STATISTICS FOR GEOLOGY COURSE

G230

By Alaa M. Atiaa

Ph. D.

LESSON THREE

Data Description

3-1 Introduction

Lesson two states that statisticians use samples taken from populations; however, when populations are small, it is necessary to use samples since the entire population can be used to gain information. For example, suppose a geologist wanted to know the average weekly earthquakes around the world. It is an impossible to collect all information concerning the number of earthquake; he would have to use a sample and make an inference to the average (mean) of the number of earthquakes.

3-2 Measures of Central Tendency

The mean

The mean, also known as the arithmetic average, if found by adding the values of the data and dividing by the total number of values. It is written mathematically as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

x = values of variable

n = the total number of values in the sample

Example 3-1

The data represent the number of days off per year for a sample of individuals selected from nine different countries. Find the mean

20, 26, 40, 36, 23, 42, 35, 24, 30

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7\ days$$

Example 3-2

The data represent the annual chocolate sales (in billions of dollars) for a sample of seven countries in the world. Find the mean.

2.0, 4.9, 6.5, 2.1, 5.1, 3.2, 16.6

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{2 + 4.9 + 6.5 + 2.1 + 5.1 + 3.2 + 16.6}{7} = \frac{40.4}{7} = 5.77\ \$bilion$$

Example 3-3

The procedure for finding the mean of grouped data uses the midpoints of the classes. This procedure is shown next.

For the following frequency distribution find the mean

A **Class boundaries**	B **Frequency**	C **Midpoint (X_m)**	D **$f \cdot X_m$**
5.5 – 10.5	1		
10.5 – 15.5	2		
15.5 – 20.5	3		
20.5 – 25.5	5		
25.5 – 30.5	4		
30.5 – 35.5	3		
35.5 – 40.5	2		
	n = 20		

Step 1: Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13 \quad etc$$

Step 2: For each class multiply the frequency by the midpoint, and place the product in column D

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad etc$$

The completed table is shown here

A Class boundaries	B Frequency	C Midpoint (X_m)	D $f \,.\, X_m$
5.5 – 10.5	1	8	8
10.5 – 15.5	2	13	26
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	90
35.5 – 40.5	2	38	76
	$n = 20$		$\sum f \cdot X_m = 490$

Step 3: Find the sum of column D

Step 5: Divided the sum by *n* to get the mean

$$\bar{x} = \frac{\sum f \cdot X_m}{n} = \frac{490}{20} = 24.5$$

The median

The median is the halfway point in a data set. Before one can find this point, the data must be arranged in order. When the data is ordered, it is called a **data array.** The median either will be a specific value in the data set or will fall between two values.

Steps in computing the median of a data array

STEP 1 : Arrange the data in order

STEP 2 : Select the middle point.

Example 3-4

The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

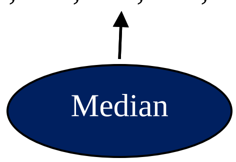
Solution

1. Arrange the data in order:

292, 300, 311, 401, 595, 618, 713

2. Select the middle value

292, 300, 311, 401, 595, 618, 713

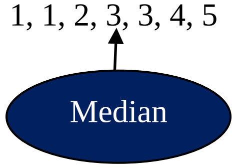


Example 3-5

Find the median for the ages of seven preschool children. The ages are 1, 3, 4, 2, 3, 5, and 1.

Solution

1, 1, 2, 3, 3, 4, 5



Examples 3-4, and 3-5 had an **odd** number of values in the data set; hence, the median was an actual data value. When there is an **even** number of values in the data set, the median will fall between two given values as illustrated in the following examples.

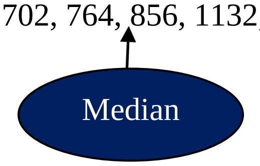
Example 3-6

The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median

684, 764, 656, 702, 856, 1133, 1132, 1303

Solution

656, 684, 702, 764, 856, 1132, 1133, 1303



Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2

$$MD = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

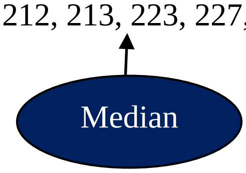
Example 3-7

The number of cloudy days for the top ten cloudiest cities is shown. Find the median

209, 223, 211, 227, 213, 240, 240, 211, 229, 212

Solution

209, 211, 211, 212, 213, 223, 227, 229, 240, 240



$$MD = \frac{213 + 223}{2} = 218$$

The mode

The third measure of average is called the mode. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

A data set can have more than one mode or no mode at all. These situations will be shown in some of the examples that follow

Example 3-8

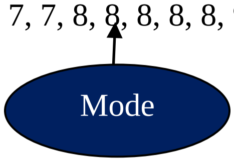
The following data represent the duration (in days) of US Space Shuttle voyages for the years 1992-1994. Find the mode.

