

Health Sciences M.Sc. Programme
Applied Biostatistics
Mean and Standard Deviation

The mean

The median is not the only measure of central value for a distribution. Another is the **arithmetic mean** or **average**, usually referred to simply as the **mean**. This is found by taking the sum of the observations and dividing by their number. The mean is often denoted by a little bar over the symbol for the variable, e.g. \bar{x} .

The sample mean has much nicer mathematical properties than the median and is thus more useful for the comparison methods described later. The median is a very useful descriptive statistic, but not much used for other purposes.

Median, mean and skewness

The sum of the 57 FEV1s is 231.51 and hence the mean is $231.51/57 = 4.06$. This is very close to the median, 4.1, so the median is within 1% of the mean. This is not so for the triglyceride data. The median triglyceride is 0.46 but the mean is 0.51, which is higher. The median is 10% away from the mean. If the distribution is symmetrical the sample mean and median will be about the same, but in a skew distribution they will not. If the distribution is skew to the right, as for serum triglyceride, the mean will be greater, if it is skew to the left the median will be greater. This is because the values in the tails affect the mean but not the median.

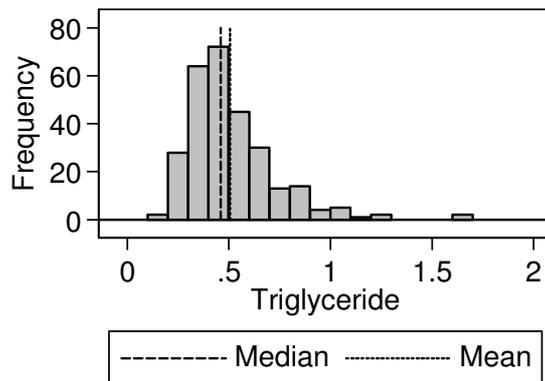
Figure 1 shows the positions of the mean and median on the histogram of triglyceride. We can see that increasing the skewness by making the observation above 1.5 much bigger would have the effect of increasing the mean, but would not affect the median. Hence as we make the distribution more and more skew, we can make the mean as large as we like without changing the median. This is the property which tends to make the mean bigger than the median in positively skew distributions, less than the median in negatively skew distributions, and equal to the median in symmetrical distributions.

Variance

The mean and median are measures of the central tendency or position of the middle of the distribution. We shall also need a measure of the spread, dispersion or variability of the distribution.

The most commonly used measures of dispersion are the variance and standard deviation, which I will define below. We start by calculating the difference between each observation and the sample mean, called the **deviations from the mean**. Some of these will be positive, some negative.

Figure 1. Histogram of serum triglyceride in cord blood, showing the positions of the mean and median



If the data are widely scattered, many of the observations will be far from the mean and so many deviations will be large. If the data are narrowly scattered, very few observations will be far from the mean and so few deviations will be large. We need some kind of average deviation to measure the scatter. If we add all the deviations together, we get zero, so there is no point in taking an average deviation. Instead we square the deviations and then add them. This removes the effect of plus or minus sign; we are only measuring the size of the deviation, not the direction. This gives us the **sum of squares about the mean**, usually abbreviated to **sum of squares**. In the FEV1 example the sum of squares is equal to 25.253371.

Clearly, the sum of squares will depend on the number of observations as well as the scatter. We want to find some kind of average squared deviation. This leads to a difficulty. Although we want an average squared deviation, we divide the sum of squares by the number of observations minus one, not the number of observations. This is not the obvious thing to do and puzzles many students of statistical methods. The reason is that we are interested in estimating the scatter of the population, rather than the sample, and the sum of squares about the sample mean is not proportional to the number of observations. This is because the mean which we subtract is also calculated from the same observations. If we have only one observation, the sum of squares must be zero. The sum of squares cannot be proportional to the number of observations. Dividing by the number of observations would lead to small samples producing lower estimates of variability than large samples from the same population.

In fact, the sum of squares about the sample mean is proportional to the number of observations minus one. If we divide the sum of squares by the number of observations minus one, the measure of variability will not be related to the sample size.

