Applied Biostatistics

Mean and Standard Deviation

Martin Bland

Professor of Health Statistics
University of York

http://www-users.york.ac.uk/~mb55/

The mean

The **arithmetic mean** or **average**, usually referred to simply as the **mean** 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:

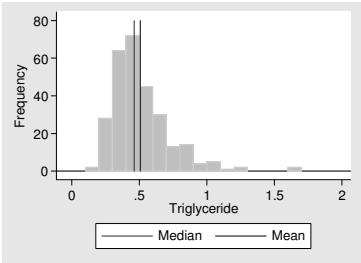
Mean FEV1 = 4.06. Median FEV1 = 4.1, so the median is within 1% of the mean.

Mean triglyceride = 0.51. Median triglyceride = 0.46. 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 usually be different.

If the distribution is skew to the right, as for serum triglyceride, the mean will usually be greater, if it is skew to the left the median will usually be greater.

This is because the values in the tails affect the mean but not the median.



Increasing the largest observation will pull the mean higher.

It will not affect the median.

Variability

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.

Variability

For use in the analysis of data, range and IQR are not satisfactory. Instead we use two other measures of variability: variance and standard deviation.

These both measure how far observations are from the mean of the distribution.

Variance is the average squared difference from the mean.

Standard deviation is the square root of the variance.

Variance

Variance is an average squared difference from the mean.

Note that if we have only one observation, we cannot do this. The mean is the observation and the difference is zero. We need at least two observations.

The sum of the squared differences from the mean is proportional to the number of observations minus one, called the **degrees of freedom**.

Variance is estimated as the sum of the squared differences from the mean divided by the degrees of freedom.

Variance

FEV1: variance = 0.449 litres^2

Gestational age: variance = 5.24 weeks^2

Variance is based on the squares of the observations and so is in squared units.

This makes it difficult to interpret.

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.

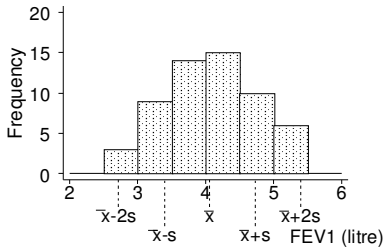
We take the square root, which will then have the same units as the observations and the mean.

The square root of the variance is called the standard deviation, usually denoted by *s*.

FEV1: $s = \sqrt{0.449} = 0.67$ litres.

Standard deviation

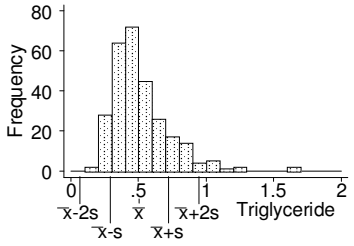
FEV1: $s = \sqrt{0.449} = 0.67$ litres.



Majority of observations within one SD of mean (usually about 2/3). Almost all within about two SD of mean (usually about 95%).

Standard deviation

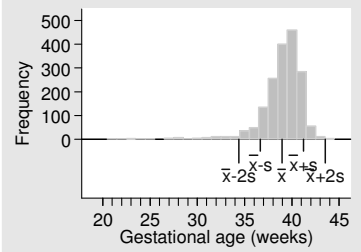
Triglyceride: $s = \sqrt{0.04802} = 0.22$ mmol/litre.



Majority of observations within one SD of mean (usually about 2/3). Almost all within about two SD of mean (usually about 95%), but those outside may be all at one end.

Standard deviation

Gestational age: $s = \sqrt{5.242} = 2.29$ weeks.



Majority of observations within one SD of mean (usually about 2/3). Almost all within about two SD of mean (usually about 95%), but those outside may be all at one end.

Spotting skewness

If the mean is less than two standard deviations, 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.

Spotting skewness

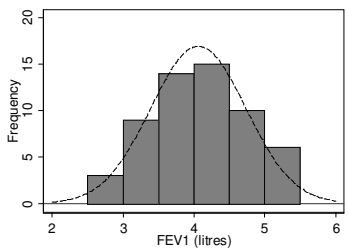
Triglyceride: median = 0.46, mean = 0.51, SD = 0.22, range = 0.15 to 1.66, IQR = 0.35 to 0.60 mmol/l.

