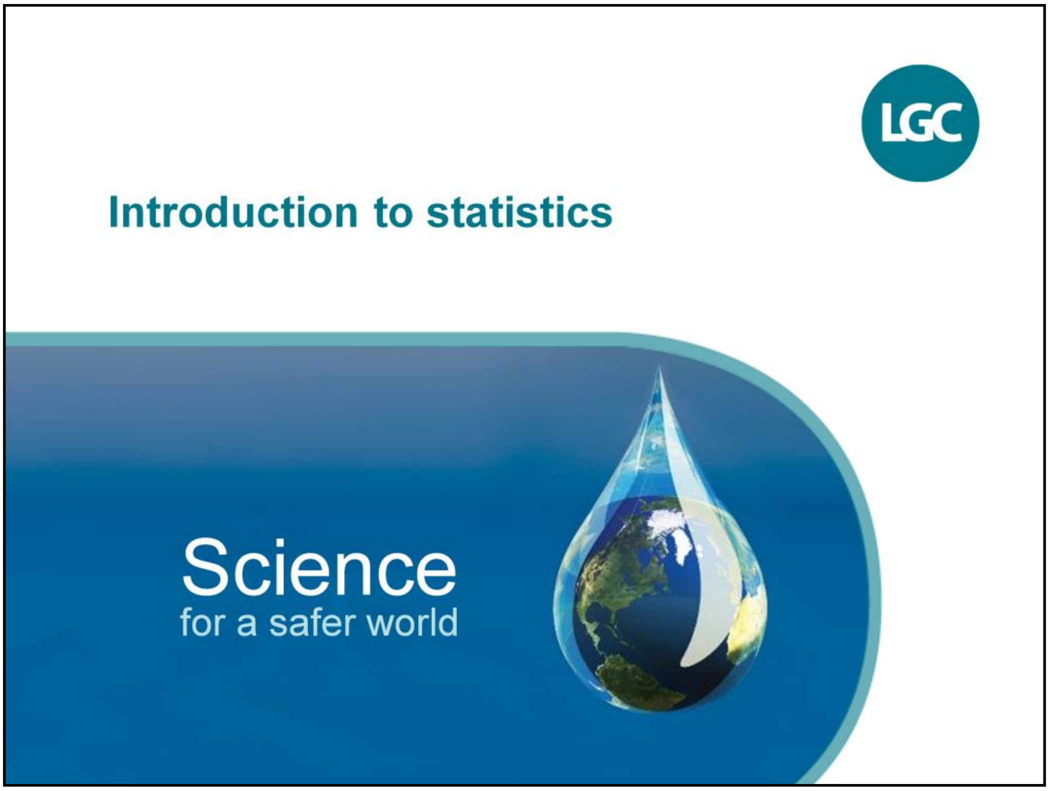
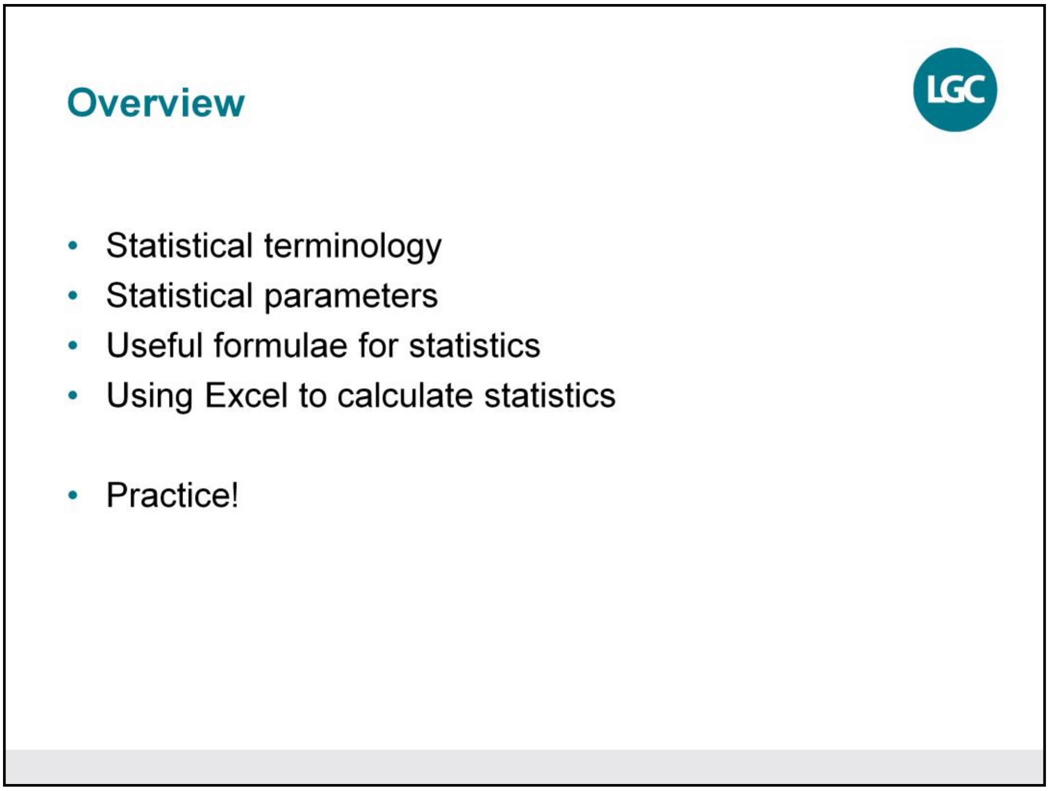


Introduction to statistics

Science
for a safer world



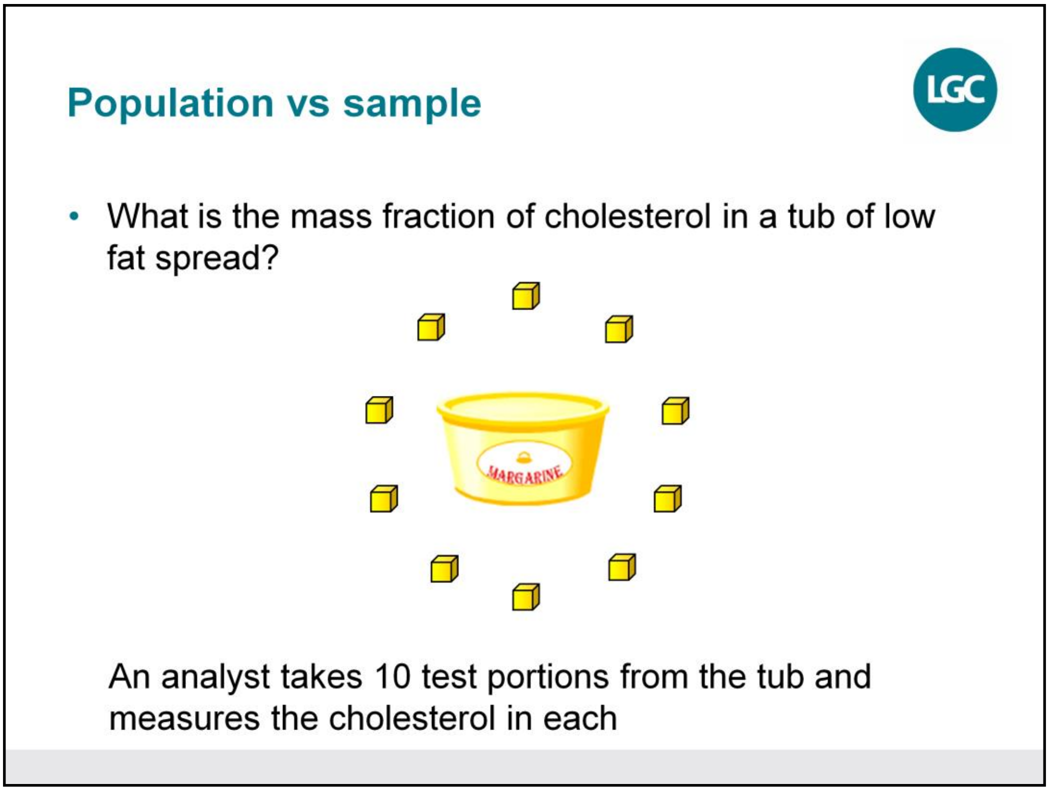
Overview

- Statistical terminology
- Statistical parameters
- Useful formulae for statistics
- Using Excel to calculate statistics

- Practice!

Analytical science uses a wide range of statistical principles. This lecture and the associated practice session form a short introduction to statistics and cover:

- the most important concepts and terms used on the course;
- the calculation of the most common statistical parameters.



