Descriptive Statistics

Monica Franzese and Antonella Iuliano, Institute for Applied Mathematics "Mauro Picone", Napoli, Italy

Introduction

Statistics is a mathematical science for collecting, analyzing, interpreting and drawing conclusions from a set of data. The first instance of descriptive statistics was given by the "bills of mortality" collected by John Graunt's in 1662. Around 1749, the meaning of statistics was limited to information about "states" to design the systematic collection of demographic and economic data. In the early 19th century, accumulation of information intensified, and the definition of statistics extended to include disciplines concerned with the biology and the biomedicine. The main important areas of statistics are the descriptive statistics and the inferential statistics. The first statistical method gives numerical and graphic procedures to summarise a collection of data in a clear and understandable way without assuming any underlying structure for such data, the second one provides procedures to draw inferences about a population on the basis of observations obtained from samples. Therefore, the use of descriptive and inferential methods enables researchers to summarize findings and conduct hypothesis testing. The descriptive statistics is the primary step in any applied scientific investigation to simplify large amounts of data in a sensible way. Indeed, the goal of descriptive statistics is to give a clear explanation and interpretation of the information collected during an experiment. For instance, in medicine and biology, the observations obtained by a phenomenon are large in number, dispersed, variable and heterogeneous preventing the researcher from directly understanding it. To have a full knowledge of the under investigated phenomenon, it is first necessary to arrange, describe, summarize and visualize the collected data (Spriestersbach *et al.*, 2009).

In this article, we present two statistical descriptive methods: graphical and numerical. The graphs and tables are used to organize and visualize the collected data. Numerical values are computed to summarize the data. Such numbers are called parameters if they describe population; they are called statistics if they describe a sample. The most useful numerical value or statistics for describing a set of observations are the measures of location, dispersion and symmetry. Generally, graphical methods are better suited than numerical methods for identifying patterns in the data, although the numerical approaches are more precise and objective. Since the numerical and graphical approaches complement each other, it is wise to use both. In the following sections, we first introduce the statistical data types (quantitative or qualitative), and then, the way to organize and visualize collected data using tables and graphs. Several kinds of statistics measures (location, dispersion and symmetry) are also discussed to provide a numerical summary of data (Manikandan, 2011a,b,c; Wilcox and Keselman, 2003). Finally, clinical data are elaborated and discussed as an illustrative example.

Statistical Data Types

The goal of descriptive statistics is to gain understanding from data. Population and sample are two basic concepts of statistics. Population can be defined as the set of individuals or objects in a statistical study. While, a sample is a subset of the population from which information is collected. In other words, a statistical population is the set of measurements corresponding to the entire collection of units for which inferences are to be perform, a statistical sample is the set of measurements that are collected in the course of an investigation from the statistical population. Each measurement is defined as statistical unit. This denomination inherited from demography that was the first application field of statistics. A statistical variable is each aspect or characteristic of the statistical unit and it varies from one individual member of the population to another. A statistical variable can be qualitative or quantitative, depending on whether their nature is countable or not. Examples of variables for humans are height, weight, sex, status, and eye color. The first two variables are quantitative variables, the last three are qualitative variables. Quantitative variables can be classified as either discrete or continuous, while qualitative variables can be divided into categorical and ordinal. We define a discrete variable as a finite or countable number of values, while, a continuous variable as a measurement that can take any value in an interval of the real line.

Organization and Visualization of Data

We define each individual or object of data as observation. The collection of all observations for particular variables is called data set or data matrix. Data set is composed by the values of variables recorded for a set of sampling units. Note that in the case of qualitative variable, we assign numbers to the different categories, and thus transform the categorical data to numerical data in a trivial sense. For example, cancer grade can be coded using the values 1,2 and 3 depending on the amount of abnormality (see **Table 1**).

Frequency Distribution

Let N be the number of individuals in the population and let X be a variable assuming the values x_i, $i=1, 2,\ldots, k$. We denote with n_i the number of times the value x_i appears in the data set. This value is called absolute frequency of the observed value

 doi:10.1016/B978-0-12-809633-8.20354-3

Table 1 Tumor data table, with different classification grades

Unit	*Gender*	*Age*	*Tumor*	*Grade*	*Year first diagnosis*
1	M	34	Prostate cancer	II	2000
2	M	45	Brain cancer	II	2003
3	F	50	Breast cancer	III	2005
...	...	...	...	...	...
90	F	62	Ovarian cancer	I	2001

Table 2 Frequency distribution

Statistical units i	*Values of characteristics* x_i	*Absolute frequency* n_i	*Relative frequency* f_i	*Cumulative absolute frequency* N_i	*Cumulative relative frequency* F_i
1	x_1	n_1	$f_1 = \frac{n_1}{N}$	$n_1 = N_1$	$f_1 = F_1$
2	x_2	n_2	$f_2 = \frac{n_2}{N}$	$n_1 + n_2 = N_2$	$f_1 + f_2 = F_2$
...	...	...	...	...	...
i	x_i	n_i	$f_i = \frac{n_i}{N}$	$n_1 + n_2 + \ldots + n_i = N_i$	$f_1 + f_2 + \ldots + f_i = F_i$
...	...	...	...	...	
k	x_k	n_k	$f_k = \frac{n_k}{N}$	$n_1 + n_2 + \ldots + n_k = \mathrm{N}$	$f_1 + f_2 + \ldots + f_k = 1$
Total		N	1		

Table 3 Frequency distribution of tumor grades for 60 patients

Tumor	n_i	f_i
I	36	0,6
II	9	0,15
III	15	0,25
Total	60	1,00

with x_i, and the ratio

$$f_i = \frac{n_i}{N} \tag{1}$$

indicates the relative frequency of the observed value with n_i. In other words, f_i is the proportion on the total population N of individuals presenting the value with x_i. In particular, Eq. (1) satisfies the following conditions:

$$\sum_{i=1}^{k} n_i = N, \ \sum_{i=1}^{k} f_i = 1$$

The observed values of each variable, their absolute and relative frequencies are usually organized in a table called frequency distribution (see **Table 2**). Sometimes, we are often interested in the percentage frequency that is obtained by dividing the relative frequency by the total number of observations N and multiplying the result by 100. The sum of absolute (or relative) frequencies of all the values equal to or less than the considered value is called cumulative absolute (or relative) frequency. This is represented as N_i (or F_i). If cumulative frequencies are represented in a table then it is called as cumulative frequency distribution (see **Table 2**). The frequency distribution is useful when data sets are large and the number of different values is not too large. For example, if we consider the brain tumor stage in a sample of 60 patients, then we can use the frequency distribution to compute the absolute and relative frequency (see **Table 3**).