The estimate of variability found in this way is called the **variance**. The quantity is called the **degrees of freedom** of the variance estimate, often abbreviated to **df** or **DF**. We shall come across this term several times. It derived from probability theory and we shall accept it as just a name. We often denote the variance calculated from a sample by s^2 .

For the FEV data, $s^2 = 25.253371/(57 - 1) = 0.449$. Variance is based on the squares of the observations. FEV1 is measured in litres, so the squared deviations are measured in square litres, whatever they are. We have for FEV1: variance = 0.449 litres². Similarly, gestational age is measured in weeks and so the gestational age: variance = 5.24 weeks². A square week is another quantity hard to visualise.

Variance is based on the squares of the observations and so is in squared units. This makes it difficult to interpret. For this reason we often use the standard deviation instead, described below.

Standard deviation

The variance is calculated from the squares of the observations. This means that it is not in the same units as the observations, which limits its use as a descriptive statistic. The obvious answer to this is to take the square root, which will then have the same units as the observations and the mean. The square root of the variance is called the **standard deviation**, usually denoted by s . It is often abbreviated to **SD**.

For the FEV data, the standard deviation = $\sqrt{0.449} = 0.67$ litres. Figure 2 shows the relationship between mean, standard deviation and frequency distribution for FEV1. Because standard deviation is a measure of variability about the mean, this is shown as the mean plus or minus one or two standard deviations. We see that the majority of observations are within one standard deviation of the mean, and nearly all within two standard deviations of the mean. There is a small part of the histogram outside the mean plus or minus two standard deviations interval, on either side of this symmetrical histogram.

For the serum triglyceride data, $s = \sqrt{0.04802} = 0.22$ mmol/litre. Figure 3 shows the position of the mean and standard deviation for the highly skew triglyceride data. Again, we see that the majority of observations are within one standard deviation of the mean, and nearly all within two standard deviations of the mean. Again, there is a small part of the histogram outside the mean plus or minus two standard deviations interval. In this case, the outlying observations are all in one tail of the distribution, however.

For the gestational age data, $s = \sqrt{5.242} = 2.29$ weeks. Figure 4 shows the position of the mean and standard deviation for this negatively skew distribution. Again, we see that the majority of observations are within one standard deviation of the mean, and nearly all within two standard deviations of the mean. Again, there is a small part of the histogram outside the mean plus or minus two standard deviations interval. In this case, the outlying observations are almost all in the lower tail of the distribution.

In general, we expect roughly $2/3$ of observations or more to lie within one standard deviation of the mean and about 95% to lie within two standard deviations of the mean.

Figure 2. Histogram of FEV1 with mean and standard deviation marked.

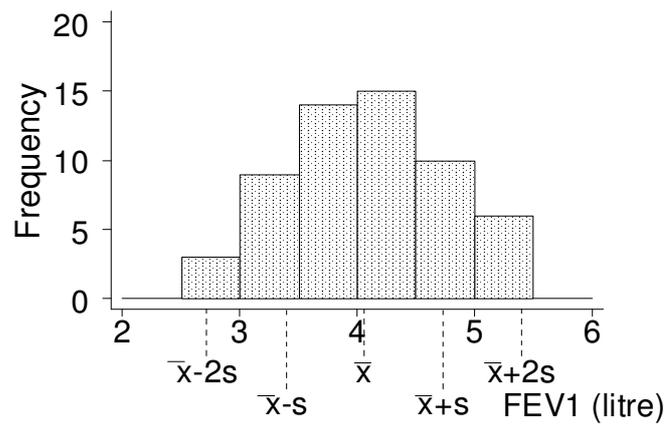


Figure 3. Histogram of serum triglyceride with positions of mean and standard deviation marked