Population vs sample

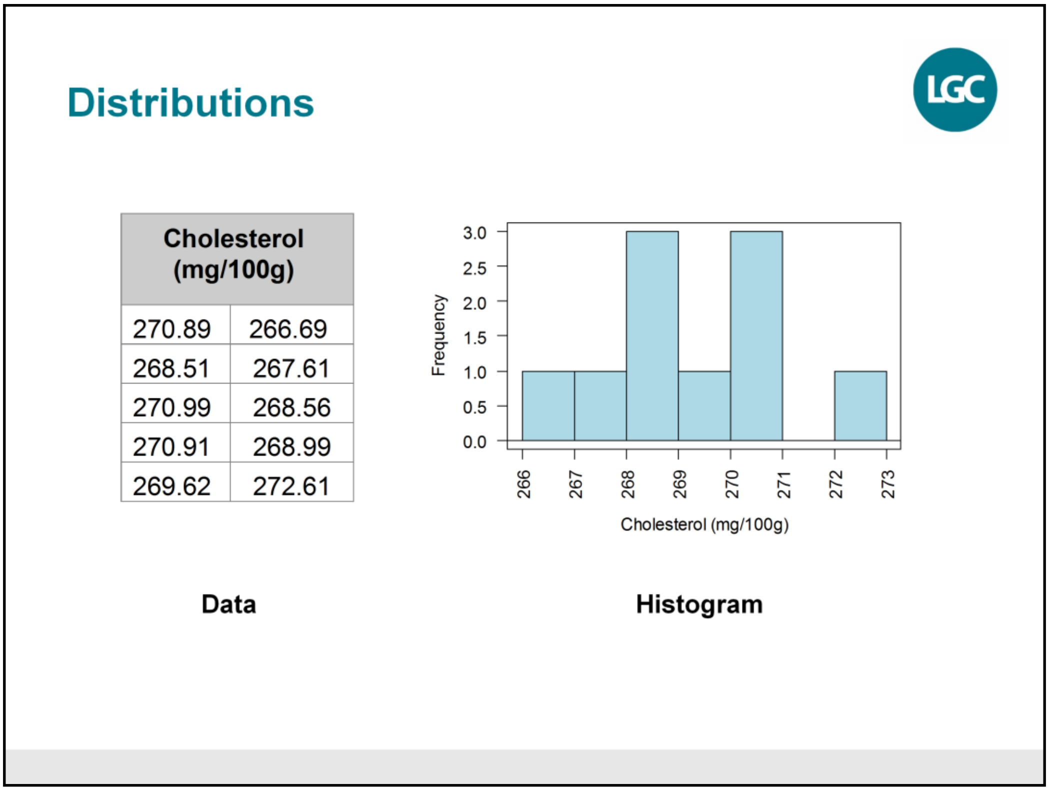
- What is the mass fraction of cholesterol in a tub of low fat spread?

An analyst takes 10 test portions from the tub and measures the cholesterol in each

In general, we don’t have access to the entire population of measurements. When we are asked to measure the amount of an analyte (e.g. mass fraction or concentration), we usually make a relatively small number of measurements on test portions and use the results as our best estimate of the true amount of the analyte present.

In the example on the slide, an analyst has been asked to determine the mass fraction of cholesterol in a tub of low fat spread. It is not practical to make a very large number of measurements using the entire tub of low fat spread, so the analyst takes 10 random test portions from the tub and determines the mass fraction of cholesterol in each. These 10 results represent a *sample* of the total *population* of results. The results from the 10 test portions are used to estimate the true cholesterol content of the entire tub of low fat spread.

The distinction between samples and populations of data is important, as it affects how some statistical parameters (in particular the standard deviation) are calculated. This issue is discussed further in this lecture, however, for the rest of the course we shall be concerned only with statistical parameters relating to samples from a population.



Distributions

Cholesterol (mg/100g)	
270.89	266.69
268.51	267.61
270.99	268.56
270.91	268.99
269.62	272.61

Data

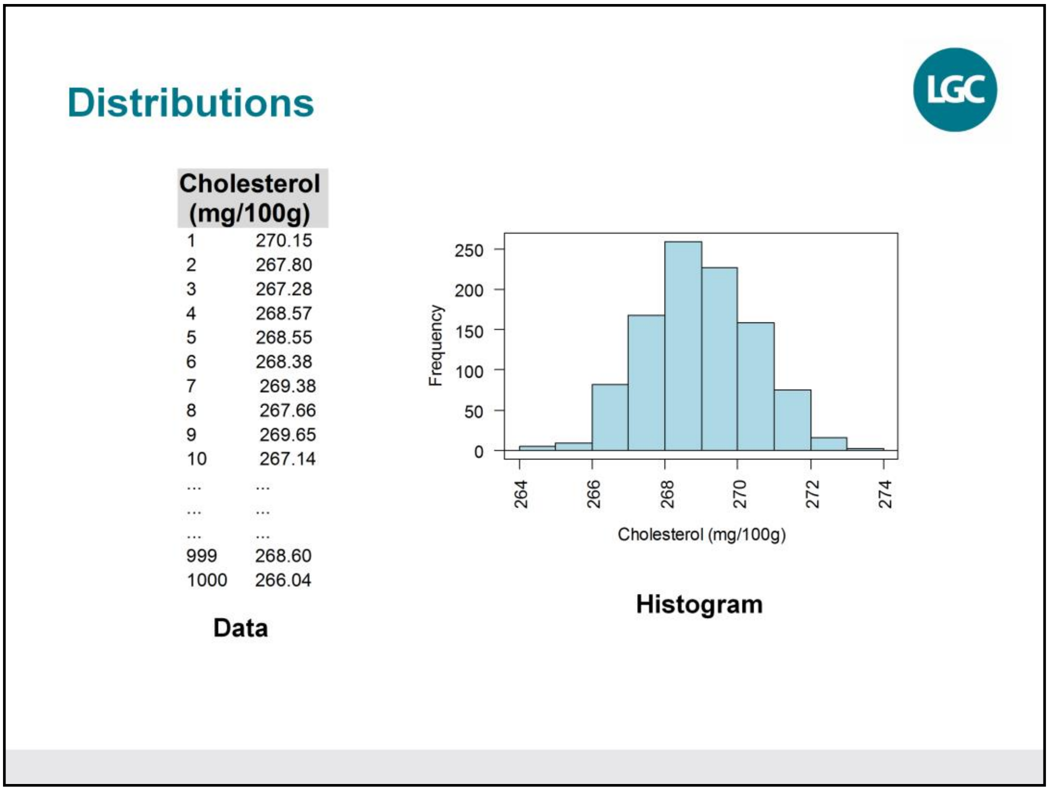
Histogram

The table shows a typical set of analytical data - a series of repeated determinations of the mass fraction of cholesterol in a sample of low fat spread, taken in the same laboratory. As expected, random variation results in a set of slightly different values.

The figure on the right shows a histogram of the same data. Each bar shows the number of determinations falling in a given range - the *frequency* of occurrence. The chart shows the *distribution* of the data.

The chemist needs to know a number of things about such data. In most cases, it is important to know:

- an estimate of the 'true value' - usually an average of the results;
- something about the spread of observations;
- how the estimate of the 'true value' might vary from one set of results to another.



Distributions

	Cholesterol (mg/100g)
1	270.15
2	267.80
3	267.28
4	268.57
5	268.55
6	268.38
7	269.38
8	267.66
9	269.65
10	267.14
…	…
…	…
…	…
999	268.60
1000	266.04

