3.2 Measures of Dispersion for Ungrouped Data

- Range
- Variance and Standard Deviation
- Population Parameters and Sample Statistics

Range

Finding Range for Ungrouped Data

Range = Largest value – Smallest Value

Example 3-11

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Find the range for this data set.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Example 3-11: Solution

Range = Largest value – Smallest Value

= 267,277 – 49,651

= **217,626 square miles**

Thus, the total areas of these four states are spread over a range of 217,626 square miles.

Range

Disadvantages

- The range, like the mean has the disadvantage of being influenced by outliers. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.
- Its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range.

Variance and Standard Deviation

- The standard deviation is the most used measure of dispersion.
- The value of the standard deviation tells how closely the values of a data set are clustered around the mean.

Variance and Standard Deviation

- In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean.

- In contrast, a large value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively large range around the mean.

Variance and Standard Deviation

- The Variance calculated for population data is denoted by σ^2 (read as sigma squared), and the variance calculated for sample data is denoted by s^2.

- The standard deviation calculated for population data is denoted by σ, and the standard deviation calculated for sample data is denoted by s.

Table 3.5

x	$x - \bar{x}$
82	$82 - 84 = -2$
95	$95 - 84 = +11$
67	$67 - 84 = -17$
92	$92 - 84 = +8$
	$\Sigma(x - \bar{x}) = 0$

Variance and Standard Deviation

Short-cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

Variance and Standard Deviation

Short-cut Formulas for the Variance and Standard Deviation for Ungrouped Data

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Example 3-12

The following table gives the 2008 market values (rounded to billions of dollars) of five international companies. Find the variance and standard deviation for these data.

Company	Market Value (billions of dollars)
PepsiCo	75
Google	107
PetroChina	271
Johnson & Johnson	138
Intel	71

Example 3-12: Solution

Let x denote the 2008 market value of a company. The value of Σx and Σx^2 are calculated in Table 3.6.

x	x^2
75	5625
107	11,449
271	73,441
138	19,044
71	5041
$\Sigma x = 662$	$\Sigma x^2 = 114{,}600$

Example 3-12: Solution

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{114{,}600 - \frac{(662)^2}{5}}{5-1} = \frac{114{,}600 - 87{,}648.80}{4} = 6737.80$$

$$s = \sqrt{6737.80} = 82.0841 = \$82.08 \text{ billion}$$

Thus, the standard deviation of the market values of these five companies is $82.08 billion.

Two Observations

1. The values of the variance and the standard deviation are never negative.

2. The measurement units of variance are always the square of the measurement units of the original data.

Example 3-13

Following are the 2009 earnings (in thousands of dollars) before taxes for all six employees of a small company.

88.50 108.40 65.50 52.50 79.80 54.60

Calculate the variance and standard deviation for these data.

Example 3-13: Solution

Let x denote the 2009 earnings before taxes of an employee of this company. The value of Σx and Σx^2 are calculated in Table 3.7.

x	x^2
88.50	7832.25
108.40	11,750.56
65.50	4290.25
52.50	2756.25
79.80	6368.04
54.60	2981.16
$\Sigma x = 449.30$	$\Sigma x^2 = 35{,}978.51$

Example 3-13: Solution

$$\sigma^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} = \frac{35{,}978.51 - \frac{(449.30)^2}{6}}{6} = 388.90$$

$$\sigma = \sqrt{388.90} = \$19.721 \text{ thousand} = \$19{,}721$$

Thus, the standard deviation of the 2009 earnings of all six employees of this company is $19,721.

Population Parameters and Sample Statistics

- A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a population parameter, or simply a **parameter**.

- A summary measure calculated for a sample data set is called a sample statistic, or simply a **statistic**.

3.3 Mean, Variance, and Standard Deviation for Grouped Data

- Mean for Grouped Data
- Variance and Standard Deviation for Grouped Data

Mean for Grouped Data

Calculating Mean for Grouped Data

Mean for population data: $\mu = \dfrac{\sum mf}{N}$

Mean for sample data: $\bar{x} = \dfrac{\sum mf}{n}$

where m is the midpoint and f is the frequency of a class.

Example 3-14

Table 3.8 gives the frequency distribution of the daily commuting times (in minutes) from home to work for *all* 25 employees of a company.

