Chapter 2: Using Numerical Measures to Describe Data

Department of Mathematics
Izmir University of Economics

Week 2
2014-2015

Introduction

In this chapter we will focus on

- the measures of central tendency and location: *mean*, *median*, *mode*, and *quartiles*,
- the measures of variability: *range*, *interquartile range*, *variance*, and *standard deviation*, and
- the measures of relationships between variables: *covariance* and *correlation coefficient*.

Introduction

In Chapter 1, we described the data **graphically** using different graphs for categorical and numerical variables.

In this chapter we will describe data **numerically**.

Measures of Central Tendency and Location

Recall, from Chapter 1, that a *parameter* refers to a specific population characteristic while a *statistic* refers to a specific sample characteristic. Measures of central tendency are usually computed from sample data rather than from population data.

Arithmetic Mean (or Mean or Average)

The *(arithmetic) mean* of a set of data is the sum of the data values divided by the number of observations.

Arithmetic Mean (or Mean or Average)

If the data set is the entire population, then the *population mean*, μ , is a parameter given by

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N},$$

where N is the population size.

If the data set is from a sample, then the *sample mean*, $\bar{x}$, is a statistic given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where n is the sample size.

Median

The *median* is the middle observation of a set of observations that are arranged in increasing (or decreasing) order.

If the sample size, *n*,

- is an odd number, the median is the middle observation,
- is an even number, the median is the average of the two middle observations.

$$\text{median } (\tilde{x}) = \text{value at } 0.50(n+1)\text{th } \textbf{ordered} \text{ position}$$

Mode

The *mode* (if one exists) is the most frequent occurring value.

Example. Find the mode(s) if there is (are) any:

a) $3, 2, 1, 2, 1$

b) $1, 3, 5, 4, 6$

c) $1, 3, 7, 3, 2$

d) $3, 7, 5, 5, 3, 7$

Example. Given the following data:

$$60, 84, 65, 67, 75, 72, 80, 85, 63, 82, 70, 75$$

a) Find the mean.

b) Find the median.

c) Find the mode(s) if there is (are) any.

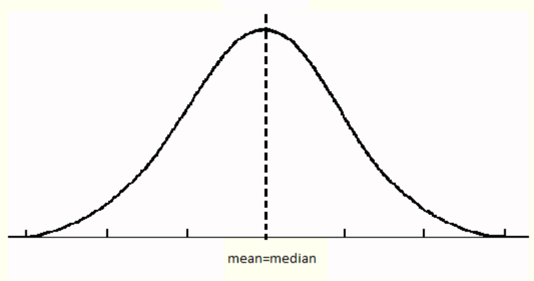
The decision as to whether the mean, median, or mode is the appropriate measure to describe central tendency of data is context specific.
One factor that influences our choice is the type of data:

- if data are categorical, we generally use median or mode,
- if data are numerical, we generally use mean.

Another factor is the presence of outliers: The median is not affected by the outliers but the mean is.

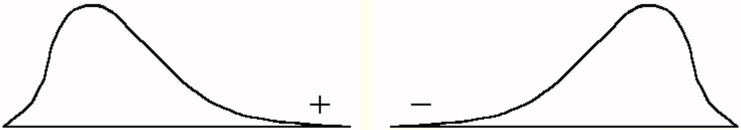
Shape of a Distribution

The shape of a distribution is *symmetric* if the observations are balanced about its center.



Shape of a Distribution

The shape of a distribution is *skewed* or *asymmetric* if the observations are not symmetrically distributed on either side of the center.



skewed-right (positively skewed) distribution
usually mean>median

skewed-left (negatively skewed) distribution
usually mean<median

Example. Consider the following data:

$$10.2, 3.1, 5.9, 7.0, 3.7, 1.9, 6.8, 7.3, 8.2, 4.3, 9.5$$

a) Calculate the mean.

b) Calculate the median.

c) Find the mode(s) if there is (are) any.

d) Comment on the symmetry.

