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 and Variables

A variable is a characteristc of a population which can take different values. The population might be a human population, for example, the Colombian population as recorded in the 2005 census; variables of interest in this population might be age, birthrates, mortality rates, income, migration rates, number of children and education level. One could envisage a population of teaching hospitals and perhaps be interested in such variables as annual budget, number of medical students and number of X-rays taken per year. The Social Security Minister might be interested in a population of small chemical processing plants and wish to measure variables related to the health of employees - for example, fresh air recirculation times, or number of emergency dousing showers per factory. Whatever the specifics, a population embraces all the elements that we might wish to measure - even if actually doing so might be logistically impossible.

The variables we will deal with in statistics are called random variables because the values **they** take are generated by chance events. Sometimes the randomness is obvious, as when the data are generated by rolling a dice and observing the variable we call "face value" which can take values 1, 2, 3. 4, 5 or 6. **But** as long as there is a random component to the measurement we can make use of the variable in a statistical sense. **So** a researcher might be interested in the rise in blood pressure (the variable) in patients who are fed a controlled high salt diet. The blood pressure response in patients can be thought of as the result of two components: a constant non-random response in all patients due to the preselected salt dosage, and a random unpredictable response componenf due to each individual's physiological makeup, personality and perhaps scores of other unknown factors.

Data are measurements collected on a variable as a result of taking observatíons. Often, data will have associated units of measurement, for example, data collected by observing a patient's blood pressure will have units of millimetres of mercury (mmHg). Most often, due to time and resource constraints, we are dealing with data collected on a subset or sample of a population. A counter-example to **this** is the census carried out by the Colombian Bureau of Statistics (DANE) in 2005, that aimed to collect information on the entire population of Colombia.

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

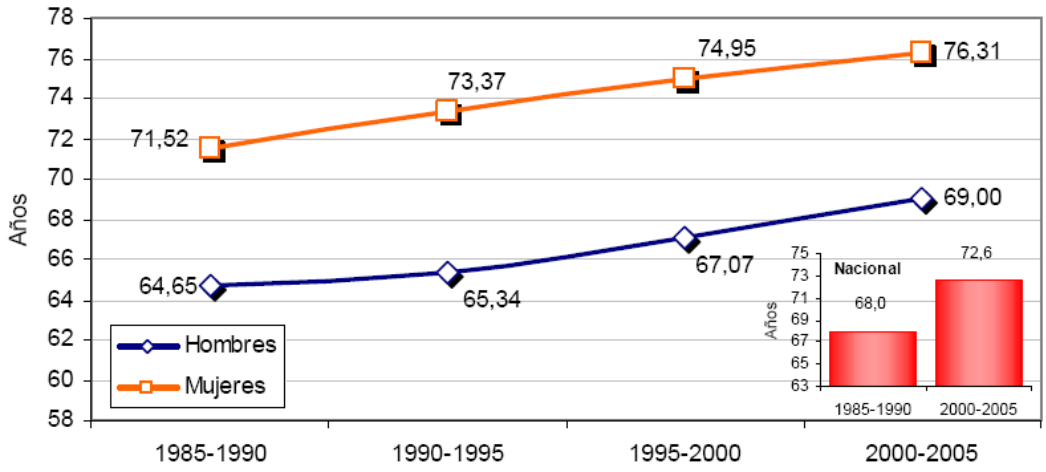
different aspect of the data, are easier to understand than a single large table that contains many details or variables. A table should be self-explanatory. **If** a table is taken out of its original context, **it** should still convey all the information necessary for the reader to understand the 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

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.