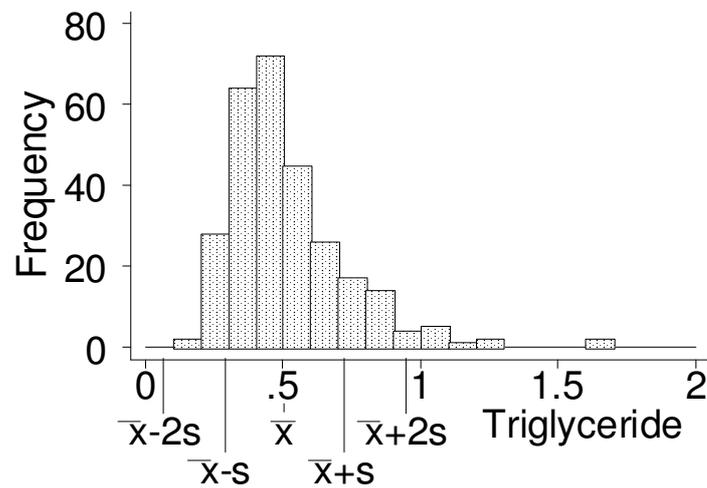


Figure 4. Histogram of gestational age with mean and standard deviation marked.

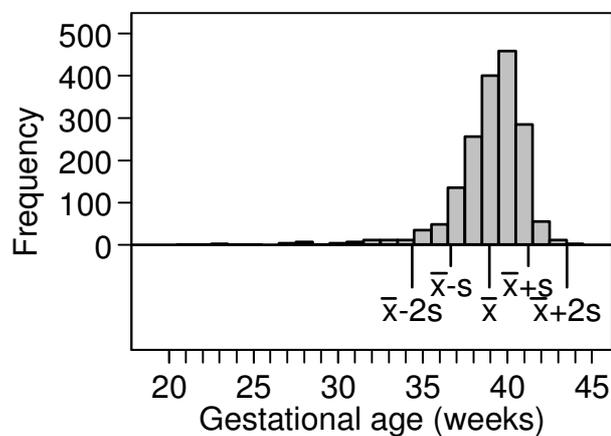
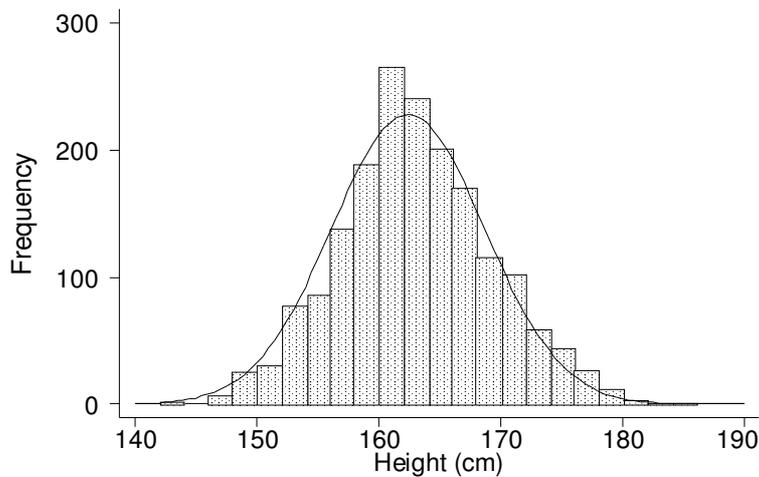


Figure 5. Distribution of height in a sample of pregnant women, with the corresponding Normal distribution curve



Spotting skewness

Histograms are fairly unusual in published papers. Often only summary statistics such as mean and standard deviation or median and range are given. We can use these summary statistics to tell us something about the shape of the distribution.

If the mean is less than two standard deviations, then any observation less than two standard deviations below the mean will be negative. For any variable which cannot be negative, this tells us that the distribution must be positively skew.

If the mean or the median is near to one end of the range or interquartile range, this tells us that the distribution must be skew. If the mean or median is near the lower limit it will be positively skew, if near the upper limit it will be negatively skew.

