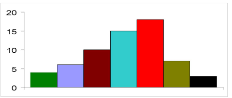
Chapter Three:
Data Description

Data Summarization

Numerical Measures of the Data



Outline

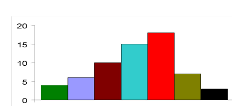
Introduction

3-1 Measures of Central Tendency

3-2 Measures of Variation

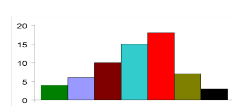
3-3 Measures of Position

3-4 Exploratory Data Analysis



Objectives

1. Summarize data using the measures of central tendency, such as the mean, median, mode, and midrange.
2. Describe data using the measures of variation, such as the range, variance, and standard deviation.
3. Identify the position of a data value in a data set using various measures of position, such as percentiles, and quartiles.
4. Use the techniques of exploratory data analysis, including stem and leaf plots, box plots, and five-number summaries to discover various aspects of data.



3-1 Measures of Central tendency

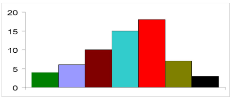
We will compute two means: one for the sample and one for a finite population of values.

The symbol $\overline{X}$ represents the **sample mean**

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X}{n}.$$

The Greek symbol μ represents the population mean. The symbol μ is read as "mu".
N is the size of the finite population.

$$\mu = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum X}{N}.$$



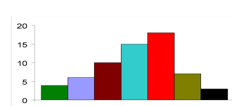
Example:- (Sample Mean)

The ages of a random sample of seven students at a certain school are 11, 10, 12, 13, 7, 9, 15

Find the average (Mean) age of this sample

$$\bar{X} = \frac{\sum X}{n} = \frac{11+10+12+13+7+9+15}{7}$$

$$= \frac{77}{7} = 11 \text{ years.}$$



Example:- population mean

A small company consists of the owner, the manager, the salesperson, and two technicians. The salaries are listed as \$5000, 2000, 1200, 900 *and* 900 *respectively. (Assume this is the population.)*

Then the population mean will be

$$\mu = \frac{\sum X}{N}$$

$$= \frac{5000 + 2000 + 1200 + 900 + 900}{5}$$

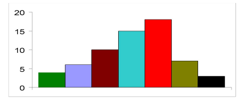
$$= \$2000.$$

The Sample Mean for an Ungrouped Frequency Distribution

The mean for an ungrouped frequency distributu ion is given by

$$\bar{X} = \frac{\sum (f \cdot X)}{n}.$$

Here f *is the frequency for the corresponding value of* X, *and* $n = \sum f$.



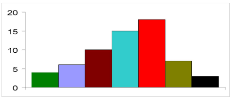
The Sample Mean for an Ungrouped Frequency Distribution –

Example

The scores for 25 students on a 4 – point quiz are given in the table. Find the mean score.

Score	Frequency	f.X
0	2	0
1	4	4
2	12	24
3	4	12
4	3	12

$$\overline{X} = \frac{\sum f \cdot X}{n} = \frac{52}{25} = 2.08.$$



The Sample Mean for a Grouped Frequency Distribution

The mean for a grouped frequency distribution is given by :

$$\overline{X} = \frac{\sum (f \cdot X_m)}{n}$$

Here X_m is the corresponding class midpoint

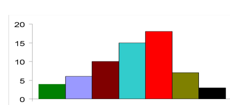
Given the table below, find the mean.

Class	Frequency	X_m	$f.X_m$
15.5 - 20.5	3	18	54
20.5 - 25.5	5	23	115
25.5 - 30.5	4	28	112
30.5 - 35.5	3	33	99
35.5 - 40.5	2	38	76

$$\sum f \cdot X_m = 54+115+112+99+76$$
$$= 456$$

and $n = 17$. *So*

$$\overline{X} = \frac{\sum f \cdot X_m}{n}$$
$$= \frac{456}{17} = 26.82.$$



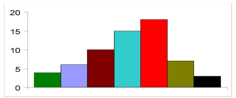
Important remark :

- In some situations the mean may not be representative of the data.
- As an example, the annual salaries of five vice presidents at AVX, LLC are \$90,000, \$92,000, \$94,000, \$98,000, and \$350,000. The mean is:

$$\mu = \frac{\Sigma X}{N} = \frac{(\$90,000 + \$92,000 + \$94,000 + \$98,000 + \$350,000)}{5}$$

$$= \frac{\$724,000}{5} = \$144,800$$

- Notice how the one extreme value (\$350,000) pulled the mean upward. Four of the five vice presidents earned less than the mean, raising the question whether the arithmetic mean value of \$144,800 is typical of the salary of the five vice presidents.



Properties of the mean

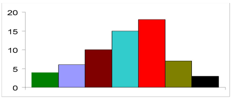
- As stated, the mean is a widely used measure of central tendency . It has several important properties.

1. Every set of interval level and ratio level data has a mean.
2. All the data values are included in the calculation.
3. A set of data has only one mean, that is, the mean is unique.
4. The mean is a useful measure for comparing two or more populations.
5. The sum of the deviations of each value from the mean will always be zero, that is $\sum(X-\overline{X})=0$
6. The mean is highly affected by extreme data .

Note: Illustrating the fifth property

Consider the set of values: 3, 8, and 4. The mean is 5.

$$\Sigma(X-\overline{X})=[(3-5)+(8-5)+(4-5)]=0$$



- **Median :** The median splits the ordered data into halves

the symbol used to denote the **median** is m_e

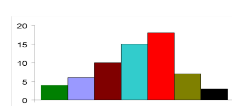
Example:- The weights (in pounds) of seven army recruits are 180, 201, 220, 191, 219, 209, and 186. Find the median.

Arrange the data in order and select the middle point.

Data array: 180, 186, 191, 201, 209, 219, 220.

The median, = 201.

In the previous example, there was an odd number of values in the data set. In this case it is easy to select the middle number in the data array.



When there is an **even number** of values in the data set, the median is obtained by taking the **average of the two middle numbers**.