Quartiles

Quartiles are descriptive measures that separate large data sets into four quarters:

First quartile, Q_1, separates approximately the smallest 25% of the data from the rest of the data.
Second quartile, Q_2, separates approximately the smallest 50% of the data from the rest of the data and is, actually, the median.
Third quartile, Q_3, separates approximately the smallest 75% of the data from the rest of the data.

$$\begin{aligned} Q_1 &= \text{ value at } 0.25(n+1)\text{th } \textbf{ordered} \text{ position} \\ Q_2 &= \text{ value at } 0.50(n+1)\text{th } \textbf{ordered} \text{ position} \\ Q_3 &= \text{ value at } 0.75(n+1)\text{th } \textbf{ordered} \text{ position} \end{aligned}$$

Five-Number Summary

The five-number summary refers to the five descriptive measures:

$$\text{minimum} < Q_1 < Q_2 = \tilde{x} < Q_3 < \text{maximum}$$

Note: Do not forget to sort the data (from the smallest to the largest) while calculating elements of five-number summary!

Example. Construct the five-number summary for the following data:

$$60, 84, 65, 67, 75, 72, 80, 85, 63, 82, 70, 75$$

Measures of Variability

The mean alone does not provide a complete or sufficient description of data. Consider the following two sets of sample data:

Sample A:	1	2	1	36
Sample B:	9	8	10	13

Although $\bar{x}_A = \bar{x}_B = 10$, the data in sample A are farther from 10 than the data in sample B.

We need descriptive numbers like range, interquartile range, variance, and standard deviation to measure this spread.

Range and Interquartile Range

Range is the difference between the largest and the smallest observations. For the given samples A and B, we have $\text{range}_{\text{A}} = 36 - 1 = 35$ and $\text{range}_{\text{B}} = 13 - 8 = 5$.

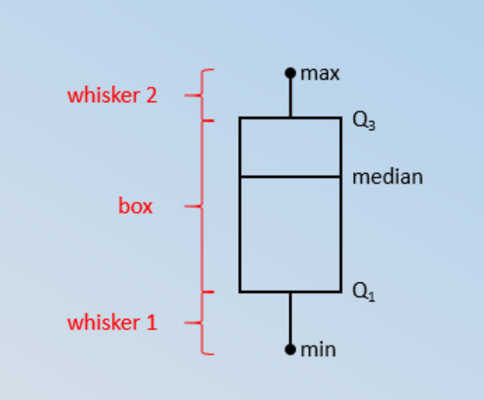
The *interquartile range (IQR)* measures the spread in the middle 50% of the data, that is,

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

In the previous example we obtained $Q_1 = 65.5, Q_2 = 73.5,$ and $Q_3 = 81.5$. So, the IQR for the data in that example is $\text{IQR} = 81.5 - 65.5 = 16$.

Box-and-Whisker Plot

A *box-and-whisker plot* is a graph that describes the shape of a distribution in terms of the five-number summary.



Example. The following table gives the weekly sales (in hundreds of dollars) from a random sample of 10 weekdays from two different locations of the same cafeteria.

Location-1:	6	8	10	12	14	9	11	7	13	11
Location-2:	1	19	2	18	11	10	3	17	4	17

Find mean, median, mode, range, IQR for each location and graph the data with a box-and-whisker plot.

Variance and Standard Deviation

The *population variance*, σ^2, is the sum of the squared differences between each observation and the population mean divided by the population size:

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

The *sample variance*, s^2, is the sum of the squared differences between each observation and the sample mean divided by the sample size minus 1:

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

Variance and Standard Deviation

The *population standard deviation*, σ, is the (positive) square root of the population variance:

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

The *sample standard deviation*, s, is the (positive) square root of the sample variance:

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

Example. A bacteriologist found $6, 12, 9, 10, 6, 8$ microorganisms of a certain kind in 6 cultures. Calculate the sample standard deviation.