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

Solution

Arrange the data in order

6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14



Since 8-day voyages occurred 5 times – a frequency larger than any other number – the mode for the data set is 8.

Example 3-9

Find the mode for the number of coal employees per county for 10 selected counties in southwestern Pennsylvania.

110, 713, 1031, 84, 20, 118, 1162, 1977, 103, 752

Solution

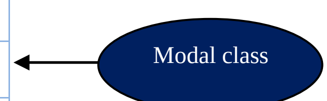
Since each value occurs only once, there is no mode.

The mode for grouped data is the modal class. The modal class is the class with the largest frequency.

Example 3-10

Find the modal class for the frequency distribution of miles that 20 runners ran in one week as described below

Class boundaries	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2



The midrange

The midrange is a rough estimate of the middle. It is found by adding the lowest and highest values in the data set and dividing by 2. It is very rough estimate of the average and can be affected by one extremely high or low value.

$$MR = \frac{lowest\ value + highest\ value}{2}$$

EXERCISES

For Exercises 1 through 8, find (a) the mean, (b) the median, (c) the mode, and (d) the midrange.

1. The average undergraduate grade-point average (GPA) for the 25 top-ranked medical schools are listed below.

3.80	3.77	3.70	3.74	3.70
3.86	3.76	3.68	3.67	3.57
3.83	3.70	3.80	3.74	3.67
3.78	3.74	3.73	3.65	3.66
3.75	3.64	3.78	3.73	3.64

Source: *U.S. News & World Report Best Graduate Schools.*

2. The heights (in feet) of the 20 highest waterfalls in the world are shown here. (*Note:* The height of Niagara Falls is 182 feet!)

3212	2800	2625	2540	2499	2425	2307	2151	2123	2000
1904	1841	1650	1612	1536	1388	1215	1198	1182	1170

Source: *N.Y. Times Almanac.*

3. The following data are the number of burglaries reported for a specific year for nine western Pennsylvania universities. Which measure of average might be the best in this case? Explain your answer.

61, 11, 1, 3, 2, 30, 18, 3, 7

Source: *Pittsburgh Post Gazette.*

4. The number of hospitals for the five largest hospital systems is shown here.

340, 75, 123, 259, 151

Source: *USA TODAY.*

5. The lengths of service (in years) of the Chief Justices of the Supreme Court are

7, 1, 5, 35, 28, 10, 15, 22, 11, 10, 12, 6, 8, 14, 18, 16

Source: *Columbia Encyclopedia.*

6. The salaries for the 12 highest paid mayors in the U.S. are listed below.

170,000	157,300	146,891	125,000
165,000	147,000	127,230	115,851
160,500	147,000	125,000	115,000

Source: *N.Y. Times Almanac.*

7. Twelve major earthquakes had Richter magnitudes shown here.

7.0, 6.2, 7.7, 8.0, 6.4, 6.2,
7.2, 5.4, 6.4, 6.5, 7.2, 5.4

Source: *The Universal Almanac.*

8. The Land Trust Alliance reported these numbers of acres in trusts for each state.

22,077	178	156,747	18,751
737	4,644	27,497	125,070
1,386	44,230	1,692	35,203
0	1,946	1,722	4,180
484,271	27,273	5,299	12,357
45,419	16	176,573	0
44,220	9,163	15,665	12,569
33,062	14,528	98,896	326,616
97,197	119,052	1,067,227	9,784
7,188	66,159	65,789	47,483
947	18,928	7,116	489,381
89,266	105,318	26,909	155
15,080	44,314		

Source: *USA TODAY*.

9. Find the (a) mean, (b) median, (c) mode, and (d) midrange for the data in Exercise 17 in Section 2–2. Is the distribution symmetric or skewed? Use the individual data values.

10. Find the (a) mean, (b) median, (c) mode, and (d) midrange for the distances of the home runs for McGwire and Sosa, using the data in Exercise 18 in Section 2–2.

Compare the means. Decide if the means are approximately equal or if one of the players is hitting longer home runs? Use the individual data values.

11. These data represent the number of traffic fatalities for two specific years for 27 selected states. Find the (a) mean, (b) median, (c) mode, and (d) midrange for each data set. Are the four measures of average for fatalities for Year 1 the same as those for Year 2? (The data in this exercise will be used in Exercise 15 in Section 3–3.)