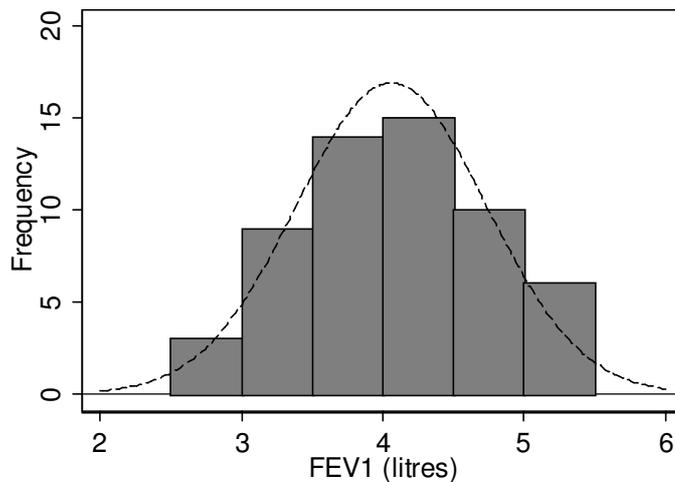
For example, for the triglyceride data, median = 0.46, mean = 0.51, SD = 0.22, range = 0.15 to 1.66, and IQR = 0.35 to 0.60 mmol/l. The median is less than the mean and the mean and median are both nearer the low end of the range than to the high end. These are both indications of positive skewness.

These rules of thumb only work one way, e.g. the mean may exceed two standard deviations and the distribution may still be skew, as in the triglyceride case. For the gestational age data, the median = 39, mean = 38.95, SD = 2.29, range = 21 to 44, and IQR = 38 to 40 weeks. Here median and mean are almost identical, the mean is much bigger than the standard deviation, and median and mean are both in the centre of the interquartile range. Only the range gives away the skewness of the data, showing that median and mean are close to the upper limit of the range.

The Normal Distribution

Many statistical methods are only valid if we can assume that our data follow a distribution of a particular type, the Normal distribution. This is a continuous, symmetrical, unimodal distribution described by a mathematical equation. I shall omit the mathematical detail. Figure 5 shows the distribution of height in a large sample of pregnant women. The distribution is unimodal and symmetrical. The Normal distribution curve corresponding to the height distribution fits very well indeed. Figure 6 shows the distribution of FEV1 in male medical students. The Normal distribution curve also fits these data well.

Figure 6. Distribution of FEV1 in a sample of male medical students, with the corresponding Normal distribution curve



For the height data in Figure 5 the mean = 162.4 cm, variance = 39.5 cm², and SD = 2.3 cm. For the FEV data in Figure 6, mean = 4.06 litres, variance = 0.45 litres², and SD = 0.67 litres. Yet both follow a Normal distribution. This can be true because the Normal distribution is not just one distribution, but a family of distributions. There are infinitely many members of the Normal distribution family. The particular member of the family that we have is defined by two numbers, called parameters. **Parameter** is a mathematical term meaning a number which defines a member of a class of things. The parameters of a Normal distribution happen to be equal to the mean and variance of the distribution. These two numbers tell us which member of the Normal family we have. So to draw the Normal distribution curve which best matches the height distribution, we find the mean and variance of height and then use these as parameters to define which member of the Normal distribution family we should draw on the graph. We shall meet several such distribution families in this module.