Example:-

Six customers purchased the following number of magazines: 1, 7, 3, 2, 3, 4. Find the median.

Arrange the data in order and compute the middle point.

Data array: 1, 2, 3, 3, 4, 7.

The median, $m_e = (3 + 3)/2 = 3.$

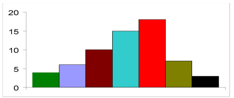
Example:-Find the median grade of the following sample

62, 68, 71, 74, 77, 82, 84, 88, 90, 94

62, 68, 71, 74, 77 82, 84, 88, 90, 94

5 on the left 5 on the right

$m_e = 79.5$



example

- Find the median grade of the following sample of students grades :

A B A D F D F A B C C C F D A F D A A B B F D A B F C

- Data array:

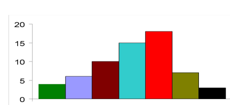
F F F F F F D D D D D C C C C B B B B B A A A A A A A

The median grade is : C

Half of the students had at least C (a grade less than or equal C.

Half of the students had at most C (a grade more than or equal C .

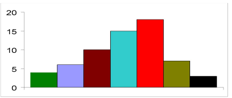
The median can be determined for ordinal level data .



Properties of the Median

- The major properties of the median are:

1. The median is a unique value, that is, like the mean, there is only one median for a set of data.
2. It is not influenced by extremely large or small values and is therefore a valuable measure of central tendency when such values do occur.
3. It can be computed for ratio level, interval level, and ordinal-level data.
4. Fifty percent of the observations are greater than the median and fifty percent of the observations are less than the median.



Mode:- is the score that occurs most frequently (denoted by M)

Example:- The following data represent the duration (in days) of U.S. space shuttle voyages for the years 1992-94. Find the mode.

Data set: 8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11.

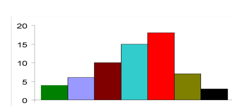
Ordered set: 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14.

Mode = 8 days.

Example:- Six strains of bacteria were tested to see how long they could remain alive outside their normal environment. The time, in minutes, is given below. Find the mode.

Data set: 2, 3, 5, 7, 8, 10.

There is no mode. since each data value occurs equally with a frequency of one.



Example:- Eleven different automobiles were tested at a speed of 15 mph for stopping distances. The distance, in feet, is given below. Find the mode.

Data set: 15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26.

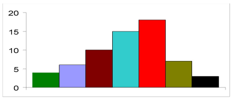
There are two modes (bimodal). The values are 18 and 24.

The Mode for an Ungrouped Frequency Distribution

Example

Values	Frequency, f
15	3
20	5
25	8
30	3
35	2

Mode



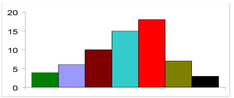
The Mode for a Grouped Frequency Distribution –

Can be approximated by the midpoint of the modal class.

Example

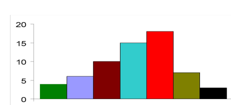
Modal Class

Class	*Frequency, f*
15.5 - 20.5	*3*
20.5 - 25.5	*5*
25.5 - 30.5	*7*
30.5 - 35.5	*3*
35.5 - 40.5	*2*



Properties of the Mode

1. The mode can be found for all levels of data (nominal, ordinal, interval, and ratio).
2. The mode is not affected by extremely high or low values.
3. A set of data can have more than one mode. If it has two modes, it is said to be bimodal.
4. A disadvantage is that a set of data may not have a mode because no value appears more than once.



The **weighted mean** is used when the values in a data set are not all equally represented.

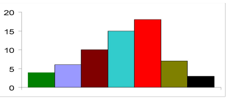
The **weighted mean of a variable** *X* is found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

The weighted mean

$$\overline{X}_w = \frac{w_1 X_1 + w_2 X_2 + \ldots + w_n X_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum wX}{\sum w}$$

where $w_1, w_2, \ldots, w_n$ *are the weights*

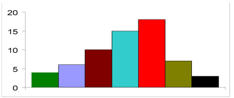
for the values $X_1, X_2, \ldots, X_n$.



Example:- During a one hour period on a hot Saturday afternoon a boy served fifty drinks. He sold five drinks for \$0.50, fifteen for \$0.75, fifteen for \$0.90, and fifteen for \$1.10. Compute the weighted mean of the the price of the drinks :

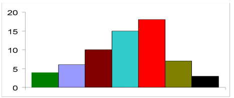
$$\overline{X}_w = \frac{5(\$0.50)+15(\$0.75)+15(\$0.90)+15(\$1.15)}{5+15+15+15}$$

$$= \frac{\$44.50}{50} = \$0.89$$



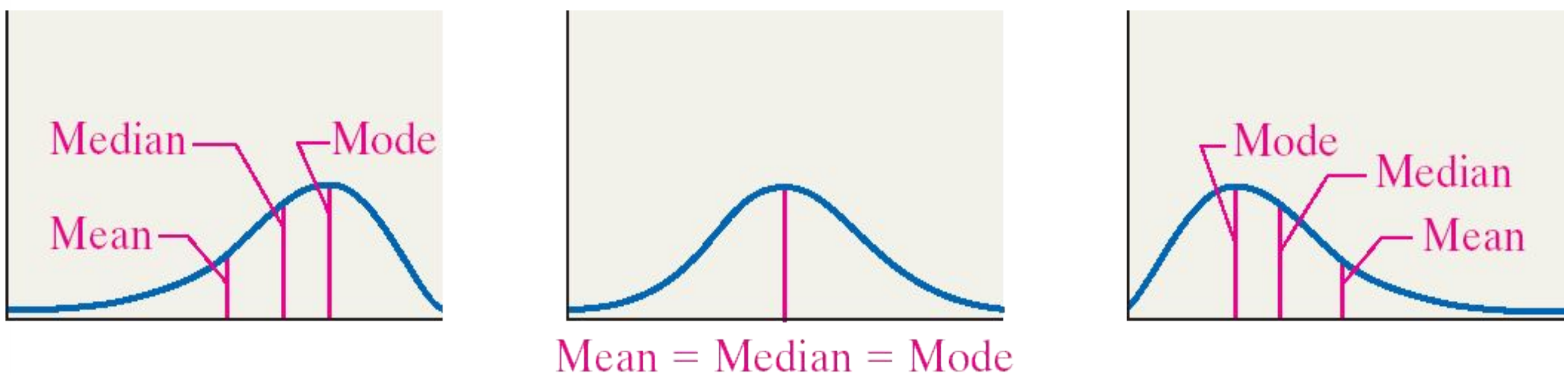
Best measure of central tendency

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median



Relationship between mean , median and mode and the shape of the distribution

- Symmetric – the mean =the median=the mode
- Skewed left – the mean will usually be smaller than the median
- Skewed right – the mean will usually be larger than the median