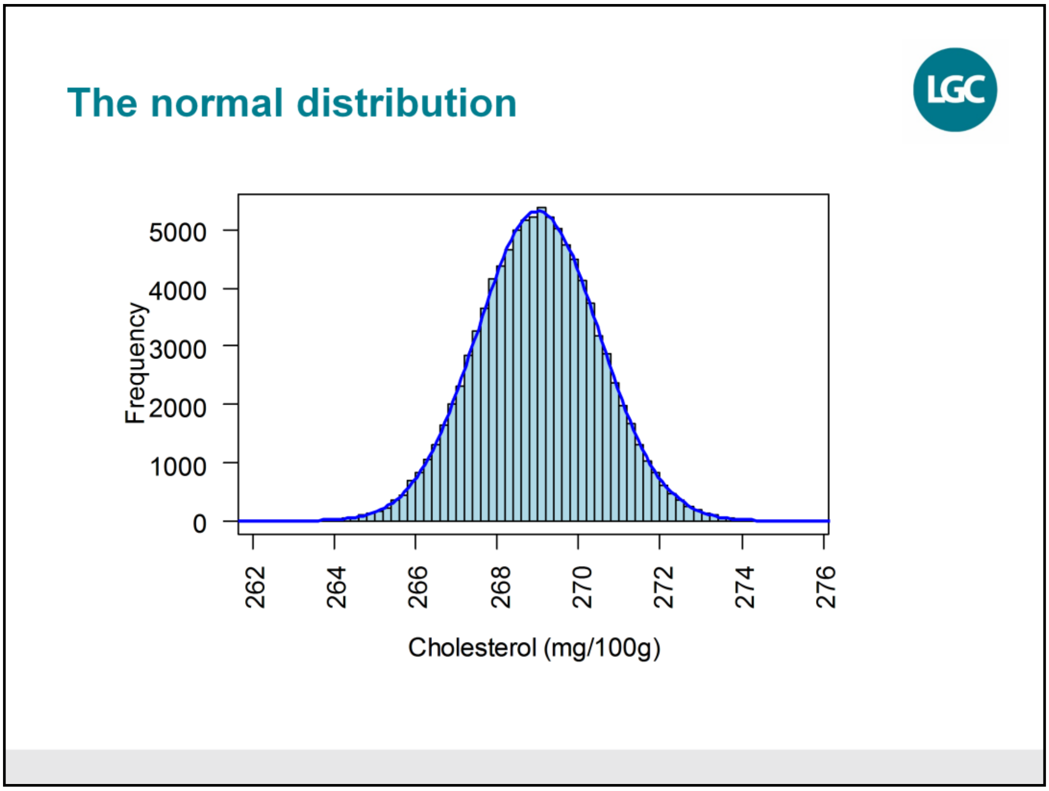
Data

Histogram

Here, part of a simulated data set is shown. The simulated data are based on the previous data from the determination of cholesterol in a sample of low fat spread. This time the histogram conveys a more definite impression of the distribution of the data.

Two points to note here are:

- data are concentrated in the central region of the histogram;
- the distribution is roughly symmetrical.



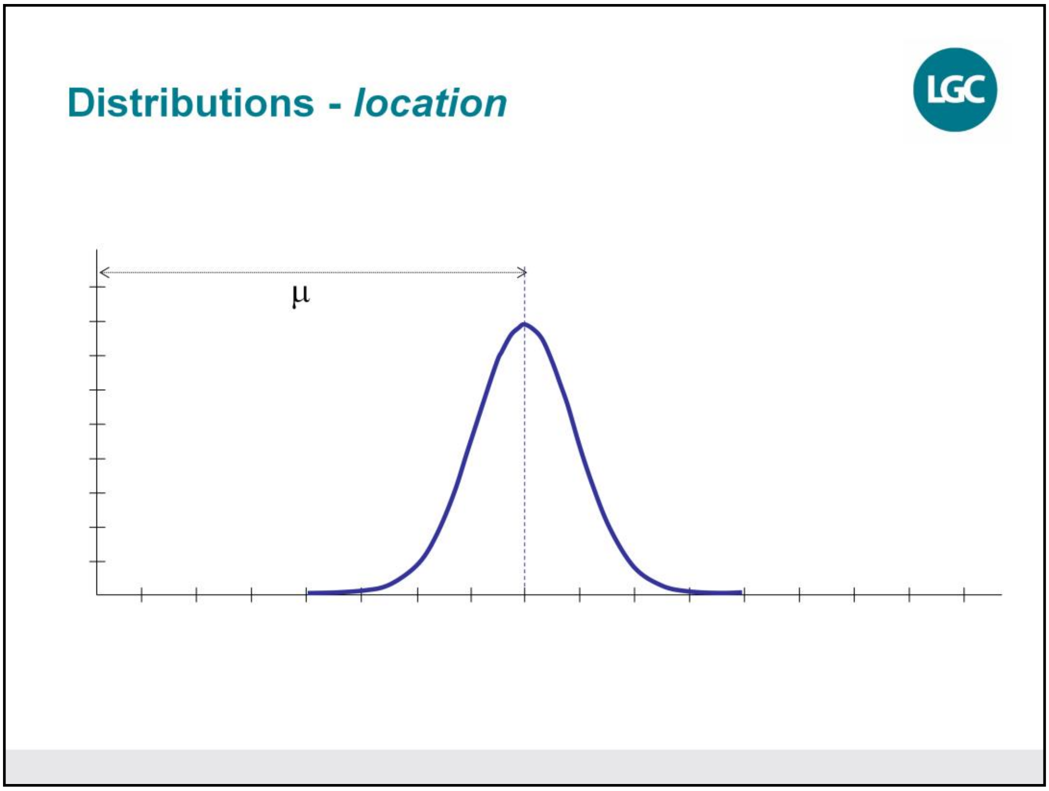
The normal distribution

If we took an even larger amount of data and distributed it between an even larger number of bins, a histogram such as the one above would result. This time we can see that the shape of the underlying population is clearly tending to be perfectly symmetrical about a central peak.

Finally, with a very large amount of data and a large number of bins, the shape of the underlying population becomes clear. We can think now of the population distribution as being described not by a histogram but by a smooth curve the equation of which we could, in principle, determine.

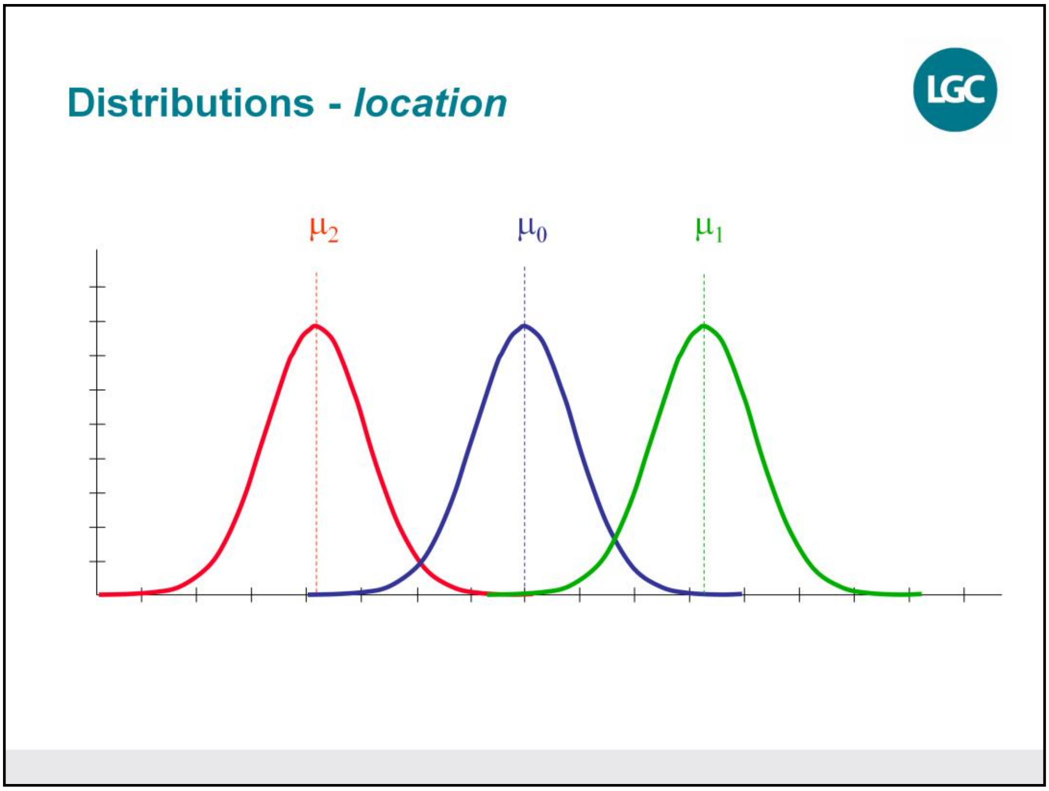
The Normal Distribution

As the name implies, the normal distribution describes the way results are commonly distributed. The very large majority of measurements subject to several different effects (environment, reagent variation, instrument 'noise' etc.) will, repeated frequently, fall into a normal distribution, with most results clustered around a central value and a decreasing number at greater distance. Note that the distribution has potentially infinite *range* - values may turn up at great distances from the centre of the distribution.



