BNAD/ECON/MGMT 276
Lecture 2
Descriptive Statistics II

Phuong Ho

May 11, 2017

Outline

Outline

Introduction

- In the last lecture, we studied how to summarize a data set with tabular and graphical presentations.
- In this lecture, we will learn how to summarize data with numerical measures, which is convenient if the data set contains a lot of variables.

- We first recall that almost every data set is a sample from population.
- Thus, keep in mind that we are working with sample when we summarize data with any numerical measures in this lecture $\rightarrow$ sample measures.

Outline

- When we get a data set or data, the first thing we may want to know is the central location for that data.
- The central location for data offers us useful information.
- For example, the central location of meal prices tells us how much meal price is on average in Tucson.
- There are two most frequently used measure of the central location for data.

1. Mean
2. Median

Notations

- A variable will be denoted by an alphabet letter without subscript.

  e.g. x, y or z

- A specific value of a variable will be denoted by an alphabet letter with subscript.

  e.g. $x_1, x_2, \ldots, x_n$ for x
  e.g. $y_1, y_2, \ldots, y_n$ for y

Example

x				
12	15	18	20	33

- We don't know what x is for. It could be any variable, meal price or the number of commercials. Anyway we decide call it x.
- We will have data for x as in the above. When we say x_1, we means the first *value* in the above data.
- Thus, $x_1 = 12$, $x_5 = 33$.
- Keep in mind that whenever we use x_i (or y_i) it indicates i^{th} observation of variable x (or y) in a data set.

Mean

Sample Mean

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

where $\sum x_i$ means

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n,$$

and n is the total number of observations.

So, the sample mean equals the sum of all values of observations divided by the number of observations.

Example

x				
12	15	18	20	33

The mean can be calculated as follows.

1. Sum up all values in data.

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 98$$

2. Divide the sum by the number of observations. In this example, we have 5 observations. Thus,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{98}{5} = 19.6$$

- The sample mean of the above data (or variable) is 19.6.
- The central location of the above data is around 19.6.

Median

- The median is another useful measure of central location of the data.

Median

The median is the value in the middle when all values in data is arranged in ascending order.

- The median can be found by the following procedure.

Median

Arrange data in ascending order (from smallest to largest value)

(a) For an odd number of observations, the median is the middle value.

(b) For an even number of observations, the median is the mean of the two middle values.

Example 1: An odd number of observations

32	20	30	23	25

1. Arrange numbers in ascending order.

20	23	25	30	32

2. The median is 25.

Example 1: An even number of observations

4	6	8	12	15	100

1. Find two values in the middle.
2. These two values in the above data are 8 and 12.
3. Calculated the mean of these two.

$$\text{Median} = \frac{8 + 12}{2} = 10$$

Mean vs. Median

- The mean is good since it is easy to calculate. However, this measure is really sensitive to an extreme values.
- Consider the following example.

4	6	8	12	15	100

- The median of the above data is 10.
- What is the mean of the above data? The mean is 29.
- In this case, median gives us better information about the central location of data.

Outline

Mode

Mode

The mode is the value that occurs with the greatest frequency

- The mode is merely tells us what value happens most frequently.

Example

32	42	46	46	54

- What is the mode of the above data?

32	32	42	43	45	45	50	51	100

- What is the mode of the above data?

Percentiles

Percentile

The p_{th} percentile is a value such that *at least* p percent of the observations are less than or equal to this value and *at least* (100-p) percent of the observations are greater than equal to this value.

- Suppose we have data which has 35 as p^{th} percentile.
- It means that p% of observations in the data are less than 35 and $(100 - p)$ % of observations are greater than 35.

Calculating the p^{th} percentile

Step 1 Arrange values in data in the ascending order

Step 2 Compute an index i

$$i = \left(\frac{p}{100}\right) n$$

Step 3

(a) If i is not integer, round up. The rounded-up number will tell the position of the p^{th} percentile.

(b) If i is an integer, the p^{th} percentile is the mean of i^{th} and $(i+1)^{th}$ value.

Example

Participant's Age											
32	33	34	36	36	39	41	43	43	47	48	50

- What is 85^{th} percentile?
- What is 50^{th} percentile? Compare it to the median.