Qualitative variable

The number of observations that fall into a particular class (or category) of a qualitative variable indicates the frequency (or count) of the class. In this case, cumulative frequencies make sense only for ordinal variables, not for nominal variables. The qualitative data can be represented graphically either as a pie chart or as a horizontal or vertical bar graph. A pie chart is a disk divided into pie-shaped pieces proportional to the relative frequencies of the classes. To obtain angle for any class, we multiply the relative frequencies by 360°, which corresponds to the complete circle. A horizontal bar graph displays the classes on the horizontal axis and the absolute frequencies (or relative frequencies) of the classes on the vertical axis. The absolute frequency (or relative frequency) of each class is represented by vertical bar whose height is equal to the absolute frequency (or relative frequency) of the

Table 4 Frequency distribution based on class intervals. The symbol -| means that only the superior limit is included into the class interval

Class intervals $x_i \dashv x_{i+1}$	*Absolute frequency* n_i	*Relative frequency* f_i	*Cumulative absolute frequency* N_i	*Cumulative relative frequency* F_i
$x_1 \dashv x_2$	n_1	$f_1 = \frac{n_1}{N}$	$n_1 = N_1$	$f_1 = F_1$
$x_2 \dashv x_3$	n_2	$f_2 = \frac{n_2}{N}$	$n_1 + n_2 = N_2$	$f_1 + f_2 = F_2$
...	...	...	...	...
$x_{i-1} \dashv x_i$	n_i	$f_i = \frac{n_i}{N}$	$n_1 + n_2 + ... + n_i = N_i$	$f_1 + f_2 + ... + f_i = F_i$
...	...	...	...	
$x_{k-1} \dashv x_k$	n_k	$f_k = \frac{n_k}{N}$	$n_1 + n_2 + ... + n_k = \mathrm{N}$	$f_1 + f_2 + ... + f_k = 1$
Total	N	1		

class. In a bar plot, its bars do not touch each other. In a vertical bar graph, the classes are displayed on the vertical axis and the frequencies (absolute and relative) of the classes on the horizontal axis. Nominal data is best displayed by pie chart and ordinal data by horizontal or vertical bar graph.

Quantitative variable

Quantitative variable can also presented by a frequency distribution. The data of a discrete variable can be summarised in a frequency distribution as in **Table 2**. Differently, continuous data are first grouped into classes (or categories) and, then collected into a frequency distribution (see **Table 4**). The main steps in a process of grouping quantitative variable into classes are:

1. Find the minimum x_{min} and the maximum x_{max} values into the data set.
2. Construct intervals of equal length that cover the range between x_{min} and x_{max} without overlapping. These intervals are called class intervals, and their end points are defined class limits. The magnitude of class interval depends on the range and the number of classes. The range is the difference between x_{max} and x_{min}. A class interval is generally in multiples of 5, 10, 15 and 20. The magnitude of class is given by

$$d = \frac{x_{max} - x_{min}}{K}$$

   where K is the number of classes equal to $K = 1 + 3.322 \log_{10}(k)$.
3. Count the number of observations in the data that belongs to each class interval. The count in each class is the absolute class frequency.
4. Calculate the relative frequencies of each class by dividing the absolute class frequency by the total number of observations into the data.

In this case, the information contained in **Table 4** can be illustrated graphically using histograms (or vertical bar graph) and cumulative frequency curves. A histogram is a graphical representation of the absolute or relative frequencies for each value of the variables. It is like a horizontal bar graph where the bars are closed each other. In particular, a histogram is composed by rectangles over each class where the area of each rectangle is proportional to its frequency. The base of rectangles are the range of each class, i.e. the difference between x_{i+1} and x_i, for $i = 1,\ldots,k$, while the height of the rectangle, called class intensity, is given by

$$h_i = \frac{n_i}{x_{i+1} - x_i}, i = 1, \ldots, k$$

A cumulative frequency curve is a plot of the number or percentage of individuals falling in or below each value of the characteristic. If quantitative data is discrete, then the variable should graphically be presented by a bar graph. Also in the case in which the frequency table for quantitative variable is composed by unequal class intervals, the variable can be represented by a bar graph.

Statistical Measures

In this section we present three types of statistical measures that describe and summarise the observed data: the measures of central tendency, the measures of dispersion, and the measures of symmetry. They are called statistics, i.e., functions or modifications of the obtained data. These descriptive measures are a direct consequence of the frequency distribution of the data. In fact, they provide a numerical summary of the frequency distribution.

Measures of Central Tendency

The central tendency of a frequency distribution is a statistical measure that identifies a single value as representative of an entire distribution (Manikandan, 2011a). It aims to provide an accurate description of the data and it is a numerical value that is most representative of the collected data. The mean, median, quartiles and mode are the commonly used measures of central tendency. The mean is the most used measure of central tendency. The (arithmetic) mean is the sum of all the values $x_1, x_2,\ldots,x_k$ in the data

set divided by the number of observations k ($k \leq N$). Denoting the sample mean by $\bar{x}$, it is given by the formula

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i \tag{2}$$

Sometimes the prefix "sample" is dropped, but it is used to avoid the confusion of $\bar{x}$ with the population mean μ on the entire population. In this case, the mean is computed by adding all the values in the data set divided by the number of observations N. The formula is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3}$$

In terms of frequency distribution, the sample mean is given by

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i n_i \tag{4}$$

where n_i is the absolute frequency. In the case of class frequency distribution, we first calculate the central value $c_i = \frac{x_{i+1}+x_i}{2}$ and, then, the sample mean $\bar{x}$ with the following formulas

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} c_i n_i \tag{5}$$

The median is a measure of location that separates the set of values in half, so that the data in one half are less than or equal to the median value and the data in the other half are greater or equal to the median value (Manikandan, 2011b). It divides the frequency distribution exactly into two halves. Fifty percent of observations in a distribution have scores at or below the median. Hence, median is the 50th percentile. To compute the median, we first collect the data in increasing order and then determine the middle value in the ordered list. In particular, we can distinguish two cases:

1. If the number of observations is odd, then the median is the value that occupies the middle position into the data. If we let k denote the number of observations, then the sample median is at position $\frac{k+1}{2}$ into the ordered list of data.
2. If the number of observations is even, then the median is the mean of the two middle observations into the ordered data, i.e., the mean of the values at position $\frac{k}{2}$ and $\frac{k}{2}+1$ into the ordered data.