Year 1			Year 2		
1113	1488	868	1100	260	205
1031	262	1109	970	1430	300
4192	1586	215	4040	460	350
645	527	254	620	480	485
121	442	313	125	405	85
2805	444	485	2805	690	1430
900	653	170	1555	1160	70
74	1480	69	180	3360	325
158	3181	326	875	705	145

Source: *USA TODAY*.

For Exercises 12 through 21, find the (a) mean and the (b) modal class.

12. For 108 randomly selected college students, this exam score frequency distribution was obtained. (The data in this exercise will be used in Exercise 18 in Section 3–3.)

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9

13. The scores for the LPGA—Giant Eagle were

Score	Frequency
202–204	2
205–207	7
208–210	16
211–213	26
214–216	18
217–219	4

Source: LPGA.com

14. Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained. (The data in this exercise will be used in Exercise 20 in Section 3–3.)

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

15. These numbers of books were read by each of the 28 students in a literature class.

Number of books	Frequency
0–2	2
3–5	6
6–8	12
9–11	5
12–14	3

16. Find the mean and modal class for the two frequency distributions in Exercises 8 and 18 in Section 2–3. Are the "average" reactions the same? Explain your answer.

17. Eighty randomly selected lightbulbs were tested to determine their lifetimes (in hours). This frequency distribution was obtained. (The data in this exercise will be used in Exercise 23 in Section 3–3.)

Class boundaries	Frequency
52.5–63.5	6
63.5–74.5	12
74.5–85.5	25
85.5–96.5	18
96.5–107.5	14
107.5–118.5	5

18. These data represent the net worth (in millions of dollars) of 45 national corporations.

Class limits	Frequency
10–20	2
21–31	8
32–42	15
43–53	7
54–64	10
65–75	3

19. The cost per load (in cents) of 35 laundry detergents tested by a consumer organization is shown. (The data in this exercise will be used for Exercise 19 in Section 3–3.)

Class limits	Frequency
13–19	2
20–26	7
27–33	12
34–40	5
41–47	6
48–54	1
55–61	0
62–68	2

20. This frequency distribution represents the commission earned (in dollars) by 100 salespeople employed at several branches of a large chain store.

Class limits	Frequency
150–158	5
159–167	16
168–176	20
177–185	21
186–194	20
195–203	15
204–212	3

21. This frequency distribution represents the data obtained from a sample of 75 copying machine service technicians. The values represent the days between service calls for various copying machines.

Class boundaries	Frequency
15.5–18.5	14
18.5–21.5	12
21.5–24.5	18
24.5–27.5	10
27.5–30.5	15
30.5–33.5	6

22. Find the mean and modal class for the data in Exercise 12 in Section 2–2.

23. Find the mean and modal class for the data in Exercise 13 in Section 2–2.

24. Find the mean and modal class for the data in Exercise 14 in Section 2–2.

25. Find the mean and modal class for the data in Exercise 15 in Section 2–2.

26. Find the weighted mean price of three models of automobiles sold. The number and price of each model sold are shown in this list.

Model	Number	Price
A	8	$10,000
B	10	12,000
C	12	8,000

27. Using the weighted mean, find the average number of grams of fat in meat or fish that a person would consume over a 5-day period if he ate these:

Meat or fish	Fat (grams/oz)
3 oz fried shrimp	3.33
3 oz veal cutlet (broiled)	3.00
2 oz roast beef (lean)	2.50
2.5 oz fried chicken drumstick	4.40
4 oz tuna (canned in oil)	1.75

Source: *The World Almanac and Book of Facts.*

28. A recent survey of a new diet cola reported the following percentages of people who liked the taste. Find the weighted mean of the percentages.

Area	% Favored	Number surveyed
1	40	1000
2	30	3000
3	50	800

29. The costs of three models of helicopters are shown here. Find the weighted mean of the costs of the models.

Model	Number sold	Cost
Sunscraper	9	$427,000
Skycoaster	6	365,000
High-flyer	12	725,000

30. An instructor grades exams, 20%; term paper, 30%; final exam, 50%. A student had grades of 83, 72, and 90, respectively, for exams, term paper, and final exam. Find the student's final average. Use the weighted mean.

31. Another instructor gives four 1-hour exams and one final exam, which counts as two 1-hour exams. Find a student's grade if she received 62, 83, 97, and 90 on the 1-hour exams and 82 on the final exam.

32. For these situations, state which measure of central tendency—mean, median, or mode—should be used.

a. The most typical case is desired.
b. The distribution is open-ended.
c. There is an extreme value in the data set.

