STAT 2361
STATISTICS FOR BUSINESS AND ECCONOMICS
STAT 2311
STATISTICS 1
LECTURE NOTES

PREPEARED BY MOHAMMAD MADIAH

CHAPTER 3
DESCRIPTIVE STATISTICS
NUMERICAL MEASURES

3.1 Measures of Location (Central Tendency)

The central tendency of a distribution is an estimate of the **"center"** of a distribution of values. There are three major types of estimates of central tendency the **mean**, the **median** and the **mode**.

1. The Mean

- The (arithmetic) mean (الوسط الحسابي) or the average is probably the most commonly used method of describing central tendency.
- The mean of a **sample** is denoted by $\bar{x}$ (it is read "x bar"), it is an example of **statistic.**
- The mean of a population is denoted by μ (it is read "mu"), it is an example of **parameter**

- The mean of a sample data with n items is given by:

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

- The mean for a population data with N items is given by:

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

Example 1

Find the mean of the following sample data:

4 6 8 10 12 14 16

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{4+6+8++10+12+14+16}{7}$$
$$= \frac{70}{7}$$
$$= 10.$$

2. The Median (الوسيط)

- Another measure of central tendency is the median.
- The **median** is the value found at the **middle** of the set of values that has been arranged in an increasing order.

- The median is most easily computed by arranging the data in an **increasing order** (from the smallest to the largest). The median is the middle value in the distribution.

- If the number of items is **odd** n, the median is the middle item in the list, that is, the value in the position $\frac{n+1}{2}$.

- If the number of items is **even** m, the median is the mean of the two values in positions $\frac{m}{2}$, and $\frac{m}{2}+1$.

Example 2

Find the median for the following data set

11, 2, 3, 7, 2, 7, 6, 4, 7

Solution:

Arrange the values in an increasing order

2, 2, 3, 4, 6, 7, 7, 7, 11

The number of values is 9. So, the median position is $\frac{9+1}{2}=5$

The median is the value in position 5. That is, the median is 6.

Example 3

Find the median for the following data set

5, 7, 6, 7, 8, 2, 10, 13, 12, 15

Solution:

Arrange the values in an increasing order

2, 5, 6, 7, 7, 8, 10, 12, 13, 15

The number of values is 10. So, the median is the mean of the two values in positions $\frac{10}{2}=5,$ and $\frac{10}{2}+1=6$.

The median is the average of the two values in positions 5 and 6. That is, the median is $\frac{7+8}{2}=7.5$.

3. The Mode (المنوال)

- The **mode** of a data set is the value that occurs most frequently.
- The mode is more useful in describing a qualitative data rather than quantitative data. In some data sets there is more than one modal value.

Example 4

Find the mode for the each set of data

(a) 15, 20, 21, 20, 30, 15, 30, 25, 15, 40.

The value 15 occurs three times, so, it is the mode of the data.

(b) 20, 35, 14, 15, 20, 30, 35, 40, 25, 20, 35, 15.

Both 20 and 35 occur three times. That is, the data is bimodal.

(c) 10, 15, 12, 20, 25, 23, 24, 30.

No value of this data set occurs more than once. This data has no mode.

Some observations about the mean, the median, and the mode

- The mean and the median are always used to describe quantitative data, while the mode may be used to describe a qualitative data.
- The mean and the median of a set of data are unique, but the mode is not unique. A set of data may have more than one mode, but it should have only one mean and one median.
- The mean influenced by extreme values (very small or very large data values), but the median does not influenced. In such cases, the median is a more useful measure of central tendency rather than the mean.

Example 5

Consider the following salaries (in dollars).

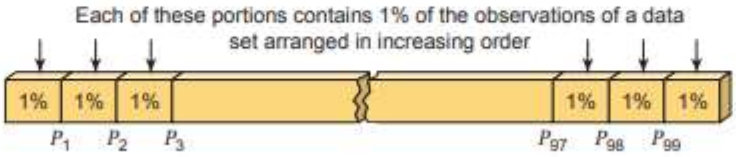
800, 1000, 1200, 1500, 5500

The mean of the data is $2000 while the median is $1200). It is clear that in this data set, the median is a good representative measure of central tendency rather than the mean. The mean is influenced by the extreme value 5500. But, the median represents the center of the data.

4. Percentiles

If you are interested in where a data value stands compared to the rest of the data values, you need a statistic that reports *relative position*, and that statistic is called a **percentile**.

- ❖ The p^{th} percentile is a value in a data set that splits the data into two parts: The lower part contains p percent of the data, and the upper piece contains the rest of the data (which amounts to [100 – *p* percent, because the total amount of data is 100%].
- ❖ A **percentile** is a measure that tells us what percent of the total frequency scored at or below that measure.
  For example the 97th percentile is a value such that 97% of the data values are less than or equal to that value. It is denoted by P_{97}


Each of these portions contains 1% of the observations of a data set arranged in increasing order
1%
1%
1%
1%
1%
1%
P_1
P_2
P_3
P_{97}
P_{98}
P_{99}

- ❖ **To calculate the p^{th} percentile** (where p is any number between zero and one hundred), do the following steps:
  1. Order all the values in the data set from smallest to largest.
  2. Compute the index (number) **i**.

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

  3. If the index obtained in Step 2 is not integer, round it up to the next integer, this integer is the position of the p^{th} percentile.
  4. If the index obtained in Step 2 is an integer **k**, then the p^{th} percentile is the average of the two values in positions **k** and **k+1**.