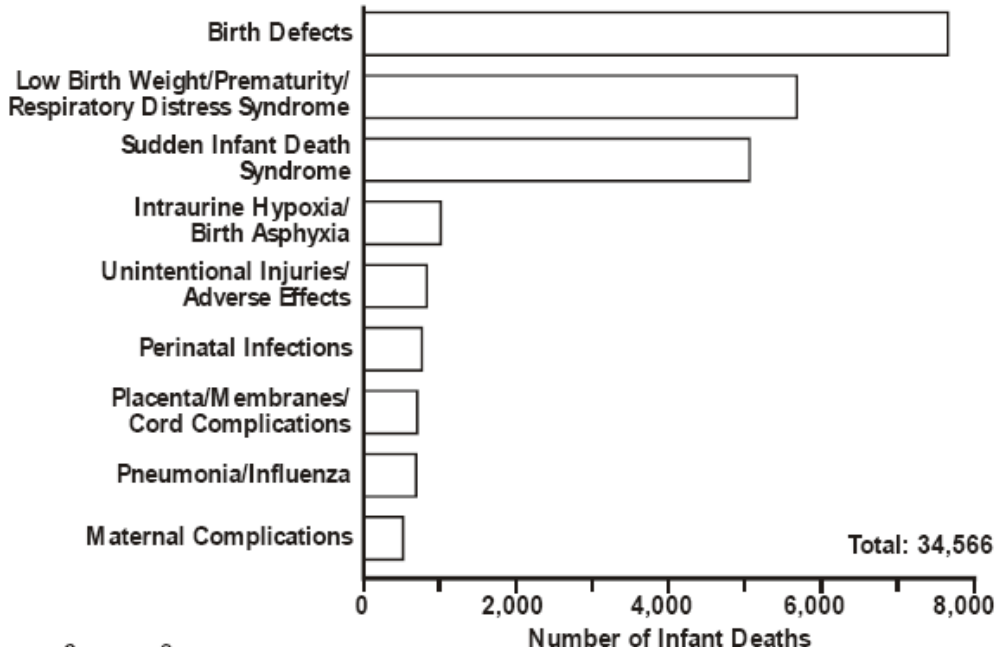


Life expectancy for men and women in Colombia

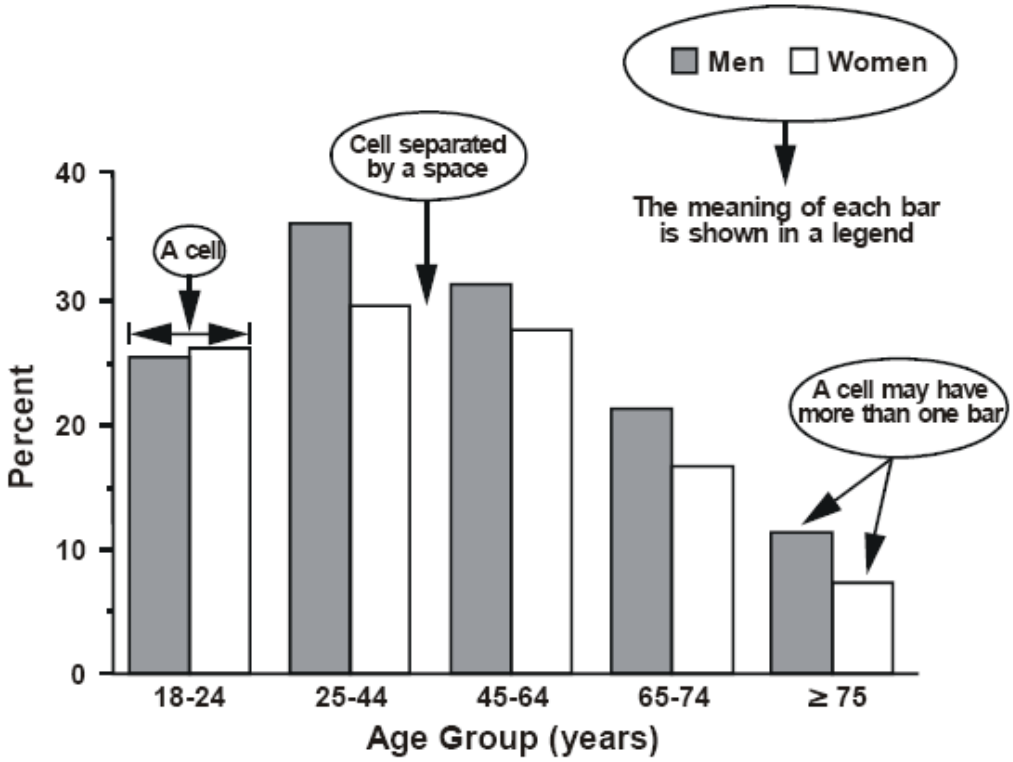
Bar Charts

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 horizontal bar chart:
Number of infant deaths by leading causes, United States, 1983



