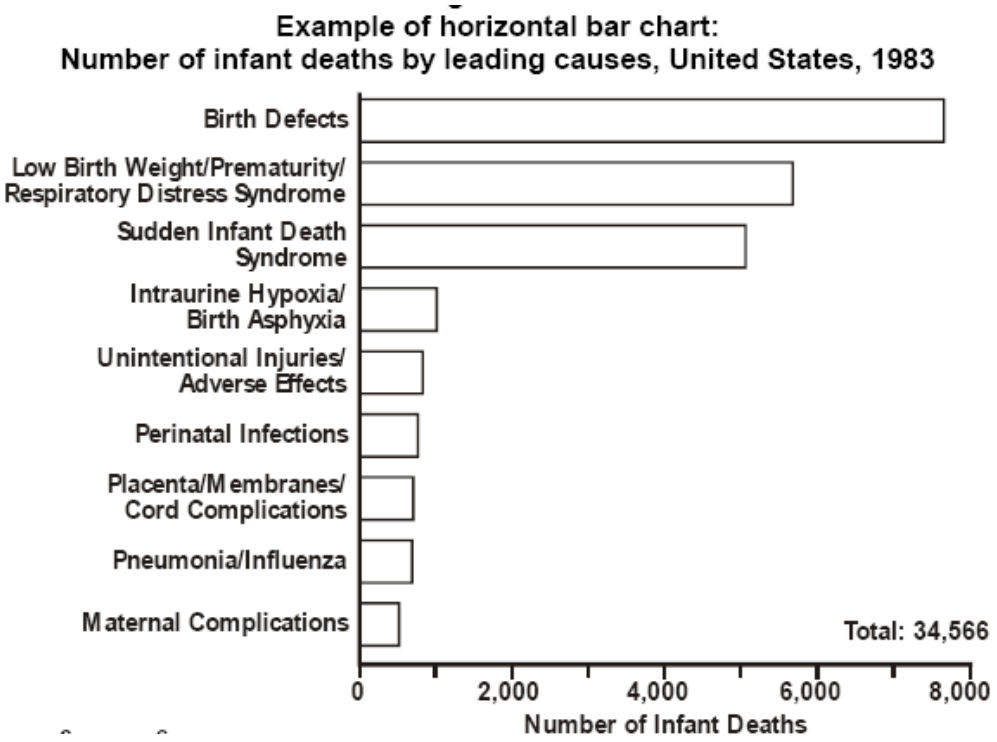


Life expectancy for men and women in Colombia

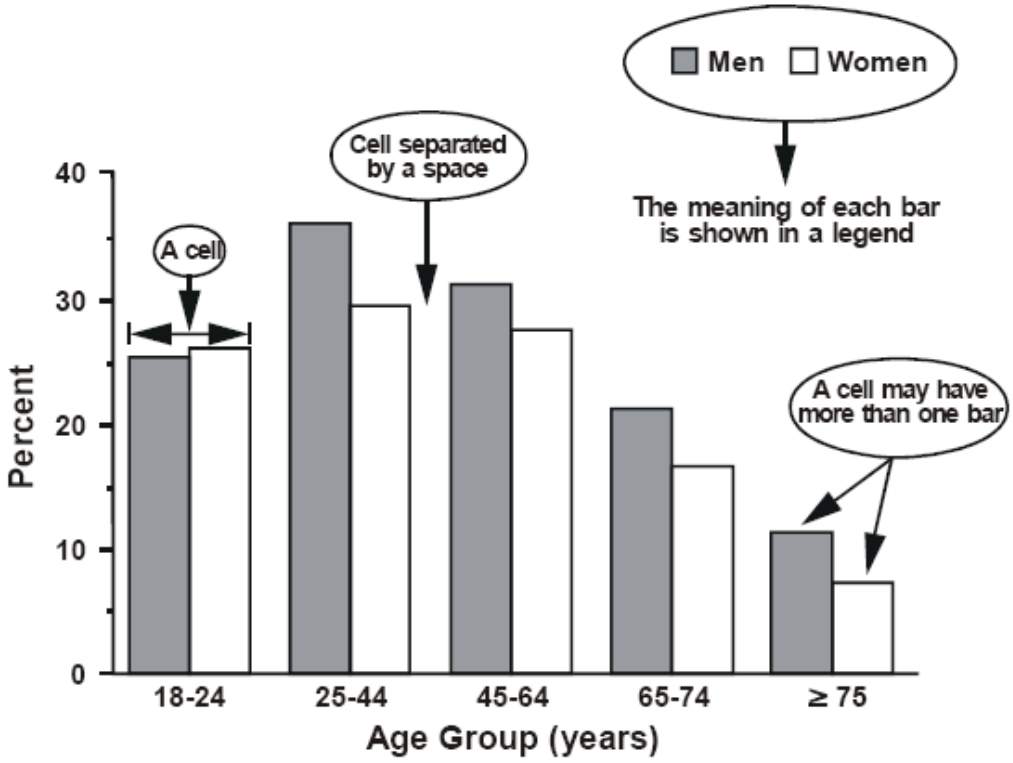
A line graph is a way to show quantitative data visually, using a system of coordinates. **It** is a kind of statistical snapshot that helps us see patterns, trends, aberrations, similarities, and differences in the data. **Also**, a line graph is an ideal way of presenting data to others. The important aspects of a set of data are better remembered from a graph line than from a table. **It** is common to use rectangular coordinate graphs, which have two lines, one horizontal and one vertical, that intersect at a right angle. **These lines** are referred as the horizontal axis (or ***x***-axis), and the vertical axis (or ***y***-axis). The horizontal axis is usually used to show the values of the independent (or ***x***) variable, which is the method of classification, such as time. The vertical axis is used to show the dependent (or ***y***) variable, which is usually a frequency measure, such as number of cases or rate of disease. Each axis is labeled to show what it represents (both the name of the variable and the units in which it is measured) and mark a scale of measurement along the line. Line graphs are made by joining up points plotted on a graph.

Bar Charts

Charts are methods of illustrating statistical information using only one coordinate. They are most appropriate for comparing data with discrete categories other than place, but have many other uses as well. The simplest bar chart is used to display the data from a one-variable table. Each value or category of the variable is represented by a bar. The length of the bar is proportional to the number of persons or events in that category. The figure below shows the number of infant deaths by cause in the United States. This presentation of the data makes **it** very easy to compare the relative size of the different causes and to see that birth defects are the most common cause of infant mortality. Variables shown in bar charts are either discrete and noncontinuous (e.g., race; sex) or are treated as though they were discrete and noncontinuous (e.g., age groups rather than age intervals along an axis). Bars can be presented either horizontally or vertically:



Example of vertical bar chart with annotation: Percentage of adults who were current cigarette smokers (persons ≥18 years of age who reported having smoked at least 100 cigarettes and who were currently smoking) by sex and age, United States, 1988