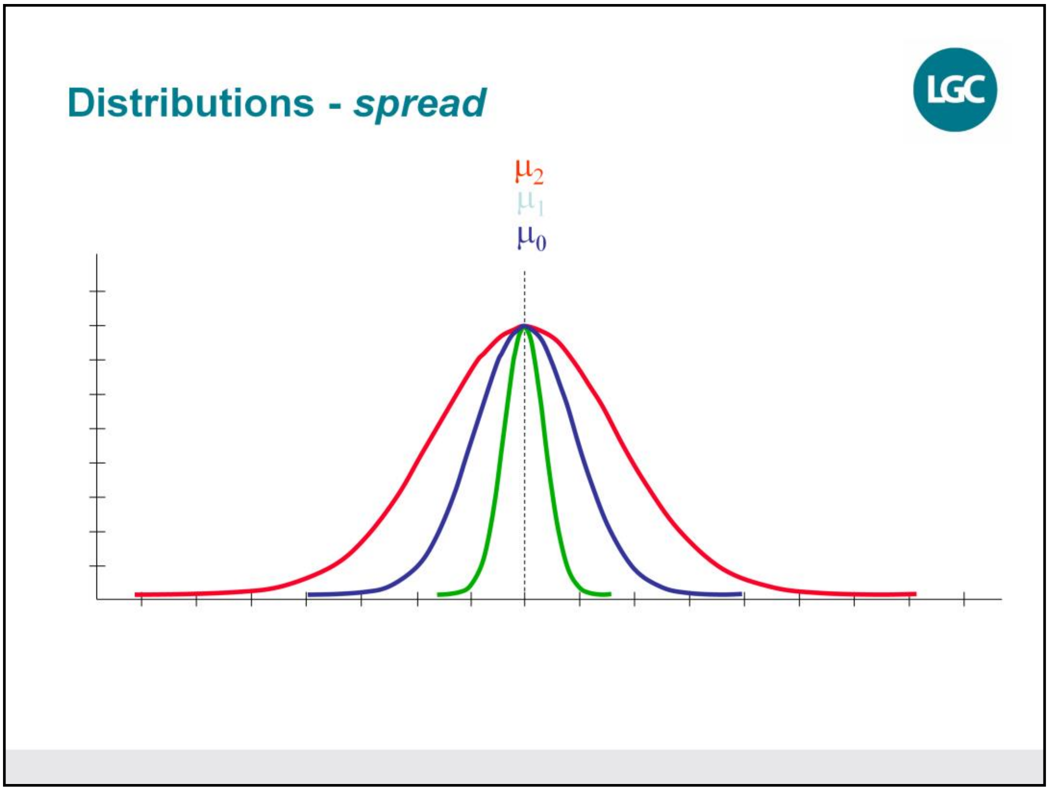
Distributions - *location*

We can characterise a distribution by means of a parameter μ (pronounced ‘mu’) which describes the centre, or location, of the distribution.



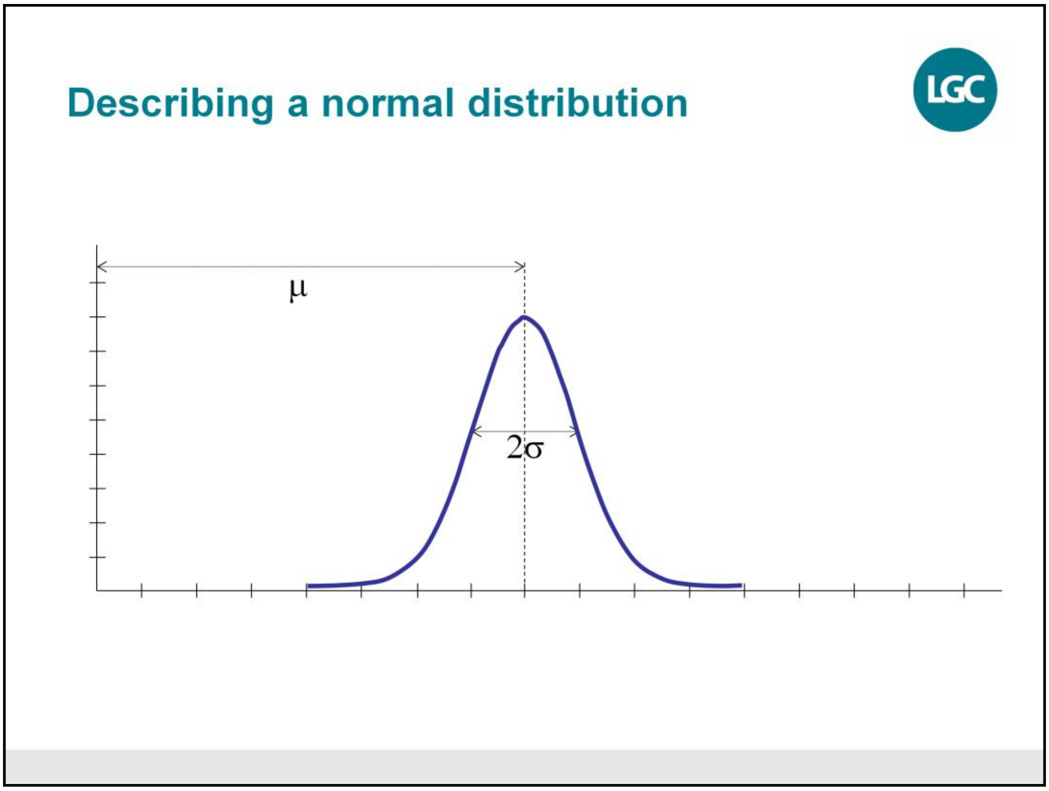
Distributions - *location*

The parameter μ allows us to distinguish between different distributions in terms of whereabouts on our measurement scale they are located.



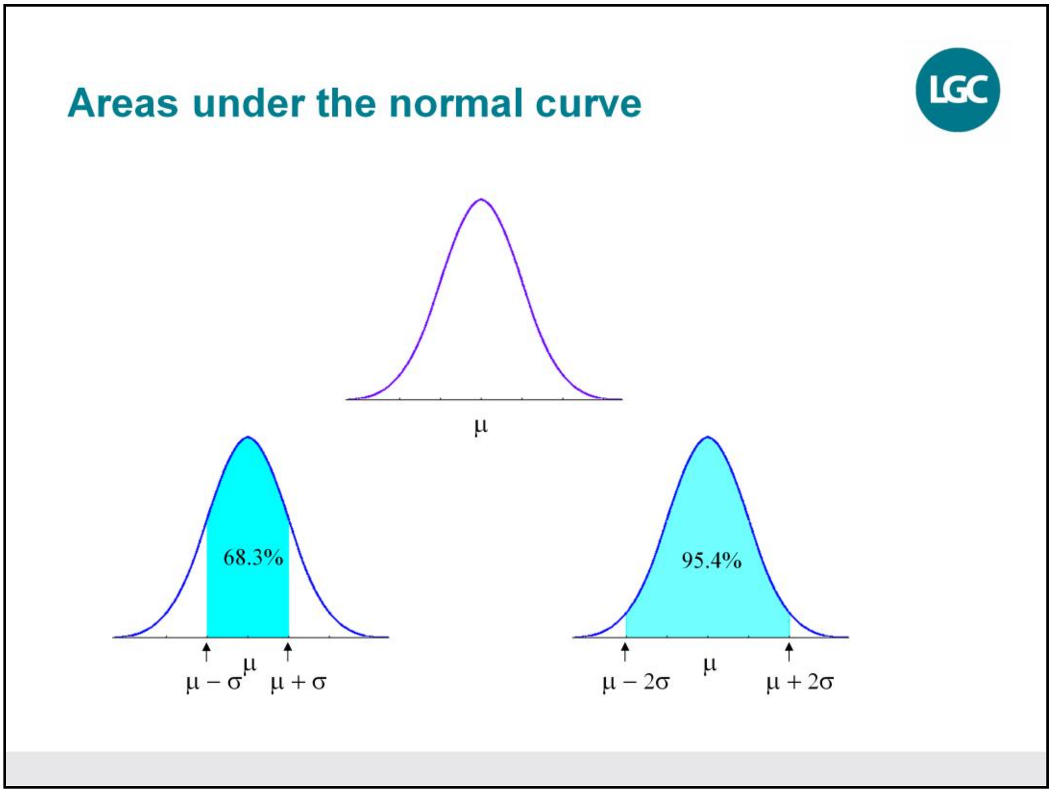
Distributions - *spread*

However, μ is not sufficient on its own to completely characterise a population, since several different populations could be located at the same point.



Describing a normal distribution

We therefore need a second parameter σ to measure the spread, or dispersion, of the population.