Quartiles: Special Percentiles

- Quartiles divides data into four parts as its name implies.

$$
\begin{aligned}
Q_1 &= \text{the first quartile, or } 25^{th} \text{ percentile} \\
Q_2 &= \text{the second quartile, or } 50^{th} \text{ percentile} \\
Q_3 &= \text{the third quartile, or } 75^{th} \text{ percentile}
\end{aligned}
$$

Example

Participant's Age											
32	33	34	36	36	39	41	43	43	47	48	50

- What are Q_2 and median?
- What is Q_1 and Q_3?

Outline

- Providing just a mean or median of data does not offer enough information.
- Another important property about data is the variability of data (i.e. how much the data vary).
- For example, when we look at monthly returns from an asset, how much the return monthly varies is also important information.
- We will see three different measures of variability.

Range

Range

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

- The range gives us general information about how values in data vary.
- The higher range is, the more variation of values in data is.
- The range is very sensitive to (highly influenced by) extreme values.

Example: Range

2	50	79	150	200	300

- The range of the above data is $300 - 2 = 288$.
- One can see that the values in data are actually varying much.

2	58	59	60	62	300

- The range of the above data is $300 - 2 = 288$.
- Note that the values in the above data are not actually varying much.
- It motivates us to use the other measure of variability.

Interquartile Range

Interquartile Range

$$\text{IQR} = Q_3 - Q_1$$

- This measure overcomes the dependency on extreme values.

Variance

Sample Variance

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

- Variance measure utilizes all the values of the data.
- Sample variance measures how far each value in data is from the mean of the data.
- $(x_i - \bar{x})$ is **deviation about the mean**. We square these deviations and sum up for the numerator.
- If we have many x_is that are far from the center of data, it means there is high variation of values in data.
- Thus, higher variance means higher variation of values in data.

Example

1	5	7	15	20	30

1. At first we need to calculate the mean, $\bar{x}$, to get $(x_i - \bar{x})$ for each x_i. $\bar{x} = 13$.

2. Now calculate $(x_i - \bar{x})$ for each i.

$x_1 - \bar{x}$	$x_2 - \bar{x}$	$x_3 - \bar{x}$	$x_4 - \bar{x}$	$x_5 - \bar{x}$	$x_6 - \bar{x}$
1-13	5-13	7-13	15-13	20-13	30-13
-12	-8	-6	2	7	17

3. Square each $(x_i - \bar{x})$.

-12	-8	-6	2	7	17
144	64	36	4	49	289

4. Sum up all squared deviations and divide it by $n - 1$.

144	64	36	4	49	289
		586			
		$\frac{586}{6-1} = \frac{586}{5} = 117.2$			

Example

1	5	7	15	20	30
$s^2 = 117.2$					

- Consider the other example that values in data does not vary much.

1	18	19	20	22	30
$s^2 = 90$					

- Ranges of these two data is the same as $30 - 1 = 29$.
- However, variances are different. The second data has less variance than the first data.

Standard Deviation

Standard Deviation

$$\text{Sample Standard Deviation} = s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

1	18	19	20	22	30
$s^2 = 90$			$\Rightarrow s = \sqrt{s^2} = 9.5$		

- Standard deviation is a "same-unit version of variance".
- For example, the sample variance for the starting salary data is $s^2 = 27,500$ (dollars)^2. Since the standard deviation is the square root of the variance, the units of the variance, dollar squared, are converted to the dollars in the standard deviation.
- The standard deviation of the starting salary data is 165 (dollars), the same units as the origin data.

Outline

- So far we only summarize data for a single variable.
- As in the last lecture (recall the cross-tabulations), we want a measure that tells us some information about a relationship between two variables.
- Two popular measures are covariance and correlation.

Covariance

- Suppose that now we have a data set consisting of two variables, x and y.

e.g. The number of Commercials and Sales for a restaurant.

