Lab 2: Measuring dispersion. Standardizing data. Standard errors.

These labs contain exercises to be completed using a spreadsheet software.

There are multiple providers of spreadsheet software, but the University of Glasgow provides students with a license to use Microsoft 365 which contains Excel. You can use Excel either on your computer or in the cloud. The GTA is using Excel in the cloud and it is **strongly recommended** that you also use Excel cloud during the Lab sessions, because it will look similar to what the GTA has. Note that if you use Excel locally on your computer, the interface is specific to your operating system and system language.

To go to Excel online, go to https://www.office.com/launch/excel?auth=2.

Exercise 1.

Open dataset "testscores.xls" in Excel Online. We have a sample (n=200) of student test scores in Math and English (data is imaginary). Minimal text score is 0 and maximal test score is 100.

Part 1. Visualizing dispersion in data.

- Plot a histogram of English test scores.
- Plot a histogram of Math test scores.
- Based on these two histograms, which variable do you think is more dispersed? (Which has a higher variance?)

Part 2. Quantifying dispersion in data.

- Compute the mean of both test scores.
- Compute the sample variance of both test scores using the Excel formula VAR().
- Compute the sample standard deviation of both test scores using Excel formula STDEV().
- Compute the variance WITHOUT using Excel formula VARIANCE. You are only allowed to use the Excel formulas SUM() and COUNT() and standard mathematics operations.
- Interpret your observations. Based on the standard deviation and the variance, which variable is more dispersed?

Part 3. Standardizing data.

- For each observation of variable "Math" in the sample, compute the z-score. Use the mean and the standard deviation computed in Part 2.
- Compute the mean and the standard deviation of the z-scores.
- Plot the histogram of the z-scores for Math. What does the histogram look like?

Exercise 2.

In the lecture slides, we have seen that the formula for the standard error of the sample mean is given by

$$SE(\text{sample mean}) = \sigma_{\bar{x}} = \frac{\sigma_\mu}{\sqrt{n}}$$

This depends on the population standard deviation, σ_μ. In practice, however, the parameters related to the population are unknown.

Part 1. SE of sample mean (English).

We know that σ_μ for variable "English" is equal to 4.6.

- Find the sample mean for the variable "English".
- Find the standard error of the sample mean for the variable "English".
- Argue in words: What is the standard error of the sample mean useful for?

Part 2. SE of sample mean (Math).

We do **not** know σ_μ for the variable "Math".

- Find the sample mean of variable "Math".
- Find the sample standard deviation of variable "Math", denote as s.
- Calculate the following quantity:

$$\widehat{SE}(\text{sample mean}) = \frac{s}{\sqrt{n}}$$

Note - we use the hat to highlight that $\widehat{SE}(\text{sample mean})$ is different from $SE(\text{sample mean})$!

Exercise 3.

Part 1. If-functions.

- Create a new variable which equals 1 if the student has a math score above the median score.
- Create a new variable which equals 1 if the student has an English score above the median score.
- Create a new variable which equals 1 if the student has *both* English score above median and Math score above median. Call this variable *"high_performer"*.

Hint: Use the IF-function in Excel.

Part 2. Sample proportions.

- Calculate the sample proportion of students who have both English score above median and Math score above median. Interpret.
- Calculate the average of variable *"high_performer".* Interpret.

Part 3. Standard Error of Sample proportion

We know that the population proportion of students who have above median grades in both Math and in English is P =0.29.

- Using the following formula, compute the standard error of the sample proportion.

$$SE(\text{sample proportion}) = \sigma_{\hat{p}} = \frac{\sqrt{P(1-P)}}{\sqrt{n}} = \sqrt{\frac{P(1-P)}{n}}$$

- Calculate the probability that $\hat{p}$, the sample proportion of students with above median grades in both subjects is between 0.25 and 0.33.

  Hint: See Lecture slides, Unit 3, Example 3.