# Excel – Getting the basics right, computer lab

## Introduction

A spreadsheet is an interactive computer application for organization, analysis and storage of data in tabular form. Spreadsheets are developed as computerized simulations of paper accounting worksheets. The program operates on data entered in cells of a table. Each cell may contain either numeric or text data, or the results of formulas that automatically calculate and display a value based on the contents of other cells. A spreadsheet may also refer to one such electronic document.

Spreadsheet users can adjust any stored value and observe the effects on calculated values. This makes the spreadsheet useful for "what-if" analysis since many cases can be rapidly investigated without manual recalculation. Modern spreadsheet software can have multiple interacting sheets, and can display data either as text and numerals, or in graphical form.

Besides performing basic arithmetic and mathematical functions, modern spreadsheets provide built-in functions for common financial and statistical operations. Calculations such as net present value or standard deviation can be applied to tabular data with a pre-programmed function in a formula. Spreadsheet programs also provide conditional expressions, functions to convert between text and numbers, and functions that operate on strings of text.

Spreadsheets have replaced paper-based systems throughout the world. Although they were first developed for accounting and bookkeeping tasks, they are now used extensively in any context where tabular lists are built, sorted, and shared.

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

In this practical you will learn how to use MS Excel to manage, analyse and visualize data.

You are expected to read, think and then execute the instructions in this practical. Before asking a question make sure that you have carefully read the item in question. You would be surprised how often the answer we seek is at the end of the paragraph or just around the corner of your next critical thought.

This practical will guide you through an exercise meant to cover some important aspects of Excel. You will practice how to input, format and rearrange data as well as filter and sort it. You will work create different types of charts. Finally you will work advanced formulas like t-tests and fold change calculations.

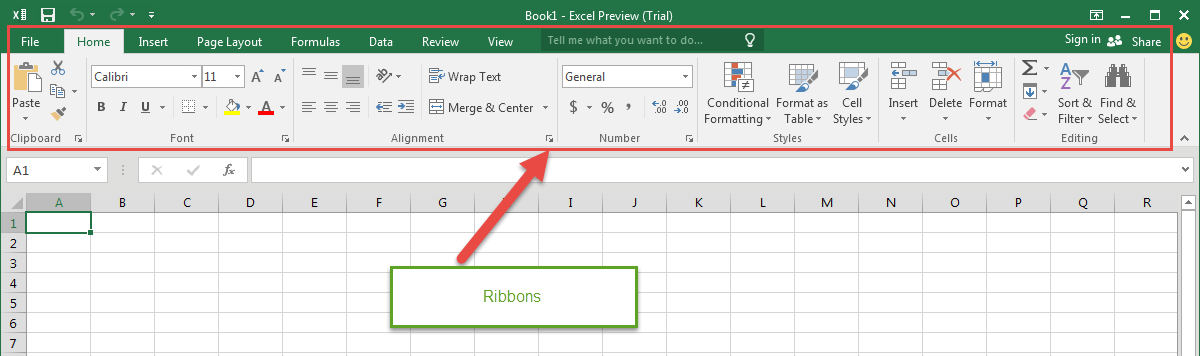
You will find several tasks in this exercise. You are required to perform them and submit them in an excel file to the following mail:

To: [luciano.fernandez@gu.se](mailto:luciano.fernandez@gu.se)

Subject: Excel Practical by [your name].

## GUI Overview

Ribbons are designed to help you quickly find the command that you want to execute in Excel. Ribbons are divided into logical groups called Tabs and each tab has its own set of unique function to perform. There are various tabs – Home, Insert, Page Layout, Formulas, Date, Review and View.



## Module 1: Data Input

### Excel provides you with different means to input data into a spreadsheet. You can enter data manually by typing text or number in cells. You can copy and paste data from other documents and use the *Text to Columns* feature to break rows into columns given a delimiter. You can use Excels import functions like *From Access* (import from a Database), *From Web* (import from HTML)*,* and *From Text* (import from text). Last but not least you can use the *From Other Sources* function to bring your data from an assorted number of sources like SQL, XML or OData.

### You can find the relevant input functions on the Data tab.

### Module Tasks

* Open Excel and start a new Workbook.
* Go to the course GitHub page:
  + https://github.com/bcfgothenburg/HT19/wiki/Excel-101-2019
* Navigate to the Module 1 page.
* Input the data into an Excel worksheet called Module 1 using two different methods.
* Make a note under each table explaining what methods you used.

### Module Achievements

After successfully completing the module tasks you should now be familiar with the different ways to enter data into Excel.

## Module 2: Arranging and formatting data

Excel has a very intuitive point-and-click interface that allows you to arrange your data as you need it. For example, you can change the size of the rows and columns to better fit your data. Additionally you can move the contents of a cell from one position to another. Moving a cell will update its position if references from other cells exits.

Excel also provides you with a plethora of functions that will help you customize the looks of your cells. You can adjust the colors of both text and background of a cell, change the type and style of a font, e.g. bold or italics. Adjust the alignment of the text in reference to a cell or multiple cells. Change the style of the numerical contents of a cell like making it a percentage, controlling the decimal display or displaying the numbers as a currency.

