ECO239 Statistics I

Week 4

Describing Data: Numerically

- Central Tendency
  - Mean
  - Median
  - Mode
- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Mean

The *sample mean*, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where x_1, x_2, ..., x_n represent the n observed values.

The *population mean* is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.

The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Calculating Mean

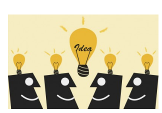
e.g. 1 {1, 2, 3, 4, 5}

X_bar= (1+2+3+4+5)/5 = 3

e.g. 2 {1, 2, 3, 4, 20}

X_bar=(1+2+3+4+20)/5 = 6

- Check raw data to detect outliers.
- Mean is affected by outliers/extreme values.



- Finalscore

Weighted Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- GPA (Grade Point Average) Calculation

A1 = 4
A2 = 3.5
B1 = 3
B2= 2.5
C1 = 2
C2 = 1.5
D1 = 1
D2 = 0.5
F = 0

	Grade	Score (X)	Credit (W)	X*W
ECO239	A1		3	
ECO336	A2		3	
ECO448	B1		3	
ING250	C2		2	
ING350	A2		2	
TOTAL			13	

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

41.5/13 = 3.19 = GPA

	Grade	Score (X)	Credit (W)	X*W
ECO239	A1	4	3	12
ECO336	A2	3.5	3	10.5
ECO448	B1	3	3	9
ING250	C2	1.5	2	3
ING350	A2	3.5	2	7
TOTAL			13	41.5

Median

The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

Median

Finding Median

Step 1: Order the data in ascending order
Step 2: Find Median Position = (n+1)/2
Step 3: Find the Median at the Median Position

If n is **odd**, Median is the middle number.
e.g. n=5 => Median position = (5+1)/2 = 3.

If n is even, Median is the average of two middle numbers.
e.g.n=12, Median position = (12+1)/2=6.5. Median is average of 6^{th} and 7^{th} values.

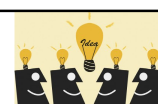
Median

- Data{ 8, 4, 3, 5, 9, 7, 8}

Find Median.

Step1: sort => 3, 4, 5, 7, 8, 8, 9
Step2: M.P. (7+1)/2 = 4.
Step3: 4^{th} value = 7. <= Median.

Median



- Data { 4, 3, 5, 7, 8, 8, 20}

Step1: 3, 4, 5, 7, 8, 8, 20
Step2: M.P. (7+1)/2=4
Step3: Median = 7.

*Median is not affected by an extreme value.

Median

Data {4, 3, 5, 7, 8, 8, 9, 20}

Step 1: 3, 4, 5, 7, 8, 8, 9, 20
Step2: M.P. = (8+1)/2 = 4.5
Step3: Median = (7+8)/2 = 7.5.

Mode

- Value that occurs most often.

- There may not be any mode.
- There may be multiple mode.

e.g.
1, 3, 4, 5, 5, 7, 9, 9, 9, 10, 12, 12, 13, 14

Mode = 9

Mode

e.g. 3, 5, 7, 4, 8, 8, 9
Mode = 8
e.g. 3, 5, 7, 4, 8, 8, 30
Mode = 8

=> Not affected by extreme values.

Mode

e.g. 0, 1, 2, 3, 5, 6
Mode = No Mode

e.g. 0, 1, 1, 1, 2, 3, 3, 3 ,4, 5, 6
Mode = 1 and 3

Practice

Housing Prices
1. $ 2,000,000
2. $ 500,000
3. $ 300,000
4. $ 100,000
5. $ 100,000

Mean = (200000+500000+300000+100000+100000)/5 = 3000000/5=600,000

Median = 300,000

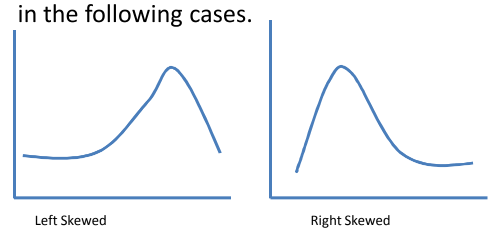
Mode = 100,000

Q: Find mean, median and mode.

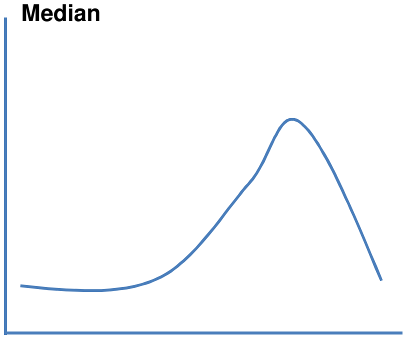
When do we use Mean and when do we better use Median???

Shape of Distribution

- We can judge skewness using Mean and Median
- Consider the relative size of Mean and Median in the following cases.

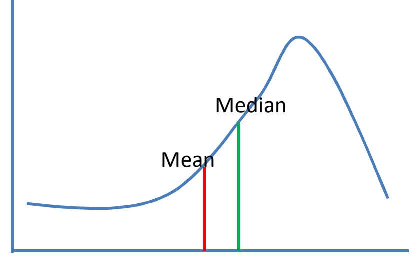


Question: For Left Skewed Distribution, the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median



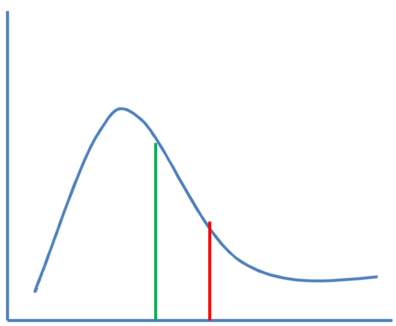
Question: For Left Skewed Distribution, the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median

(b) Mean < Median



Question: For Right Skewed Distribution,

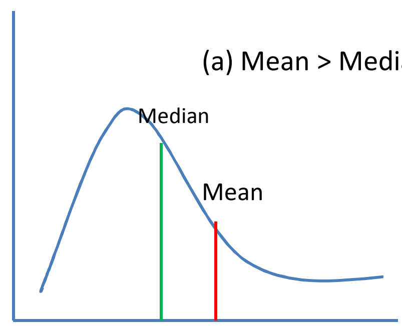
the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median

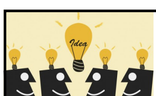


Question: For Right Skewed Distribution,

the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median

(a) Mean > Median





Practice

Data { 40, 45, 50, 51, 55, 60, 80, 99}
n= 8

Comment on skewness.