In the case of (discrete) frequency distribution, we first compute the value $\frac{k}{2}$ and, then we consider the absolute cumulative frequency that is greater than $\frac{k}{2}$, i.e. $N_i \geq \frac{k}{2}$, for $i=1,2,\ldots,k$. The corresponding value x_i is the median. In terms of grouped value, we first identify the median class as the class that includes the 50% of cumulative relative frequencies, i.e., $F_i \geq 0.50$, for $i=1,2,\ldots,k$. Then, under the assumption that the frequencies are uniformly distributed, we compute the median M_e as the following approximation

$$M_e \approx x_i + (x_{i+1} - x_i)\frac{0.5 - F_{i-1}}{F_i - F_{i-1}} \tag{6}$$

The quartiles (or percentiles) are location measures that divide the data set into four equal parts, each quartile contains the 25% of the total observations. The first quartile Q_1 (lower quartile) is the number below which lies the 25% of the bottom data. Q_1 is at position $\frac{k+1}{4}$ into the ordered sample data. The second quartile Q_2 is the median of the data. The data are divided into two equal parts, the bottom 50% and the top 50% and it is at position $\frac{k+1}{2}$ into the ordered sample data. The third quartile Q_3 (upper quartile) is the number above which lies the 75% of the top data. Q_3 is at position $\frac{3(k+1)}{4}$ into the ordered sample data. In the case of class frequency, the first quartile class Q_1 is the class that includes the 25% of cumulative relative frequencies, i.e. $F_i \geq 0.25$, for $i=1,2,\ldots,k$, while the third quartile class Q_3 is the class that includes the 75% of cumulative relative frequencies, i.e., $F_i \geq 0.75$, for $i=1,2,\ldots,k$. Under the assumption that the frequencies are uniformly distributed, the first quartile class Q_1 and the third quartile class Q_3 are approximately equal to

$$Q_1 \approx x_i + (x_{i+1} - x_i)\frac{0.25 - F_{i-1}}{F_i - F_{i-1}}, Q_3 \approx x_i + (x_{i+1} - x_i)\frac{0.75 - F_{i-1}}{F_i - F_{i-1}} \tag{7}$$

Generally, a frequency distribution can be represented constructing a graph called boxplot (or whisker diagram) which is a standardized way of displaying the data distribution based on five number summary: x_{min}, Q_1, $Q_2=M_e$, Q_3, x_{max}. The central rectangle spans the first quartile to the third quartile (the interquartile range). The segment inside the rectangle shows the median and "whiskers" above and below the box show the positions of the minimum and maximum (see **Fig. 1**).

The mode (or modal value) of a variable is the value that occurs most frequently in the data (Manikandan, 2011b). It is given by

$$M_o = \{x_i : n_i = max\}, i = 1, 2, \ldots, k \tag{8}$$

The mode may not exist, and even if it does, it may not be unique. This happens when the data set has two or more values of equal frequency that is greater than the other values. The mode is usually used to describe a bimodal distribution. In a bimodal distribution, the taller peak is called the major mode and the shorter one is the minor mode. For continuous data, the mode is the midpoint of the interval with the highest rectangle in the histogram. If the data are grouped into class intervals, than the mode is defined in terms of

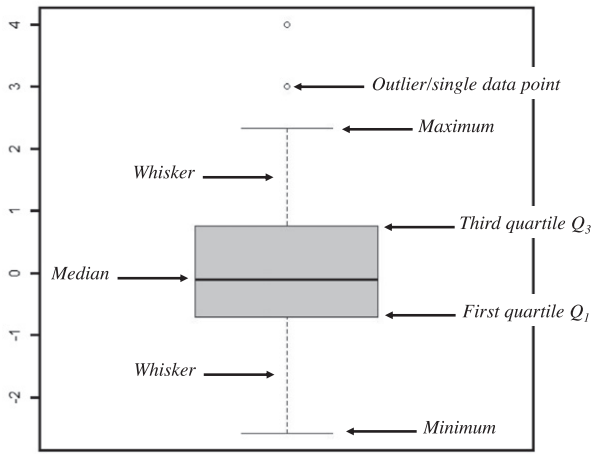

Fig. 1 Boxplot or box and whisker diagram.

class frequencies. With grouped quantitative variable, the mode class is the class interval with highest frequency and it is given by

$$m_c = \{x_i - |x_{i+1} : h_i = max\} \tag{9}$$

where $h_i = \frac{n_i}{x_{i+1} - x_i}$ is the class intensity. Under the assumption that the frequencies are uniformly distributed, the mode is approximately equal to

$$M_o \approx \frac{x_i + x_{i+1}}{2} \tag{10}$$

We observe that median and mode are not influenced by extreme values or outliers. On the contrary, the mean suffers of them. An outlier is an observation that lies an abnormal distance from other values in a sample from a population. It is a value that exceeds the third quartile by a magnitude greater than $1.5 \times (Q_3 - Q_1)$ or is less than the first quartile by a magnitude greater than $1.5 \times (Q_3 - Q_1)$. In addition, for qualitative and categorical data, the mode can be calculated, while the mean and median do not. On the other hand, if the data is quantitative one, we can use any one of the three averages presented. For symmetric data the mean, the median and the mode can be approximately equal; for skew (or asymmetric) data the median is less sensitive than the mean to extreme observations (outliers). As the mean, the mode and the median have the corresponding population median and population mode, which are all unknown. In fact, the sample mean, the sample median, and the sample mode can be used to estimate the values of these corresponding unknown population values.