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

3-2 Measures of Dispersion(variation)

- Learning objectives
  - The range of a variable
  - The variance of a variable
  - The standard deviation of a variable
  - Use the Empirical Rule
- Comparing two sets of data
- The measures of central tendency (mean, median, mode) measure the differences between the "average" or "typical" values between two sets of data
- The measures of dispersion in this section measure the differences between how far "spread out" the data values are.

Variability -- provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together.

- Tells how meaningful measures of central tendency are
- Help to see which scores are outliers (extreme scores)

Why do we Study Dispersion?

A direct comparison of two sets of data based only on two measures of central tendency such as the mean and the median can be misleading since an average does not tell us anything about the spread of the data.

See Example 3-15 page 128 of your text book

Comparison of two outdoor paints : 6 gallons of each brand have been tested and the data obtained show how long (in months) each brand will last before fading .

Brand A : 10 60 50 30 40 20

Brand B : 35 45 30 35 40 25

Calculate the mean for each brand :

Measures of dispersion are :

1. The range ,
2. The interquartile range ,
3. The variance and standard deviation ,
4. The coefficient of variation

The **range** (R) of a variable is the difference between the largest data value and the smallest data value

$$R = \text{highest value} - \text{lowest value}.$$

Properties of the range

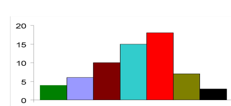
1. Only two values are used in the calculation.
2. It is influenced by extreme values.
3. It is easy to compute and understand.

Example

- Compute the range of 6, 1, 2, 6, 11, 7, 3, 3
- The largest value is 11
- The smallest value is 1
- Subtracting the two … 11 – 1 = 10 … the range is 10

Relative measure of Range called coefficient of Range

$$Coeff.of\ Range = \frac{H - L}{H + L}$$



The variance of a variable

The variance is based on the deviation from the mean

$(x_i - \mu)$ for populations

$(x_i - \bar{x})$ for samples

To treat positive differences and negative differences, we square the deviations

$(x_i - \mu)^2$ for populations

$(x_i - \bar{x})^2$ for samples

The **population variance** of a variable is the sum of the squared deviations of the data values from the mean divided by the number in the population

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

where

$X = individual\ value$

$\mu = population\ mean$

$N = population\ size$

The population variance is represented by $\boldsymbol{\sigma^2}$

Standard deviation: The square root of the variance.

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

i.e. the square root of the arithmetic mean of the squares of deviations from arithmetic

Properties of the variance and standard deviation

1. it is the typical or approx. average distance from the mean
2. if it is small, then scores are clustered close to mean; if it is large, they are scattered far from mean
3. it describes how variable or spread out the scores are.
4. it is very influenced by extreme scores
5. The measurement units of the variance are square of the original units. While the measurement of the SD is same as the original data
6. All values are used in the calculation.
7. Variance and St. dev are always greater than or equal to zero. They are equal zero only if all observations are the same.

The **sample variance** of a variable is the sum of the squared deviations of data values from the mean divided by one less than the number in the sample

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

The sample variance is represented by s^2

Sample standard deviation (s)

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

We say that this statistic has $n - 1$ degrees of freedom

Example;- Find the variance and standard deviation for the following sample: 16, 19, 15, 15, 14.

$\Sigma X = 16 + 19 + 15 + 15 + 14 = 79.$

$\Sigma X^2 = 16^2 + 19^2 + 15^2 + 15^2 + 14^2 = 1263.$

Using the short cut formula (without calculating the mean)

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

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

$$s = \sqrt{3.7} = 1.9235$$

Symbols for Standard Deviation

	Sample	Population	
Textbook	s	σ	Book
Some graphics calculators	Sx	σx	Some graphics calculators
Some non-graphics calculators	$x\sigma_{n-1}$	$x\sigma_n$	Some non-graphics calculators

Articles in professional journals and reports often use SD for standard deviation and VAR for variance.

Sample Variance for Grouped and Ungrouped Data

For grouped data, use the class midpoints for the observed value in the different classes.

For ungrouped data, use the same formula with the class midpoints, X_m, replaced with the actual observed X value.

Example:-

Find the variance and SD for the following data set

2,3,4,5,2,2,2,3,2,4,3,2,5,2,3,3,4,2,5,4,4,3,3,2,5,2

Step one put the data I ungrouped frequency

Value (x)	Frequency f	x^2	$f.x$	$f.x^2$
2	10	4	20	40
3	7	9	21	63
4	5	16	20	80
5	4	25	20	100
Total	**26**		**81**	**283**

$$s^2 = \frac{n\Sigma f(x)^2 - (\sum fx)^2}{n(n-1)} = \frac{26(283) - 81^2}{26(26-1)}$$

$$= \frac{797}{650} = 1.2262$$

$$s = \sqrt{1.2262} = 1.1073$$

Example:- find the variance and SD for the frequency distribution of the data representing number of miles that 20 runners run during one week

Class	Freq. f	Midpoint	$f.x_m$	x_m^2	$f.x_m^2$
5- 11	1	8	8	64	64
11-17	2	14 x_m	28	196	392
17-23	3	20	60	400	1200
23-29	5	26	130	676	3380
29-35	4	32	128	1024	4096
35-41	3	38	114	1444	4332
41-47	2	44	88	1936	3872
total	20		556		17336

$$s^2 = \frac{n\Sigma f(x)^2 - (\sum fx)^2}{n(n-1)} = \frac{20(17336) - 556^2}{20(20-1)}$$

$$= \frac{37584}{380} = 98.905$$

$$s = \sqrt{98.905} = 9.95$$

Interpretation and Uses of the Standard Deviation

The standard deviation is used to measure the spread of the data. A small standard deviation indicates that the data is clustered close to the mean, thus the mean is representative of the data. A large standard deviation indicates that the data are spread out from the mean and the mean is not representative of the data.