Example 6

Consider the following test scores (sorted data)

43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77,
78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99

To find the 20th percentile. Start by computing the index i

$$i = \left(\frac{20}{100}\right) * 25 = 5 \quad ;$$

The 20th percentile is the average of the 5th and 6th values in the ordered data set (62 and 66). The 20th percentile then comes to (62 + 66) ÷ 2 = 64.

This mean that 20% of the scores are less than or equal to 64 and 80% are greater than or equal to this value

To find the 90th percentile for these scores,

$$i = \left(\frac{90}{100}\right) * 25 = 22.5 \Rightarrow \text{Round up to 23}$$

Counting from left to right you go until you find the 23rd value in the data set. That value is 98, and it's the 90th percentile for this data set.

This mean that 90% of the scores are less than or equal to 98 and 10% are greater than or equal to this value

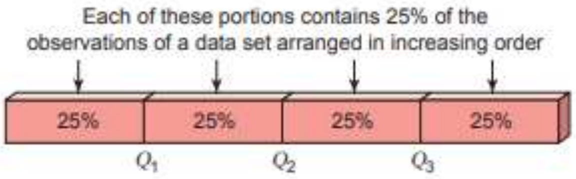
Quartiles

Quartiles break the data set into 4 equal parts. If 100% is broken into four equal parts, we have subdivisions at 25%, 50%, and 75% creating the:

First quartile Q_1 (lower quartile) to be at the 25th percentile.

Second quartile (Median) Q_2 to be at the 50th percentile.

Third quartile Q_3 (upper quartile) to be the 75th percentile.


Each of these portions contains 25% of the observations of a data set arranged in increasing order
25%
25%
25%
25%
Q_1
Q_2
Q_3

3.2 Measures of Variation (Dispersion)

This section concerned with another numerical method of describing data, namely, measures of dispersions or variation.

- Dispersion refers to the spread of the values around the central tendency.
- There are three common measures of dispersion, the range, interquartile range (IQR), and the standard deviations.

1. The Range

The simplest measure of dispersion and the easiest to compute and understand is the range. It is based on the largest and smallest values in the data set. The range is simply the highest value minus the lowest value.

Range = largest data value- smallest data value

Example 1

The range of the data

135,145, **100**, 134, 124, 152, 156, 168, 112, **196**, 115

is the largest value – smallest value = 196 – 100

= 96.

2. The interquartile range

The interquartile range IQR = $Q_3 - Q_1$

Example 2

1, 3, 3, 4, 5, 6, 6, 7, 8, 20

The numbers are already in order

Q1= 3, Q3 = 7

IQR = 7 – 3 = 4

3. The Standard Deviation

The most important measure of dispersion is the standard deviation; it is based on the deviation from the mean. The sample standard deviation is denoted by **s**, while the population standard deviation is denoted by σ (read sigma).

The standard deviation **s** of sample of size **n** is given by the following formula

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

$x-\bar{x}$**: The deviation of the value x from the mean** $\bar{x}$

The population standard deviation σ is given by the following formula

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

The variance

When the standard deviation is squared, the resulting measure of dispersion is called the variance. That is, the sample variance is

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

Example 3

Find the standard deviation of the following sample data

9 6 8 10 12 14 16 13 7 15

Solution

First, calculate $\bar{x}$, the sample mean

$$\bar{x} = \frac{\sum x}{n} = \frac{110}{10} = 11.$$

X	$x-\bar{x}$	$(x-\bar{x})^2$
9	-2	4
6	-5	25
8	-3	9
10	-1	1
12	1	1
14	3	9
16	5	25
13	2	4
7	-4	16
15	4	16
Total	**0**	110

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$
$$= \sqrt{\frac{110}{9}} = 3.49$$

The standard deviation for s the given data is 3.49.

** Notice that the sum of the deviations from the mean is equal to zero (see the second column of the table).

Some observations about the measure of dispersion

- The standard deviation is the most widely reported measures of variation
- The range influenced by extreme values.
- The range, IQR, variance and standard deviation are always positive.
- A relatively small standard deviation indicates that the data values tend to cluster close to the mean, and a relatively large standard deviation shows that the data values are widely scattered from the mean.

Coefficient of Variation

A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean.
The coefficient of variation expresses the standard deviation as a percentage of the mean and is computed as follows.

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad (\mathrm{CV} = \frac{S}{\bar{x}} \times 100\%)$$

If we wish to compare the variability of two or more data sets, we can use the coefficient of variation. The data set for which the coefficient of variation is large indicates that the group is more variable and it is less stable or less uniform. If a coefficient of variation is small it indicates that the group is less variable and it is more stable or more uniform.

Example 4

A company has two sections with 40 and 65 employees respectively. Their average weekly wages are \$450 and \$350. The standard deviation is 90 and 80.