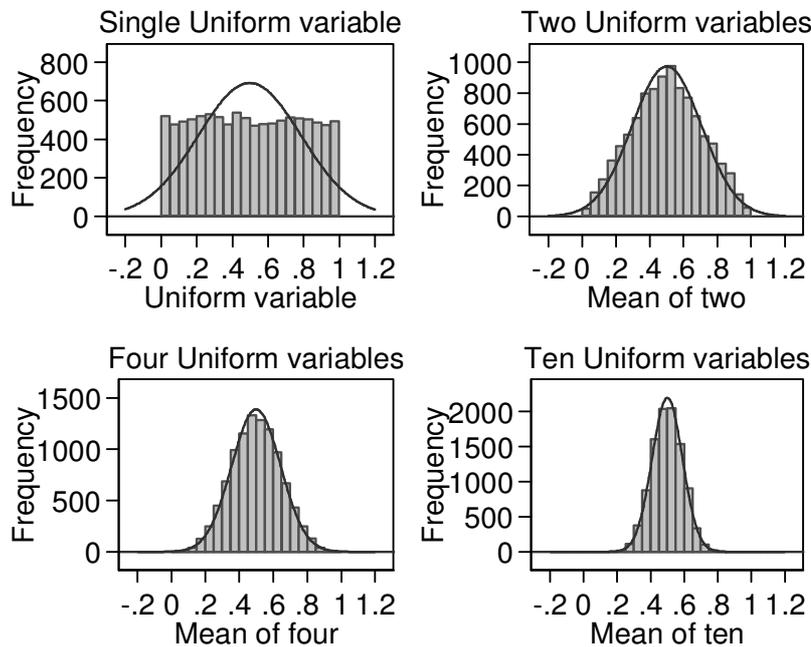
The particular member of the Normal family for which mean = 0 and variance = 1 is called the **Standard Normal distribution**. If we subtract the mean from a variable which has a Normal distribution, then divide by the standard deviation (square root of the variance) we will get the Standard Normal distribution.

The Normal distribution, also known as the **Gaussian distribution**, may be regarded as the fundamental probability distribution of statistics. The word 'normal' here is not used in its common meaning of 'ordinary or common', or its medical meaning of 'not diseased'. The usage relates to its older meaning of 'conforming to a rule or pattern'. It would be wrong to infer that most variables are Normally distributed, though many are.

The Normal distribution is important for two reasons.

1. Many natural variables follow it quite closely, certainly sufficiently closely for us to use statistical methods which require this. Such methods include t tests, correlation, and regression.
2. Even when we have a variable which does not follow a Normal distribution, if we take the mean of a sample of observations, such means will follow a Normal distribution.

Figure 7. Simulation study showing the effect of taking the average of several variables



The second comment deserves a bit more explanation. To be more precise, if we have any series of independent variables which come from the same distribution, then their sum tends to be closer and closer to a Normal Distribution as the number of variables increases. This is known as the **central limit theorem**. As most sets of measurements are observations of such a series of random variables, this is a very important property. From it, we can deduce that the sum or mean of any large series of independent observations follows a Normal Distribution.

Such a claim deserves some evidence to back it up. For example, consider the **Uniform or Rectangular Distribution**. This is the distribution where all values between two limits, say 0 and 1, are equally likely and no other values are possible. Figure 7 shows the histogram for the frequency distribution of 10000 observations from the Uniform Distribution between 0 and 1. It is quite different from the Normal Distribution with the same mean and variance, which is also shown. Now suppose we create a new variable by taking two Uniform variables and finding their average. The shape of the distribution of the mean of two variables is quite different to the shape of the Uniform Distribution. The mean is unlikely to be close to either extreme, and observations are concentrated in the middle near the expected value. The reason for this is that to obtain a low mean, both the Uniform variables forming it must be low; to make a high mean both must be high. But we get a mean near the middle if the first is high and the second low, or the first is low and second high, or both first and second are moderate. The distribution of the mean of two is much closer to the Normal than is the Uniform Distribution itself. However, the abrupt cut-off at 0 and at 1 is unlike the corresponding Normal Distribution. Figure 7 also shows the result of averaging four Uniform variables and ten Uniform variables. The similarity to the Normal Distribution increases as the number averaged increases and for the mean of ten the correspondence is so close that the distributions could not easily be told apart.