These rules of thumb only work one way, e.g. mean may exceed two SD and distribution may still be skew.

Gestational age: median = 39, mean = 38.95, SD = 2.29, range = 21 to 44, IQR = 38 to 40 weeks.

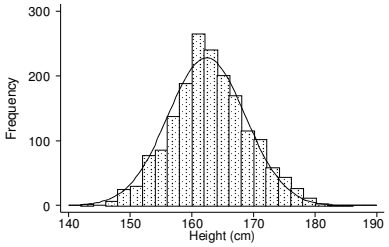
The Normal distribution

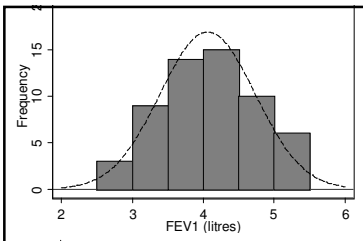
Many statistical methods are only valid if we can assume that our data follow a distribution of a particular type, the Normal distribution. This is a continuous, symmetrical, unimodal distribution described by a mathematical equation, which we shall omit.



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, which we shall omit.

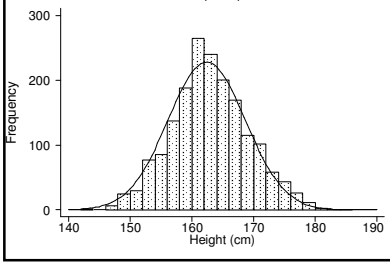




Mean = 4.06 litres

Variance = 0.45 litres^2

SD = 0.67 litres



Mean = 162.4 cm

Variance = 39.5 cm^2

SD = 2.3 cm

The Normal distribution is not just one distribution, but a family of distributions.

The particular member of the family that we have is defined by two numbers, called parameters.

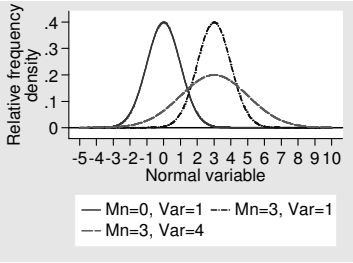
Parameter is a mathematical term meaning a number which defines a member of a class of things.

The parameters of a Normal distribution happen to be equal to the mean and variance.

These two numbers tell us which member of the Normal family we have.

The parameters (mean and variance) of a Normal distribution happen to be equal to the mean and variance.

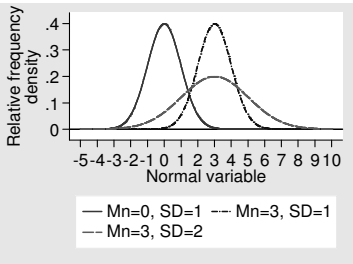
These two numbers tell us which member of the Normal family we have.



Mean=0, variance=1 is called the **Standard Normal distribution**.

The parameters (mean and variance) of a Normal distribution happen to be equal to the mean and variance.

These two numbers tell us which member of the Normal family we have.

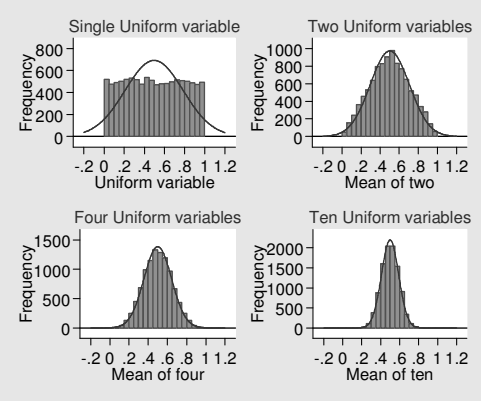


The distributions are the same in terms of standard deviations from the mean.

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.
2. Even when we have a variable which does not follow a Normal distribution, if we the take the mean of a sample of observations, such means will follow a Normal distribution.

An illustration of the Central Limit Theorem



There is no simple formula linking the variable and the area under the curve.

Hence we cannot find a formula to calculate the 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 were used to produce extensive tables of the Normal distribution.

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.

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.