Mean = 60
Median = (51+55)/ 2 = 53
Since Mean > Median, Right Skewed.

Measures of Variation

- Range
- Interquartile Range (Discussed in week3)
- Variance
- Standard Deviation
- Coefficient of Variation

Range

- Range = X_largest – X_smallest

- E.g. {7, 8, 9, 11, 12}
- Range = 12-7 = 5

Interquartile Range

- Discussed under Box plot in week 3.

- IQR depends only on a few points of the entire data set.

IQR

Compare the following two cases.

Data{ 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21}
IQR = ?

Data{ 0, 2, 3, 3, 4, 4, 14, 15, 16, 16, 21}
IQR= ?

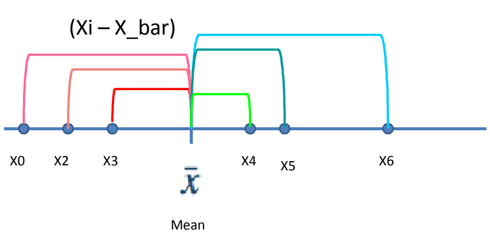
Data{ 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21}
Q1 location = (11+1)/4=3 => Q1 = 3
Q3 location = (11+1)*3/4 = 9 => Q3 = 16
IQR = 16-3 = 13

Data{ 0, 2, 3, 3, 4, 4, 14, 15, 16, 16, 21}
Q1 location = (11+1)/4 = 3 => Q1 = 3
Q3 location = (11+1)*3/4=9 => Q3 = 16
IQR = 16-3 = 13.

<= IQR depends only on a few points of the entire data set. May not be a good indicator of variation of data.

Variance

Measuring the average of the total distance between each observation and the mean.



Variance: Calculation

Step 1: Compute the distance between each data point and mean.

Step 2: Square the each distance

Step 3: Sum all the squared distances and divide by observation size (for population) OR by observation size – 1 (for sample data)

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

μ = Population Mean

X_i = each observation, i = 1,… N.

N = population size

Sample Variance

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

$\bar{x}$:Sample Mean

n : Sample Size

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

WHY???

- Do we divide by (n-1), instead of n for sample variance???

A: Sample variance is an unbiased estimator of the population variance. It's a better estimator of the population variance if divided by n-1.

<u>This will be discussed in detail in ECO240.</u>

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Data { 50, 60, 65, 70, 88}
Calculate variance.

Mean = 67
Var. = 792/4=198

NOTE: Variance

- If you forget to square the distance, the calculated value =

0

Standard Deviation

: Average spread around the mean
: Square root of the variance.
: Has the same unit as the original data

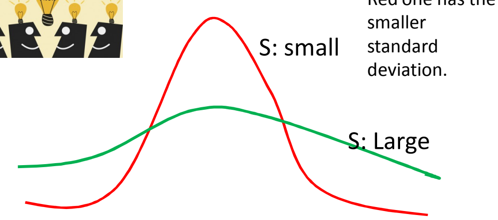
$$s = \sqrt{s^2}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Data { 50, 60, 65, 70, 88}
Calculate standard deviation.

*IF the data is exam score, for example, the unit of standard deviation is also "exam points".

Which s is smaller / larger???



Compare standard deviations

Data1 { 11, 12, 13, 16, 16, 17, 18, 21}
Data2 { 14, 15, 15, 15, 16, 16, 16, 17}
Data3 { 11, 11, 11, 12, 19, 20, 20, 20}

Calculate Mean and Standard Deviation.
Compare.

HW! Confirm these results.
Means= 15.5
Stdev1 = 3.338, Stdev2 = 0.926, Stdev3 = 4.570.

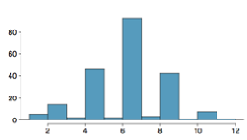
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$s = \sqrt{4.11} = 2.03\ hours$



- We can see that all of the data are within 3 standard deviations of the mean.

- finalscore

Quiz 2

Data{ 44 45 46 47 49
52 53 57 60 65
67 67 69 71 73
76 78 80 82 84 }

n= 20

Q1: Find Q1, Q2, Q3 and IQR.

Q2: Draw a Box Plot.

Quiz Answer

Q1:

Q1 location=21/4=5.25

=> Q1=49+(52-49)*0.25=49.75

Q2 location = 21/2=10.5

=>Q2=65+(67-65)*0.5=66

Q3 location = 21*3/4=15.75

=>Q3=73+(76-73)*0.75=75.25

IQR = Q3-Q1 = 75.25-49.75=25.5

Upper whisker reach = 75.25+1.5*25.5=113.5

Lower whisker reach = 49.75-1.5*25.5=11.5.

Since Max value =84 < 113.5, our maximum value is directly 84. We do not have any outliers.

Since Min value = 44 > 11.5, our minimum value is directly 44. We do not have any outliers.

I gave full points for those who calculated max = 113.5 and min=11.5 for today's quiz. However it won't be so for exams.