Measures of Dispersion

The measures of central tendency are representatives of a frequency distribution but they are not sufficient to give a complete representation of a frequency distribution (Manikandan, 2011c). Two data sets can have the same mean but they can be entirely different. Thus, to describe data, one needs to know also the extent of variability. This is given by the measures of dispersion. A measure of dispersion (called also variability, scatter, or spread) is a statistics that indicates the degree of variability of data. Range, interquartile range, variance, standard deviation and coefficient of variation are the commonly used measures of dispersion. The (sample) range is obtained by computing the difference between the largest observed value x_{max} of the variable in a data set and the smallest one x_{min}. This measures is easy to compute even if a great deal of information is ignored. In fact, only the largest and smallest values of the variable are considered while the other observed values are disregarded. In addition, the range is always increase, when additional observations are included into the data set which means that the range is overly sensitive to the sample size. The (sample) interquartile range (IQR) is equal to the difference between 75th and 25th percentiles, i.e. the upper and lower quartiles

$$\text{IQR} = \text{Q3} - \text{Q1} \tag{11}$$

The (sample) interquartile range represents the length of the interval covered by the center half of the observed values of the data. This measure is not distorted if a small fraction of the observed values are very large or very small.

The variance and the standard deviation are two very popular measures of dispersion. They measure the spread of data across the mean. The population variance σ^2 is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \tag{12}$$

where x_i is the value of the ith observation, μ is the population mean and N is the population size. The population standard deviation σ is defined as $\sigma = \sqrt{\sigma^2}$. In terms of frequency distribution, the (sample) variance s^2 is given by

$$s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \tag{13}$$

where x_i is the value of the ith observation, $\bar{x}$ is the sample mean and k is the sample size. In the last formula, the sum of the squared deviations from the mean provides a measure of total deviation from the mean of the data. If k is large the difference between the formulas (12) and (13) is minimal; if k is small, the difference is very sensitive. Generally, for calculations an easier formula is used. The equation of this formula is given by

$$s^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 n_i - \bar{x}^2 \tag{14}$$

where $\bar{x}_2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 n_i$ is called second order statistics. This computational formula avoids the rounding errors during calculation. The (sample) standard deviation SD is given by $s = \sqrt{s^2}$. The more variation there is into the data, the larger is the standard deviation. However, the standard deviation does have its drawbacks. For instance, its values can be strongly affected by a few extreme observations. We observe that factor $k-1$ is present in both formulas (13) and (14) instead of k in the denominator, this produces a more accurate estimate of standard deviation.

In the case of class frequency distribution, we first calculate the central value c_i and, then the (sample) variance s^2 given by

$$s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(c_i - \bar{x})^2 n_i \left(or\ s^2 = \frac{1}{k}\sum_{i=1}^{k} c_i^2 n_i - \bar{x}^2 \right) \tag{15}$$

When the two distributions are expressed in the same units and their means are equal or nearly equal, the variability of data can be compared directly by using the relative standard deviations. However, if the means are widely different or if they are expressed in different units of measurement, we cannot use the standard deviations as such for comparing the variability of both data. Therefore, we use as measure of dispersion the coefficient of variation (CV) that is the ratio of the standard deviation to the mean. The population CV is given by

$$CV = \frac{s}{|\bar{x}|} \tag{16}$$

Formula (16) for population is the ratio of the population standard deviation σ and the population mean μ. The CV is a unit-free measure and it is always expressed as percentage. The CV is small if the variation is small and it is unreliable if the mean is near zero. Hence, if we consider two groups, the one with less CV is said to be more consistent.

Measures of Symmetry

An important aspect of the data is the shape of its frequency distribution. Generally, we are interested to observe if the frequency distribution can be approximated by the normal distribution ($\mu = M_e$). The normal distribution is a continuous random variable and its density curve is symmetric, bell-shaped curve and characterised by its mean μ and standard deviation σ (**Fig. 2**).

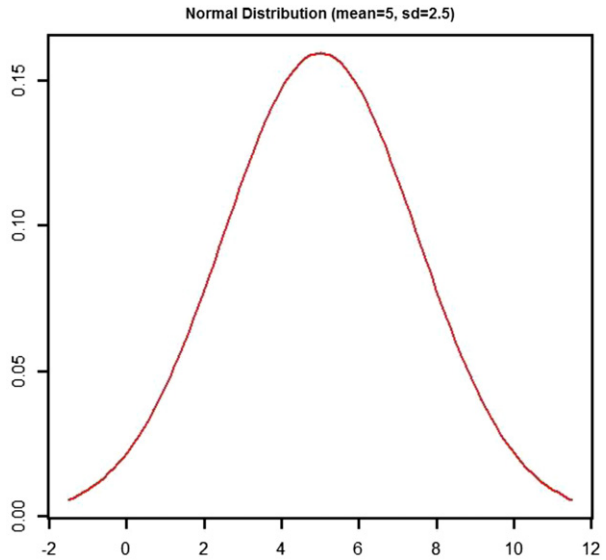

Fig. 2 Normal distribution with mean = 5 and standard deviation = 2.5.

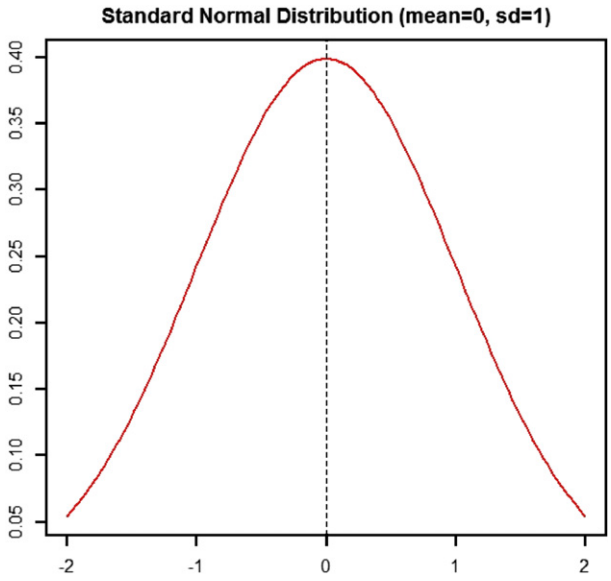

Fig. 3 Standard normal distribution with mean = 0 and standard deviation = 1.

