PC Lab 1

Revision of Basic Statistical Concepts and Using Excel for Data Analysis

Welcome to your first PC lab for PUBH620! In this class we will be going over several basic statistical concepts. Your lecturer/tutor will introduce themselves to you and you will have the chance to get to know the students in your class.

**Ice Breaker Activity**

1. Form groups of two or three and give your group a name.
2. Have one group open up Excel on their computer and setup a table like the one shown below. Your tutor will do a brief demonstration of how to do this.
3. For each member of the group, record their name, gender, age, country of birth, occupation (or previous study background) and whether they use an iPhone or Android phone.
4. Once your group has completed collecting their data, combine your data with two other groups in the class to form a larger group. You may wish to use the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Gender** | **Country** | **Age** | **Occupation** | **iPhone or Android?** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

1. Can you identify what scales of measurement each of the above categories (Name, Gender, Country, Age, Occupation, Smartphone OS) are?
2. Calculate the mean age of your large group.
3. Reorder the ages from lowest to highest and determine the median age.
4. Is there a mode age? If so, what is it?
5. Which is more popular in your large group? iPhone or Android?
6. What percentage of people in your group use iPhone? What percentage use Android?
7. Spend about 10 minutes (total) with your large group sharing about why you chose to study Public Health and what your hopes and ambitions are after completing your degree.

**Numerical and Arithmetic Exercises**

Let’s start by learning how to input formulas into Excel.

The equals “=” sign is used to initiate a formula in Excel. When entering in an equation to be solved, you must always start with the “=” sign. For example, if you wanted to compute the simple equation 3 x 7 then you would write:

=3\*7 into one of the cells on your spreadsheet.

Note here that when doing multiplication, “\*” has been used. Can you find out what symbol you would need to use to do division? Addition, subtraction and rounded brackets should be easy to find on your keyboard.

Try entering in the below equations into Excel to get some practice with this:

1. 5 + 4 – 3 x 2 – 1 =
2. 7 x 6 – 5 + 4 + 3 =

Excel also has in-built Excel functions that can be used to do simple arithmetic. Some common ones are =Product(), =Sum(), =POWER(). See if you can work out what each of these in-built functions does. Write your answers below:

=PRODUCT() –

=SUM() –

=POWER() –

What if you wanted to do squares and square roots? Or cubes and cube roots? The caret “^” symbol is used to compute powers or orders. For example, if you wanted to calculate 7 squared, then you would write:

=7^2

Same applied for other powers such as 7 cubed:

=7^3

Lastly, if you wanted to compute square roots, there are two simple ways to do this. Let’s calculate the square root of 49 (which of course we know is 7).

=sqrt(49)

OR

=49^(1/2)

A square root is the same as raising to the power of a half. How do you think you would compute a cube root? Test this out!

**Now go back to the icebreaker activity above and complete questions 6 to 10 using Excel.**

**The Standard Deviation Formula**

The following are a set of heights, in centimetres, from a small student population:

Shape

Description automatically generated178 160 180 182 165 159 172 191

Calculate their mean and standard deviation (s.d.) (Try to do this manually using Excel).

The circled numbers (e.g. ➂) and arrows are there to guide you in the order of the calculations. Compare your answer for the standard deviation with someone next to you or you can check your answer using either =stdev.s or =stdev.p, what is the difference? Which one should you use?

**Analysing Random Data**

Now we get into some of the fun we can have in Excel. Data analysis!