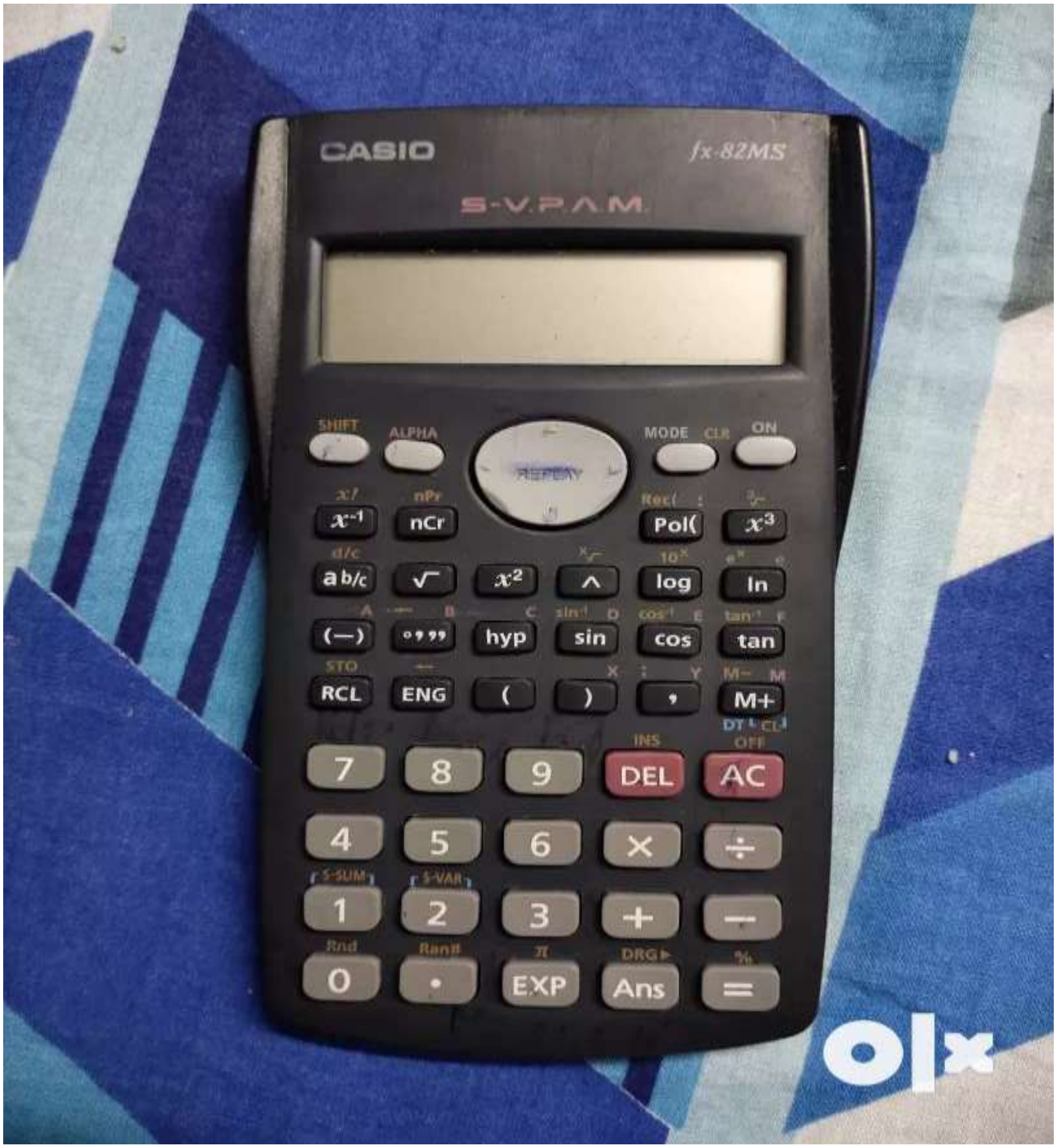
1. Which section has a larger wage bill?
2. Which section has larger variability in wages?

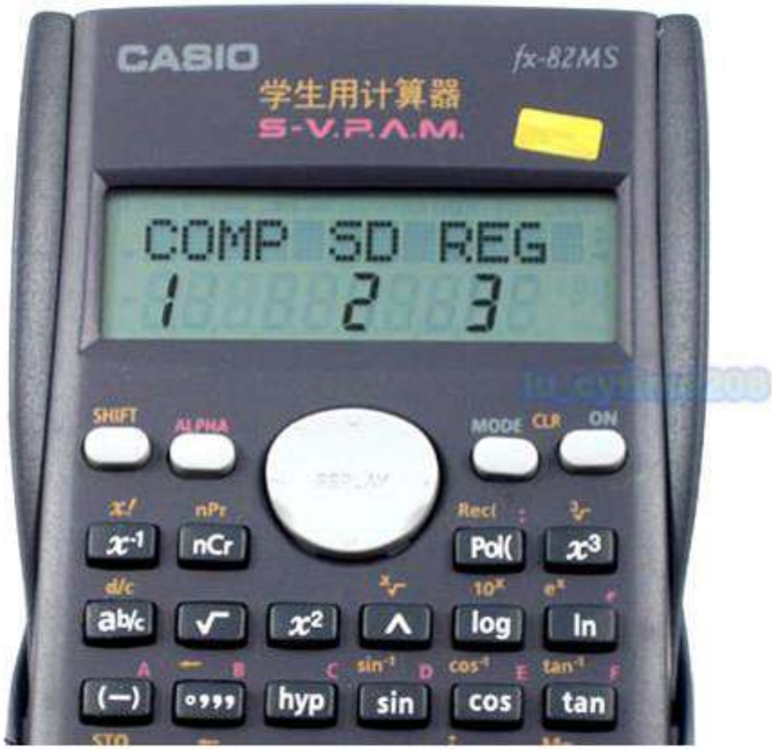
Solution:

Wage bill for section A = 40 x 450 = \$18000
Wage bill for section B = 65 x 350 = \$22750
Section B is larger in wage bill.

Coefficient of variance for Section A = 90/450 x 100 = 20 %
Coefficient of variance for Section B = 80/350 x 100 = 22.85%
Since the coefficient of variation of B is greater than that is of A, there is greater variability in the wages of section B.

