Introductory Statistics for Economics

ECON1013: LAB 2

Duong Trinh

University of Glasgow

2025

Intro

- Duong Trinh
  - PhD Student in Economics (Bayesian Microeconometrics)
  - Email: Duong.Trinh@glasgow.ac.uk

- ECON1013 LB07 - 42 BUTE GDNS:1105 LAB M
  - Tuesday 2-3 pm
  - 3 sessions (4-Feb, 18-Feb, 4-March)
- ECON1013 LB08 - 42 BUTE GDNS:1105 LAB M
  - Tuesday 3-4 pm
  - 3 sessions (4-Feb, 18-Feb, 4-March)
- ECON1013 LB13 - 42 BUTE GDNS:912A LAB A
  - Friday 11 am - 12 pm
  - 3 sessions (7-Feb, 21-Feb, 7-March)
- ECON1013 LB11 - 42 BUTE GDNS:912A LAB A
  - Friday 12-1 pm
  - 3 sessions (7-Feb, 21-Feb, 7-March)

Record Attendance

Setup

- Step 1: Download Lab materials from **Moodle** page $\rightarrow$ Extract the folder in PC.
- Step 2: Log in **Microsoft onedrive** using your student account https://onedrive.live.com/login/ and upload the folder above.
- Step 3: Launch the **Excel** online https://www.office.com/launch/excel?auth=2, which we will use for all lab sessions.

Exercise 1.

Exercise 1.

- Data set: `testscores.xls`
- About: A sample ($n = 200$) of student test scores in Math and English
    - Minimal text score is 0 and maximal test score is 100.

Part 1. Visualizing dispersion in data.

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

Part 2. Quantifying dispersion in data.

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

Part 3. Standardizing data.

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

[Review] z-score

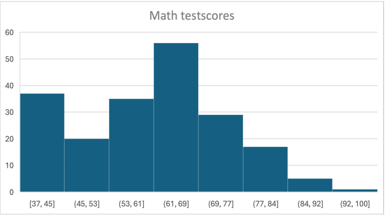
- A z-score shows the position of a value relative to the mean of the distribution.
- It indicates the number of standard deviations a value is from the mean.
- If the data set is the entire population of data and the population mean, μ, and the population standard deviation, σ, are known, then for each value, x_i, the z-score associated with x_i is

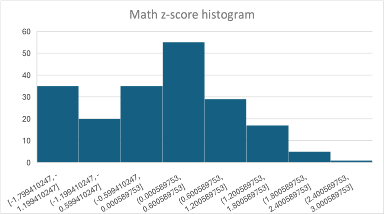
$$z_i = \frac{x_i - \mu}{\sigma}$$

Part 3. Standardizing data.

3. Plot the histogram of the z-scores for Math. What does the histogram look like?

```
## Warning: package 'knitr' was built under R version 4.2.3
```





Exercise 2.

Exercise 2. SE of sample mean

- X: student's test score.
  - Population standard deviation: σ_μ
- $\bar{X}$: sample mean test score of $n = 200$ students.
  - Standard Error of the sample mean: $\sigma_{\bar{X}}$
  - Sampling distribution of the sample mean

$$\bar{X} \sim \text{Normal}(\mu, \sigma^2_{\bar{X}}) \text{ when } n \text{ is sufficient large}$$

- Formula

$$\text{Standard Error(sample mean)} = \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}}$$

$$SE(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_\mu}{\sqrt{n}}$$

Exercise 2. SE of sample mean

- Formula

$$\text{Standard Error(sample mean)} = \frac{\text{Population standard deviation}}{\sqrt{\text{Sample size}}}$$

$$SE(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_\mu}{\sqrt{n}}$$

- When σ_μ is unknown $\rightarrow$ Use sample standard deviation s instead.

$$\widehat{SE}(\bar{X}) = \frac{s}{\sqrt{n}}$$

Part 1. SE of sample mean (English).

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

1. Find the sample mean for the variable "English".
2. Find the standard error of the sample mean for the variable "English".
3. 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 variable "Math".

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

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

Exercise 3.

Part 1. If-functions.

1. Create a new variable which equals 1 if the student has a math score above the median score.
2. Create a new variable which equals 1 if the student has an English score above the median score.
3. 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() in Excel.

Part 1. If-functions.