Calculate the mean of the daily commuting times.

Table 3.8

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Example 3-14: Solution

Daily Commuting Time (minutes)	f	m	mf
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\Sigma mf = 535$

Example 3-14: Solution

$$\mu = \frac{\sum mf}{N} = \frac{535}{25} = 21.40 \text{ minutes}$$

Thus, the employees of this company spend an average of 21.40 minutes a day commuting from home to work.

Example 3-15

Table 3.10 gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the mean.

Table 3.10

Number of Orders	Number of Days
10–12	4
13–15	12
16–18	20
19–21	14

Example 3-15: Solution

Number of Orders	f	m	mf
10–12	4	11	44
13–15	12	14	168
16–18	20	17	340
19–21	14	20	280
	$n = 50$		$\Sigma mf = 832$

Example 3-15: Solution

$$\bar{x} = \frac{\sum mf}{n} = \frac{832}{50} = 16.64 \text{ orders}$$

Thus, this mail-order company received an average of 16.64 orders per day during these 50 days.

Variance and Standard Deviation for Grouped Data

Short-Cut Formulas for the Variance and Standard Deviation for Grouped Data

$$\sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum m^2 f - \frac{\left(\sum mf\right)^2}{n}}{n-1}$$

where σ^2 is the population variance, s^2 is the sample variance, and m is the midpoint of a class.

Variance and Standard Deviation for Grouped Data

Short-cut Formulas for the Variance and Standard Deviation for Grouped Data

The standard deviation is obtained by taking the positive square root of the variance.

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Example 3-16

Table 3.8 gives the frequency distribution of the daily commuting times (in minutes) from home to work for all 25 employees of a company. Calculate the variance and standard deviation.

Table 3.8

Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Example 3-16: Solution

Daily Commuting Time (minutes)	f	m	mf	m^2f
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	$N = 25$		$\Sigma mf = 535$	$\Sigma m^2f = 14{,}825$

Example 3-16: Solution

$$\sigma^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{N}}{N} = \frac{14{,}825 - \frac{(535)^2}{25}}{25} = \frac{3376}{25} = 135.04$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{135.04} = 11.62 \text{ minutes}$$

Thus, the standard deviation of the daily commuting times for these employees is 11.62 minutes.

Example 3-17

Table 3.10 gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the variance and standard deviation.

Table 3.10

Number of Orders	Number of Days
10–12	4
13–15	12
16–18	20
19–21	14

Example 3-17: Solution

Number of Orders	f	m	mf	m^2f
10–12	4	11	44	484
13–15	12	14	168	2352
16–18	20	17	340	5780
19–21	14	20	280	5600
	$n = 50$		$\Sigma mf = 832$	$\Sigma m^2f = 14{,}216$

Example 3-17: Solution

$$s^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1} = \frac{14{,}216 - \frac{(832)^2}{50}}{50-1} = 7.5820$$

$$s = \sqrt{s^2} = \sqrt{7.5820} = 2.75 \text{ orders}$$

Thus, the standard deviation of the number of orders received at the office of this mail-order company during the past 50 days in 2.75.

3.5 QUARTILES, PERCENTILES AND DECILES

A few other measures that are allied to the median include the **quartiles**, **deciles** and **percentiles.** These measures are based on their position in a series of observations. They are .not necessarily central values and hence they are referred to as **measures of location.** Collectively, they together are called **quantiles**, **fractiles** or **partition values**. We discuss these measures below.

3.5.1 The Quartiles

There are three quartiles in a data set, usually denoted by Q_1, Q_2 and Q_3, which divide the whole distribution into four equal parts. The second quartile Q_2, is identical with the median. The first quartile, Q_1, is the value at or below which one-fourth (25%) of all observations in the set fall; the third quartile, Q_3 is the value at or below which three-fourths (75%) of the observations lie.

For ungrouped data, a quartile, as does the median, either assumes the value of one of the items or falls between two values. If n is divisible by 4, the first quartile (Q_1) has the value half-way between the $n/4$th and $(n/4 + 1)$th observation. If n is not exactly divisible by 4 (i.e. $n/4$ is not an integer), the first quartile has the value of the next higher integer. To find the third quartile Q_3, we replace $n/4$ by $3n/4$.