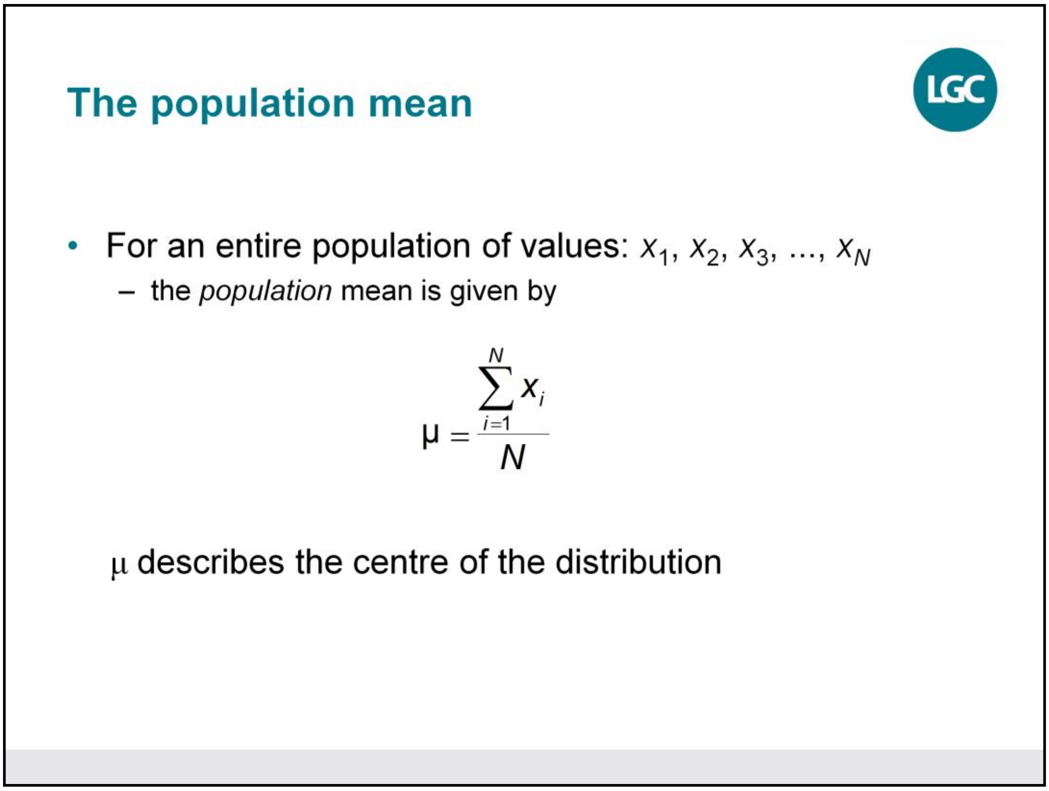


Areas under the normal curve

Areas under the Normal Curve

For any normal distribution, approximately 68% of the population values lie within ± 1 standard deviation of the mean; approximately 95% of the population values lie within ± 2 standard deviations of the mean; and approximately 99.7% of the population values lie within ± 3 standard deviations of the mean. Thus the bulk of a normal distribution is contained within ± 3 standard deviations of the mean. These values are summarised in the table below.

± σ	% population
1.00	68.3
1.64	90.0
1.96	95.0
2.00	95.4
2.57	99.0
3.00	99.7



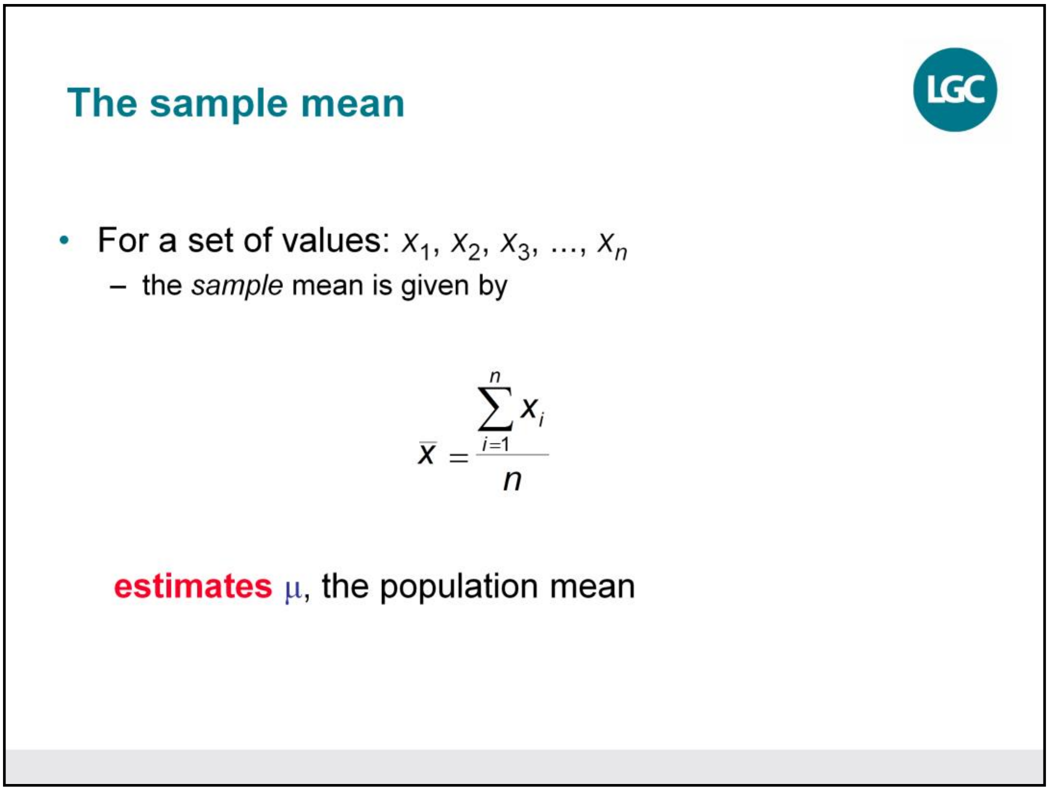
The population mean

- For an entire population of values: x_1, x_2, x_3, …, x_N
  - the *population* mean is given by

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

μ describes the centre of the distribution

The mean is simply the arithmetic average of the data - add them all up and divide by the number of data points. If we have the entire population of *N* data points, then this gives the population mean, μ.



The sample mean

- For a set of values: x_1, x_2, x_3, …, x_n
  - the *sample* mean is given by

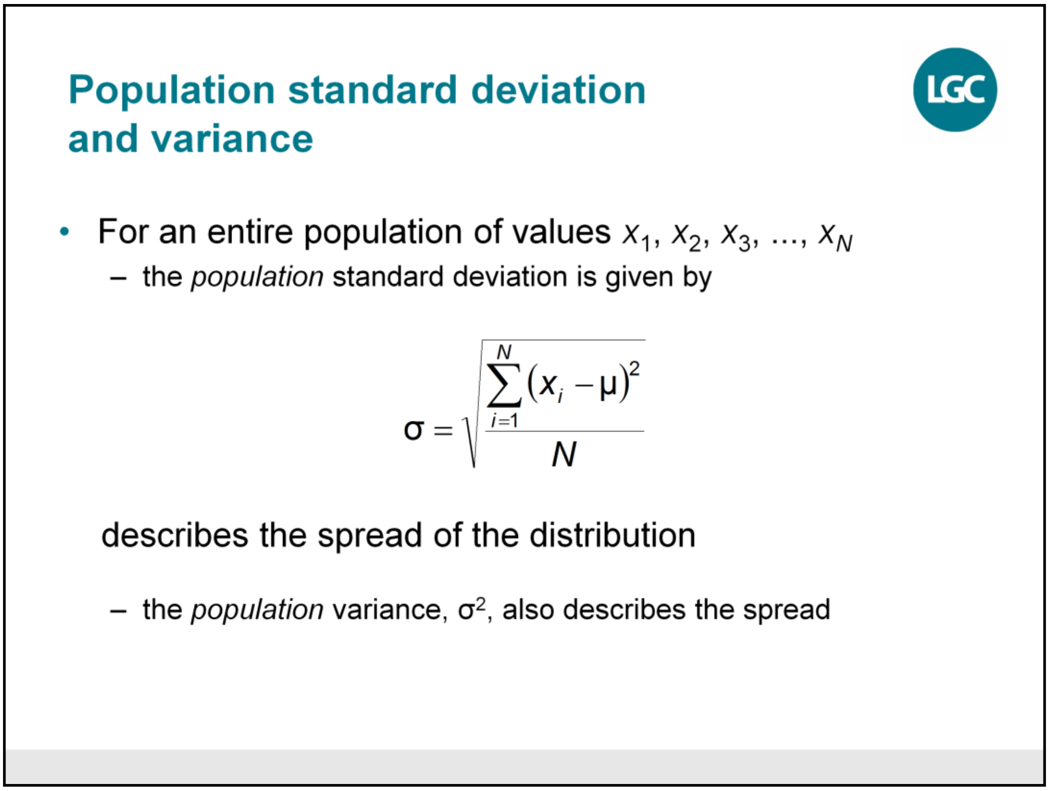
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

estimates μ, the population mean

However, as mentioned previously, we rarely have access to the entire population of values. We generally have a set of observations, *n*, which represents a *sample* of the population. Therefore, when we calculate the mean of *n* data points, we are estimating the population mean, μ. The sample mean is represented by $\bar{x}$.