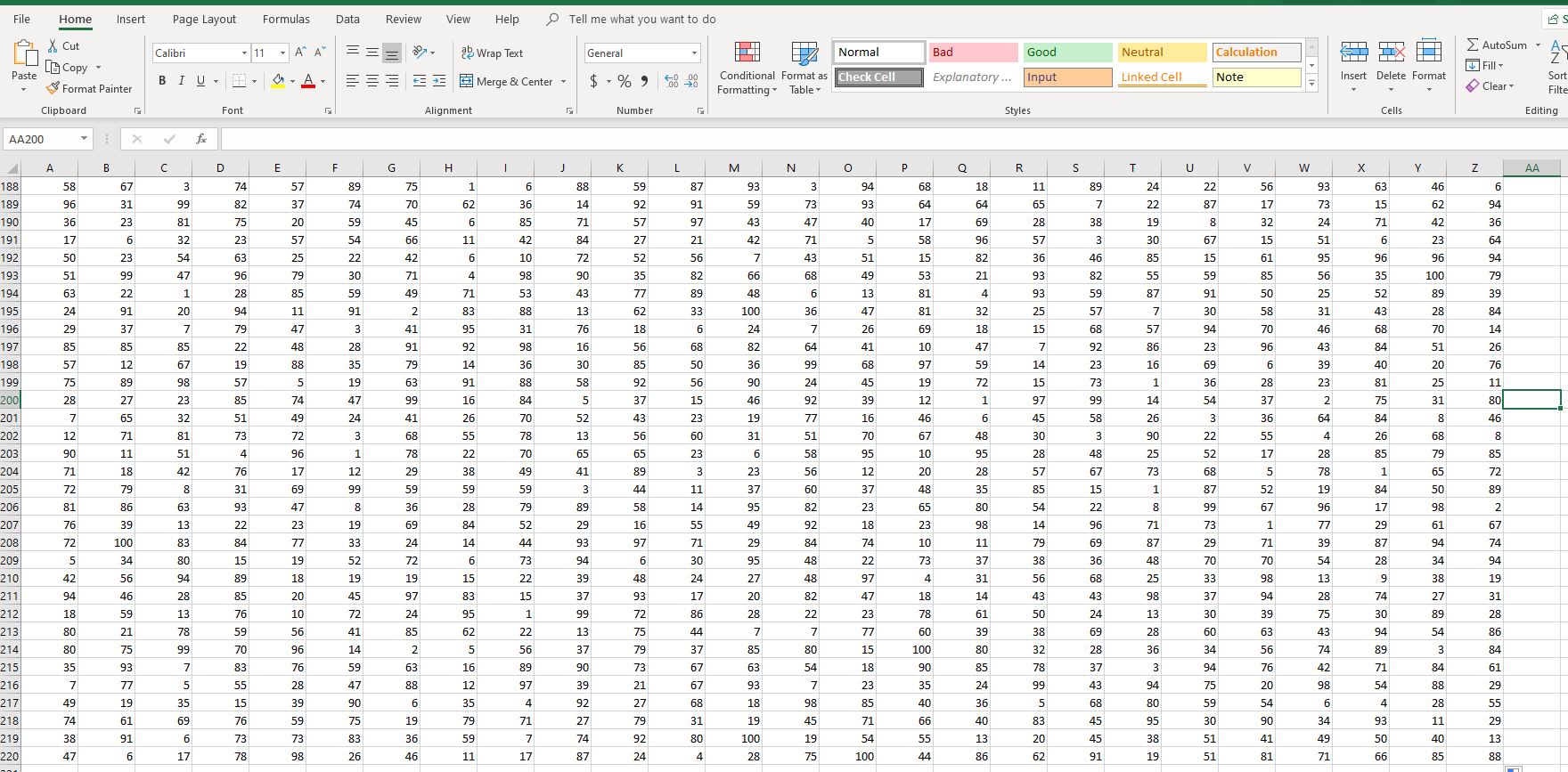
In a new tab (ask your tutor if you need help with creating new tabs in Excel), we want to create a randomly generated data set from cells A1 to Z220. Before we start, rename the new tab to “RANDOM DATA”. It is always good practice to label your work so that when you come back to it later, you will be able to understand what you have done.

Excel has in-built Excel functions to generate a set of random numbers. Select the first cell A1 and use the =RANDBETWEEN(1,100) function in just this single cell. You will see a number pop up in the cell, this number has been randomly generated and values have been chosen between 1 and 100. You can type the formula again into another cell and you will see that a new number between 1 and 100 is generated again, furthermore, the number in cell A1 would have changed also (again selecting a number between 1 and 100).

