

Descriptive statistics

Some basic concepts

A *population* is a finite or infinite collection of individuals or objects. Often it is impossible or impractical to get data on all the members of the population and a small part of it or *sample* is examined instead.

If a sample is representative of the population, a lot can be learned about the population by studying the sample.

Descriptive statistics deals with the description or simple analysis of population or sample data.

Perhaps the simplest analysis of data is to classify individual observations into two or more *categories* according to some *attribute* that they possess, e.g. people can be classified into employed/unemployed, or into several groups according to religion, etc.

A *variable* is a measurable characteristic which changes from one member of a sample or a population to another, e.g. age of a person, number of defective items in a standard box of screws, GDP of a country, etc.

A *continuous variable* is a measurable characteristic which potentially can take any value in a continuous range, without any breaks or jumps. Height and income of a person are examples of continuous variables.

A *discrete variable* is a measurable characteristic which is restricted to a specific set of values. (Which can be finite or infinite). Number of people in a family, number of defective lightbulbs in a load, score on a die, etc. are examples of discrete variables because, in these cases, only *integer* values are possible. Note that a variable would still be discrete if it could potentially take *any* integer value, however high – the important thing being there are discrete jumps in the value. The discrete jumps don't have to be integer, either; a variable that could take the values (say) 0, 0.25, 0.5, 0.75 or 1 and none other would still be discrete.

When observations are classified into a number of *class intervals* specified in terms of a variable, we have a *frequency distribution* showing e.g. that there are 2 people in a class with height below 5', 4 people with height

between 5' and 5'2", 7 people with height between 5'2" and 5'4", etc. Such distributions can be represented in the form of a table or diagram. Discrete data are often represented by *bar charts*, showing e.g. number of families with 0,1,2,... children. Continuous data are usually represented as *histograms*, showing *frequency density* per *unit* of the variable – so for example if we had 8 people with incomes between £10,000 and £14,000, and 4 people with incomes between £14,000 and £16,000, these two categories would have the same frequency density (twice as many people, but twice as wide a category), and therefore the same height in a histogram. The first category though would have twice the *width* in the histogram.

It is important to distinguish between *exact* limits of a class interval and *grouping* limits, e.g. incomes are often recorded to the nearest pound, and for grouping purposes a class interval may be specified as containing (say) income between £50 and £99 p.w. (including both values). The exact limits for the interval are £49.50 and £99.50 p.w. because of rounding. Exact limits are important for analytic and graphical purposes.

Measures of Central Tendency

For comparative purposes, it is often desirable to represent (summarise) a frequency distribution by a single value. *Mean, media and mode* are the most commonly used measures of *central tendency*. The *mean* (arithmetic mean) or average is defined as:

$$\bar{X} = (\sum_{i=1}^n X_i) / n$$

Where X_i , $i=1, \dots, n$ are individual observations, there being n of them. (n is called the *sample size*). For grouped data we have

$$\bar{X} = \frac{\sum_{i=1}^r f_i X_i}{\sum_{i=1}^r f_i}$$

Where X_i are midpoints of exact class intervals (there being r of them) and f_i are class frequencies.

The mean is easy to understand and has a number of properties which are important for analytical and empirical work. For example, if \bar{X} is the average income of n people, their total income is given by $n * \bar{X}$. So \bar{X} can be seen as the income everyone would have if all the income in the group were distributed equally. Also, $\sum_{i=1}^n (X_i - \bar{X}) = 0$, i.e. deviations around the mean sum to zero. A disadvantage of the mean is that it is greatly influenced by extreme observations.

The median is defined as the middle value when observations are arranged in an ascending or descending order. When the number of observations is even, it is the average of the two middle values.

For grouped data (frequency table), the median is given by

$$Me = L_m + c_m \left[\frac{n}{2} - \sum_{i=1}^{m-1} f_i \right] / f_m$$

Where L_m = exact lower limit of class interval, c_m = width of the median class interval, $(n/2)$ = rank order of the median, $\sum_{i=1}^{m-1} f_i$ = cumulative frequency up to but not including median category, f_m = frequency of the median class interval.

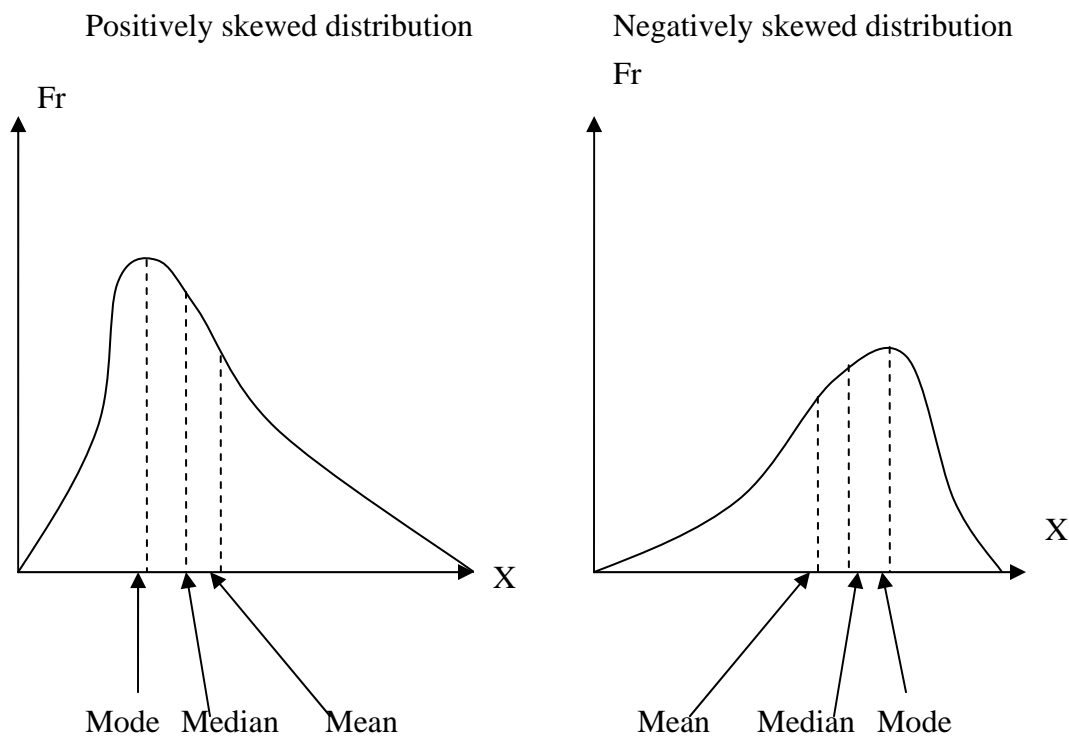
The median is easy to understand, and it is not greatly affected by extreme observations. It is often used in preference to the mean when extreme observations are present or when a distribution is heavily skewed, e.g. distribution of income or wealth.

