Introductory Statistics for Economics

ECON1013: TUTORIAL 1

Duong Trinh

University of Glasgow

Jan 2024

Intro

- Duong Trinh
  - PhD Student in Economics (Bayesian Microeconometrics)
  - Email: Duong.Trinh@glasgow.ac.uk
- ECON1013-TU08
  - Monday 3-4 pm
  - 4 sessions (23-Jan, 6-Feb, 20-Feb, 5-March)

Record Attendance

Scan the QR code below or use the password listed below to take your attendance

7zg8ld

Exercise 1

A group of 11 former college students are interviewed 10 years after their graduation. Their incomes are as follows (in $1,000$ pounds):

$$\{20, 22, 23, 23, 25, 28, 28, 30, 30, 34, 160\}$$

For this sample, we have calculated the following summary statistics:

- Sample average ≈ 38.5
- Sample median 28
- Sample standard deviation 40.5
- Interquartile range 7 (from 23 to 30)

Exercise 1

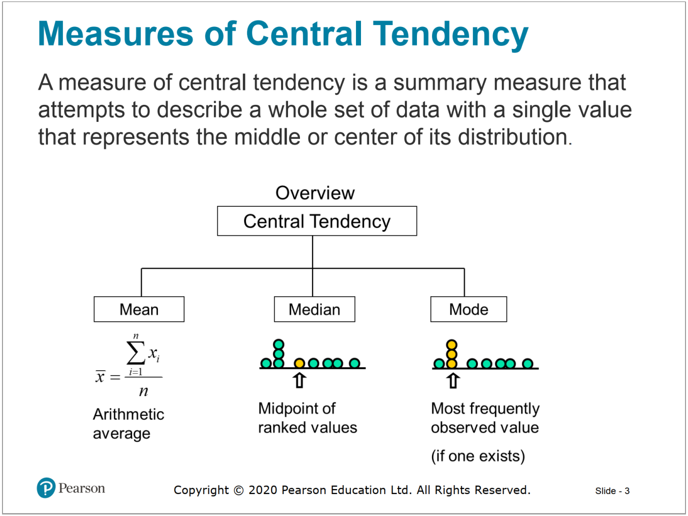
A group of 11 former college students are interviewed 10 years after their graduation. Their incomes are as follows (in $1,000$ pounds):

$$\{20, 22, 23, 23, 25, 28, 28, 30, 30, 34, 160\}$$

a. We wish to measure the central tendency in this sample. Which measure is the most appropriate? (You can use the summary statistics provided above or other measures.) Argue.
b. We wish to measure the variability in this sample. Which measure is the most appropriate? (You can use the summary statistics provided above or other measures.) Argue.
c. Construct a box-and-whisker plot of the data.
d. There was a reporting mistake in the data set - the largest value is actually 360 instead of 160. How do the summary statistics change? Do your answers to questions 1 and 2 change?

(a) We wish to measure the central tendency in this sample. Which measure is the most appropriate?

(a) We wish to measure the central tendency in this sample. Which measure is the most appropriate?



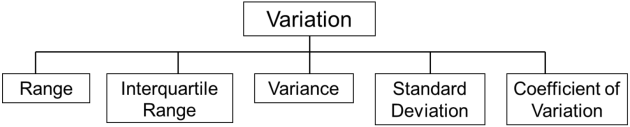
(a) We wish to measure the central tendency in this sample. Which measure is the most appropriate?

- The distribution is characterised by a cluster of observations around 30 and a single large outlier.
- Because of this outlier, sample mean is significantly larger than the values in the main cluster of observations.
- Sample median gives a more accurate description of the main cluster of observations.

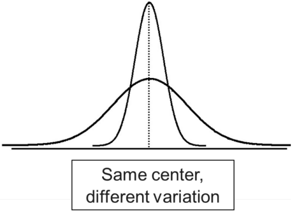
(b) We wish to measure the variability in this sample. Which measure is the most appropriate?

(b) We wish to measure the variability in this sample. Which measure is the most appropriate?

Measures of Variability



Measures of variation give information on the spread or variability of the data values.



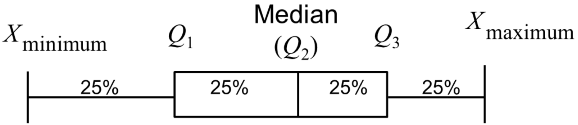
(b) We wish to measure the variability in this sample. Which measure is the most appropriate?

- ◇ The sample standard deviation depends on the sample mean. Since the sample mean is affected by the single outlier, so is the sample standard deviation. As a consequence, the sample standard deviation is high, which indicates a high degree of variability in the data.
- ◇ However, in the main cluster of observations, values are relatively close to each other.
- ◇ This is more informatively summarized by the interquartile range (7), which indicates a relatively low degree of variability in the data.

(c) Construct a box-and-whisker plot of the data.

(c) Construct a box-and-whisker plot of the data.

- The box-and-whisker plot usually displays 5 values:
  The minimum, the 25^{th} percentile, the median, the 75^{th} percentile, and the maximum.