Covariance

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

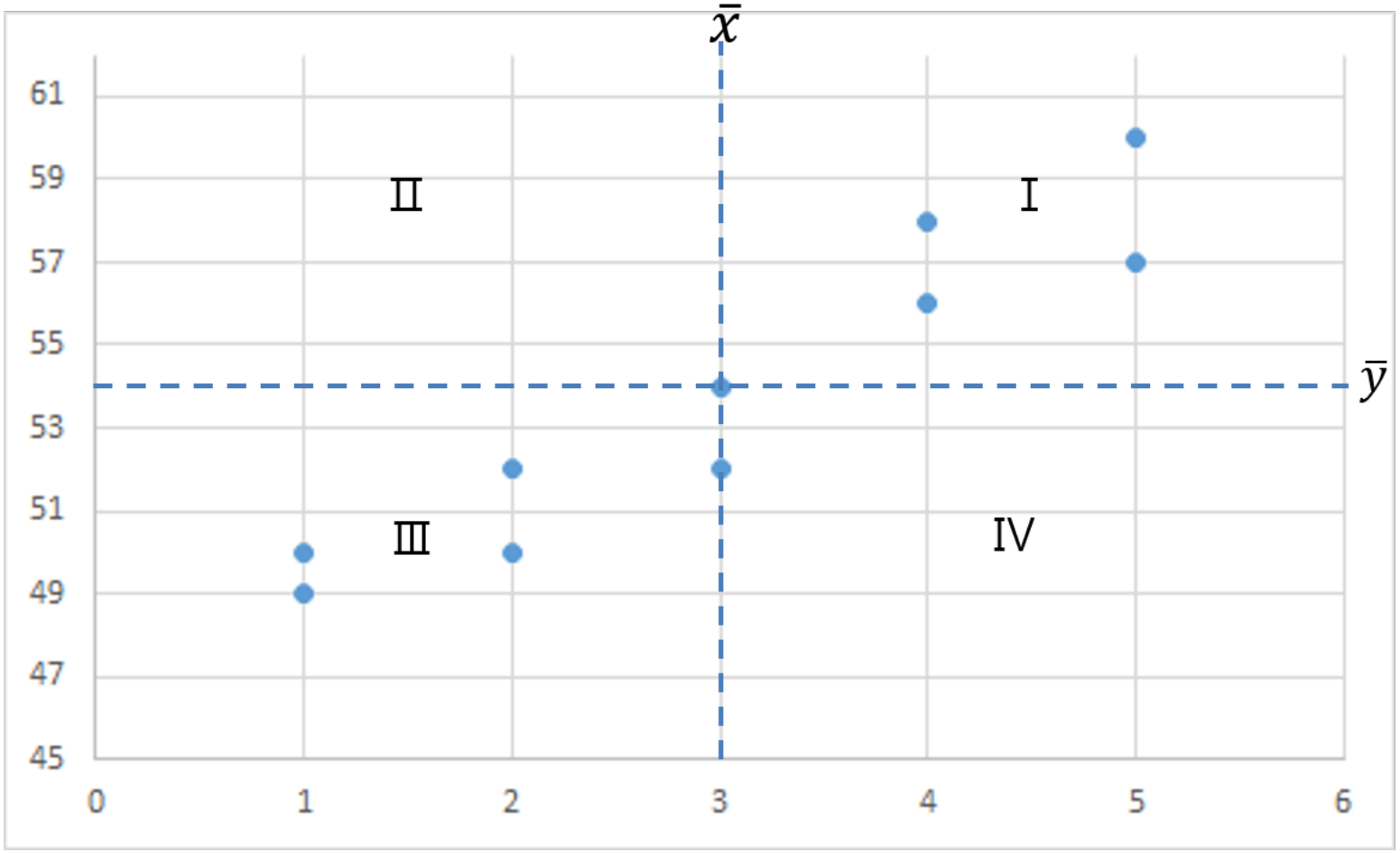
- $x_i - \bar{x}$ (or $y_i - \bar{y}$) measures the deviation about the mean of x_i (or y_i).
- We pair $x_i - \bar{x}$ with $y_i - \bar{y}$ for the same i and multiply these two.
- Then, sum up these numbers, $(x_i - \bar{x})(y_i - \bar{y})$, across all observations, $i = 1, 2, ..., n$.
- As we did for the sample variance, we divide the sum by $n - 1$.