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



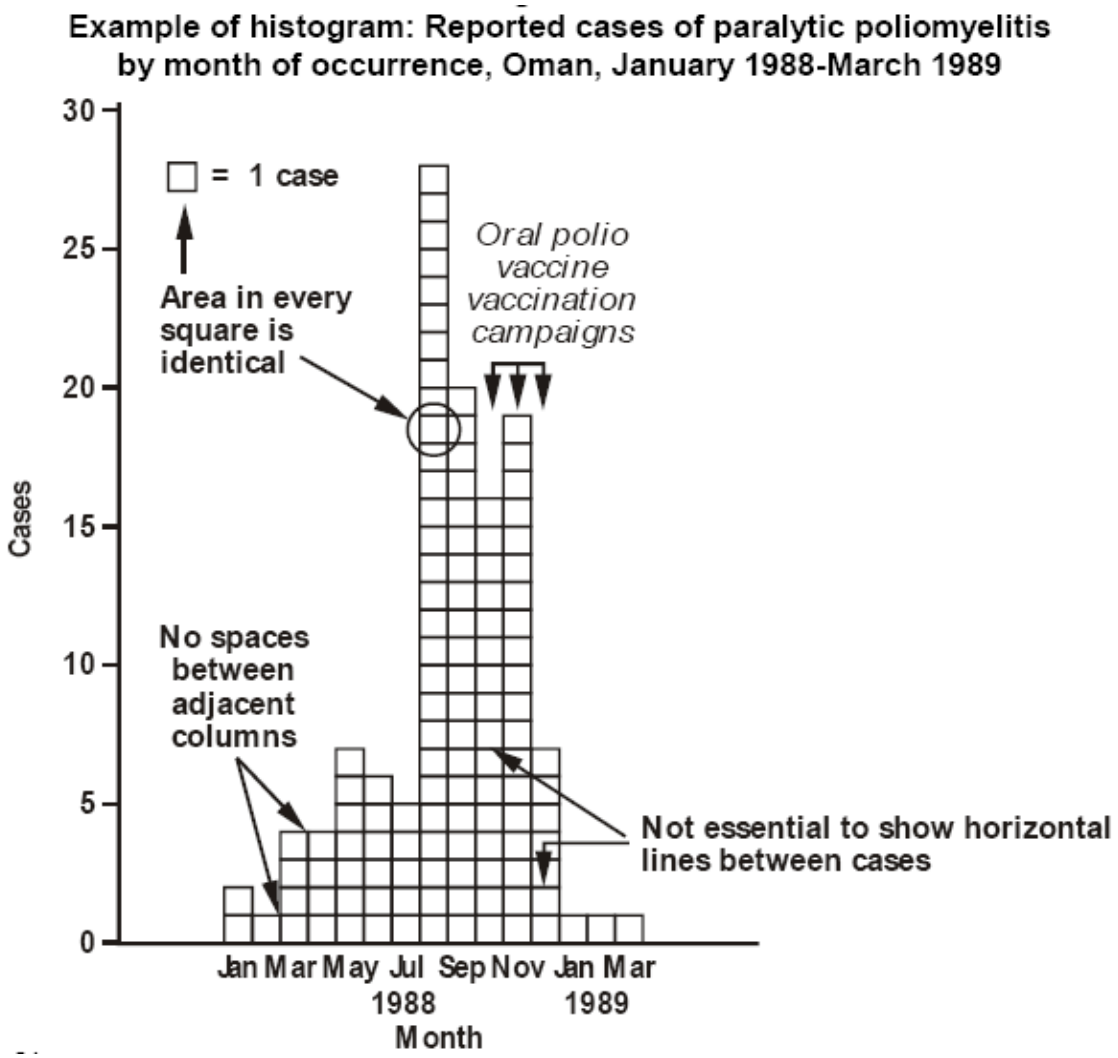
The length or height of each bar is proportional to the frequency of the event in that category. For **this reason**, a scale break should not be used with a bar chart since this could lead to misinterpretation in comparing the magnitude of different categories.