Figure 8. Three members of the Normal distribution family, specified by their parameters, mean and variance

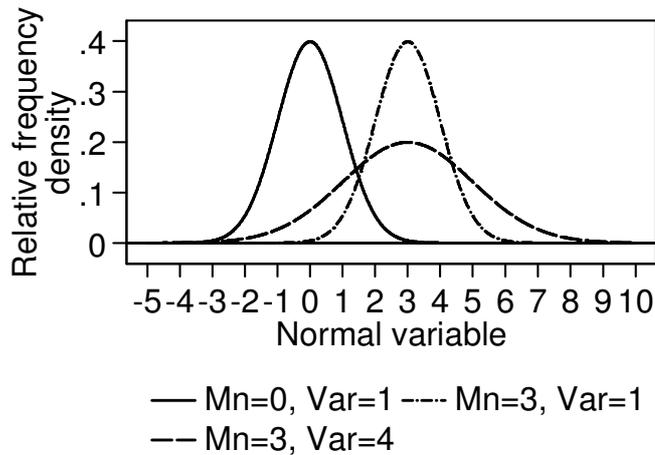
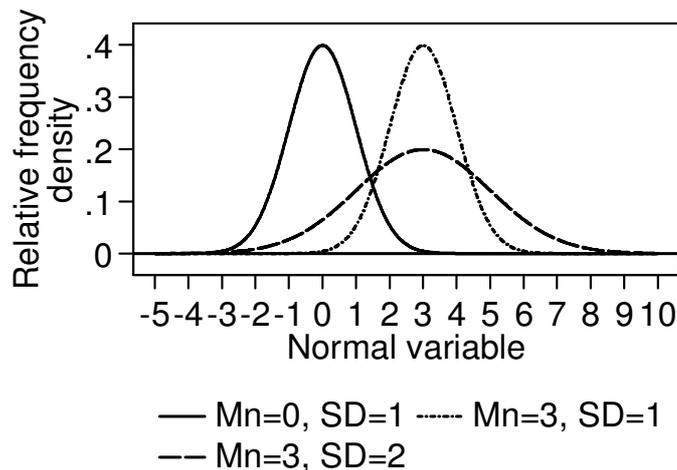


Figure 9. Three members of the Normal distribution family with by their means and standard deviations



We are often able to assume that averages and some of the other things we calculate from samples will follow Normal distributions, whatever the distribution of the observations themselves.

Figure 8 shows three members of the Normal distribution family, the Standard Normal distribution with mean = 0 and variance = 1 and those with mean = 3 and variance = 1 and with mean = 3 and variance = 4. We can see that if we keep the standard deviation the same and change the mean, this just moves the curve along the horizontal axis, leaving the shape unchanged. If we keep the mean constant and increase the standard deviation, we stretch out the curve on either side of the mean.

As variance is a squared measure, it is actually easier to talk about these distributions in terms of their standard deviations, so Figure 9 shows the same distributions with their standard deviations. If we look at the curve for mean = 0 and SD = 1, we can see that almost all the curve is between -3 and +3. (In fact, 99.7% of the area beneath the curve is between these limits.) If we look at the curve for mean = 3 and SD = 1, we can see that almost all the curve is between 0 and +6. This is within 3 on either

side of the mean (which is also = 3). If we look at the curve for mean = 3 and SD = 2, we can see that almost all the curve is between -3 and $+9$. This is within 6 on either side of the mean, which is equal to 3 standard deviations. In terms of standard deviations from the mean, the Normal distribution curve is always the same.

To get numbers from the Normal distribution, we need to find the area under the curve between two values of the variable, or equivalently below any given value of the variable. This area will give us the proportion of observations below that value. Unfortunately, there is no simple formula linking the variable and the area under the curve. Hence we cannot find a formula to calculate the relative frequency between two chosen values of the variable, nor the value which would be exceeded for a given proportion of observations.

Numerical methods for calculating these things with acceptable accuracy have been found. These were used to produce extensive tables of the Normal distribution. Now, these numerical methods for calculating Normal frequencies have been built into statistical computer programs and computers can estimate them whenever they are needed. Two numbers from tables of the Normal distribution:

1. we expect 68% of observations to lie within one standard deviation from the mean,
2. we expect 95% of observations to lie within 1.96 standard deviations from the mean.

This is true for all Normal distributions, whatever the mean, variance, and standard deviation.

Martin Bland
10 August 2006