Consider the following set of 12 values arranged in ascending order:

14, 17, 19, 23, 27, 32, 40, 49, 54, 59, 71, 80.

Here n=12, which is divisible by 4. The quotient is 3. Thus the first quartile will be the average value of the 3rd and the 4th observations:

$$Q_1 = \frac{19 + 23}{2} = 21$$

We add a new value 94 to the set so that $n/4$ is not an integer. Here $n/4$ =13/4= 3.25. The next higher integer is 4. Thus the 4th value will be the first quartile. Observe that the 4th value in the set is 23.

For n=12, the third quartile is the value mid-way between 3n/4th and (3n/4+1)th observation, since $3n/4 = 9$ is an integer. Thus Q_3 is the average of the 9th and 10th observations. That is

$$Q_3 = \frac{54 + 59}{2} = 56.5$$

If we add a new value 94 to the set, n becomes 13 and the third quartile now is the 10th observation, since $3n/4$ =39/4=9.75, which is not an integer. The next higher integer is 10. Thus Q_3 is the tenth value, which is equal to 59. That is Q_3=59.

For large data sets, the computation of quartile values simply by inspection of the data set is practically impossible. In such situations the quartile values can be calculated by first forming the cumulative frequency distribution and then locating the desired quartile from the given values

based on n values. The following is an example, which illustrates how to compute median from such ungrouped data.

Example 3.19: Distribution of 70 students according to their weight in kg is as follows:

Weight	No. of students	Cumulative frequencies
40	6	6
43	11	17
51	19	36
55	17	53
60	13	66
63	4	70
Total	70	-

To obtain Q_1 and Q_3, we cumulate the frequencies as shown above. Since $n/4=70/4=17.5$ is not an integer, the first quartile will be the 18th observation (next higher integer of the fraction 17.5), which in this case corresponds to 51. Since $3n/4 = 52.5$, the Q_3 is the 53rd value which equals 55.

With grouped data, the method of estimating the first and third quartile is similar to that of estimating the median. This can be accomplished through the following general formula:

$$Q_r = l_r + \frac{h}{f_r}\left(\frac{rn}{4} - F_{(r)-1}\right), \quad r = 1,2,3 \quad \ldots(3.18)$$

where

n = total number of observations in the distribution
h = Class width
$F_{(r-1)}$ = Cumulative frequency of the class prior to rth quartile class
f_r = Frequency of rth quartile class
l_r = Lower limit of the rth quartile class

Example 3.20: Compute the first quartile and third quartile from the data presented in Table 3.2.

Solution: For ready reference, we reproduce the table and construct an extra column (column 3) displaying the cumulative frequency to which the quartile values relate:

Age (in years)	Frequency (f_i)	Cumulative frequency
24.5–29.5	3	3
29.5–34.5	9	12
34.5–39.5	15	27
39.5–44.5	12	39
44.5–49.5	7	46
49.5–54.5	4	50
Total	50	–

Following (3.18) for r=1, the formula for computing Q_1 is

$$Q_1 = l_1 + \frac{h}{f_1}\left(\frac{n}{4} - F_{(1)-1}\right)$$

Here for r=1

$$n/4=50/4=12.5, l_1 = 34.5, , h=5, f_1=15 \text{ and } F_{(1)-1}=12$$

so that

$$Q_1 = 34.5 + \frac{5}{15}(12.5 - 12) = 34.67$$

To compute the third quartile Q_3, r=3 and the other values required are $3n/4 = 37.5$, $l_3 = 39.5$, $F_{(3)-1} = 27, f_3 = 12$, $h = 5$. Thus from (3.18)

$$Q_3 = l_3 + \frac{h}{f_3}\left(\frac{3n}{4} - F_{(3)-1}\right) = 39.5 + \frac{5}{12}(37.5 - 27) = 43.87$$

A value of 34.67 for Q_1 implies that 25 percent of the workers are below age 34.67. Similarly, there are 75 percent workers in the company who are below 43.87 years of age and only 25 percent of them are above this age as implied by the value of Q_3.