Example. Calculate the variance and standard deviation of the following sample data:

$$3, 0, -2, -1, 5, 10$$

z-Score

We know consider another measure called a z-score that examines the location or position of a value relative to the mean of the distribution. A *z-score* is a standardized value that indicates the number of standard deviations a value is from the mean.

z-Score

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 corresponding *z-score* associated with x_i is defined as follows:

$$z = \frac{x_i - \mu}{\sigma}.$$

A z-score greater than zero indicates that the value is greater than the mean, a z-score less than zero indicates that the value is less than the mean, and a z-score of zero indicates that the value is equal to the mean.

Example. A company produces lightbulbs with a mean lifetime of 1200 hours and a standard deviation of 50 hours.

a) Find the z-score for a lightbulb that lasts only 1120 hours.

b) Find the z-score for a lightbulb that lasts 1300 hours.

Example. Suppose that the mean score on the mathematics section of the SAT is 570 with a standard deviation of 40.

a) Find the z-score for a student who scored 600.

b) A student is told that his z-score is -1.5. What is his actual SAT math score?

Measures of Relationships Between Variables

To describe the relationship between two variables we can use

- cross tables (for categorical data),
- scatter plots (graphical way for numerical data),
- covariance and correlation coefficient (numerical way for numerical data).

Covariance

Covariance is a measure of the linear relationship between two variables.

If the covariance is greater than zero, the variables move together; if it is less than zero, the variables move inversely.

Covariance

A *population covariance*, is

$$Cov(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N},$$

where x_i and y_i are the observed values, μ_x and μ_y are the population means, and N is the population size.
A *sample covariance*, is

$$Cov(x, y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

where x_i and y_i are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

Correlation Coefficient

The correlation coefficient is computed by dividing the covariance by the product of the standard deviations of the two variables.

A *population correlation coefficient*, ρ, is

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

A *sample correlation coefficient*, r, is

$$r = \frac{s_{xy}}{s_x s_y}.$$

$-1 \leq r \leq 1$ always holds while s_{xy} depends on the unit of the variables.

- If $0 < r \leq 1$, then there is a positive linear relationship between x and y.
- If $-1 \leq r < 0$, then there is a negative linear relationship between x and y.
- If $r = 0$, then there is no linear relationship between the variables x and y.

There exists a strong linear relationship if $|r| \geq \frac{2}{\sqrt{n}}$

Example. Compute the covariance and correlation coefficient for

$$(12, 20) \quad (15, 27) \quad (14, 21) \quad (17, 26) \quad (18, 30) \quad (19, 32)$$

Alternative Formulae

We can calculate population and sample variances and covariances using the following alternative formulae:

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y}{N} \qquad s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Example. Following is a random sample of price per piece of plywood, X, and quantity sold, Y (in thousands):

Price per piece (x)	Thousands of pieces sold (y)
6	80
10	0
8	70
9	40
7	60

a) Compute the covariance.

b) Compute the correlation coefficient.

c) Discuss the relationship between price per piece and quantity sold.

Optional Material: Weighted Mean

The *weighted mean* of a set of data is

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

where w_i is the weight of the ith observation.

Example. Calculate the GPA of a student who got AA, CB, CC, DC, and BB from 2, 3, 3, 4, and 2 credit courses, respectively. (Assume that AA stands for 4.0, BB for 3.0, CB for 2.5, CC for 2.0, and DD for 1.5.

Solution.

x_i	w_i
AA$= 4.0$	2
CB$= 2.5$	3
CC$= 2.0$	3
DC$= 1.5$	4
BB$= 3.0$	2

$$GPA = \bar{x} = \frac{\sum_{i=1}^{5} w_i x_i}{\sum_{i=1}^{5} w_i} = \frac{2(4.0)+3(2.5)+3(2.0)+4(1.5)+2(3.0)}{2+3+3+4+2} = \frac{33.5}{14} = 2.39$$