CALCULATOR: SD MODE




CASIO
fx-82MS
学生用计算器
S-V.P.A.M.
COMP SD REG
1 2 3
SHIFT
ALPHA
REPLAY
MODE
CLR
ON
nCr
Pol(
log
ln
hyp
sin
cos
tan

Example (Review)

Given the following data set

5, 10, 14, 16, 18, 22, 28, 30, 30, 42, 55

Find the following statistics:

The mean, the median, the standard deviation, coefficient of variation, Q1, Q3, IQR, P_{90}

Solution:

- Q1 = 14

  Q2 = median = 22

  Q3 = 30

  IQR = 16

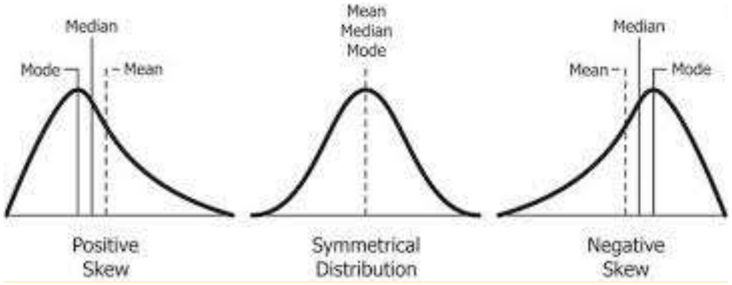
  $P_{90} = 42$
- To find the mean and the standard deviation we will use calculator
  - Reset calculator (clear) Shift mode 3 = =
  - Mode 2 (SD mood)
  - Data entry: 5M+, 10M+, …, 55M+
  - To find the mean: shift 2 1= , $\bar{x} = 24.75$
  - To find the population standard deviation: shift 2 2 =, $o = 13.92$
  - To find the sample standard deviation: shift 2 3 =, s = 14.6
- The sample coefficient of variation $CV = \frac{s}{\bar{x}} \times 100\% = 59\%$

3.3 Measures of distribution shape, relative location, and detecting outliers

Skewness

The histogram discussed in section 2.2 is a graphical method used to show the shape of the distribution (the skewness).

Left Skewed curves (negative skewness), the curve has a longer tail towards the left, and usually in such curves the mean is less than the median.
Right Skewed curves (positive skewness), the curve has a longer tail towards the right, and usually in such curves the mean is greater than the median.
Symmetric curves: In symmetric curves the mean, median and mode are equal.



The z – score (the standardized value)

- A ***z*-score** specifies the location of each x- value within a distribution.
- The sign of the *z*-score (+ or -) signifies whether the score is above the mean (positive) or below the mean (negative).
- The numerical value of the *z*-score specifies the distance from the mean by counting the number of standard deviation units between a value and the mean.
- A **z-score** gives an idea of how far is a data value from the mean.
- The value of the z-score tells you **how many standard** deviations you are away from the mean.
- The z- score measures the relative location of a data value in a data set
- Given a data set with mean $\bar{x}$ and a standard deviation s, the z – score for any data value x is given by

$$z = \frac{x - \bar{x}}{s}$$

Example 1

Given a data with mean of 50 and standard deviation of 6

The z –score for x = 62 is 2

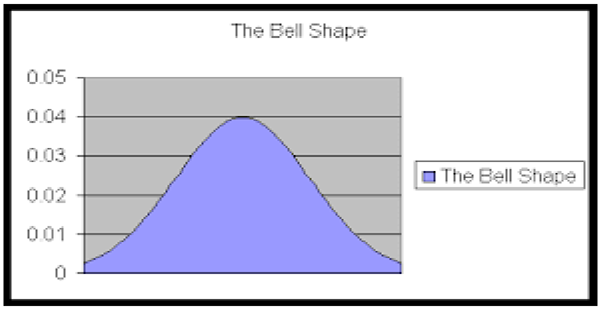
The z –score for x = 50 is 0

The z –score for x = 38 is -2

The z –score for x = 30 is – 3.33

The Empirical Rule

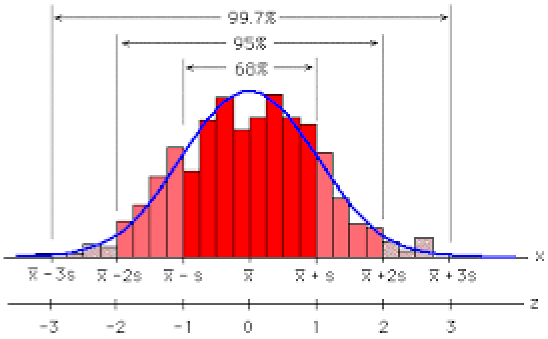
The **Empirical Rule** is a statement about **bell-shaped distributions.** **Bell-shaped distribution:** A symmetric distribution with mean = median = mode.



The Empirical Rule can be used to estimate the percentage of observations that should fall **within** the intervals of one, two, and three standard deviations of the mean.