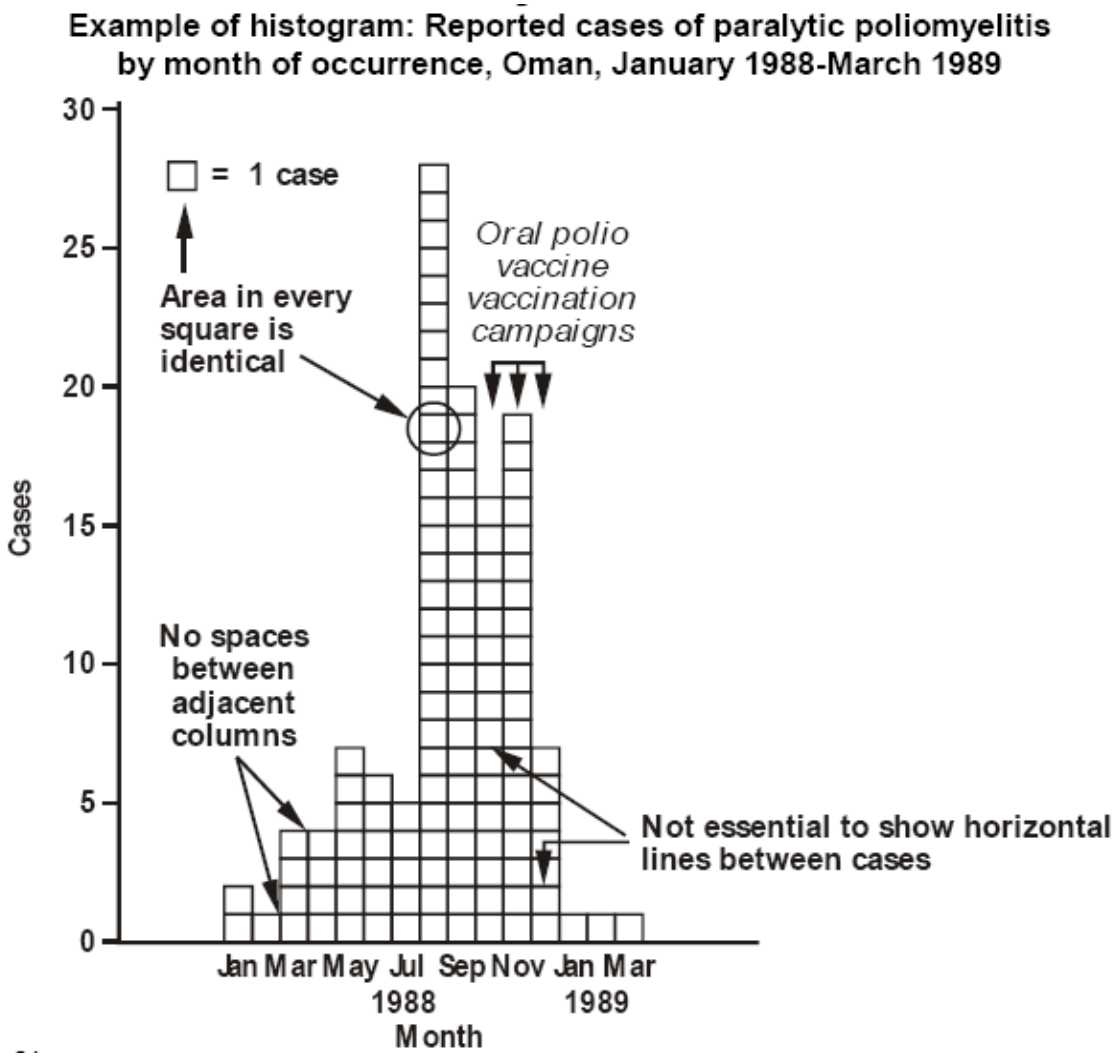
The length or height of each bar is proportional to the frequency of the event in that category. For **this reason**, a scale break should not be used with a bar chart since this could lead to misinterpretation in comparing the magnitude of different categories.

A vertical bar chart differs from a histogram (see below) in that the bars of a bar chart are separated while the bars of a histogram are joined. **This distinction** follows from the type of variable used on the ***x***-axis. A histogram is used to show the frequency distribution of a continuous variable such as age or serum cholesterol or dates of onset during an epidemic. A bar chart is used to show the frequency distribution of a variable with discrete, noncontinuous categories such as sex or race.

The vertical bar chart above represents three variables: age, sex, and current smoking status. Current smoking status is the outcome variable and has two categories: yes or no. The bars represent the 10 age-sex categories. The height of each bar is proportional to the percentage of current smokers in each age-sex category.

The histogram

A histogram is a graph of the frequency distribution of a continuous variable. It uses adjoining columns to represent the number of observations for each class interval in the distribution. The area of each column is proportional to the number of observations in that interval.