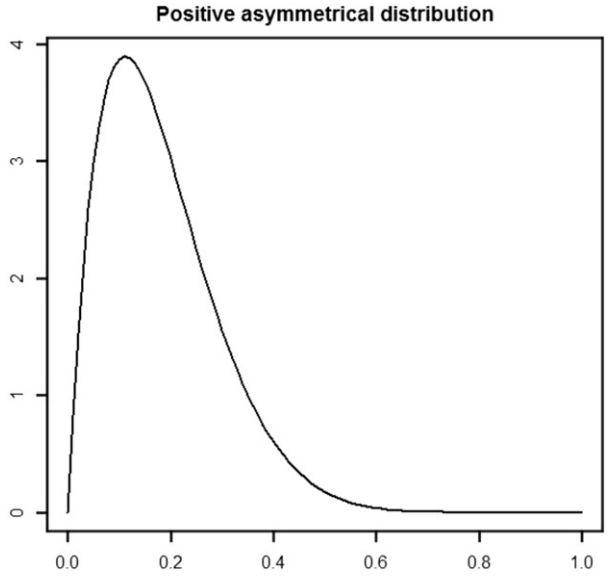

Fig. 4 Positively skewed distribution (to the right).

A continuous random variable follows a standard normal distribution if the variable is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$ (see Fig. 3). Two important measures of shape are skewness and kurtosis. The first measure is the deviation of the distribution from symmetry (departure from horizontal symmetry), the second one measures the peakedness of the distribution (how tall and sharp the central peak is, relative to a standard bell curve). In particular, if the skewness is different from zero, then that distribution is asymmetrical, while normal distributions are perfectly symmetrical. If the kurtosis is not equal to zero, then the distribution is either flatter or more peaked than normal. We define the moment coefficient of skewness of a sample as

$$\gamma_1 = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^3}{\left(\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2\right)^{3/2}} \tag{17}$$

where $\bar{x}$ is the mean and k is the sample size, as usual. The numerator is called the third moment of the variable x. The skewness can also be computed as the average value of z^3, where z is the familiar z-score, i.e. $z = \frac{x-\bar{x}}{\sigma}$. In terms of frequency distribution, the

skewness is

$$\gamma_1 = \frac{\sqrt{k}}{(k-1)\sqrt{k-1}s^3}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i \tag{18}$$

where s is the sample standard deviation and k the sample size. Similarly, for class frequency distribution, the skewness is

$$\gamma_1 = \frac{\sqrt{k}}{(k-1)\sqrt{k-1}s^3}\sum_{i=1}^{k}(c_i - \bar{x})^3 n_i \tag{19}$$

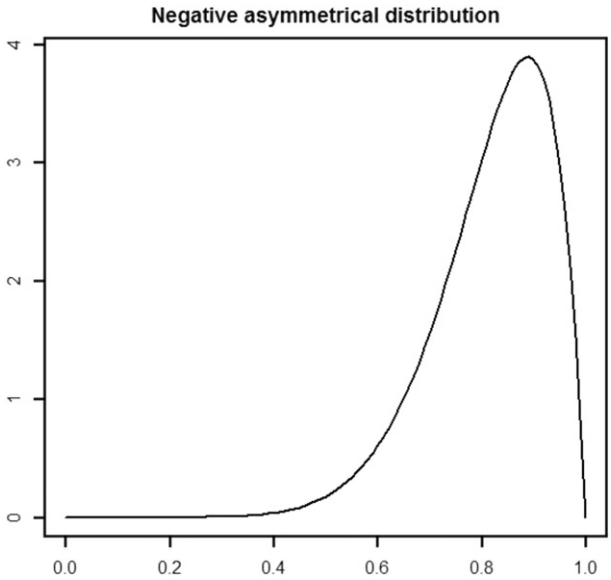

Fig. 5 Negatively skewed distribution (to the left).

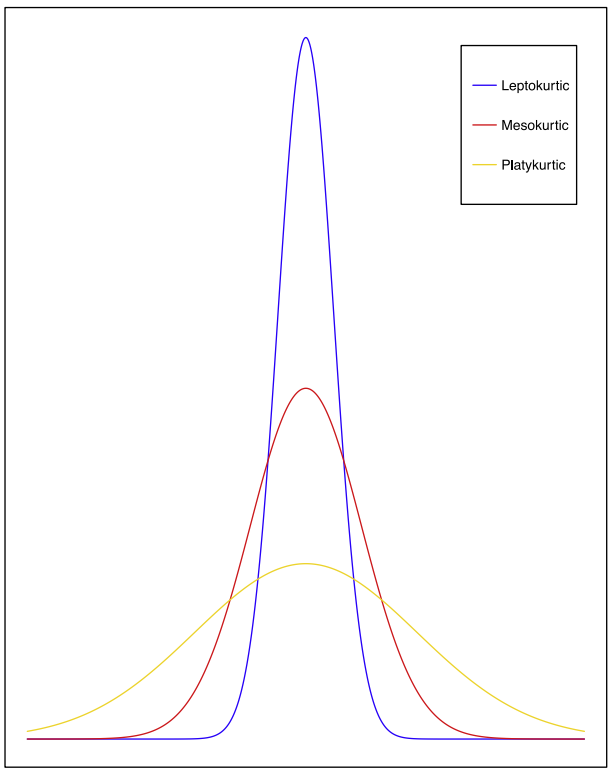

Fig. 6 Kurtosis distributions: a distribution with kurtosis equal to zero is called mesokurtic or mesokurtotic (line red); a distribution with positive kurtosis is called leptokurtic, or leptokurtotic (line blue); a distribution with negative kurtosis is called platykurtic, or platykurtotic (line gold).

Negative values of γ_1 indicate that the data are skewed left, positive values of γ_1 show that the data are skewed right. By skewed left, we mean that the left tail is long relative to the right tail (see **Fig. 4**). Similarly, skewed right means that the right tail is long relative to the left tail (see **Fig. 5**). If the data are multi-modal, then this may affect the sign of the skewness. An alternative formula is the Galton skewness (also known as Bowley's skewness)

$$\gamma_1 = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} \tag{20}$$

where Q_1 is the lower quartile, Q_3 is the upper quartile, and Q_2 is the median.