Empirical Rule: For bell-shaped distributions:

1. About 68% of the data will be within one standard deviation of the mean.
2. About 95% will be within two standard deviations of the mean
3. Almost (about 99.7%) will be within three standard deviations of the mean



Example 2

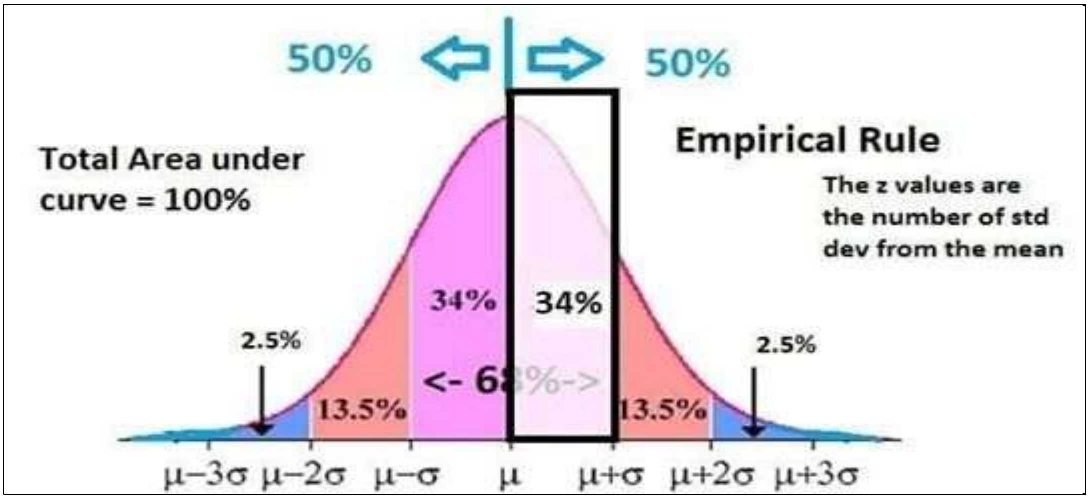
The distribution of IQ scores is bell-shaped with a mean of 100 and standard deviation of 10.

- About 68% of individuals have IQ scores in the interval
  100 ± 1(10) = [90,110]
- About 95% of individuals have IQ scores in the interval

  100 ± 2(10) = [80,120]

- About 99.7% of individuals have IQ scores in the interval

  100 ±3 (10) = [70,130]


50%
50%
Total Area under curve = 100%
Empirical Rule
The z values are the number of std dev from the mean
34%
34%
2.5%
2.5%
<- 68%->
13.5%
13.5%
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ

Example 3

A survey indicates that the weight of BZU students has a bell shaped distribution with mean of 68 kg and standard deviation 10 kg.

1) What is the percentage of student whose weighs between 58 and 78 kg?

The z –score for 58 is (58 – 68)/10 = - 1
The z –score for 78 is (78 – 68)/10 = 1
According to empirical rule 68% of the students have weights between 58 and 68.

2) What is the percentage of student whose weighs between 48 and 88 kg?

The z –score for 48 is - 2
The z –score for 88 is 2
According to empirical rule 95% of the students have weights between 58 and 68.

3) What is the percentage of student whose weighs between 58 and 88 kg?

The z –score for 58 is – 1
68% of the data values are between z = - 1 and z = 1, so 68% /2 = 34 % is between) z = -1 and z = 0 (the mean) (because of the symmetry)
The z –score for 88 is 2
95 % of the data values are between z = -2 and z = 2, so 95%/2 = 47.5 % is between z = 0 (the mean) and z =2 (because of the symmetry)
According to empirical rule 34. % + 47.5 % = 81.5% of the students have weights between 58 and 88.

4) What is the percentage of student whose weighs greater than 88 kg?

The z –score for 88 is 2
95% of the data values are between z = -2 and z =2, so 5% of the data values are outside this range, which means 5%/2 = 2.5% of the data values is greater than z = 2 (weight 88).

5) What is the percentage of student whose weighs is greater than 100 kg?

The z –score for 100 is 3.2
Almost all the data values are between z = – 3 and z = 3,
so the percentage of data greater 100 (z = 3.2) is zero.

Identifying Outliers (extreme values)

Some observations within a set of data may fall outside the general scope of the other observations, that is, some data values are unusually large or unusually small (extreme). Such observations are called **outliers**.

An outliers may be:

- A data value that has been **incorrectly recorded** --- correct this value.
- A data value that has been **incorrectly included** --- remove this value.
- An **unusual** data value ----- Chance.

We have two different methods for identifying outliers: the interquartile range (IQR) method (**The box plot method section 3.4**) and the **z-score method**.

z - Score method

It is unusual for an observation to fall more than 3 standard deviations from the mean. Thus, any observation with a z score less than -3 or greater than +3 is considered an outlier

A data value x is an outlier if its z-value is less than - 3 or greater than 3
That is, it is outside the range $\mu \pm 3\sigma$.

3.4 Exploratory Data Analysis

Five number summary

A data set can be summarized by the following numbers.

1. Smallest data value
2. First Quartile (Q1)
3. Median (Q2)
4. Third Quartile (Q3)
5. Largest data value

Example 1

A five number summary for a data set is given by: **10, 28, 37, 40, and 65**

- The range of the data is 65 – 10 = 55
- The IQR is Q3 – Q1 = 40 – 28 = 12
- The median is 37
- The percentage of data between 28 and 40 is 50% (the percentage of data between Q1 and Q3)
- The percentage of data greater than or equal to 40 is 25% (the percentage of data between Q1 and Q3)

The box plot

- A box plot is a **graphical summary** of data based on five number summary.
- Using the box plot method, we set up a "**fence**" outside of Q1 and Q3. Any value(s) that fall outside of this fence is (are) considered outlier(s).
- A box is constructed with ends located at Q1 and Q3.
- A vertical line is drawn in the box at the location of the median
- To build this fence we compute 1.5 times the IQR and then **subtract** this value from Q1 and **add** this value to Q3.
- The lower limit of the fence is Q1 – 1.5 IQR
- The upper limit of the fence is Q3 + 1.5 IQR
- We draw the **whiskers**, whiskers are two lines drawn from the ends of the box constructed, one in is to the left of the box from Q1 to the smallest data value within the fence, the other is to the right of the box from Q3 to the largest data value within the fence.
- Any value(s) greater than the upper fence limit or less than the lower limit is (are) considered to be an outlier(s) and it is (are) recognized by drawing a ***** at the location of this value.