Coefficient of Variation :- $C.V.$

The relative measure of St. Dev. is the **coefficient of variation** which is defined to be the standard deviation divided by the mean. The result is expressed as a percentage.

$$C.V. = \frac{\sigma}{\mu}.100\%$$

Or

$$C.V. = \frac{s}{\bar{x}}.100\%$$

<u>Important note:</u>

The coefficient of variation should only be computed for data measured on *a ratio scale*.

See the following example

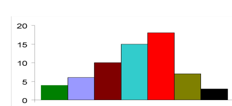
Example :

- To see why the coefficient of variation should not be applied to interval level data, compare the same set of temperatures in Celsius and Fahrenheit:

  Celsius: [0, 10, 20, 30, 40]

  Fahrenheit: [32, 50, 68, 86, 104]

- The CV of the first set is 15.81/20 = 0.79. For the second set (which are the same temperatures) it is 28.46/68 = 0.42
- **So the coefficient of variation does not have any meaning for data on an interval scale.**



Advantages

The *coefficient of variation* is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a unitless (***dimensionless)number***. So when comparing between data sets with different units or widely different means, **one should use the coefficient of variation for comparison instead of the standard deviation.**

Disadvantages

When the mean value is near zero, the coefficient of variation is sensitive to small changes in the mean, limiting its usefulness.

.

Example:- Data about the annual salary (000's) and age of CEO's in a number of firms has been collected. The means and standard deviations are as foll

	Mean	SD
Salary	404.2	220.5
Age	51.47	8.92

- Which distribution has more dispersion? Is direct comparison appropriate?

Salary and age are measured in different units and the means show that there is also a significant difference in magnitude

Direct comparison is not appropriate

	Mean	SD	C.V.
Salary	404.2	220.5	54.55%
Age	51.47	8.92	17.33%

Comparing CV's we can now see clearly that the dispersion or variability relative to the mean is greater for CEO annual salary than for age.

Measure of position:

Measures of position are used to locate the relative position of a data value in the data set

1- Standard Scores

To compare values of different units a z-score for each value is needed to be obtained then compared

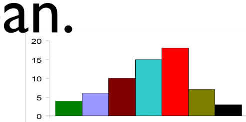
A z-score or standard score for each value is obtained by

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

$$or$$

For population
$$z = \frac{x - \mu}{\sigma}$$

The z-score represents the number SD that a data value falls above or below the mean.



Standard Scores (or z-scores) specify the exact location of a score within a distribution relative to the mean

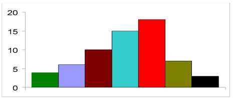
- The sign (- or +) tells whether the score is above or below the mean
- The numerical value tells the distance from the mean in terms of standard deviations

E.g., a z-score of -1.3 tells us that the raw score fell 1.3 standard deviations *below* the mean.

-

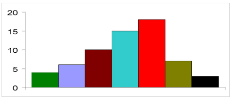
Raw score is the original, untransformed score.

To make them more meaningful, raw scores can be converted to z-scores.



Characteristics of Standard Scores

1. The shape of the distribution of standard scores is the same as the shape of the distribution of raw scores (the only thing that changes is the units on the x-axis)
2. The mean of a set of standard scores = 0.
3. The St. deviation of a set of standard scores = 1.
4. ***A standard score of greater than +3 or less than - 3 is an extreme score, or an outlier.***



Example:- A student scored 65 on a statistics exam that had a mean of 50 and a standard deviation of 10. Compute the *z*-score.

$z = (65 – 50)/10 = 1.5.$

That is, the score of 65 is 1.5 standard deviations above the mean.

Above - since the *z*-score is positive.

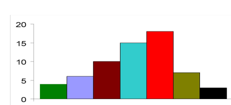
Assume that this student scored 70 on a math exam that had a mean of 80 and a standard deviation of 5 .

Compute the z-score .

$Z= (70-80)/5=-2$

That is, the score of 70 is 2 standard deviations below the mean.

below - since the *z*-score is positive.

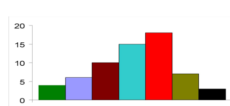


Example:- a student scored 65 on a calculus test that had a mean of 50 and a SD of 10. she scored 30 on statistics test with a mean of 25 and variance of 25, compare relative positions of the two tests.

$$z_{Cal} = \frac{x-\bar{x}}{s} = \frac{65-50}{10} = 1.5$$

$$z_{stat} = \frac{30-25}{5} = 1.0$$

Since the z-score for calculus is larger , her relative position in the calculus class is higher than her relative position in the statistics class.



2. Quartiles

Quartiles divide the data set into 4 groups.

Quartiles are denoted by Q_1, Q_2, and Q_3.

The median is the same as Q_2.

Finding the Quartiles

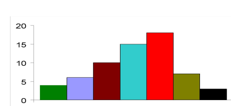
Procedure: Let Q_k be the k^{th} quartile and *n* the sample size.

Step 1: Arrange the data in order.

Step 2: Compute ***c* = ({*n*+1}·*k*)/4.**

Step 3: If *c* is not a whole number, round off to whole number. use the value halfway between x_c and x_{c+1}.

Step 4: If *c* is a whole number then the value of x_c is the position value of the required percentile.



Example:

For the following data set: 2, 3, 5, 6, 8, 10, 12

Find Q_1 and Q_3

$n = 7$, so for Q_1 we have $c = ((7+1)\cdot 1)/4 = 2$.