Table 5 Simulated data. The dataset contains the following variables: Age, Sex (M = male, F = female), Alcohol (1 = yes, 0 = no), Smoke (1 = yes, 0 = no) and Nationality (Italian and Asiatic) of 50 patients

	Age	*Sex*	*Alcohol 1 = yes, 0 = no*	*Smoke 1 = yes, 0 = no*	*Nationality*
Patient 1	50	M	1	0	Italian
Patient 2	55	F	0	1	Italian
Patient 3	25	F	0	1	Asiatic
Patient 4	30	M	1	1	Asiatic
Patient 5	37	M	0	1	Asiatic
Patient 6	52	M	0	1	Italian
Patient 7	42	M	0	1	Italian
Patient 8	60	F	1	0	Asiatic
Patient 9	41	F	1	0	Asiatic
Patient 10	38	F	1	0	Asiatic
Patient 11	62	M	0	0	Asiatic
Patient 12	65	M	0	0	Asiatic
Patient 13	32	M	0	0	Asiatic
Patient 14	33	M	0	1	Italian
Patient 15	50	M	0	1	Italian
Patient 16	44	M	1	1	Italian
Patient 17	45	M	0	0	Italian
Patient 18	50	M	0	1	Asiatic
Patient 19	60	M	0	1	Asiatic
Patient 20	37	M	1	0	Asiatic
Patient 21	56	F	0	0	Italian
Patient 22	57	F	0	0	Italian
Patient 23	48	M	0	1	Asiatic
Patient 24	50	F	1	1	Asiatic
Patient 25	27	M	0	1	Asiatic
Patient 26	43	F	1	0	Italian
Patient 27	53	F	0	0	Italian
Patient 28	30	M	0	0	Asiatic
Patient 29	25	F	0	0	Italian
Patient 30	28	M	1	1	Asiatic
Patient 31	50	M	1	1	Asiatic
Patient 32	69	F	1	1	Asiatic
Patient 33	50	M	1	1	Italian
Patient 34	50	M	1	1	Italian
Patient 35	50	M	1	0	Italian
Patient 36	50	M	1	0	Italian
Patient 37	50	M	0	0	Asiatic
Patient 38	50	F	0	1	Asiatic
Patient 39	50	F	0	1	Italian
Patient 40	39	F	0	1	Italian
Patient 41	50	M	0	1	Italian
Patient 42	46	F	0	1	Italian
Patient 43	50	F	0	1	Asiatic
Patient 44	50	F	0	0	Asiatic
Patient 45	50	M	0	0	Asiatic
Patient 46	67	M	0	0	Italian
Patient 47	70	M	1	0	Asiatic
Patient 48	61	M	0	0	Italian
Patient 49	68	F	1	1	Italian
Patient 50	45	M	0	1	Asiatic

The kurtosis can be explained in terms of the central peak. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak. As skewness involves the third moment of the distribution, kurtosis involves the fourth moment and it is defined as

$$\gamma_2 = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^4}{\left(\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2\right)^2} \tag{21}$$

Similarly, in terms of frequency distribution, the kurtosis is

$$\gamma_2 = \frac{k(k+1)}{(k-1)(k-2)(k-3)s^4}\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i \tag{22}$$

$$\gamma_2 = \frac{k(k+1)}{(k-1)(k-2)(k-3)s^4}\sum_{i=1}^{k}(c_i - \bar{x})^4 n_i \tag{23}$$

Usually, kurtosis is quoted in the form of excess kurtosis (kurtosis relative to normal distribution kurtosis). Excess kurtosis is simply kurtosis less 3. In fact, kurtosis for a standard normal distribution is equal to three. There are three different ways to define the kurtosis. A distribution with excess kurtosis equal to zero (and kurtosis exactly 3) is called mesokurtic, or mesokurtotic. A distribution with positive excess kurtosis (and $\gamma_2 > 3$) is called leptokurtic, or leptokurtotic. A distribution with negative excess kurtosis (and $\gamma_2 < 3$) is called platykurtic, or platykurtotic. For instance, see **Fig. 6**. In terms of shape, a leptokurtic distribution has fatter tails while a platykurtic distribution has thinner tails. Finally, if you have the whole population, then γ_1 and γ_2 are the measure of skewness, formula (17), and kurtosis, formula (21).

Data Analysis and Results

In this section, we describe an illustrative example and the results obtained by using the main descriptive tools and measures. We simulate a clinical random dataset composed by 50 patients and different type of variables (quantitative and qualitative). More precisely, the data matrix, in **Table 5**, shows information about age, sex (M-male; F-female), alcohol (1-yes, 0-no), smoke (1-yes, 0-no) and nationality (Italian, Asiatic) for each patient. Hence, the dataset is composed by four categorical variables (sex, alcohol, smoke, and nationality) and one numerical variable (age). In the first step of our analysis, we plot each variable in order to examine the relative distribution. In particular, we use the pie chart for the qualitative variables alcohol and smoke (see **Fig. 7**). In the graph A, we observe that the percentage of patients who alcoholic drinks is 36%, while it is 64% for patients do not consume alcoholic drinks. In the graph B, we notice that the percentage of patients who smoke is 54%, while it is 46% for patients do not smoke. **Fig. 8** shows the vertical bar plot of the patient's frequencies grouped by nationality, while **Fig. 9** displays the horizontal bar plot of the patient's frequencies grouped by gender. The first plot indicates that the community of Asiatic people is greater than of Italian one, while the second graph shows that the males are more numerous than females. In the second step of the analysis, we analyze the quantitative variable: age. We organize the relative data in class frequency. In particular, we first divide the variable age in 7 classes according to the procedure illustrated in Section Quantitative Variable, then, we compute the relative absolute frequencies. Based on this information, formula (1) is applied to calculate the relative frequencies and percentages for each class. The cumulative relative frequency is also computed. Therefore, the table frequency distribution, **Table 6**, is created. Using the data of this table, the histogram and the cumulative frequency curve of patients grouped by age are plotted (see **Figs. 10** and **11**).