You have been asked to generate numbers from cell A1 to Z220. You do not want to do this manually for each cell! What you can do instead is learn how to use the “fill adjacent cells” function. Make sure you learn how to use this feature. If you need help with this, use the Excel help file or consult your tutor.

Your data should look like Figure 1 below but due to the random generation, your numbers will be different. If you get stuck, use the Excel help file or consult your tutor.

Remember to save your file so you can come back to it later.



**Figure 1: Randomly generated dataset in Excel**

Calculate the sample mean, standard deviation, variance, range, minimum, maximum and sample size of your dataset. Familiarise yourself with typing in the cell references as opposed to highlighting the cells, this will save you time.

1. What Excel commands did you use? Write them down.
2. What happens if you use =MEAN in Excel to try to calculate the mean?
3. How do you distinguish between a sample standard deviation/variance and a population standard deviation/variance in Excel?
4. How did you determine the sample size of your dataset? Was there an Excel formula you could use? (HINT: look for an Excel formula that allows you to count the number of cell items you have).
5. Experiment with an Excel feature called ‘conditional formatting’. How might this be useful to you in the future? Use the Excel help file if you get stuck.
6. What does =RAND() do? What type of values does it give you?
7. What does hitting F9 on the keyboard do to your dataset?

**Working with Statistics Formulas**

Below are three formulas which you will encounter during the semester. Do not worry too much about the meaning of these formulas right now, the focus here is to brush up on your numerical and arithmetic skills! **You are encouraged to use Excel to complete these tasks.** Start a new sheet in your Excel file and call it ‘t-test’. Copy the data in table 1 below to your new t-test sheet.

Use the data in table 1 below:

**Table 1: Cortisol levels of two different student groups**

|  |  |
| --- | --- |
| **Stressed students’ cortisol levels (mcg/dL)** | **Non-stressed students’ cortisol levels (mcg/dL)** |
| 33 | 10 |
| 28 | 18 |
| 35 | 11 |
| 26 | 13 |
| 25 | 21 |
| 38 | 20 |
| 31 | 7 |

1. Calculate:
   1. Standard error of the mean for *stressed students* only:

A picture containing text

Description automatically generated

* 1. Z-score for the observation () *non-stressed student No. 2 (row 2)*:



* 1. The value of t is:

This will really test your ability to input mathematical formulas into Excel! Your value for t should be 6.058…did you get it? If not, make sure you have defined each of the variables (write these down) in the above formula and consider using multiple cells for each part of the fractions.

**Types of Data**

You may need to use an internet search for the below exercises.

1. In the context of the influenza virus, list five variables that are:
   1. Measured on a nominal scale –
   2. Measured on a ratio scale –
2. In the context of a postnatal care, list five variables that are:
   1. Continuous data –
   2. Categorical data –
3. Why is it that temperature measured in Kelvin is on a ratio scale and not an interval scale?
4. In the context of a public hospital, list three variables that are:
   1. Ordinal data –
   2. Discrete data –
5. Read the paper “Cardiac surgery clinical outcomes in 70 years and over patients” by Atalay et al. and complete the table below:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Abbreviation** | **Type of Variable (continuous, discrete or categorical)** |
| Age | AGE | Continuous |
| Sex | SEX | Categorical |
| Body Mass Index | BMI |  |
| Smoking status | SMOKE |  |
| Comorbidity | COM |  |
| Length of Hospital Stay | LOS |  |
| Ejection Fraction (<30%) | EF |  |
| Surgical Intervention | SURG |  |
| Mortality status | MORT |  |

**Hypothesis Testing**

For each of the scenarios below, write the null (H0) and alternate (H1) hypotheses:

1. You suspect that the average number of patients at two different hospitals is different.
2. You have collected data on the ages of individual Instagram users and the number of hours that each user spends on Instagram per month. You suspect that people under the age of 40 spend more time on Instagram than people over the age of 40.
3. You have the results of K10 (depression and anxiety test) tests for 212 patients. The results are listed by state and you suspect that the K10 results are different in different states.
4. Your spouse or loved one is in a bad mood. You suspect their dopamine levels will be different for washing the dishes than for buying them a gift.
5. You suspect that the % real meat content is different at five different kebab joints.