In Excel, the mean of a set of data is calculated using the AVERAGE function.

Note: for most statistical calculations, 'samples' are taken to be *random* samples from a population. For most purposes, experiments are accordingly designed to ensure that samples are indeed sufficiently random. Some of the basic tests discussed later are simply tests for particular kinds of non-random behaviour; e.g. *t*-tests to check for *bias*.



Population standard deviation and variance

- For an entire population of values x_1, x_2, x_3, …, x_N
  - the *population* standard deviation is given by

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

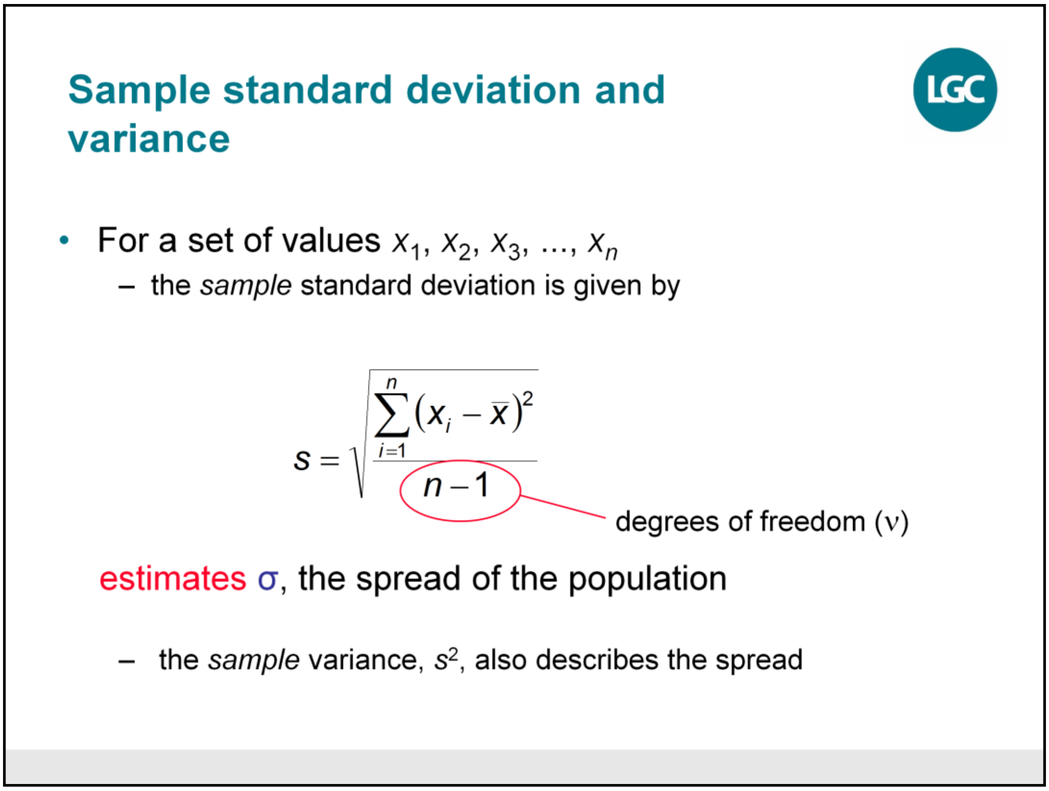
describes the spread of the distribution

  - the *population* variance, σ^2, also describes the spread

The population standard deviation σ is a measure of the spread of the entire population of data around the population mean - a *precision* measure. It is always in the same units as the mean.

Variance

The variance is simply the square of the standard deviation σ, and is also a measure of precision. It is an important parameter used in some statistical tests (e.g. the F test, discussed in a later lecture).



Sample standard deviation and variance

- For a set of values x_1, x_2, x_3, …, x_n
  - the *sample* standard deviation is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

degrees of freedom (ν)

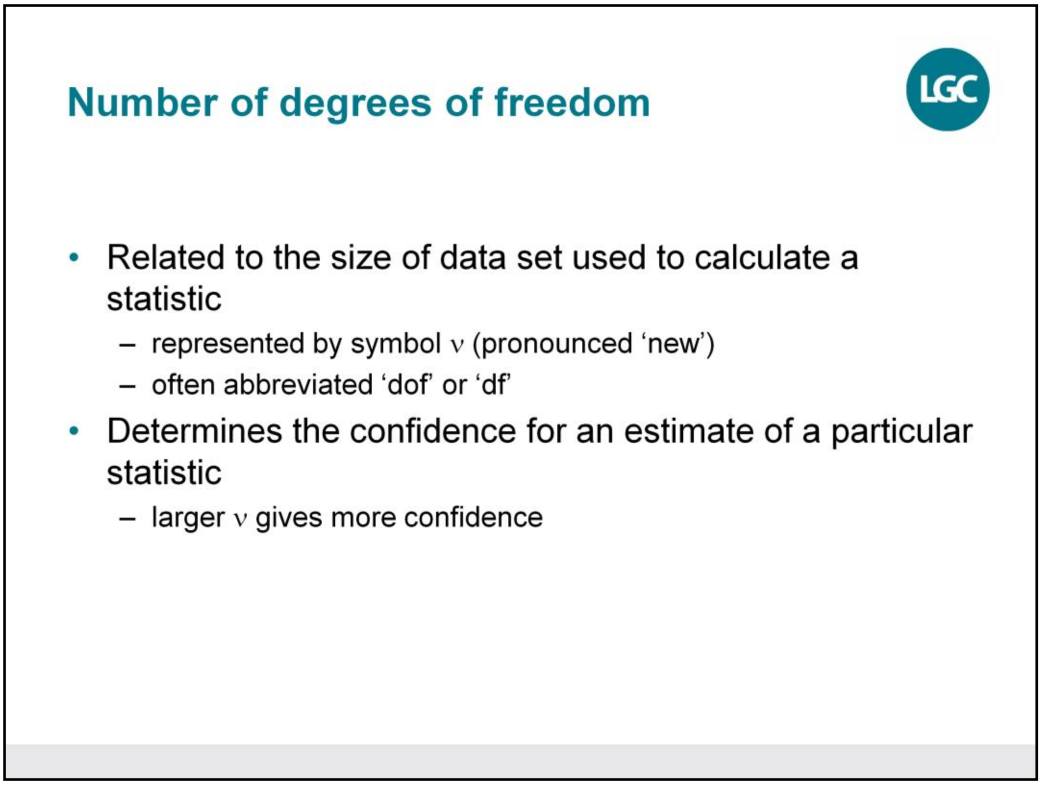
estimates σ, the spread of the population

- the *sample* variance, s^2, also describes the spread

As mentioned previously, we rarely have access to the entire population of data. Therefore, the standard deviation, *s*, calculated from a sample of *n* data points is an estimate of the population standard deviation, σ. During this course, and for most practical purposes, **we will be concerned solely with the sample standard deviation *s*.** On many calculators, *s* is denoted by $\sigma_{n\text{-}1}$. In Excel, the sample standard deviation of a set of numbers is calculated using the STDEV or STDEV.S function.

s is sometimes qualified for clarity; for example, s(*x*) would be the standard deviation associated with a value *x* or set of values x_1, x_2, ...x_n.

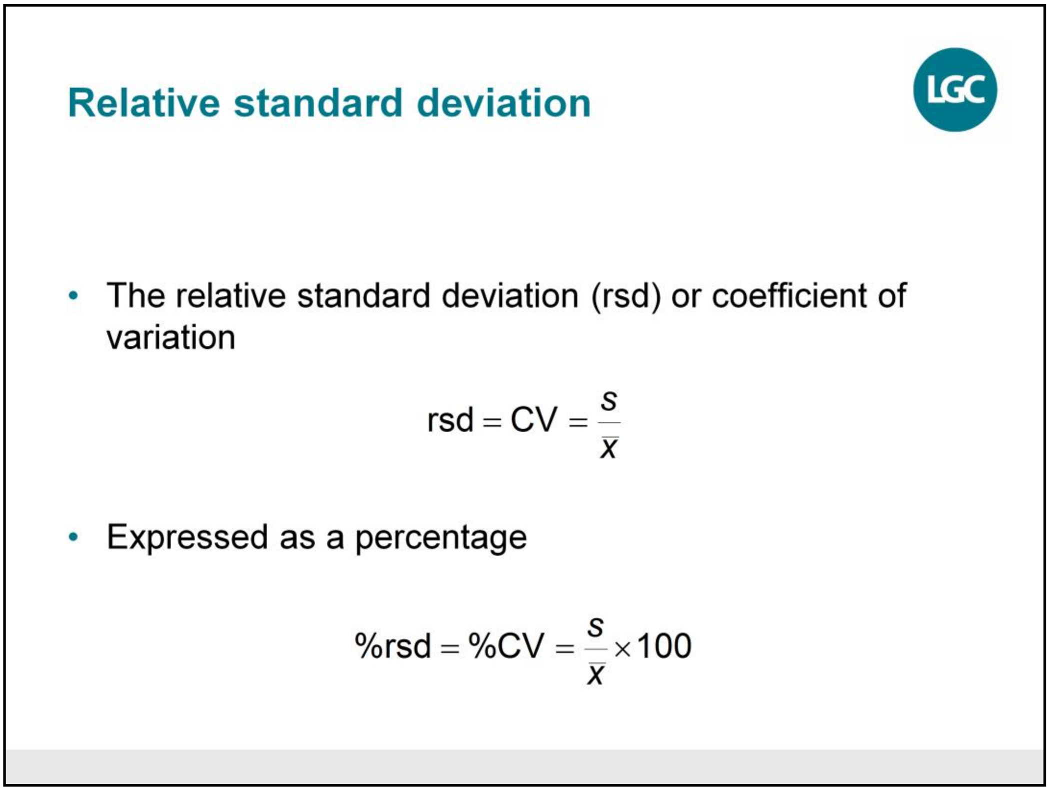
The figure *n*-1 in the formula for *s* is called the number of degrees of freedom. This is discussed on the next slide.