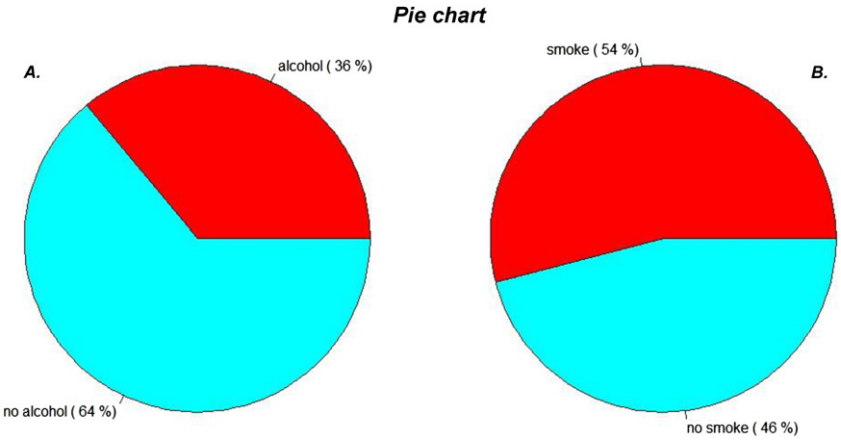

Fig. 7 Pie chart A shows the percentage of *alcohol/no alcohol* patients; pie chart B shows *smoke/no smoke* patients.

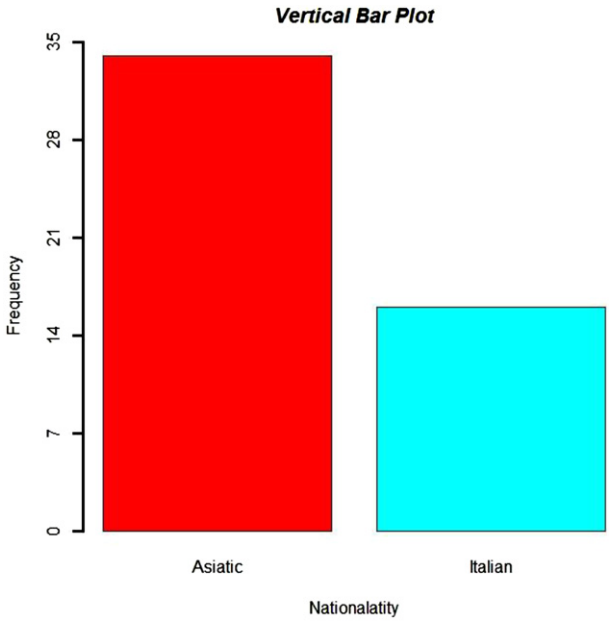

Fig. 8 Vertical bar shows the frequencyof patients grouped by nationality.

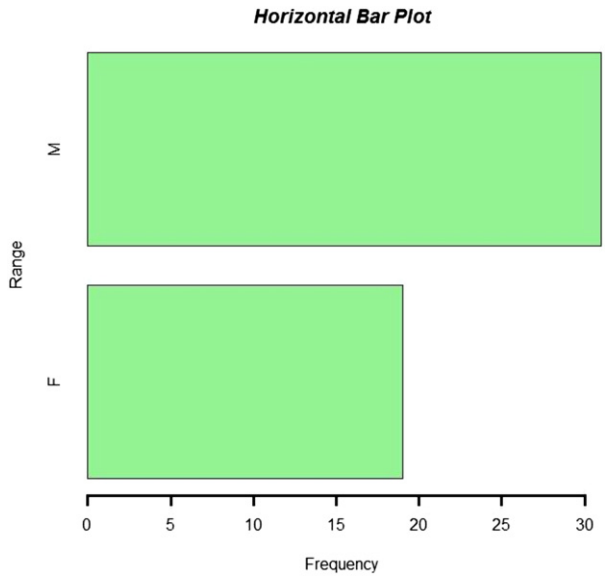

Fig. 9 Horizontal bar shows the frequency of patients grouped by gender.

In addition, the density plot of the variable age grouped in class is shown in **Fig. 12**. We note that the density shows a bimodal distribution with two different modes that are two different peaks (local maxima). Finally, to conclude our descriptive analysis the most important statistical measures are determined. By considering the information collected in **Table 6**, we construct the **Table 7**. In particular, using formula (5), the mean is equal to 47.19, which means that the center of the age distribution is around 47.19 years. The median class of the variable age is the interval (46,53] corresponding to $F_i \geq 0.50$, for $i=1,\ldots,7$. Hence, the median, in according to formula (6), is approximately to 48.21. Using formula (10), the first quartile Q_1 is approximately equal to 39.5($F_i \geq 0.25$, for $i=1,\ldots,7$), while the third quartile Q_3 is approximately equal to 52.3 ($F_i \geq 0.75$, for $i=1,\ldots,7$). The mode class is (46,53]. Hence, by using formula (9), the mode is approximately equal to 49.5. Median class and mode class coincide

Finally, we calculate the measure of dispersion for variable age. Using data collected in **Table 7**, we first compute the range, which is equal to 45. A larger range value indicates a greater spread of the data. Then, by considering formula (11), we calculate the interquartile range IQR, which is equal to 12.8. In addition, the variance and the standard deviation are equal to 132.88 and 11.52, respectively. While the coefficient of variation CV is 0.24. This means that exists a low variability among the data, hence the existing patterns can be seen clearly. To obtain these last results we use formula (15) and (16).