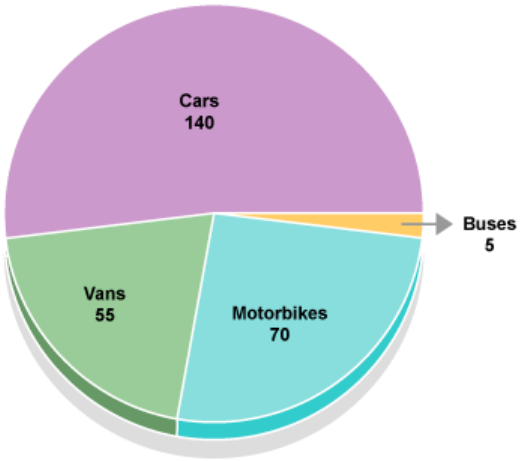
Pie Charts

A pie chart is a simple, easily understood chart in which the size of the "slices" show the proportional contribution of each component part. Pie charts are useful for showing the component parts of a single group or variable. To draw a pie chart, we need to represent each part of the data as a proportion of 360, because there are 360 degrees in a circle. For example, if 55 out of 270 vehicles are vans, we will represent this on the circle as a segment with an angle of: $(^{55}/_{270})$ x 360 = 73 degrees. **This** will give the following results:

Table 2 Traffic Survey 31 January 2008

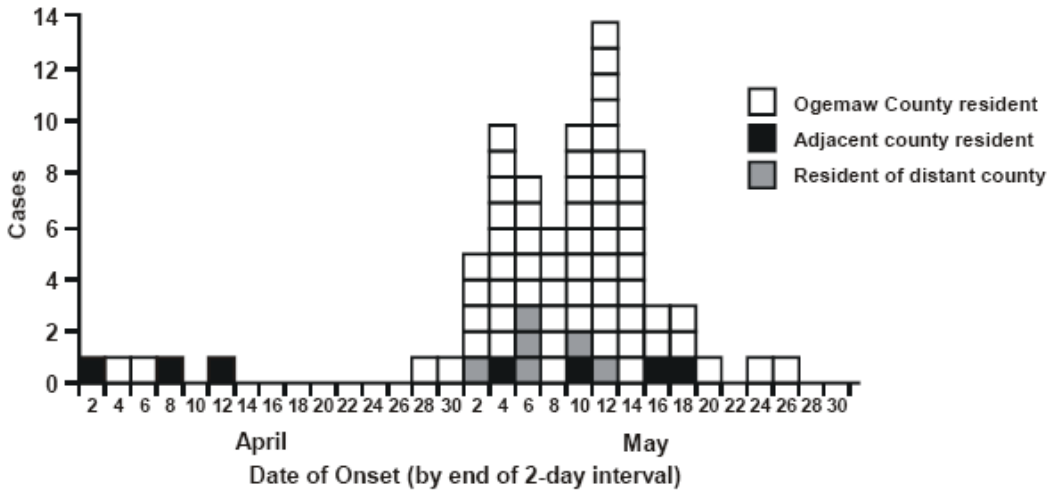
Type of vehicle	Number of vehicles	Calculation	Degrees of a circle
Cars	140	$(^{140}/_{270})$ x 360	= 187
Motorbikes	70	$(^{70}/_{270})$ x 360	= 93
Vans	55	$(^{55}/_{270})$ x 360	= 73
Buses	5	$(^{5}/_{270})$ x 360	= 7

This data is represented on the pie chart below:



ACTIVITY

1. Suppose we wish to present information on the drinking habits of Colombian males. Let us say that 1000 males were selected at ramdom from the electoral roll (list of people entitled to vote) and their drinking habits were ascertained. For convenience we might categorise the variable "drinking habits" into 6 classes of “grams of alcohol consumed per day”: 0 to 9 g/day, 10 to 19 g/day, 20 to 29 g/day, 30 to 39 g/day, 40-59 g/day and 60-99 g/day. Present this information by means of a table.
2. Present the same information by means of a bar graph (remember that the height of each bar must be proportional to the frequency of occurrence of its respective score), and a pie bar (each slice of the pie must have an area proportional to the relative frequency of the score it is representing).
3. The histogram below shows the number of cases of hepatitis A diagnosed by date of onset and residency status in Ogemaw County, April-May 1968. Draw a table detailing the number of cases diagnosed between April 27 and May 20.



4. Ninety people were asked which newspaper they read. 45 read The Times. 20 read Colombia Today. 15 read another paper. 10 do not read a paper. Calculate the number of degrees required to represent each answer in a pie chart. Then draw the pie chart.