2 Introduction to Excel

2.1 Installation

You will have access to *Microsoft Excel* through your office365 student account at TU Dublin.

2.2 Week 1

2.2.1 Prelude: Descriptive Statistics

Large amounts of data or numbers are typically not easy to use or remember! We are looking for single numbers that are useful in revealing at least some aspect or information of the dataset. Imagine your data is a sequence of numbers. Useful numbers that describe the contents of a sequence are, for example, the minimum, and maximum. Knowing the minimum min, and maximum max of a data set, you can determine the range r as

$$r = \max - \min$$

Example: You have collected the age (in years) of a group of people

$$21, 18, 23, 24, 22, 21, 20, 23, 21, 20, 19, 19, 21$$

The age range is $r = 24 - 18 = 6$ as the minimum is 18 and the maximum is 24.

Let us look at further measures that can tell us more information about an array (or sequence) of data points. Sometimes, those numbers are referred to as summary measures as they summarise information into one number. A commonly used number is the mean. The mean is an average or a measure of location, because it tells where numbers are located on the number line.
Given a sequence of n numbers, $a_1, a_2, a_3, \ldots, a_n$, the mean is calculated as

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i,$$

i.e. the sum over all elements divided by the amount of elements.

Example: You measure five small trees and collect their heights as $t_1 = 17, t_2 = 23, t_3 = 31, t_4 = 25, t_5 = 14$, where t_1 is tree number one and so on. The mean height is

$$\bar{t} = \frac{1}{5}(17 + 23 + 31 + 25 + 14) = 22.$$

Another commonly used measure is the mode. The mode is the most commonly occuring value in a dataset. This can be particularly useful when the data are integers and nearly all the same.

Example: The best answer to the question "How many fingers does a human have?", is "Ten" which is the mode. Some people might have less, some have more, so that the mean might be 9.8 which isn't a useful answer.

Finally, we will consider the median. The median is the middle number when all numbers are arranged in order. It is a robust measure of location that is insensitive to outliers.

Example: Using the tree example, the ordered data is $13, 17, 23, 25, 31$, and the median is 23. It is robust to outliers because if instead of 31, the height of the largest tree would be 100, the median wouldn't change.

Exercise: You have collected data on the height (in cm) of a group of people

$$163, 175, 180, 157, 178, 176, 182, 181, 173$$

Determine the median and the mean height.[1]

Quartiles divide the data (arranged in order) into four equal parts.

[1] Answer: 176, 173.89

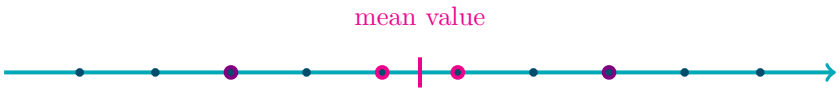
- The first quartile is the number that has 25% of the data to its left and 75% to its right.
- The second quartile is the median.
- The third quartile is the number that has 75% of the data to its left and 25% to its right.

Exercise: Generalising the definitions above, what do you think are the zeroth and fourth quartile?[2]

The definition of quartiles depend of the number of data points, in particular if the number is odd or even. For an **odd** number of discrete, ordered data points:

- The median splits the data set exactly in half
- Include the median in both halfs to determine the first and third quartiles

For an **even** number of discrete, ordered data points:



- The median is the mean value of the to middle points
- The first and third quartiles are the median of the respective left and right half

Exercise: You have data on the size of families in your area:

$$4, 2, 3, 7, 9, 4, 3, 3, 2$$

Determine the median and the third quartile.[3]

Central Tendencies

So far, we have looked at measures of location, but sometimes measures of dispersion, i.e the spread of data points, also tell us important aspects of the behaviour of our data.

Standard deviation is the typical error we make when approximating the data with the mean. It is computed as the root mean square deviation.

Example: Consider the five data points $22, 28, 25, 22, 23$ which have a mean of 24. To calculate the standard deviation, we compute:

Data	22	28	25	22	23
Deviation from the mean:	22-24 =-2	28-24 =+4	+1	-2	-1
Squared deviations:	$(-2)^2 = 4$	$4^2 = 16$	1	4	1

We then calculate the mean as the sum of all squared deviations divided by four $(4+16+1+4+1) \div 4 = 6.5$, and take the root to get $\sqrt{6.5} = 2.5495$. Looking back at the example, you might have noticed that we divided by four (and not five) to compute the mean. Why are we dividing by 4 (and not 5)?
This is because 4 is the number of independent comparisons. This also means that a standard deviation of only one data point does not make sense.

The variance of a data set is the standard deviation squared. For the example above, this would be 6.5. Just like standard deviation, variance measures dispersion. However, unlike standard deviation, variance has linear qualities, i.e. if variation arises from a number of independent sources the total variance can be computed as the sum of the individual variances.