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

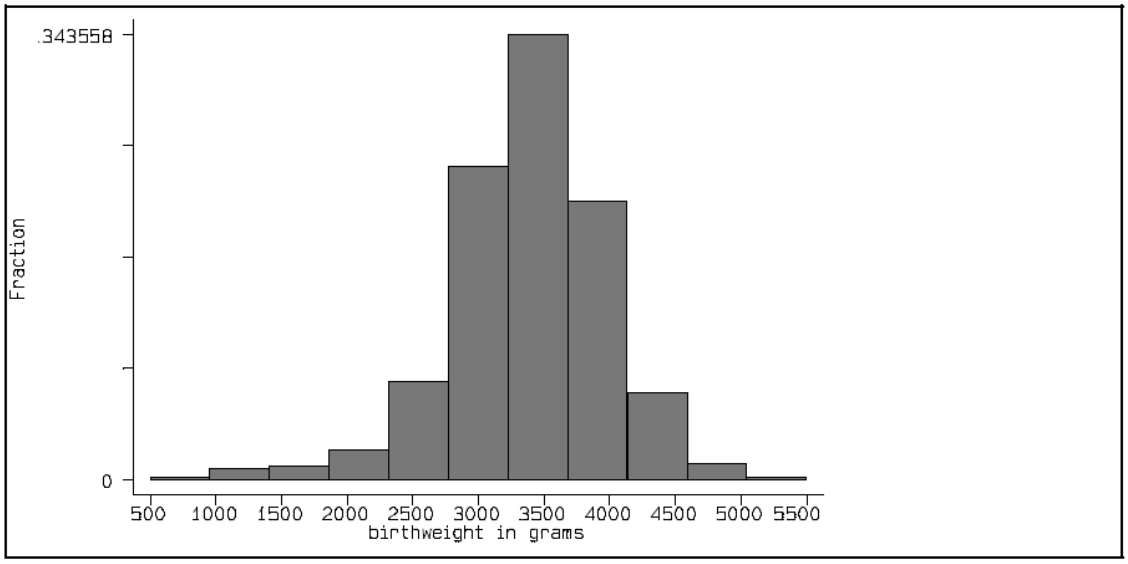
A histogram is a graph of the frequency distribution of a continuous variable. It uses adjoining columns to represent the number of observations for each class interval in the distribution. The area of each column is proportional to the number of observations in that interval.



A histogram looks a bit like a bar chart, but:

- Histograms are informative ways of presenting relative frequencies (how "dense" are the data in each particular sub-interval); bar graphs are used for frequencies and misused for relative frequencies.
- Histograms give you an idea of the shape of the relative frequency distribution. Bar charts are just ***tallies*** and can't tell you about distribution shapes (a ***tally*** is a record or account of items such as things bought or points scored).

The figure below is a histogram representing the relative frequencies (and shape) of a distribution of birthweight from a cohort of 489 infants born in a hospital during 1988. Note the negative skew of the distribution (long tail on the left) due to several atypical low birthweights.



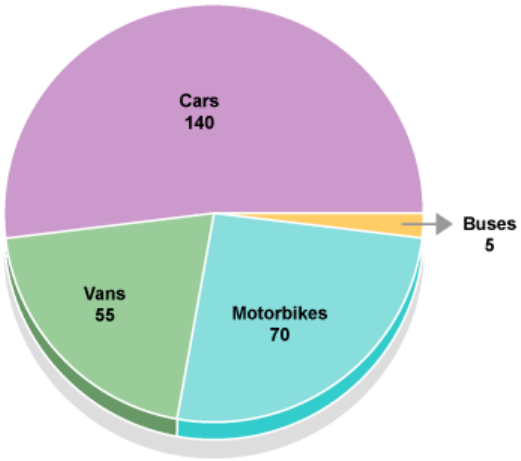
Pie Charts

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:



Frequency Distributions and Data Presentation

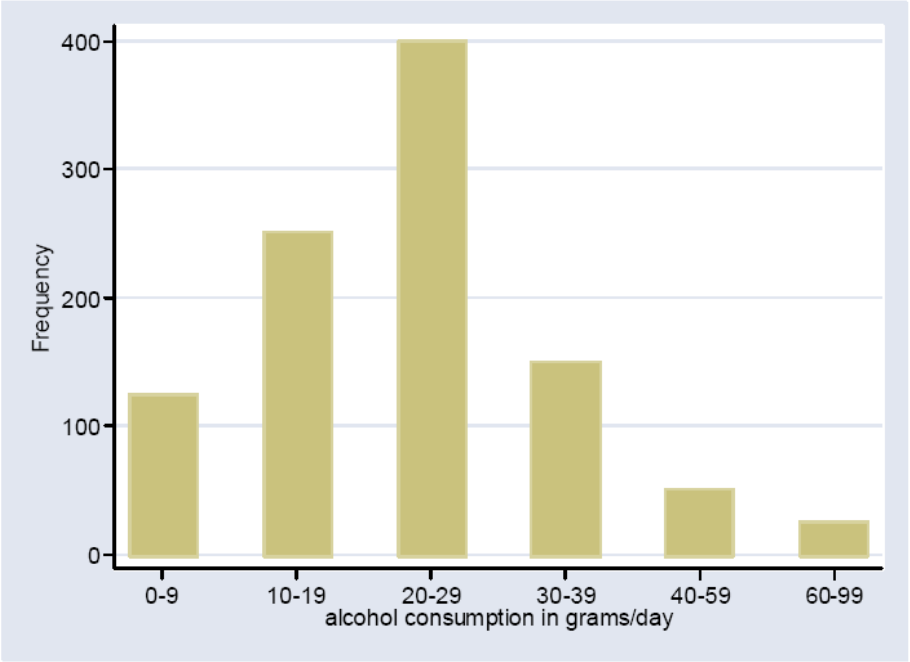
If you arrange your raw data so that the scores on a variable of ínterest are in order of magnitude, that is, you rank the data, and then indícate by means of a table or graph how often a score occurs, **then** you will have constructed a frequency distribution — a tally of the scores.

Example 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.

You might note that the classes chosen in Example 1 don't overlap; the class limits are not shared between two classes as they would be if, for exarnple, we erroneously defined the second and third classes as 10-20 and 20-30 g/day. In this latter case, **if** a person drank 20 g/day, how would we know to which class he should be assigned? **Such problems** are avoided by ensuring that classes are mutually exclusive. The limits of the intervals, e.g. 40 and 59 in the fifth interval, are called the observed class límits, **Since** we are presumably prepared to measure the consumption to the nearest gram, the true class limits are, for this interval, 39.5g to 59.5g, and similarly for the other intervals.
You might present the distribution of the observed drinking scores as a table:
For example, 250 men were recorded as consuming at least 10, but not more than 19, grams of alcohol per day. Note that the sum of the frequencies adds to the total number in the sample (that is, 1000). The first two columns are all you need to present the frequency distribution.

TABLE 3 ***Frequency and relative frequency distributions of alcohol use in 1000 Colombian males***

grams/day	frequency	relative frequency
0-9	125	125/1000 = 0.125
10-19	250	250/1000 = 0.250
20-29	400	400/1000 =0.400
30-39	150	150/1000 = 0.150
40-59	50	50/1000 = 0.050
60-99	25	25/1000 = 0,025



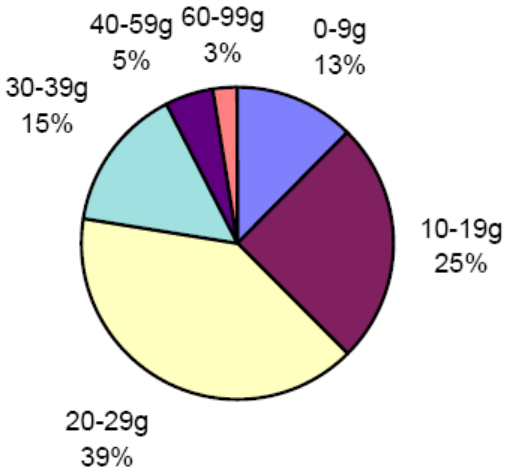
Alcohol consumption in 1000 Colombian males

The same information can be presented by means of a bar graph (see above). The height of each bar is proportional to the frequency of occurrence of its respective score.

Relative Frequencies

The third column of Table 3 defines a related distribution –the distribution or relative frequences. Note that the fractions add to 1 (the percentage would add to 100%). A pie graph might do the same job (see below): each slice of the pie would have an area proportional to the relative frequency of the score it was representing. **For example**, a consumption score of 20-29 grams per day would account for two-fifths (40%) of the pie.