Hence the value of Q_1 is the 2nd value.

Thus Q_1 for the data set is 3.

for Q_3 we have $c = ((7+1)\cdot 3)/4 = 6$.

Hence the value of Q_3 is the 6th value.

Thus Q_3 for the data set is 10.

Example: Find Q_1 and Q_3 for the following data set:

2, 3, 5, 6, 8, 10, 12, 15, 18.

Note: the data set is already ordered.

n = 9, so for Q_1 we have c = ((9+1)· 1)/4 = 2.5.

Hence the value of Q_1 is the halfway between the 2nd value and 3rd value.

$$Q_1 = \frac{3+5}{2} = 4$$

for Q_3 we have c = ((9+1)· 3)/4 = 7.5.

Hence the value of Q_3 is the halfway between the 7th value and 8th value

$$Q_3 = \frac{12+15}{2} = 13.5$$

Example:

For the following data set: 2, 3, 5, 6, 8, 10, 12

Find Q_1 and Q_3

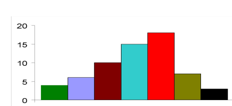
The median for the above data is ***6***

The median for the lower group of data which is less than median is ***3***

So the value of Q_1 is the 2nd value which means that Q_1 =3.

The median for the upper group of data which is grater than median is ***10***

So the value of Q_3 is the 6th value which means that Q_3 =10.



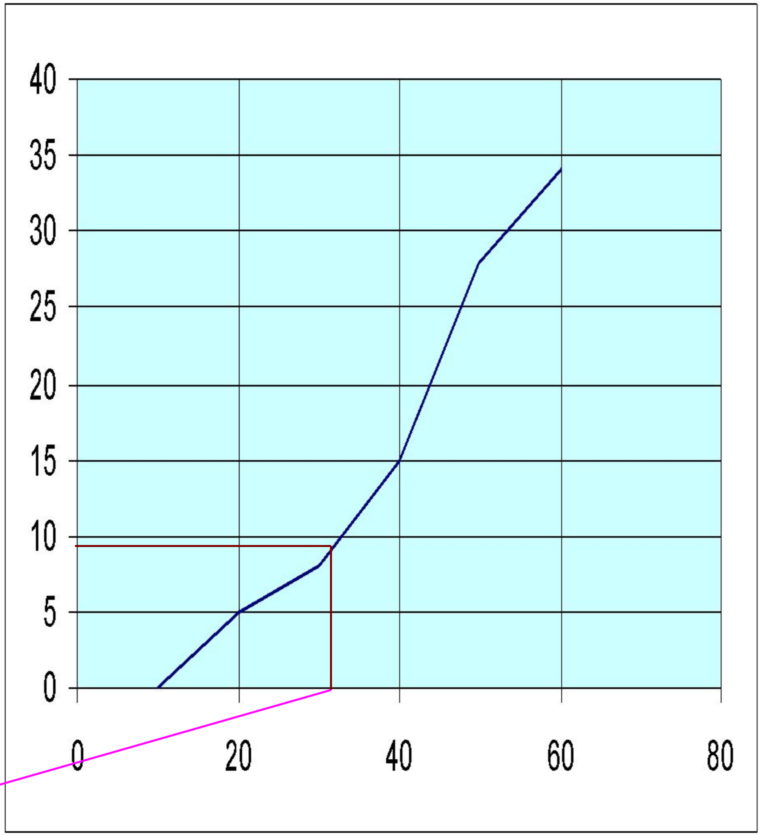
The Q_1 can be obtained graphically using the Ogive

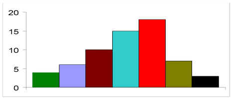
locate the point, which represent the value obtained from

(division *n* by 4; 34/4 = 8.5)

And draw a horizontal line until it intersects the Ogive then draw a vertical line until it intersects the X-axis. The intersection represent the Q_1

Value of Q_1



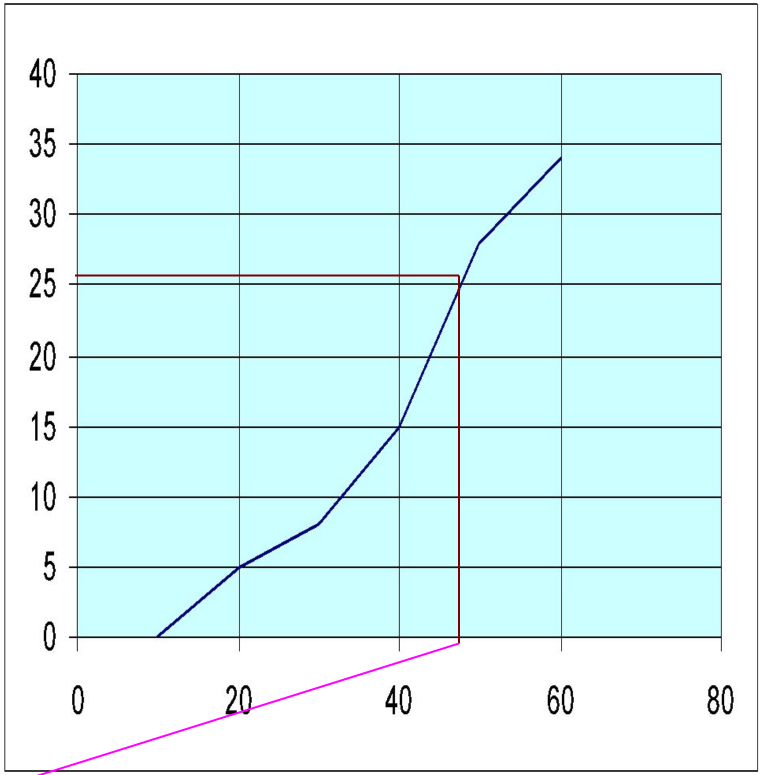


The Q_3 can be obtained graphically using the Ogive

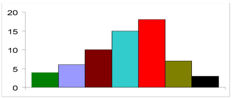
locate the point, which represent the value

(of $3n$ by 4; (3*34)/4 = 25.5)

And draw a horizontal line until it intersects the Ogive then draw a vertical line until it intersects the X-axis. The intersection represent the value of Q_3



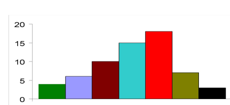
Q_3



The Interquartile Range (IQR)

The **Interquartile Range**, $IQR = Q_3 - Q_1$.

the **Interquartile Range** (**IQR**), also called the **midspread** , **middle fifty or inner 50% data range**, is a measure of statistical dispersion (variation), being equal to the difference between the third and first quartiles.