[2] Answer: The zeroth quartile is the minimum value, and the fourth quartile is the maximum value.
[3] Answer: 3, 4

2.2.2 Excel's AVERAGE, COUNTing and ROUNDing

We want to use *Microsoft Excel* to perform some elementary calculations using a given data set. In *Microsoft Excel*, entries are saved in cells of a big matrix that is indexed with letters (columns) and numbers (rows). If you want to reference a cell from a different cell, you can do so using formulae. Formulae always start with a = sign, e.g. =A2+A3, means that you calculate the sum of the entries stored in cells A2 and A3. Luckily the software offers a lot of functions, that will help you write formulae. This section will introduce you to all functions that you will be working with in the laboratory exercise.

AVERAGE(start:end)	Calculates the average over a row or column of entries starting at start, and ending at end Example: =AVERAGE(A1:A3) calculates the average of the entries A1, A2, and A3.
AVERAGEIF(start:end,condition)	Calculates the average over all entries between start and end that fulfill the condition Example: =AVERAGEIF(A1:A3,">0") calculates the average of all positive entries between A1 and A3. Note that the condition is placed in inverted commas.
COUNT(start:end)	Counts the amount of entries containing numbers between start and end Example: =COUNT(A1:A3) counts the amount of numbers in cells A1, A2, and A3.
COUNTA(start:end)	Same as COUNT but for general values instead of only numbers Example: =COUNTA(A1:A3) counts how many of the cells A1, A2, and A3 are not empty.
COUNTIF(start:end, condition)	Counts the number of entries between start and end that meet the condition. Example: =COUNTIF(A1:A3,"=0") counts the number of zero entries between A1 and A3.
MEDIAN(start:end)	Calculates the median of entries starting at start and ending at end. Example: =MEDIAN(A1:A3) returns the median of the entries between A1 and A3.
MODE(start:end)	Calculates the mode of the entries starting at start and ending at end. Example: =MODE(A1:A3) returns the mode of the entries between A1 and A3.
ROUND(cell,number)	Rounds the value in cell to number decimals. Example: =ROUND(A3,3) rounds the entry in cell A3 to three decimals.
STDEV(start:end)	Calculates the standard deviation of the entries from start to end. Example: =STDEV(A1:A3) returns the standard deviation of the entries between A1 and A3.

2.2.3 Laboratory

Attempt to solve the exercise on your own. If you get stuck, use the channels described in section 1.4.2 to ask for help. The footnote provides a link to a video showing the full solution.

Exercise: The following exercise is based on a data record of blood concentrations of a drug administered to 20 volunteer subjects. For each subject the blood concentration of this drug is recorded at 0 hours (baseline) and then at 4 hour intervals over a 24 hour period.

The data has been compiled into the file MATH4002_BloodConcentration.xls which also contains hints. Please attempt the following, using the hints in the spread sheet as needed. Take note of any problems that you encounter along the way.

1. Use the Excel function COUNTA to count the number of subjects.
2. Use the Excel function AVERAGE to compute the average for each subject (column H)
3. Use the ROUND function to round the subjects averages to the nearest whole number, i.e. with no decimal places.
4. Calculate the average concentration for (a) each time point and (b) over all time points (i.e. the global average concentration observed).
5. The trial protocol requires that any subject that has an average rounded concentration of 105 or above is labelled as "high risk" and is labelled as "normal" otherwise. Use the logical IF function to create a label for every subject

   =IF(I3>=105, "high risk","normal")

   This reads as follows: if the value in the cell I3 is greater than or equal to 105 then return the value high risk else return the value normal.

   A side note on comparison operators available:

= equals	< less than	> greater than
<= less than or equal to	>= greater than or equal to	

6. Calculate the average concentrations for: (a) all subjects with normal risk profiles, (b) all subjects with high risk profiles. This requires the use of the AVERAGEIF function

   =AVERAGEIF(I2:I5,"normal",J2:K5)

   This reads as follows: if the value in the cells I2 to I5 are normal, a corresponding row between J and K will be added to the average, e.g. IF I3 is normal, say, it adds J3 and K3 to the average.
7. The study protocol also requires that the rounded average for each subject is expressed as a percentage of the average for normal subjects. Calculate the column – don't forget to use an absolute reference for the overall average. Use the format>cell command window to format these results as percentages displayed to the nearest whole percent.
8. Now use the IF() function, so that for each subject this column has the label 'Discontinue Trial' if their average concentration is greater than 120% of the average for normal subjects.

[4]

[4] Answer: https://youtu.be/ib-MrxIlW-g