Example 2

Construct a box plot for the following data

75 69 84 112 74 104 81 90 94 144 79 98

Solution

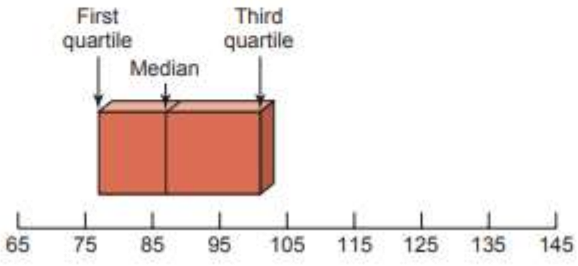
1. Arrange the data in increasing order.
   69 74 75 79 81 84 90 94 98 104 112 144.

2. Calculate the values of the median, the first quartile, the third quartile, and the interquartile range.
   To find Q1
   $i = (\frac{25}{100})(12) = 3 \rightarrow \text{positions } 3 \text{ and } 4 \rightarrow Q_1 = \frac{75+79}{2} = 77$.
   Similarly Q2 = median = 87 and Q3 = 101
   IQR = Q3 – Q1 = 101 – 77 = 24
   1.5 * IQR = 36.

3. Find the limits of the fenc
   Lower Fence limit = Q1 – 1.5 IQR = 77 -36 = 41
   Upper Fence Limit = Q3 + 1.5 IQR = 101 + 36 =137.

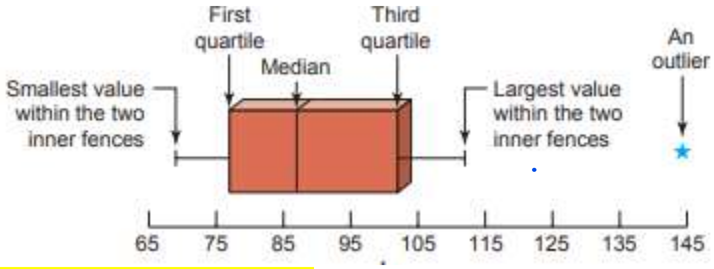
4. Determine the smallest (69) and the largest values (112) in the given data set within the fence limits.

5. Draw a horizontal number line with a suitable scale. Draw three vertical lines at Q1 (77), median (87) and Q3 (101), just above the number line. Join the lines to form a box (draw a box –rectangle- starting from Q1 ending at Q3).



6. Draw two lines, joining the points of the smallest and the largest values within the fence to the box. These values are 69 and 112. The two lines that join the box to these two values are called **whiskers.**

7. Any value that falls outside the two fence limits is shown by marking a (*) and is called an outlier.


First quartile
Third quartile
Median
An outlier
Smallest value within the two inner fences
Largest value within the two inner fences
65
75
85
95
105
115
125
135
145

The value 144 is an outlier

Example 3

Consider the following data:

20, 1, 22, 30, 7, 32, 21, 42, 21, 48, 25

Determine whether the data has (have) an outliers(s).

Solution:

- Arrange the data in ascending order.
  1, 7, 20, 21, 21, 22, 25, 30, 32. 42, 48

- Find the median, first quartile, third quartile (section 3.1)
  - Median (Q2) = 22. First quartile (Q1) = 20. Third quartile (Q3) = 32.
- Find IQR
  - IQR = Q3 – Q1 = 32 – 20 = 12
  - 1.5 IQR = 18.
- Find the fence limits
  - Lower limit = Q1 – 1.5 IQR = 2
  - Upper limit = Q3 + 1.5 IQR = 50.
- Since 1 < 2, 1 is and outlier, and since 48 < 50 it is not an outlier

3.5, 12.2: Measures of association between two variables, Covariance, Correlation, and Regression

Let x and y be two numerical variables. x be the independent variable, y is the dependent variable

Select a sample of size n, then we have the following for the two variables: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

To discuss the relationship between the two variables, the following graph and measure can be used.

I. Scatter Diagram

The first step to get some idea about possible relationship between two variables. It is an indicator "about" the relation between the two variables.

II. Covariance

Covariance provides insight into how two variables are related to one another.

III. Correlation

Correlation analysis is a group of techniques to measure the strength of the relation between two variables.

IV. Regression

In regression analysis we estimate or predict one variable called the dependent variable based on another variable (or variables) called the independent variable.

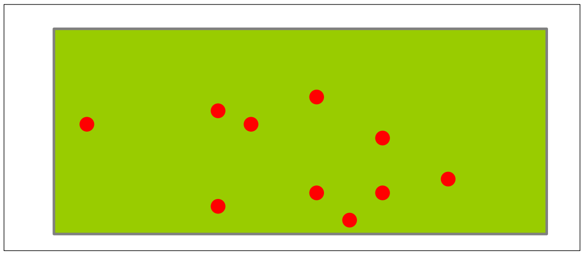
Covariance

- A covariance between two numerical variables x (independent) and y (dependent) is a measure of how the two variables in a data set will change together.
- A **positive covariance** means that the two variables at are positively related, and they move in the same direction.
- A **negative covariance** means that the variables are inversely related, or that they move in opposite directions.