Outliers

An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.

To determine whether a data value can be considered as an outlier:

Step 1: Compute $Q1$ and $Q3$.

Step 2: Find the IQR = $Q3 – Q1$.

Step 3: Compute (1.5)(IQR).

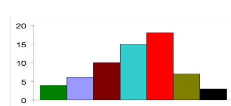
Step 4: Compute $Q1$ – (1.5)(IQR) and $Q3$ + (1.5)(IQR).

they are called lower fence and upper fence

Step 5: Compare the data value (say X) with *lower* and *upper* fences

If *X* < *lower fence* or if *X* > *upper fence ,*

then *X* is considered as an outlier.



Example

Given the data set 5, 6, 12, 13, 15, 18, 22, 50, can the value of 50 be considered as an outlier?

$Q_1 = 9$, $Q_3 = 20$, IQR = 11. *Verify.*

(1.5)(IQR) = (1.5)(11) = 16.5.

9 – 16.5 = – 7.5 and 20 + 16.5 = 36.5.

The value of 50 is outside the range (– 7.5 to 36.5), hence 50 is an outlier.

Measure of Dispersion tells us about the variation of the data set.

Skewness tells us about the direction of variation of the data set.

Definition:

Skewness is a measure of symmetry, or more precisely, the lack of symmetry.

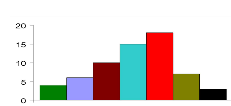
Coefficient of Skewness

Unitless number that measures the degree and direction of symmetry of a distribution

There are several ways of measuring Skewness:

Pearson's coefficient of Skewness

$$sk_2 = \frac{3(mean - median)}{s}$$



The Empirical (Normal) Rule

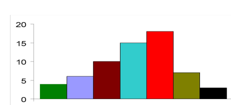
For any bell shaped distribution:

Approximately **68%** of the data values will fall within one standard deviation of the mean.

Approximately **95%** will fall within two standard deviations of the mean.

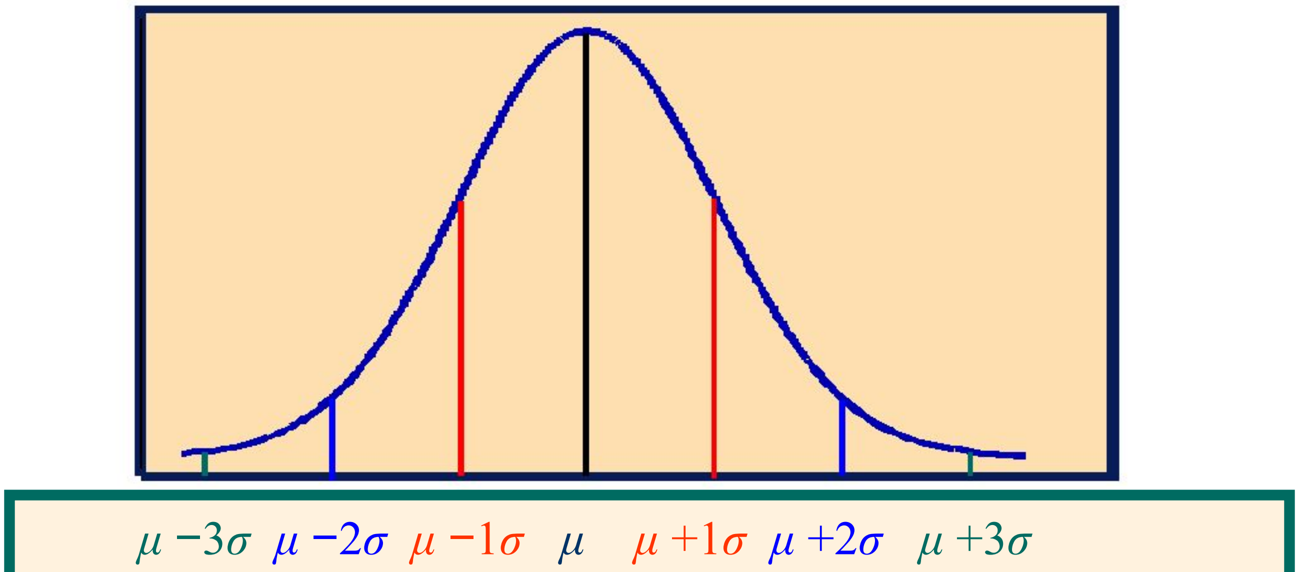
Approximately **99.7%** will fall within three standard deviations of the mean.

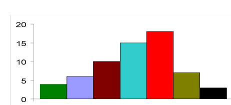
$$\mu \pm 1\sigma = 68\% \qquad \mu \pm 2\sigma = 95\% \qquad \mu \pm 3\sigma = 99.7\%$$



The Empirical (Normal) Rule

$\mu \pm 1\sigma = 68\%$ $\mu \pm 2\sigma = 95\%$ $\mu \pm 3\sigma = 99.7\%$





What is a Box Plot

To construct a box plot, first obtain the 5 number summary

$$\{ \quad Min, \quad Q_1, \quad M, \quad Q_3, \quad Max \quad \}$$

The **box-plot** is a graphical representation of data When the data set contains a small number of values, a box plot is used to graphically represent the data set. These plots involve five values: the **minimum value (the smallest value which is not an outlier), the first quartile, the median, the third quartile, and the maximum value (the largest value which is not an outlier).**

The **box plot** is useful in analyzing small data sets that do not lend themselves easily to histograms. Because of the small size of a box plot, it is easy to display and compare several box plots in a small space.