Table 6 Class Frequency Distribution of patients grouped by age. The absolute, relative and cumulative relative frequency are computed

Class intervals	n_i	f_i	$f_i(\%)$	F_i
24 -| 32	7	0.14	14	0.14
32 -| 39	5	0.10	10	0.24
39 -| 46	7	0.14	14	0.38
46 -| 53	19	0.38	38	0.76
53 -| 60	5	0.10	10	0.86
60 -| 67	4	0.08	8	0.94
67 -| 74	3	0.06	6	1
Total	50	1	100	

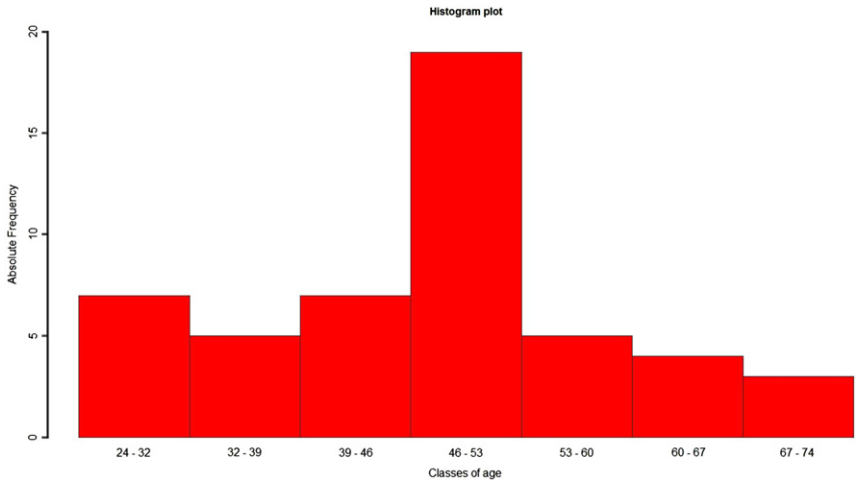

Fig. 10 Histogram shows the absolute frequency of patients grouped for classes of age.

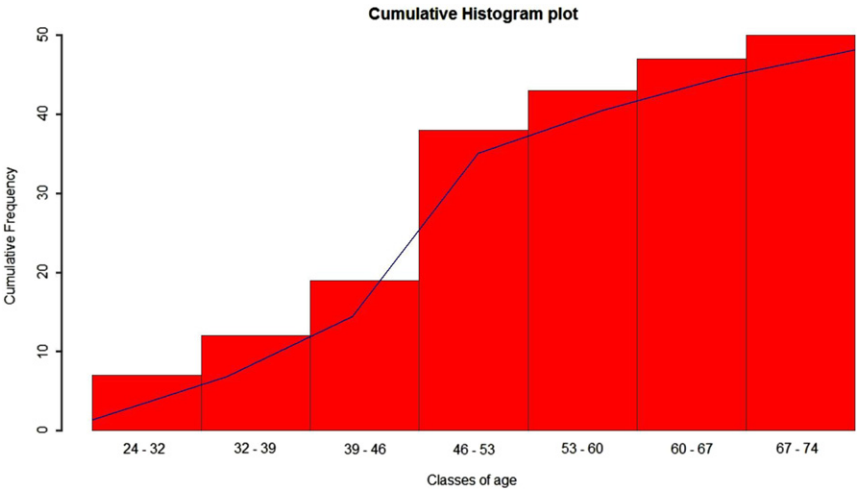

Fig. 11 Cumulative histogram shows the cumulative frequency of patients grouped for classes of age.

Software

We use the R statistical software (see Section Relevant Website) to plot the graphs and to compute the descriptive statistics. In particular, we apply the common used statistical packages in R.

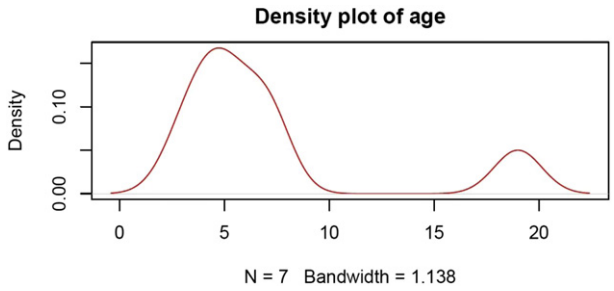

Fig. 12 Density plot of patients grouped by age.

Table 7 Data synthesis of patients grouped by age. The interval (46,53) is the median class. This interval coincides with the mode class

Class intervals	n_i	f_i	$f_i(\%)$	F_i	c_i	$c_i n_i$	$c_i^2\ n_i$
24 -I 32	7	0.14	14	0.14	28	196	5,488
32 -I 39	5	0.10	10	0.24	35.5	177.5	6,301.25
39 -I 46	7	0.14	14	0.38	42.5	297.5	12,643.75
46 -I 53	19	0.38	38	0.76	49.5	940.5	46,554.75
53 -I 60	5	0.10	10	0.86	56.5	282.5	15,961.25
60 -I 67	4	0.08	8	0.94	63.5	254	16,129
67 -I 74	3	0.06	6	1	70.5	211.5	14,910.75
Total	50	1	100			2359.5	117,988.8

Conclusion

The description of data is the first step for the understanding of statistical evaluations. In fact, if the data are of good quality and well presented, we can draw valid and important conclusions. In this work different descriptive statistical procedures are explained. These include the organization of data, the frequency distribution and the graphical presentations of data. The concepts of central tendency, dispersion and symmetry, called summary statistics, are deeply investigated for a complete exploratory data analysis.

See also: Deep Learning. Introduction to Biostatistics. Natural Language Processing Approaches in Bioinformatics

References

Manikandan, S., 2011a. Measures of central tendency: The mean. Journal of Pharmacology and Pharmacotherapeutics 2.2, 140.
Manikandan, S., 2011b. Measures of central tendency: Median and mode. Journal of Pharmacology and Pharmacotherapeutics 2.3, 214.
Manikandan, S., 2011c. Measures of dispersion. Journal of Pharmacology and Pharmacotherapeutics 2 (4), 315–316.
Spriestersbach, Albert, *et al.*, 2009. Descriptive statistics: The specification of statistical measures and their presentation in tables and graphs. Part 7 of a Series on Evaluation of Scientific Publications. Deutsches Ärzteblatt International 106.36, 578–583.
Wilcox, R.R., Keselman, HJ., 2003. Modern robust data analysis methods: Measures of central tendency. Psychological Methods 8 (3), 254.

Further Reading

Box, G.E., Hunter, W.G., Hunter, J.S., 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, vol. 1. New York: Wiley.
Daniel, W.W., Cross, C.L., 2013. Biostatistics: A Foundation for Analysis in the Health Sciences, 10th edition John Wiley & Sons.
Dunn, O.J., Clark, V.A., 2009. Basic Statistics: A Primer for the Biomedical Sciences. John Wiley & Sons.

Relevant Website

https://www.r-project.org
The R Project for Statistical Computing.