Example

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	-1	-3.8	3.8
5	57	2	3.2	6.4
1	49	-2	-4.8	9.6
3	52	0	-1.8	0
4	58	1	4.2	4.2
1	50	-2	-3.8	7.6
2	52	-1	-1.8	1.8
3	54	1	0.2	0
4	56	2	2.2	2.2
5	60	0	6.2	12.4
$\bar{x} = 3$	$\bar{y} = 53.8$	0	0	48

What is s_{xy}?

Interpretation of the Covariance



Interpretation of the Covariance

1. Positive Sign of Covariance $\rightarrow$ Positive Linear Relation
2. Negative Sign of Covariance $\rightarrow$ Negative Linear Relation
3. A Covariance close to zero $\rightarrow$ More likely to have No Linear Relation.

Note A covariance only measures a linear relationship between two variables.

Note How big a covariance is does not tell us anything.

Note Only the sign of covariance (and whether it is close to zero) matters.

Correlation Coefficient

Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

where

$$\begin{aligned} s_{xy} &= \text{Covariance} \\ s_x &= \text{Standard Deviation of } x \\ s_y &= \text{Standard Deviation of } y \end{aligned}$$

- One property we need to know is (we always have)

$$-1 \leq r_{xy} \leq 1$$

Example

- We have calculated $s_{xy} = 5.3$, $s_x = 1.5$, and $s_y = 3.8$.
- The correlation coefficient in this example is merely

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{5.3}{1.5 \times 3.8} = 0.93$$

Interpretation of Correlation Coefficient

- Let's consider the following examples.

Data Set 1		Data Set 2	
x	y	s	t
5	10	5	10
10	20	10	23
15	30	15	30

- If all the points in a data set fall on a positively sloped straight line, the correlation coefficient will be exactly +1.
- Otherwise, the correlation coefficient will be less than 1.
- As a correlation coefficient is closer to 1, we have a stronger positive linear relation.

- If all the points in a data set fall on a negatively sloped straight line, the correlation coefficient will be exactly -1.
- Otherwise, the correlation coefficient will be greater than -1.
- As a correlation coefficient is closer to -1, we have a stronger negative linear relation.

Covariance vs. Coefficient Correlation

- Recall that only the sign of covariance matters but the magnitude of covariance does not.
- Correlation coefficient tells us two things.
  1. Whether two variables have a positive or negative linear relationship $(+$ or $-)$ or no linear relationship (≈ 0),
  2. How strong the association (relation) is (close to 1 or close to -1)

- One important caveat is that the coefficient correlation does NOT implies any causation (causal relationship/interpretation.)

e.g. Relation between the number of commercials (x) and sales (y).

- Suppose we got positive covariance, $s_{xy} = 5.3$, and correlation coefficient, $r_{xy} = 9.3$.
- We know that these two variables has a strong positive linear relation. We say, *an increase in the number of commercials is associated with an increase in sales.*
- However, we can't give causal interpretation by having these numbers.
- i.e. we can't say that the higher number of commercials causes higher sales or vice versa.
- There may be the causal relationship between them but we need to investigate more, rather only based on the correlation measure.

Outline

Exercise 1

The Dow Jones Travel Index reported what business travellers pay for hotel rooms per night in major U.S. cities(*The Wall Street Journal*, January 16, 2004). The average hotel room rates for 20 cities are as follows:

Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	192	Washington, D.C.	207

a) What is the mean hotel room rate?

b) What is the median hotel room rate?

c) What is the mode?

d) What is the first quartile?

e) What is the third quartile?

Exercise 2

Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the range, variance, and standard deviation.

Exercise 3

A department of transportation's study on driving speed and miles per gallon for midsize automobiles resulted in the following data:

Speed(Miles per Hour)	30 50 40 55 30 25 60 25 50 55
Miles per Gallon	28 25 25 23 30 32 21 35 26 25

a) Compute the (sample) correlation coefficient.

b) Does two variables exhibit linear relation? If it does, determine if these two variables have positive relation or negative relation.

c) Is the linear relation strong?