- A **zero covariance** means that there is no relation between the two variables.
- The sample covariance is denoted by s_{xy}.
- The population covariance is denoted by σ_{xy}.
- $$s_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - n\bar{x}\bar{y}}{n-1}.$$
- $$\sigma_{xy} = \frac{\sum (x-\mu_y)(y-\mu_y)}{N}.$$

Correlation

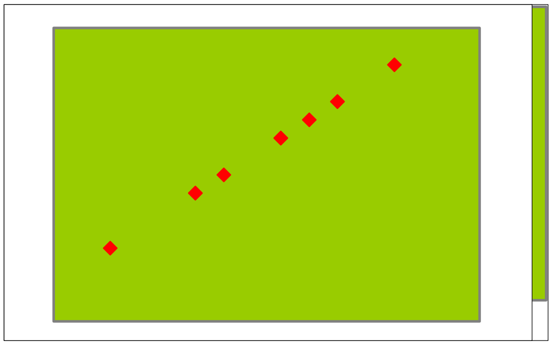
- A **correlation coefficient** measures the **strength** of the linear association between two variables x and y.
- The linear correlation coefficient measures how closely the points in a scatter diagram are spread about the **trend line.**
- The linear correlation coefficient for a sample is denoted by r_{xy}.
- The linear correlation coefficient for a population is denoted by ρ_{xy}.
- $r_{xy} = \dfrac{s_{xy}}{s_x . s_y}$, s_x is the standard deviation for the variable x and s_y is the standard deviation for the variable y.
- $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x . \sigma_y}$
- The value of r is always in the range -1 to 1($-1 \le r_{xy} \le 1$.)
- A value of correlation coefficient equals to zero indicates no linear relationship between the two variables. See the following scatter diagram.



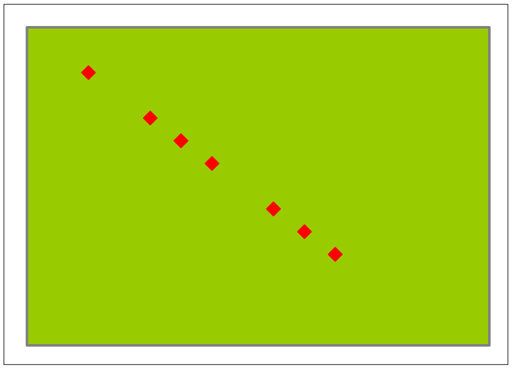
r = 0

- A correlation coefficient of r = 1 or r = - 1 indicates that the trend line goes exactly through all the points.

- A value of r = 1 corresponds to a **perfect positive** linear relationship between the two variables



- A value of r = -1 corresponds to a **perfect negative** linear relationship between the two variables.



- Values of r near zero indicate a **weak** relationship.
- Values of r near 1 (or -1) indicate a **strong** positive (or negative_ relationship.

Linear Regression

- After a scatter diagram has been graphed, we can draw a curve so that it is the best curve representing all the data values that are scattered.
- The equation that estimated the regression equation for a population is called the **estimated regression equation;** this equation is developed from sample data by using a method called the **least square method**.

- For a simple linear regression, the estimated regression equation is

$$\hat{y} = b_0 + b_1 x$$
$$\hat{y} = A + Bx$$

- The resulting line is called **the least square line** (the regression line.)
- $\hat{y}$ is the estimated or predicted value of y for a given value of x. The constants b_0, b_1 are computed using the following formulas:

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{s_{xy}}{s_x^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

Example

Consider the following distribution

	x	Y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1.	2	3	-2	4	-3	9	**6**
2.	3	5	-1	1	-1	4	**1**
3.	5	7	1	1	1	1	**1**
4.	6	9	2	4	3	9	**6**
Total	**16**	**24**	**0**	**10**	**0**	**20**	**14**

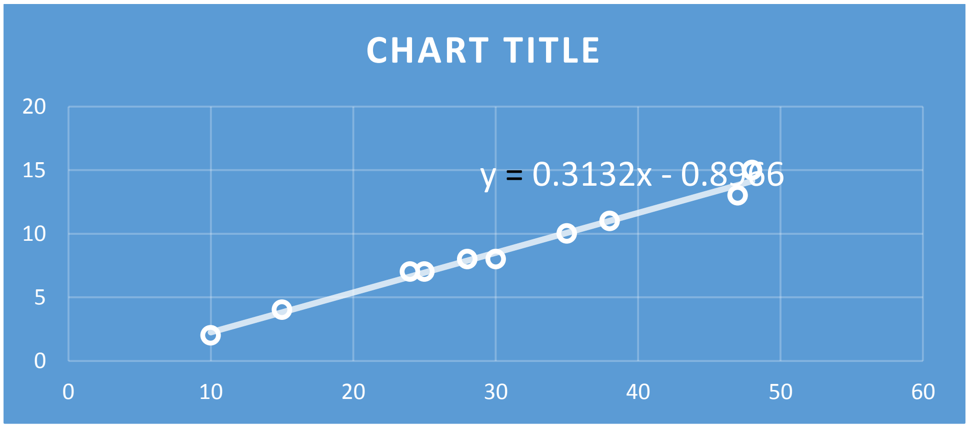
$$\bar{x} = 4,$$
$$\bar{y} = 6,$$
$$s_x = 1.83$$
$$s_y = 2.58$$
$$s_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = 4.67$$
$$r_{xy} = \frac{s_{xy}}{s_x . s_y} = \frac{4.67}{(1{,}83)(2.58)} = 0.99$$
$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{4.67}{(1{,}83)^2} = 1.4$$
$$b_0 = \bar{y} - b_1 \bar{x} = 0.4$$