Measures of variation

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider Example 3-11

Example 3-11:

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved. These two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

SOLUTION

The mean for brand A is

$$\mu = \frac{\sum X}{N} = \frac{210}{6} = 35\, months$$

The mean for brand B is

$$\mu = \frac{\sum X}{N} = \frac{210}{6} = 35\, months$$

Since the means are equal in the previous example, one might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See figure below

Even though the means are the same for both brands, the spread, or variation, is quite different. The figure shows that brand B performs more consistently, it is less variable. For the spread of variability of a data set, three measures are commonly: range, and standard deviation. Each measure will be discussed in this section.

Range

The range is the highest value minus the lowest value. The symbol R is used for the range.

R = highest value – lowest value

One extremely high or one extremely low data can affect the range markedly.

Variance and Standard Deviation

Before the variance and standard deviation are defined formally, the computational procedure will be shown.

Example 3-12:

Find the variance and standard deviation for the data set for brand A paint in Example 3-11.

10, 60, 50, 30, 40, 20

SOLUTION

1. **Find the mean for the data**
2. **Subtract the mean from each data value**
3. **Square each result**
4. **Find the sum for the squares**
5. **Divided the sum by N to get the variance**
6. **Take the square root of the variance (S^2) to get the standard deviation (S)**

A values (X)	B $X - \bar{X}$	C $(X - \bar{X})^2$
10	-25	625
60	+25	625
50	+15	225
30	-5	25
40	+5	25
20	-15	225
$\sum X = 210$		$\Sigma \quad =1750$
$\bar{X} = \frac{210}{6} = 35$		$s^2 = (1750/6) = 291.7$ $s = \sqrt{291.7} = 17.1$

For populations

Variance is σ^2

Standard deviation σ

Formula for the Sample Standard Deviation

The standard deviation of a sample (denoted by s) is

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

where

x = individual value

$\bar{x}$ = sample mean

n = sample size

Variance and Standard Deviation for Grouped Data

To find the sample variance and standard deviation for grouped data make table as shown below

Class boundaries	Frequency	Class midpoint (X_m)	X_m^2	$f.\,X_m$	$f.\,X_m^2$
	n =			Σ =	Σ =

To get variance and standard deviation use the following equations

$$s^2 = \frac{\sum f.X_m^2 - \left[\frac{\left(\sum f \cdot X_m\right)^2}{n}\right]}{n-1}$$

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

Example 3-13

Find the variance and the standard deviation for the frequency table shown below

Class boundaries	Frequency	Class midpoint (X_m)
5.5 – 10.5	1	8
10.5 – 15.5	2	13
15.5 – 20.5	3	18
20.5 – 25.5	5	23
25.5 – 30.5	4	28
30.5 – 35.5	3	33
35.5 – 40.5	2	38

Solution

Class boundaries	Frequency	Class midpoint (X_m)	X_m^2	$f.X_m$	$f.X_m^2$
5.5 – 10.5	1	8	64	8	64
10.5 – 15.5	2	13	169	26	338
15.5 – 20.5	3	18	324	54	972
20.5 – 25.5	5	23	529	115	2645
25.5 – 30.5	4	28	784	112	3136
30.5 – 35.5	3	33	1089	99	3267
35.5 – 40.5	2	38	1444	76	2888
	n = 20			Σ = 490	Σ = 13310

$$s^2 = \frac{\sum f.X_m^2 - \left[\frac{\left(\sum f \cdot X_m\right)^2}{n}\right]}{n-1} = \frac{13310 - \left[\frac{(490)^2}{20}\right]}{20-1} = 68.7$$

$$s = \sqrt{s^2} = \sqrt{68.7} = 8.3$$

Coefficient of Variation

The coefficient of variation C_{var} is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$C_{var} = \frac{s}{\bar{x}} \cdot 100\%$$

For populations,

$$C_{var} = \frac{\sigma}{\mu} \cdot 100\%$$

Chebyshev’s Theorm

The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $(1-1/k^2)$, where k is a number greater than 1.

This thermo states that at least 75% of the data values will fall within 2 standard deviation of the mean of the data set. This result is found by substituting k = 2 in the expression.

$$1-\frac{1}{k^2}=1-\frac{1}{2^2}=\frac{3}{4}=75\%$$

Example 3-14

The mean price of houses in a certain neighborhood is \$50000, and the standard deviation is \$10000. Find the price range for which at least 75% of the houses will sell.

Solution

Chebyshev’s theorem states that three-fourth, or 75% of the data values will fall within 2σ of the mean. Thus,

$$\bar{x} \pm 2\sigma$$

$$50000 + 2 \cdot 10000 = 70000$$

$$50000 - 2 \cdot 10000 = 30000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30000 to \$70000.