Number of degrees of freedom

- Related to the size of data set used to calculate a statistic
  - represented by symbol ν (pronounced ‘new’)
  - often abbreviated ‘dof’ or ‘df’
- Determines the confidence for an estimate of a particular statistic
  - larger ν gives more confidence

The number of degrees of freedom is important in helping to decide the confidence we have in estimates of statistics such as the standard deviation. It appears in many calculations. In general, the degrees of freedom is the number of data points (*n*) less the number of parameters already estimated from the data. In the case of the sample standard deviation, for example, $\nu = n$-1 as the mean (which is used in the calculation of *s*) has already been estimated from the same data.



Relative standard deviation

- The relative standard deviation (rsd) or coefficient of variation

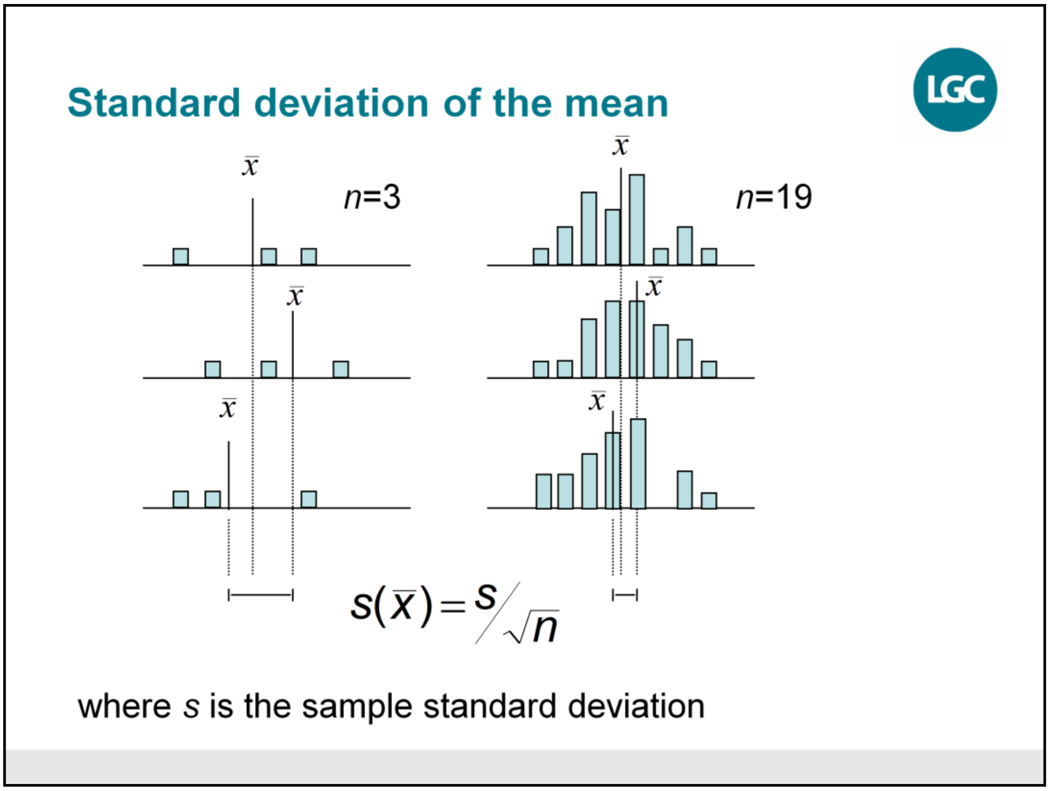
$$\text{rsd} = \text{CV} = \frac{s}{\bar{x}}$$

- Expressed as a percentage

$$\%\text{rsd} = \%\text{CV} = \frac{s}{\bar{x}} \times 100$$

Relative measures of spread are often used in chemistry, particularly where, for example, the spread of results seems to increase with concentration. In these circumstances, it is easier to quote a single relative measure than to calculate estimates for each different concentration.

The relative standard deviation (rsd) is a measure of the spread of data in comparison to the mean of the data. It is simply the standard deviation divided by the mean value. The rsd is also known as the *coefficient of variation,* CV. Either can be expressed as a percentage; the abbreviations are then usually %rsd or %CV.



Standard deviation of the mean

$$s(\bar{x}) = s/\sqrt{n}$$

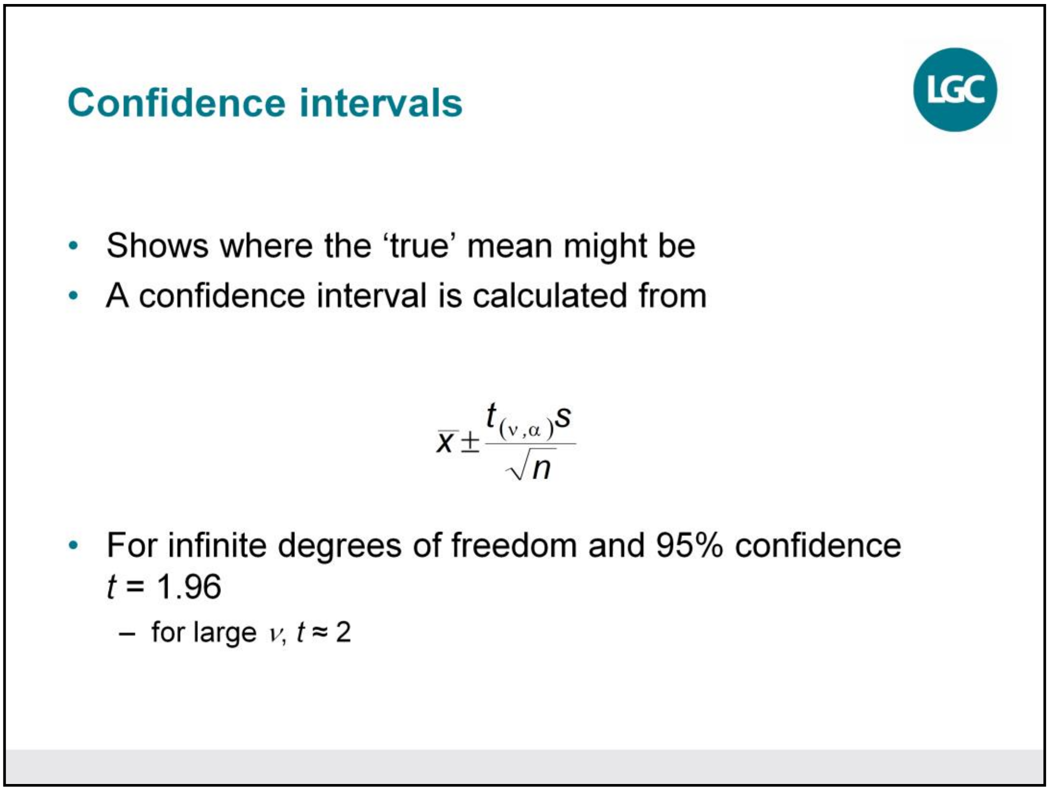
where *s* is the sample standard deviation

Given the same underlying spread of data (standard deviation *s*), as more data are gathered, one becomes more confident of the mean value being an accurate reflection of the ‘right answer’. This intuition is backed up by statistics; as the number of observations in each ‘sample’ increases, so the standard deviation of mean values becomes smaller. In fact, it reduces by a factor of $\sqrt{n}$. The value $s/\sqrt{n}$ is called the *standard deviation of the mean** sdm, or $s(\bar{x})$.