A **box plot** is a good alternative or complement to a histogram and is usually better for showing several simultaneous comparisons.

How to use it:

Collect and arrange data. Collect the data and arrange it into an ordered set from lowest value to highest.

Calculate the median. M = median= Q_2

Calculate the first quartile. (Q_1)

Calculate the third quartile. (Q_3)

Calculate the interquartile rage (IQR). This range is the difference between the first and third quartile vales. (Q_3 - Q_1)

Obtain the maximum. This is the largest data value that is less than or equal to the third quartile plus 1.5 X IQR.

$Q_3 + [(Q_3 - Q_1) \times 1.5]$

Obtain the minimum. This value will be the smallest data value that is greater than or equal to the first quartile minus 1.5 X IQR.

$$Q_1 - [(Q_3 - Q_1) \times 1.5]$$

Draw and label the axes of the graph. The scale of the horizontal axis must be large enough to encompass the greatest value of the data sets.

Draw the box plots. Construct the box, insert median points, and attach maximum and minimum. Identify outliers (values outside the upper and lower fences) with asterisks.

The box plot can provide answers to the following questions:

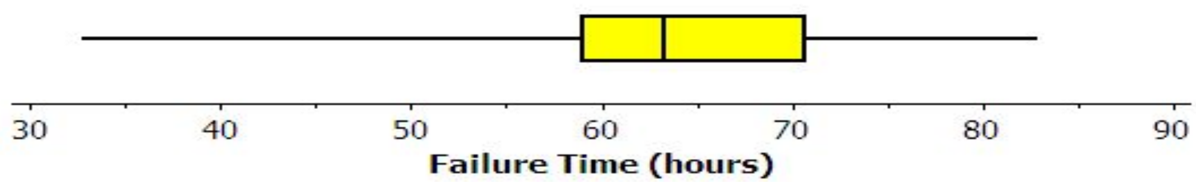
1. Does the location differ between subgroups?
2. Does the variation differ between subgroups?
3. Are there any outliers?

Example 1:- Failure times of industrial machines (in hours)

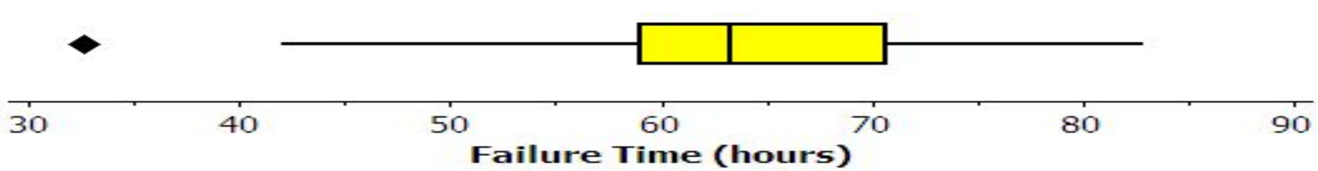
32.56	42.02	47.26	50.25	59.03	60.17	61.56	62.16
62.84	63.29	63.52	65.52	66.54	68.71	70.60	71.27
76.33	80.37	82.87					

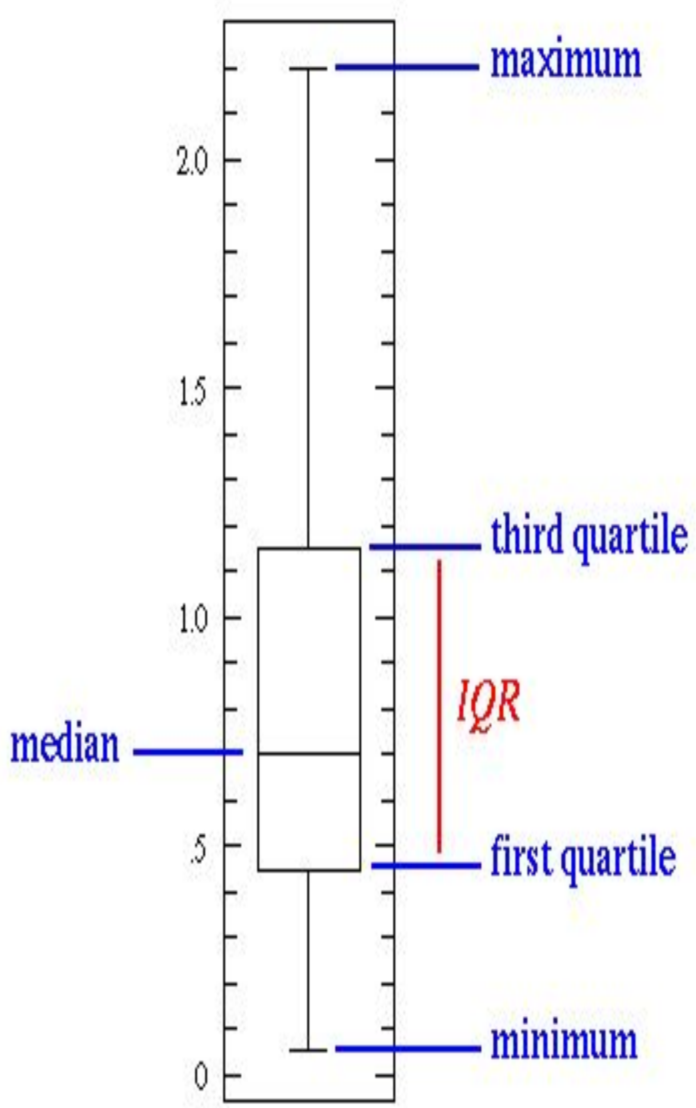
5 # summary: **{ 32.56 , 59.03 , 63.29 , 70.60 , 82.87 }**

The final product: A Simple Box-plot. Only quartile information is displayed.

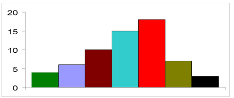


A mathematical rule designates "outliers." These are plotted using special symbols.