1. Create a new variable which equals 1 if the student has a math score above the median score.
   $\texttt{english above median}_i = 1$ **if** $\texttt{english}_i > \textit{median}(\texttt{english})$

Part 1. If-functions.

1. Create a new variable which equals 1 if the student has a math score above the median score.
   $\texttt{english above median}_i = 1 \textbf{ if } \texttt{english}_i > \textit{median}(\texttt{english})$
2. Create a new variable which equals 1 if the student has an English score above the median score.
   $\texttt{math above median}_i = 1 \textbf{ if } \texttt{math}_i > \textit{median}(\texttt{math})$

Part 1. If-functions.

1. Create a new variable which equals 1 if the student has a math score above the median score.
   $\texttt{english above median}_i = 1$ **if** $\texttt{english}_i > \textit{median}(\texttt{english})$
2. Create a new variable which equals 1 if the student has an English score above the median score.
   $\texttt{math above median}_i = 1$ **if** $\texttt{math}_i > \textit{median}(\texttt{math})$ Hint: Use the IF() in Excel.

```
= IF(logical_test,[value_if_true],[value_if_false])
```

Part 1. If-functions.

3. 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".

$$\texttt{high\_performer}_i = 1 \textbf{ if } \{\texttt{english}_i > \textit{median}(\texttt{english}) \textbf{ and } \texttt{math}_i > \textit{median}(\texttt{math})\}$$

equivalently,

$$\texttt{high\_performer}_i = 1 \textbf{ if } \{(\texttt{english above median}_i = 1) \textbf{ and } (\texttt{math above median}_i = 1)\}$$

Part 2. Sample proportions.

1. Calculate the sample proportion of students who have both English score above median and Math score above median. Interpret.
2. 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 $p = 0.29$.

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

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

2. 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.

Part 3. Standard Error of Sample proportion

1. Compute the standard error of the sample proportion.

- high_performer are students who have above median grades in both Math and in English.
- Population proportion of high_performer: $p = 0.29$
- Sample proportion of high_performer: $\hat{p}$
- Verify that sample size n is large enough $n \cdot p \cdot (1 - p) > 5$, which is correct as $200 \cdot (0.29) \cdot (1 - 0.29) \approx 41.18$.
- Hence, asymptotically,

$$\hat{p} \sim \text{Normal}\left(p, \frac{p(1-p)}{n}\right)$$

where the Standard Error of the sample proportion is $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$

Part 3. Standard Error of Sample proportion

2. $P(0.25 \leq \hat{p} \leq 0.33) = ?$

Transform $\hat{p}$ into the standard normal random variable Z

$$Z = \frac{\hat{p} - p}{SE(\hat{p})} = \frac{\hat{p} - 0.29}{SE(\hat{p})}$$

Rewrite the probability

$$\begin{aligned}
P(0.25 \leq \hat{p} \leq 0.33) &= P\left(\frac{0.25 - 0.29}{SE(\hat{p})} \leq \frac{\hat{p} - 0.29}{SE(\hat{p})} \leq \frac{0.33 - 0.29}{SE(\hat{p})}\right) \\
&\overset{\text{step 1}}{=} P\,(\text{lower bound} \leq Z \leq \text{upper bound}) \\
&\overset{\text{step 2}}{=} P\,(Z \leq \text{upper bound}) - P(Z \leq \text{lower bound}) \\
&\overset{\text{step 3}}{=} \text{probability (area under curve)}
\end{aligned}$$

Part 3. Standard Error of Sample proportion

2. $P(0.25 \leq \hat{p} \leq 0.33) = ?$

$$
\begin{aligned}
P(0.25 \leq \hat{p} \leq 0.33) &= P\left(\frac{0.25-0.29}{SE(\hat{p})} \leq \frac{\hat{p}-0.29}{SE(\hat{p})} \leq \frac{0.33-0.29}{SE(\hat{p})}\right) \\
&\overset{\text{step 1}}{=} P\left(\text{lower bound} \leq Z \leq \text{upper bound}\right) \\
&\overset{\text{step 2}}{=} P\left(Z \leq \text{upper bound}\right) - P(Z \leq \text{lower bound}) \\
&\overset{\text{step 3}}{=} \text{probability (area under curve)}
\end{aligned}
$$

Hint: Use the NORMSDIST() in Excel to return the standard normal cumulative distribution

```
= NORMSDIST(z)
```

Warning: package 'ggplot2' was built under R version 4.2.3