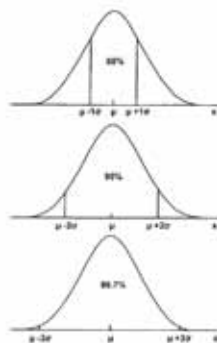
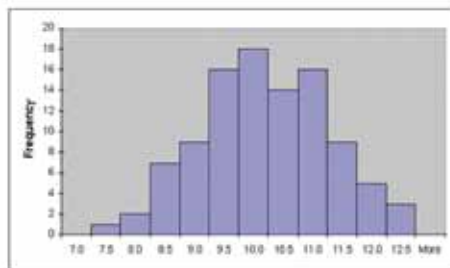


# Basic statistics

in QC-laboratory environment

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



$$SEM = \frac{s}{\sqrt{n}}$$

$$\bar{x} - z * \sigma / \sqrt{n} \leq u \leq \bar{x} + z * \sigma / \sqrt{n}$$

**Demo**

**© Lundén/Ello ab**

A self study course with  
examples and exercises

by Dr. Morgan Emtner

## Contents

0.	Getting started	5
1.	Concepts of basic statistics	7
2.	Probability	25
3.	Normal Distribution	37
4.	Samples	44
5.	Confidence Intervals	51
6.	T-test	59
7.	F-test	69
Annex A	MS Excel, Analysis Toolpak	74
Annex B	Progress table	75

**Demo**

**© Lundén/Ello ab**

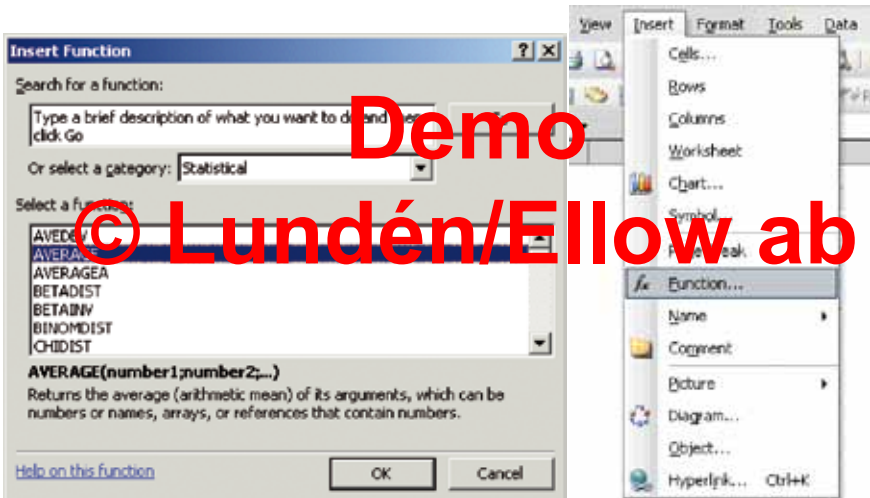


## Procedure for calculating the mean in MS Excel

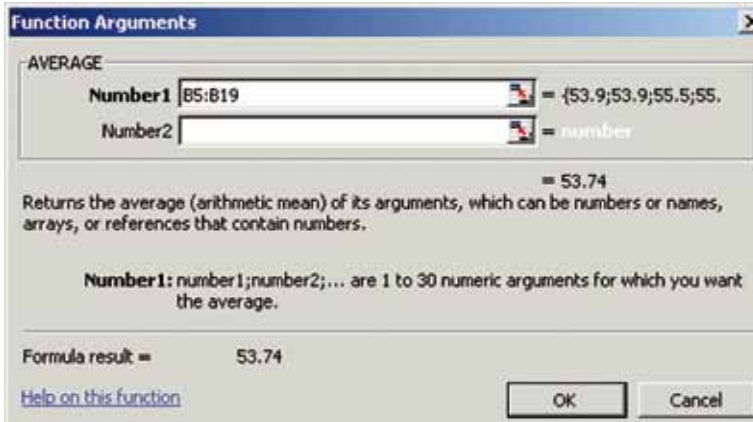
Using MS Excel to calculate the mean is very easy.

**!** **Remarks:** Before you start up MS Excel and go ahead with the first example, it might be advisable to create a new folder on your computer for saving all the files from this training course. You will find the files at [www.....](http://www.....)

- Open the file “Examples.xls” and select the spreadsheet “ExA1”.
- The results for HPLC column 1 are listed in column B on the sheet.
- Move your cursor to the cell where you wish to input the average value (e.g. cell B21).
- Go to the menu “Insert” and select “Function”.



- Select the category:”Statistical” and then function: “Average”. Click OK.



- In the “Number 1” field of the next dialog box, select the cells you want to calculate the average of (B5:B19). MS Excel usually automatically suggests a cell source. Please check that the suggested cells are the correct ones. (You can edit the source cells by moving the cursor onto the dialog box and left-clicking on the source cell.)
- Click OK and the result 53.74 (rounded off), will be displayed in cell B21.

(Don't close your Excel file yet because we will use it again soon to calculate the median.)

**© Lundén/Ello ab**

The average value is a good central measure for symmetrical distributions, but less suited to asymmetrical distributions or if one or two abnormal values exist.

If you take a closer look at the average value for HPLC column 2, you will observe what happens if the data contain deviated values. 12 of the 15 values are lower than the average value 53.7 and therefore it is not an appropriate measurement of where most of our observations are located.