Example

The income and food expenditures of ten households is given in the following table (in $100)

Household	Income X	Food expenditure Y
1.	35	10
2.	47	13
3.	25	7
4.	38	11
5.	15	4
6.	28	8
7.	30	8
8.	48	15
9.	10	2
10.	24	7

1. Graph the scatter diagram



2. Use the **REG mode** to find the following

- Shift 3 = = **(RESET)**
- Mode 3 1 **(REG MODE)**
- 35 ~~shift~~ , 10 M+, 47 ~~shift~~ , 13 M+, …, 24 ~~shift~~, 7 M+ **(DATA ENTRY)**
  Shift, =;
- x : Shift 2 **(x – statistics)**
- y : Shift 2 **replay**⇒**(y – statistics)**
- A, B, r : Shift 2 **replay**⇒ ⇒ **(Measures of Association)**

$$\bar{x} = 30,$$

$$\bar{y} = 8.5,$$

$$s_x = 12.44$$

$$s_y = 3.92$$

$$r_{xy} = 0.99$$

$$s_{xy} = (r_{xy})(s_x)(s_y) = 48.27$$

$$b_0 = -0.9$$

$$b_1 = 0.31$$

3. Comment on the strength of the relation between the two variables

The correlation coefficient ($r_{xy} = 0.99$) is close to 1, so there is a strong positive relationship between the income and the expenditure.

4. Write the regression equation

$$\hat{\mathbf{y}} = \mathbf{b_0} + \mathbf{b_1 x}$$
$$= \mathbf{-\ 0.9 + 0.31x}$$

5. Using the estimated equation; we can find the expected value of y for a given value of x. For example, the predicted food expenditure for a family with of $4000 (x = 40).

$$\hat{\mathbf{y}} = \mathbf{-\ 0.9 + 0.31(40)}$$
$$= \mathbf{(13.3)(100) = \$1330}$$

3.6 The Weighted mean and Working with Grouped Data

Weighted mean

A special case of arithmetic mean is the weighted mean. It occurs when there are several observations of the same value. The weighted mean of n values

$x_1, x_2, ..., x_n$ that are weighted by the respective weights $f_1, f_2, ..., f_n$ is given by

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Example 1

The salaries for 20 employees of a company are as shown in the below table. Find the mean salary

Salary in 1000 ILS x	Number of Employees f	Xf
6	4	24
8	8	64
10	6	60
12	2	24
Total	20	172

Solution:

Add a column to the table that represents the product of salary and number of employees.

The mean salary is:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{172}{20} = 8.6\ (1000) = 8600 \text{ ILS}$$

The mean for a grouped data

To find the mean for a data that organized in a frequency distribution (grouped data), we start by adding a column to the table that represents the class midpoint of each class, another column in which the class midpoint is multiplied by the frequency of each class is added to the table. The mean for a grouped data is given in the following formula

$$\bar{x} = \frac{\sum m_i f_i}{\sum f_i} = \frac{\sum m_i f_i}{n} \qquad \text{(n = sample size)}$$

Where:

$\mathrm{f_i}$ is the frequency of the ith class

$\mathrm{m_i}$ is the class midpoint of the ith class = $\frac{\text{upperlimit} + \text{lower limit}}{2}$

Example 2

Find the mean of the following frequency distribution

Class	Frequency	$\mathrm{m_i}$	$\mathrm{f_i m_i}$
3 – 7	4	5	20
8 – 12	6	10	60
13 – 17	8	15	120
18 – 22	2	20	40
Total	20		240

Add two columns for the above table, one for the class midpoint $\mathrm{m_i}$ and the other for the multiplication of class midpoint and frequency $\mathrm{f_i m_i}$

So, $\bar{x} = \frac{\sum m_i f_i}{\sum f_i} = \frac{240}{20} = 12$

The standard deviation for a grouped data

To find the standard deviation of a data that is grouped into a frequency distribution table, the following formulas can be used;

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

Example 3

Determine the standard deviation for the distribution in example 2

Class	Frequency	m	fm	$\mathrm{m} - \bar{\mathrm{x}}$	$(\mathrm{m} - \bar{\mathrm{x}})^2$	$(\mathrm{m} - \bar{\mathrm{x}})^2 \mathrm{f}$
3 – 7	4	5	20	-7	49	196
8 – 12	6	10	60	-2	4	24
13 – 17	8	15	120	3	9	72
18 – 22	2	20	40	8	64	128
Total	20		240			420

$$s = \sqrt{\frac{\sum (m_i - \overline{x})^2 f_i}{(\sum f_i) - 1}} = \sqrt{\frac{420}{19}} = 4.7$$

CALCULATOR

x	f
35	10
20	15
36	19
40	21
27	25

- ❖ Shift 3 = = **(RESET)**
- ❖ Mode 2 **(SD MODE)**
- ❖ 35 shift , 10 M+, 20 shift , 15 M+, …17, shift , 7 M+ **(DATA ENTRY)**
- ❖ Shift 2 1 = **(mean)**
- ❖ Shift 2 3 **(sample sd)**