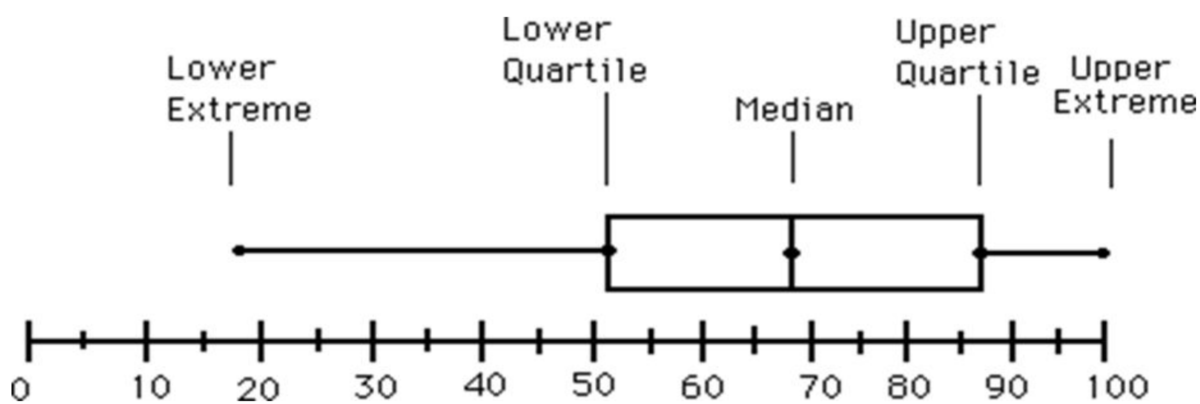
maximum
2.0
1.5
third quartile
1.0
IQR
median
.5
first quartile
minimum
0

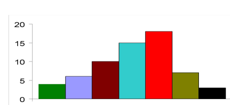


Now find the *interquartile range (IQR).* The interquartile range is the difference between the upper quartile and the lower quartile. In this case the IQR = 87 - 52 = 35. The IQR is a very useful measurement. It is useful because it is less influenced by extreme values, it limits **the range to the middle 50% of the values.**

35 is the interquartile range

begin to draw Box-plot graph.





Example 2

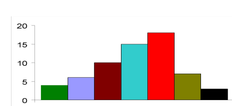
Consider two datasets:

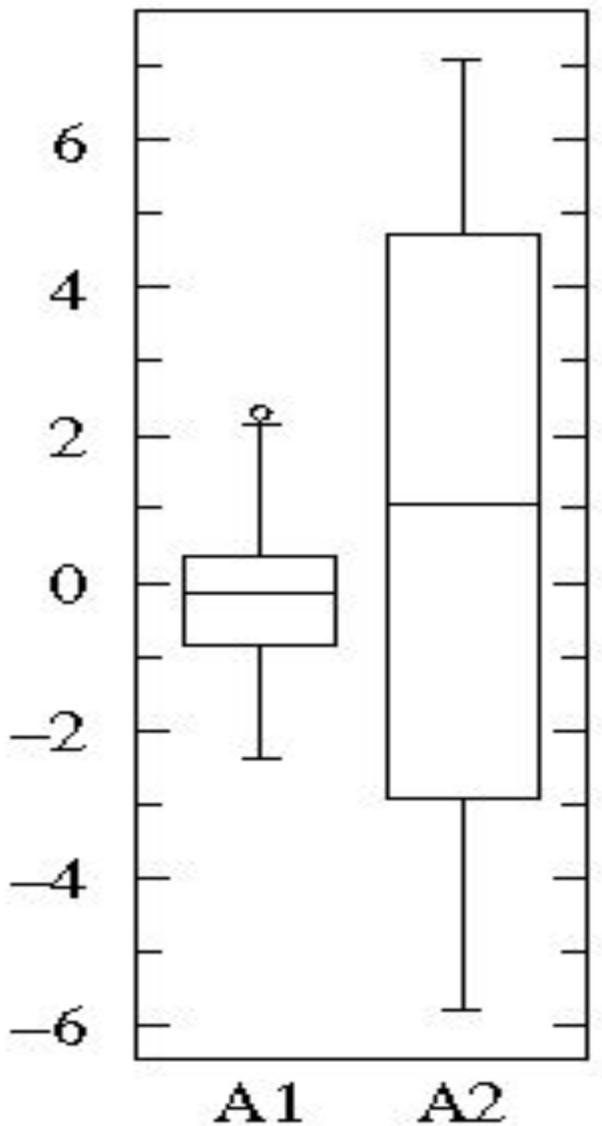
A1={0.22, -0.87, -2.39, -1.79, 0.37, -1.54, 1.28, -0.31, -0.74, 1.72, 0.38, -0.17, -0.62, -1.10, 0.30, 0.15, 2.30, 0.19, -0.50, -0.09}

A2={-5.13, -2.19, -2.43, -3.83, 0.50, -3.25, 4.32, 1.63, 5.18, -0.43, 7.11, 4.87, -3.10, -5.81, 3.76, 6.31, 2.58, 0.07, 5.76, 3.50}

Notice that both datasets are approximately balanced around zero; evidently the mean in both cases is "near" zero. However there is substantially more variation in ***A2*** which ranges approximately from -6 to 6 whereas ***A1*** ranges approximately from -2½ to 2½.

Below find box plots. Notice the difference in scales: since the box plot is displaying the full range of variation, the *y*-range must be expanded.




6
4
2
0
−2
−4
−6
A1
A2


Summary of Customer Waiting Times
0
20
40
60
80
100
Times (seconds)

Information Obtained from a Box Plot

1. If the median is near the center of the box, the distribution is approximately symmetric.
2. If the median falls to the left of the center of the box, the distribution is positively skewed.
3. If the median falls to the right of the center of the box, the distribution is negatively skewed

Similarly :

1. If the lines are about the same length, the distribution is approximately symmetric.
2. If the right line is larger than the left line, the distribution is positively skewed.
3. If the left line is larger than the right line, the distribution is negatively skewed.