You can find all the formatting functions in the Home Tab arranged as different groups in the ribbon.

### Module Tasks

* Create a new worksheet and name it Module 2.
* Copy the table from the Module 1 sheet into this sheet.
* Add a header to your table describing the origin of your data (Framingham Heart Study).
* Make sure that the header spans multiple columns so that it covers the entire table.
  + Hint: Look at the *Merge and Center* function.
* Change the order of the columns in the table so that **Entr** and **exit** **dates** are last.
* Make sure the header of the table is bold and color it to distinguish it from the data.
* Increase the size of the font for the headers so that is larger than the data itself.
* Change the color of every other row in the data portion of the table as to make it more readable.
* Add decimals to the **SBP** and **DBP** columns.
* Remove the cause of death column.

### Achievements

After successfully completing the module tasks you should now be familiar moving entire columns or rows as needed and formatting your cells to add distinction and better present your data.

## Module 3: Conditional formatting

Spreadsheets are the go-to information hub for individuals, teams, and organizations of all sizes to store data. However, large data sets can make it difficult to identify trends, pick out key pieces of information and track deadlines. Luckily, Excel has a feature called conditional formatting that alleviates some of these challenges. Conditional formatting quickly highlights important information in a spreadsheet. It enables you to apply special formatting to cells in your spreadsheet that meet certain criteria. Excel has a sizable library of pre-set conditions that you can apply fairly simply, or you can create your own conditional formatting rules using Excel formulas.

Conditional formatting enables spreadsheet users to do a number of things. First and foremost, it calls attention to important data points such as deadlines, at-risk tasks, or budget items. It can also make large data sets more digestible by breaking up the wall of numbers with a visual organizational component. Finally, conditional formatting can transform your spreadsheet (which previously only stored data) into a dependable “alert” system that highlights key information and keeps you on top of your workload.

You can find the conditional formatting functions on the Home tab.

### Module Tasks

* Create a new worksheet and name it Module 3.
* Copy the table from the Module 1 sheet into this sheet.
* Using conditional formatting highlight all the cells where the patients are dead.
* Use color scales to color the Systolic blood pressure entries using red for highest values, blue for the lowest and white for the middle values.
* Use data bars to show the Diastolic blood pressure.

### Achievements

After successfully completing the module tasks you should now be familiar in using conditional formatting to dynamically highlight important data based on its contents.

## Module 4: Filtering and sorting

A Microsoft Excel spreadsheet can contain a great deal of information. Excel gives you the ability to analyse and work with an enormous amount of data. To most effectively use this data, you may need to manipulate it in different ways. Sorting and filtering data in various ways enables you to most effectively and efficiently use spreadsheets to locate and analyse your information.

You can find the sort and filter group on the Home tab.

### Module Tasks

* Create a new worksheet and name it Module 4.
* Go to the course GitHub page:
  + https://github.com/bcfgothenburg/HT19/wiki/Excel-101-2019
* Navigate to the Module 4 data page.
* Input the data into the worksheet.
  + Hint: CVS stands for Comma Separated Values
* From this new dataset we would like to know how many male subjects there are.
  + Use the filtering function to find this information.
    - Insert a new row before the table and record this.
* Sort the unfiltered table by **age** and **BMI**
  + Hint: Make sure the fields you are sorting by are properly formatted as numeric values and not text, otherwise you will get different results.
  + Hint: A filtered dataset can only sort one column at a time, you would need a different function if you want your dataset sorted by two or more criteria.

### Achievements

After successfully completing the module tasks you should now be familiar with filtering and sorting your dataset by using the filter and the custom sort.

## Module 5: Charting

A simple chart in can say more than a sheet full of numbers. It can often be difficult to interpret Excel workbooks that contain a lot of data. Charts allow you to illustrate your workbook data graphically, which makes it easy to visualize comparisons and trends.

A chart is a powerful tool that allows you to visually display data in a variety of different chart formats such as Bar, Column, Pie, Line, Area, Doughnut, Scatter, Surface, or Radar charts. With Excel, it is easy to create a chart.

You can find the Charts group on the Insert tab.

### Module Tasks

* Create a new worksheet and name it Module 5.
* Copy the table from the Module 1 sheet into this sheet.
* Using this table as data
  + Create a bar chart showing the Systolic blood pressure.
    - The x axis should use the patient ID as labels.
    - The y axis should use the **SBP**.
    - Title the chart with Systolic blood pressure.
  + Create a scatter chart plotting the Systolic blood pressure against the Diastolic blood pressure
    - Add a trend line and show the R2 value.
      * A linear trend line will do for this case.
    - Indicate under the chart as a note if there is a correlation between the two variables.
    - Add axis tiles to both axis showing the variables they represent.
    - Tile this chart as Correlation between **SBP** and **DBP**.
* Copy the table from the Module 4 sheet into this sheet.
  + Using this table as data
    - Create a scatter chart plotting the Systolic blood (**sysBP**) pressure against the Diastolic blood pressure (**diaBP**).
      * Add a trend line and show the R2 value.
      * Indicate under the chart as a note if there is a correlation between the two variables.