## Median

Another central measure that can be useful for asymmetrical distributions, or if there are deviating values, is the median. If you take all values in order (ascending) and then determine the exact value that is in the middle you will get the result of the median. If the number of values is even, there is no value exactly in the middle. The median is then the average value of the two values in the middle.

## Summary

This is a summary of the most important items that we have covered in this chapter.

The **population** is all the objects you want to study.

The **sample** is a portion of the objects in the population.

A **variable** is a feature that describes the objects.

**Statistic inference** is used when drawing conclusions about a population with the help of a random sampling.

The two most common **measures of central tendency** are:

**Mean** - the most common one

**Median** - used when you have asymmetrical data or strongly deviating values.

The most common **measures of dispersion** are:

**Standard deviation** is the average of the deviations of individual values from the mean.

**Variance** is the square root of the standard deviation.

**Relative standard deviation** is the standard deviation expressed as a percentage of the mean.

**Demo**  
**© Lundén/Elbow ab**

## Diagnostic Test A:



Now it is time to test your knowledge on the concepts of basic statistics.

You will find the test at [www.....](http://www.....)

Click on “Diagnostic test A” to start the test.

Mark the right answer(s) of the presented alternatives for all questions on and you will then have the result on the screen.

You are welcome to use this book and MS Excel when answering the questions. Good Luck!!!

When you have passed your diagnostic test, you can move on to the next chapter: “Probability”. If you were “unlucky” when you performed the diagnostic test you can always review some portions of the material and then carry out the diagnostic test again.

## 2. Probability

Probability is a fundamental and useful concept in statistics, but if something is very likely it is not the same as if it were true. The fact that something is improbable does not mean that it has never occurred. You can calculate the probability of an event taking place, but you need some information in advance.



One sort of data you can use is the theoretical knowledge you have of a system. If you roll a dice and want to calculate the probability of getting a four, we can take advantage of the fact that all six sides of the dice have exactly the same probability of coming up. This means that a four will be come up once every six times. The calculated probability is then  $1/6 \approx 0.167$ , i.e. 16.7 %.

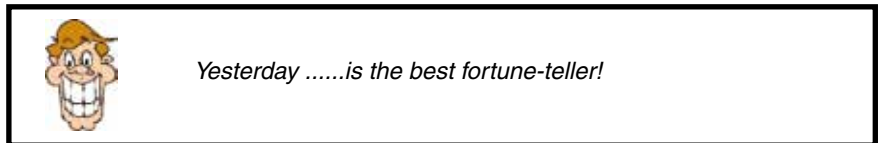
You can calculate the probability with the formula:  $p = m / N$

**m** is the outcome of interest (there is one side with four spots) and **N** is all results (there are six sides on the dice).

The **probability**  $p=0.167$  means that in approximately 17% of cases a four will come up when you roll a die over and over again (e.g. 1,000 times). The calculated probability is not as reliable in the short term. That means if you roll a dice six times it is not definite that you will get a four. The side that comes up is not influenced by the results that you have in any way so far. Even if you get 1 fours in a row, the probability of getting a four is still approx. 17%.

The probability is always a value between 1.00 (100% probability) and 0.00 (no probability).

In the case of dice, we already know all the conceivable outcomes, and they all have the same chances of occurring. In a lab, as in most other situations, the cases are more complex, so we need to use a different procedure. Here is an old adage:



The meaning of this is that if you want to find out the probability of an event, it is wise to study what has happened before.

**Example B1: Playing Darts**

John was an enthusiastic darts player. He was also an enthusiastic bettor and liked to bet whether he could get 8 points or more on his first throw.

Dart	Point	Dart	Point	Dart	Point	Dart	Point	Dart	Point
1	8	21	0	41	10	61	6	81	9
2	0	22	5	42	9	62	5	82	0
3	10	23	8	43	8	63	10	83	8
4	7	24	7	44	4	64	8	84	5
5	9	25	6	45	6	65	4	85	9
6	0	26	9	46	9	66	0	86	10
7	2	27	0	47	0	67	8	87	6
8	7	28	10	48	8	68	7	88	7
9	9	29	9	49	2	69	9	89	5
10	5	30	1	50	0	70	8	90	0
11	6	31	0	51	10	71	6	91	0
12	3	32	9	52	7	72	5	92	4
13	5	33	7	53	5	73	3	93	5
14	9	34	3	54	6	74	8	94	9
15	8	35	7	55	7	75	2	95	10
16	4	36	8	56	9	76	9	96	4
17	10	37	6	57	1	77	10	97	3
18	6	38	7	58	2	78	1	98	7
19	7	39	9	59	7	79	3	99	7
20	9	40	7	60	8	80	7	100	5



Table B1:  
Result table of 100 throws.

Would you enter that kind of a bet?

**Demo**

Before you did, you would have to find out how good John was at darts.

Take a closer look at the results table and see how he performed.

**© Lundén/Elow ab**

As mentioned before, it is good to plot the values from a table onto a graph.

It is often illustrative to draw a line chart with values on the x-axis and time on the y-axis.



**Procedure for creating a *line chart* in MS Excel**

Open the file “Examples.xls” on your computer and select spreadsheet “ExB1”. Here you will find the points John achieved on his last 100 throws.

Mark the area containing his point results (B7:B107).

Go to the “Insert” menu and then select “Chart”.

Select “line” and click “finish”.

- Mark the area containing his point results (B7:B107).
- Go to the “Insert” menu and then select “Chart”.
- Select “line” and click “finish”.