(d) There was a reporting mistake in the data set - the largest value is actually 360 instead of 160. How do the summary statistics change? Do your answers to questions 1 and 2 change?

(d) There was a reporting mistake in the data set - the largest value is actually 360 instead of 160. How do the summary statistics change? Do your answers to questions 1 and 2 change?

Our summary statistics change as follows:

- Sample average ≈ 56.6
- Sample median 28
- Sample standard deviation 100.7
- Interquartile range 7 (from 23 to 30)

Sample median and interquartile range are not affected by the value of the outlier.

Exercise 2

Consider the sample $\{X_1, X_2, \ldots, X_n\}$.

a. Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem:

$$\min_c \sum_{i=1}^{n} (X_i - c)^2 \tag{1}$$

b. What is the interpretation of the function $\sum_{i=1}^{n} (X_i - c)^2$?

(The aim of this exercise is justify the use of the sample average).

(a) Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem (1)

(a) Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem (1)

- Let $h(c) = \sum_{i=1}^{n}(X_i - c)^2 = \sum_{i=1}^{n}\left(X_i^2 - 2X_i c + c^2\right)$

(a) Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem (1)

- Let $h(c) = \sum_{i=1}^{n} (X_i - c)^2 = \sum_{i=1}^{n} \left(X_i^2 - 2X_i c + c^2\right)$
- To find the minimum, we take the first-order condition w.r.t. c:

$$h'(c) = \sum_{i=1}^{n} (-2X_i + 2c) = -2\sum_{i=1}^{n} (X_i - c)$$

(a) Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution of minimization problem (1)

- Let $h(c) = \sum_{i=1}^{n} (X_i - c)^2 = \sum_{i=1}^{n} \left(X_i^2 - 2X_i c + c^2\right)$
- To find the minimum, we take the first-order condition w.r.t. c:

$$h'(c) = \sum_{i=1}^{n} (-2X_i + 2c) = -2\sum_{i=1}^{n} (X_i - c)$$

- We wish to solve for c such that $h'(c) = 0$:

$$\begin{aligned}
h'(c) = 0 &\Leftrightarrow \sum_{i=1}^{n} (X_i - c) = 0 \\
&\Leftrightarrow (X_1 - c) + (X_2 - c) + \ldots + (X_n - c) = 0 \\
&\Leftrightarrow -nc + \sum_{i=1}^{n} X_i = 0 \\
&\Leftrightarrow c = \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

(b) What is the interpretation of the function $\sum_{i=1}^{n}(X_i - c)^2$

- The term $X_i - c$ is the difference between each observation and some constant c. These differences are sometimes positive and sometimes negative - but if we square them, they are always positive.
- The summation $\sum_{i=1}^{n}(X_i - c)$ gives us the sum of squared differences. When we solve the minimization, we are looking for the value of c such that the sum of squared differences between each observation and c is the smallest possible.
- In other words, we are looking for c such that c is closest possible to different values of X_i, when "closest possible" is measured by the squared differences.
- Thus, the sample mean is the quantity which minimizes the sum of squared differences in the data.

(b) What is the interpretation of the function $\sum_{i=1}^{n} (X_i - c)^2$

Remark. Notice that the sample median is the solution of the problem:

$$\min_c \sum_{i=1}^{n} \mid X_i - c \mid \tag{2}$$

- Thus, while sample mean minimizes sum of squared differences, sample median minimizes the sum of absolute differences.
- The "best" measure of central tendency depends on the measure of distance that we want to minimize.

Exercise 3

a. Show that $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$.
What is this estimator?

b. Show that $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right) = \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right) Y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}_n \bar{Y}_n$.
What is this estimator?

(a) Show that $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$.

(a) Show that $\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$.

We are showing a property related to *sample variance*.

(a) Show that $\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2$.

We are showing a property related to *sample variance*.

$$\begin{aligned}
\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 &= \sum_{i=1}^{n} \left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right) \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \sum_{i=1}^{n} X_i + n\bar{X}_n^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2\bar{X}_n \cdot n\bar{X}_n + n\bar{X}_n^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2
\end{aligned}$$

proving the result.

(b) Show that $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right) = \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)Y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n$.

(b) Show that $\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right) = \sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)Y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}_n\bar{Y}_n$.

We are showing a property related to *sample covariance*.

(b) Show that $\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\left(Y_i-\bar{Y}_n\right)=\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)Y_i=\sum_{i=1}^{n}X_iY_i-n\bar{X}_n\bar{Y}_n$.

We are showing a property related to *sample covariance*.

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\left(Y_i-\bar{Y}_n\right)&=\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)Y_i-\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)\bar{Y}_n\\&=\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)Y_i\end{aligned}$$

because $\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)=\sum_{i=1}^{n}X_i-n\bar{X}_n=0$.

It follows that

$$\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)Y_i=\sum_{i=1}^{n}X_iY_i-\bar{X}_n\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}X_iY_i-n\bar{X}_n\bar{Y}_n.$$