The standard deviation of the mean is always less than the sample standard deviation *s*. It forms an estimate of the uncertainty of the mean value.

When used in significance testing, the standard deviation of the mean has the same number of degrees of freedom as the estimate *s* on which it is based (usually *n*-1).

*Note: The standard deviation of the mean of *n* samples is often referred to as the *standard error* or the *standard error of the mean*. We will use ‘standard deviation of the mean’ throughout.



Confidence intervals

- Shows where the ‘true’ mean might be
- A confidence interval is calculated from

$$\bar{x} \pm \frac{t_{(\nu,\alpha)} s}{\sqrt{n}}$$

- For infinite degrees of freedom and 95% confidence t = 1.96
  - for large ν, $t \approx 2$

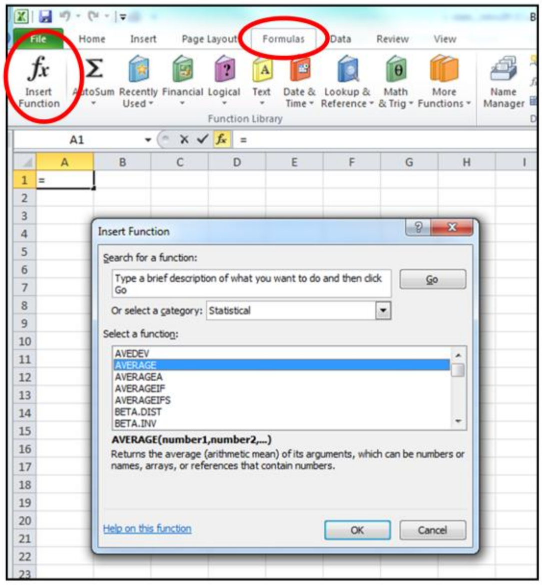
A confidence interval gives the range in which we believe, with a given level of confidence, that the true mean lies. It is calculated by multiplying the standard deviation of the mean by the appropriate value of $t_{(\nu,\alpha)}$.

$t_{(\nu,\alpha)}$ is the Student *t* value for ν degrees of freedom and a level of confidence given by α*, and is obtained from statistical tables. For a confidence level of 95%, α = 0.05 (see footnote).

Knowing the confidence interval, we can say, “The observed mean might have come from a true mean within this range in a fraction (1-α) of cases”. Equally, a mean value *outside* the same range would be expected only in a fraction α of similar experiments.

* In this notation, common in statistical tables, the confidence level is given in terms of α, the probability of a value being *outside* the range. For a confidence level in more familiar percentage terms, use 100.(1-α). For example, for 95% confidence, α is 0.05.

Excel 2007/2010 - Functions



Insert Function

Search for a function:

Type a brief description of what you want to do and then click Go

Go

Or select a category: Statistical

Select a function:

AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS
BETA.DIST
BETA.INV

AVERAGE(number1,number2,...)

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Help on this function

OK

Cancel

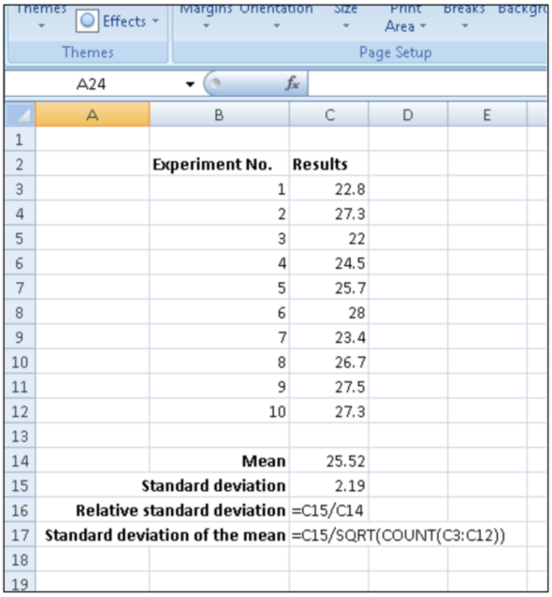
Excel contains a number of functions for calculating statistics such as the mean and standard deviation. You should first click on the cell where you want the answer to appear. The functions are then accessed by clicking on Insert Function under the Formulas ribbon.

In the Insert Function window, select Statistical from the dropdown menu. Scroll down the list in the Select a function window until you locate the required function. Click on the function, then click on OK. Alternatively you can use Search for a function to find the required function.

The functions required for the first workshop are:

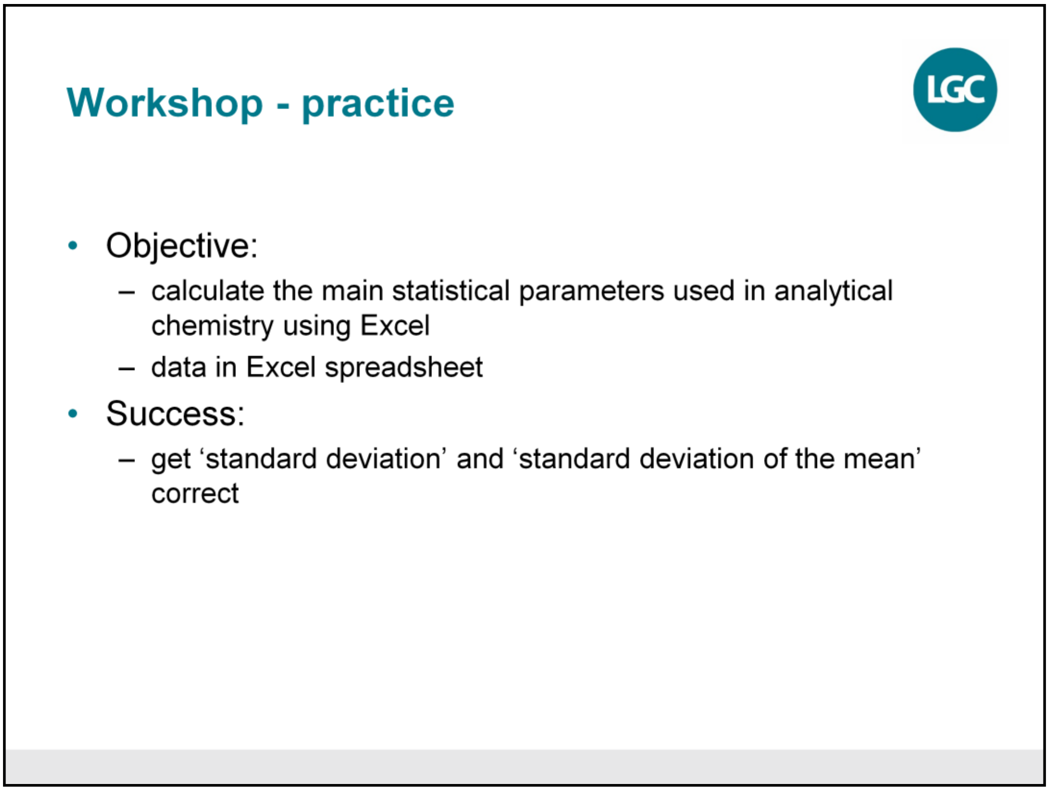
- AVERAGE – to calculate the mean of a set of data;
- STDEV (or STDEV.S) – to calculate the sample standard deviation of a set of data;
- COUNT – counts the number of cells, within a specified range, that contain numbers;
- SQRT – to calculate the square root of a number. This function is located under Math & Trig in the Function category window.

Simple calculations in Excel



	A	B	C	D	E
1					
2		**Experiment No.**	**Results**		
3		1	22.8		
4		2	27.3		
5		3	22		
6		4	24.5		
7		5	25.7		
8		6	28		
9		7	23.4		
10		8	26.7		
11		9	27.5		
12		10	27.3		
13					
14		**Mean**	25.52		
15		**Standard deviation**	2.19		
16		**Relative standard deviation**	=C15/C14		
17		**Standard deviation of the mean**	=C15/SQRT(COUNT(C3:C12))		
18					
19					

Some of the statistical parameters you will need to evaluate during the course cannot be calculated automatically by Excel. You will therefore need to enter the required calculation into the spreadsheet. To enter an equation, click on the cell where you want the answer to appear then type = followed by the equation. The equation can contain numbers, cell names or ranges, and other Excel functions (e.g. SQRT). The arithmetic operators used in Excel are +, -, * (for multiplication), / (for division) and ^ (to raise a number to the power of the number following the ^). Once you have entered the complete equation, press the return key.



Workshop - practice

- Objective:
  - calculate the main statistical parameters used in analytical chemistry using Excel
  - data in Excel spreadsheet
- Success:
  - get ‘standard deviation’ and ‘standard deviation of the mean’ correct