Measure of position

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard score, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set.

Z Score or Standard Score

A z score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z and is written mathematically as:

$$z = \frac{value - mean}{standard\ deviation}$$

For sample, the formula is

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

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

Example 3-15

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

For calculus the z score is

$$z = \frac{X - \bar{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

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

Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

Percentiles divide the data set into 100 equal groups

Percentiles are symbolized by P_1, P_2, P_3, …, P_{99}

Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$Percentile = \frac{(number\ of\ values\ below\ X) + 0.5}{total\ number\ of\ values} \cdot 100\%$$

Example 3-16

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentiles rank of a score of 12

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula

$$Percentile = \frac{(number\ of\ values\ below\ X) + 0.5}{total\ number\ of\ values} \cdot 100\%$$

Since there are six values below a score of 12, the solution is

$$Percentile = \frac{6 + 0.5}{10} \cdot 100\% = 65th\ percentile$$

Thus, a student whose score was 12 did better than 65% of the class.

Example 3-17

Using the data in Example 3-16, find the percentile rank for a score of 6.

Solution

There are three values below 6, thus

$$Percentile = \frac{3 + 0.5}{10} \cdot 100 = 35th\ percentile$$

Example 3-18

Using the scores in Example 3-16, find the value corresponding to the 25^{th} percentile

1. Arrange the data in order from lowest to highest

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

2. Compute

$$c = \frac{n \cdot p}{100}$$

where

n = total number of values

p = percentile

Thus

$$c = \frac{10 \cdot 25}{100} = 2.5$$

3. If c is not a whole number, round it up to the next whole number; in this case, c = 3.

Quartiles and Deciles

Quartiles divide the data set into four groups, separated by Q_1, Q_2, Q_3

Note that $Q_1 = 25^{th}$ percentile

$Q_2 = 50^{th}$ percentile

$Q_3 = 75^{th}$ percentile

Procedure Table

Finding data values corresponding to Q_1, Q_2, Q_3

1. Arrange data in order from lowest to highest
2. Find the median of the data values. this is the value for Q_2
3. Find the median of the data values that fall below Q_2. This is the value for Q_1
4. Find the median of the data values that fall above Q_2. This is the value for Q_3

Example 3-19

Find Q_1, Q_2, and Q_3 for the data set

15, 13, 6, 5, 12, 50, 22, 18

Solution

1. Arrange the data in the order

   5, 6, 12, 13, 15, 18, 22, 50

2. Find the median (Q_2)

5, 6, 12, 13, 15, 18, 22, 50

↓

MD

$$MD = \frac{13+15}{2} = 14$$

3. Find the median of the data values less than 14

5, 6, 12, 13

$$MD = \frac{6 + 12}{2} = 9$$

So Q_1 is 9.

4. Find the median of the data values greater than 14

$$MD = \frac{18 + 22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

Deciles

Deciles divided the distribution into 10 groups. They are denoted by D_1, D_2, D_3, etc.

$D_1, D_2, D_3, \ldots, D_9 = P_{10}, P_{20}, P_{30}, \ldots, P_{90}$

Exploratory Data Analysis

The purpose of **exploratory data analysis (EDA)** is to examine data to find out what information can be discovered about the data such as the center and the spread. In EDA data can be organized using stem and leaf plot. The measure of central tendency used in EDA is the median. The measure of variation used in EDA is the interquartile range ($Q_3 – Q_1$). In EDA the data are represented graphically using a **boxplot** (sometimes called a box – and – whisker plot).

A boxplot can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data (i.e., minimum)
2. Q_1
3. The median
4. Q_3
5. The highest value of the data set (i.e., maximum)

These values are called a **five – number summary** of the data set.

Example 3-20

A stockbroker recorded the number of clients she saw each day over an 11-day period. The data are shown below. Construct a boxplot for the data

33, 38, 43, 30, 29, 40, 51, 72, 42, 23, 31

Solution

- Arrange the data in order

23, 27, 29, 30, 31, 33, 38, 40, 42, 43, 51

- Find the median

23, 27, 29, 30, 31, 33, 38, 40, 42, 43, 51

- Find Q_1

23, 27, 29, 30, 31

- Find Q_3

38, 40, 42, 43, 51

- Draw a scale for the data on the x axis
- Locate the lowest value, Q_1, the median, Q_2, and the highest value on the scale

Draw a box around Q_1 and Q_2, draw a vertical line through the median and connect the upper and lower values.