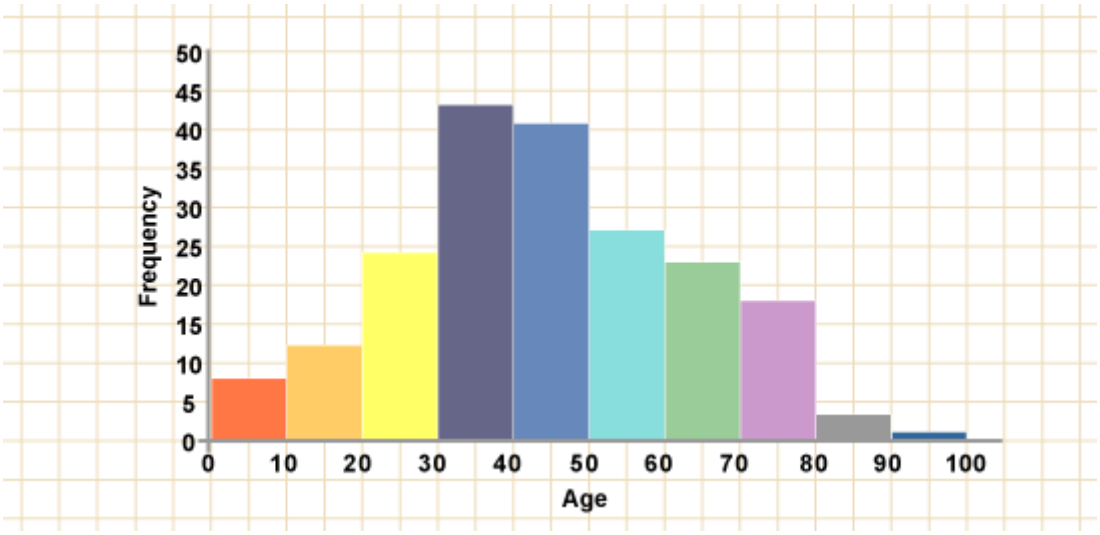
Beware pie charts that show slices of acute angles to the plane of the page, or "exploded" pie charts (slices cut away from the rest of the pie), or those giving a 3-dimensional effect. Such charts can deceive the eye and lead to misinterpretation. Usually the simply the chart, the better.



Pie graph ***Daily consumption of alcohol in 1000 Colombian males***

Activity

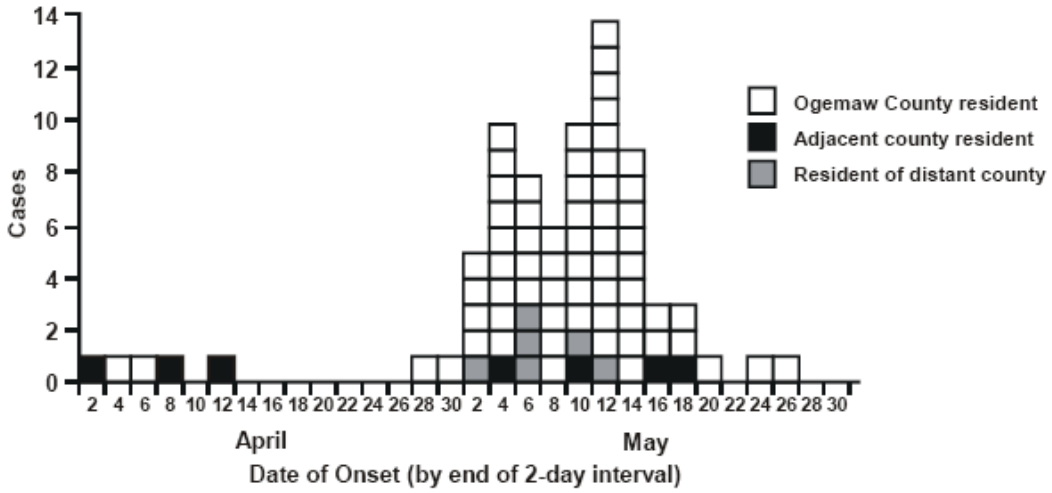
1. From the figure below, draw a table representing the ages of 200 people entering a mall on a Saturday afternoon. The ages have been grouped into the classes 0-9, 10-19, 20-29, and so on. Is the figure a bar chart or a histogram? Explain



2. The table below contains a set of data about birthrates in Colombia. Using this information, draw a bar chart or a histogram (you must decide).

Year	1985-1990	1990-1995	1995-2000	2000-2005
Birthrates	3.34	3.14	2.86	2.60

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.