The mode is defined as the most common value of individually recorder observations and as the value of the variable for which the frequency density is greatest for grouped data.

Median and Mean

In skewed distributions, these two *statistics* will differ. For example, in the case of incomes, a few people with very high incomes push up the mean, but don't affect the median, so that the mean is higher. In other words, *most people have below average incomes!* On the other hand if we consider average life expectancy, a small number of people who die in infancy (compared to very, very few indeed who live to, say, 140) push the mean down. (This will especially be true in developing countries). So, *most people live longer than the average life expectancy*.

For symmetric distributions, mean, median and mode all coincide, assuming middle values are more likely than extreme values. For skewed distributions, we can see the relationship between the three measures graphically:



2.3

Measures of variability

Variability of measurements is often more important in statistical analysis than central tendency. Three measures of variability are discussed below.

Mean deviation is defined as

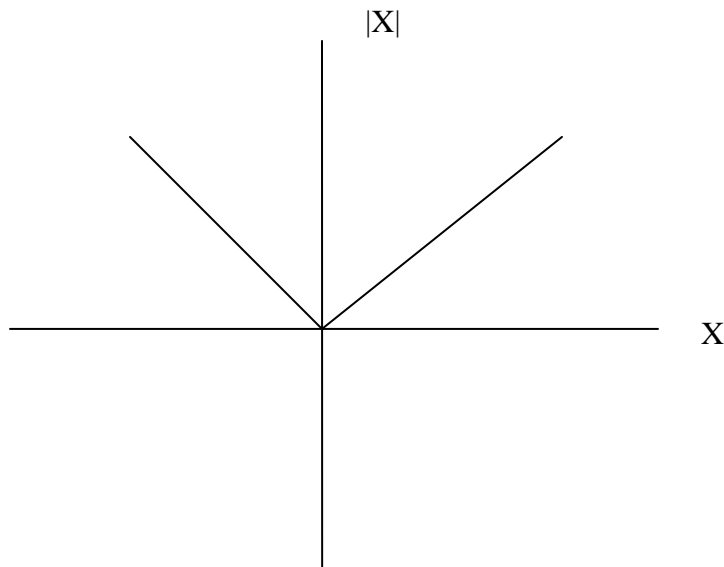
$$MD = \left(\sum_{i=1}^n |X_i - \bar{X}| \right) / n$$

(For ungrouped data), where the $|\cdot|$ sign means absolute value, that is ignoring minus signs. This measure, while intuitively natural, is not much use for analytical work because of the modulus signs – it is not a *smooth* measure. (see below)

Standard deviation is defined as:

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (\text{for ungrouped data})$$

S^2 is called the variance. Standard deviation avoids the modulus operator, and is measured in the same units as the data itself, and as the mean. It can be interpreted as a typical deviation from the mean.



The variance is the *average squared deviation from the mean*. The standard deviation is its square root.

Example

Ten students in a class have test scores of 35, 42, 46, 51, 54, 59, 64, 66, 73 and 78.

The mean score is $(35+42+46+51+54+59+64+66+73+78)/10=56.8$.

The deviations from the mean are therefore -21.8,-14.8,-12.8,-5.8,-2.8,+2.2,+7.2,+9.2,+16.2,+21.2.

To find the variance, we square each of these values (to get 475.24,219.04,...), add them up and divide by 10, to get an average of **170.56**.

Finally to find the standard deviation, we take the square root of the variance, giving a value of **13.06**.

Coefficient of variation, defined as S/\bar{X} , is a relative measure of dispersion used to compare variability of two or more distributions whose means and standard deviations differ a great deal; for example, the standard deviation of incomes, measured in £GB, in the UK would be a lot higher than in, say, a poor African country. However, if we scale this by the mean to get the CV, this could be greater or smaller in either country.

Skewness

Many distributions are not symmetrical and their degree of skewness can vary quite considerably, e.g. the distribution of wealth is more skewed than the distribution of income. Asymmetry can be measured in a number of ways. Only one measure will be mentioned here. The Pearson coefficient of skewness is defined as

$$SK = (\bar{X} - Mo) / S \approx 3(\bar{X} - Me) / S$$

The difference $\bar{X} - Mo$, which is approximately equal to $\bar{X} - Me$ in a not-too-skewed distribution, increases with skewness. Division by S ensures that it is not affected by changes in units of measurement and variability of different distributions. Note that for positively skewed distributions, $\bar{X} - Mo$ is positive, and for a negatively skewed distribution it is negative. The absolute value of SK can be as large as 3, but in practice values larger than 1 are rare.