### Achievements

After successfully completing the module tasks you should now be familiar in creating different types of charts. You should be able to modify them and to add new chart elements if needed.

## Module 6: Formulas and functions – Calculating new variable in your data

One of the most useful capabilities of Excel is the ability to generate or calculate new variables from the original data. Excel allows you to create new formulas and use functions. A formula is an expression which calculates the value of a cell. Functions are predefined formulas and are already available in Excel.

Excel offers a plethora of functions. They can be displayed and selected by clicking the *fx* symbol at the left end of the formula bar (between the ribbon and the spreadsheet) after you select a cell. You can also type a known function directly into the formula bar or the cell. A function always begins with an equal (=) sign followed by the function name with the arguments for the function (indicated by cell coordinates such as C2) in parentheses. E.g. =SUM(C2:C18).

### Module Tasks

* Create a new worksheet and name it Module 6.
* Copy the table from the Module 4 sheet into this sheet.
* We want to calculate difference between systolic (**sysBP**) and diastolic (**diaBP**) blood pressure and call it **BPdiff**.
  + Insert a new column after the DBP and name it **BPdiff**.
  + Create a formula to calculate this difference and apply it to all the rows
* The current table has a field called male where *1* means the record belongs to a male and *0* means it does not.
  + We want to create a new variable named “gender” between the male and the age columns. This new variable will show a “M” for male records and an “F” for female records.
    - Hint: The *IF* function will be very helpful for this task

### Achievements

After successfully completing the module tasks you should be familiar in calculating new variables from your data by making use of formulas and functions from Excel.

## Module 7: Fold change

Fold change is a measure describing how much a quantity changes going from an initial to a final value. For example, an initial value of 30 and a final value of 60 corresponds to a fold change of 1 (or a change to 2 times), or in common terms, a one-fold increase. Fold change is calculated simply as the ratio of the difference between final value and the initial value over the original value. Thus, if the initial value is A and the final value is B, the fold change is **(B - A)/A** or equivalently **B/A - 1**. As another example, a change from 80 to 20 would be a fold change of -0.75, while a change from 20 to 80 would be a fold change of 3 (a change of 3 to 4 times the original).

### Module Tasks

* Create a new worksheet and name it Module 7.
* Go to the course GitHub page:
  + https://github.com/bcfgothenburg/HT19/wiki/Excel-101-2019
* Navigate to the Module 7 data page.
* Input the data of the two datasets into the worksheet.
* Notice that both datasets contain expression values for the same gene IDs. Align them in your worksheet so that you have 3 columns, ID\_REF, Mock, and H1N1.
* Create a new variable called fold change and calculate the fold change between the mock and the infected sample.
* Highlight all the samples that have at least a twofold change.
  + Sort the rows so that you can easily spot the most affected genes.
  + Hint: If you sort by color, you will be able to cluster both the positive and negative changes.

### Achievements

After successfully completing the module tasks you should now be familiar with how to calculate the fold changes between two variables and finding rows that fulfil specific criteria.

## Module 8: Statistical functions

### Research involves summarizing data (descriptive statistics) and testing hypotheses (inferential statistics). Although simple statistical calculations can be done in Excel, you usually need to export data entered into Excel to a statistical software package to do more sophisticated analyses.

#### Descriptive statistics on continuous variables

Continuous variables (variables with many numerical values) are often summarized by calculating the mean (and/or median) and measures of dispersion such as the minimum and maximum values (defining the range) and the standard deviation.

Excel provides you with many functions allowing you to perform such calculations like *AVERAGE, STDEV, MIN, MAX, etc.*

#### Frequency counts on categorical variables

For variables with a few categories (categorical variables), frequency counts of each value of the variable are the most common way of summarizing that data. Excel provides functions like *COUNTIF, FREQUENCY, DCOUNT etc.*.

You can also do inferential statistics. Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population. For example, there are *T.TEST* and *CHISQ.TEST* functions. The T-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.

### Module Tasks

* Create a new worksheet and name it Module 8.
* Copy the table from the Module 4 sheet into this sheet.
* We want to calculate the Min, Max, Mean and Standard Deviation of the systolic (**sysBP**) and diastolic (**diaBP**) blood pressure
  + Hint: For the standard deviation, calculate it using the sample rather than the population variant of the function.
  + Create a summary table on top the data table where these statistics are annotated.
* We want to expand the summary table above with the total number of patients that have **diabetes**, **prevalentStroke** and **prevalentHyp**.
  + Hint: Use the COUNTIF function to achieve this goal.
* We want to test the null hypothesis that the mean value between the BMI’s of male smokers that have prevalent hypertension vs the male smokers that don’t have prevalent hypertension is the same.
  + Prove or disprove this hypothesis using a student T-test.
    - We know that the BMI’s of this dataset is normally distributed and the variance between the samples is unequal.
  + Hint: You may find it useful to filter the dataset for the desired conditions and copy the filtered set to a new sheet

### Achievements

After successfully completing the module tasks you should be familiar with many statistical functions Excel proves you with. In addition, you would have demonstrated your capacity to filter a dataset to extract specific subsets to perform advanced tests.