3.5.2 The Percentiles

Like quartiles, the statistical measure referred to as **percentile** offers a means for identifying the location of values in the data set that are not necessarily central values. Percentiles are the values, which divide the distribution into 100 equal parts. Thus there are 99 percentiles in a distribution, which are conventionally denoted by P_1, P_2, ...P_{99}.

Recall that in the discussion of the median, we found that the median divides the items arranged in order of magnitude into two equal parts. Thus in terms of percentiles, the median is the 50-th percentile. This means that $P_{50}=Q_2=\tilde{m}$. At times the 25-th percentile and/or the 75-th percentile may

be of particular interest. These two percentiles are in fact the first quartile (Q_1) and third quartile (Q_3) respectively, which we discussed earlier.

3.5.3 Percentiles for Ungrouped Data

With ungrouped data, the percentile either takes on the value half-way between the two observations or the value of one of the observations, depending on whether n is divisible by 100 or not. Consider the ordered observations

11, 14, 17, 23, 27, 32, 40, 49, 54, 59, 71, 80

To determine the 29th percentile, P_{29} say, we note that $(29\times12)/100 = 3.48$, which is not an integer. Thus the next higher integer 4 here will determine the 29th percentile value. On inspection, $P_{29} = 23$. Similarly P_{75} will be the average of the 9th and 10th observations, since $(75\times12)/100 = 9$, which is an integer. Thus the 75th percentile value is

$$P_{75} = \frac{1}{2}(9\text{th value} + 10\text{th value}) = \frac{1}{2}(54 + 59) = 56.5$$

If the percentile values are required for an ungrouped frequency distribution, the same procedure may be followed. Consider the distribution of Example 3.19. To compute 35th percentile, say, we obtain P_{35}. Here $rn/100 = (35\times\times70)/100 = 24.5$, which is not an integer. The next higher integer is 25, so that the value that corresponds to this integer is the 35th percentile. Looking at the cumulative frequency column, $P_{35}=51$. Based on this P-value, we can assert that 35% of the students scored 51 or less.

3.5.4 Percentile for Grouped Data

For a grouped frequency distribution, a formula similar to those determining the median and the quartiles may be used to determine the percentiles. In general, the ith percentile of a grouped distribution for n observations may be arrived at by using the following formula:

$$P_r = l_r + \frac{h}{f_r}\left(\frac{rn}{100} - F_{(r)-1}\right) \quad \ldots (3.19)$$

where

l_r = Lower limit of the rth percentile class

$F_{(r)-1}$ = Cumulative frequency of the pre-percentile class

f_r = Frequency of the rth percentile class

h= Width of the i-th percentile class interval

As an illustration, we compute the 30th percentile for the distribution presented in Table 3.2, which we reproduce here for computational convenience.

Age (in years)	Frequency (f_i)	Cumulative frequency
24.5–29.5	3	3
29.5–34.5	9	12
34.5–39.5	15	27
39.5–44.5	12	39
44.5–49.5	7	46
49.5–54.5	4	50
Total	50	–

The required percentile class is determined from $rn/100=(30\times50)/100=15$. Looking at the cumulative frequency in the table, we find that this value falls in the range 34.5–39.5. The other required values are:

$$r = 30,\ l_{30} = 34.5,\ h = 5,\ f_{30} = 15 \text{ and } F_{(30)-1} = 12.$$

so that

$$P_{30} = 34.5 + \frac{5}{15}(15 - 12) = 35.5$$

This implies that of all the workers, 30% were under age 35.5 years.

3.5.5 Percentile Rank

The percentile rank of any score or observation is defined as the percentage of cases in a distribution that falls at or below that score. Percentile ranks are simple to calculate if the entire collection of raw scores is available. Consider the following collection of test scores as obtained by 20 applicants in a test arranged in ascending order.

19, 22, 25, 30, 38, 39, 41, 43, 44, 47, 48, 49, 51, **54**, 56, 59, 61, 65, 67, 70

An applicant who received a score of 54 might ask himself "how does his score rank him among all the students who took part in the test?" The answer is that he scored the same as or better than 70% of the entire group, indicating that his **percentile rank** is 70%, or in other words, 70% of the students have scored 54 or below. Note that the score 54 ranks 14th from the bottom of the 20 scores so that the percentile rank of the score (π_r) 54 is

$$\pi_r = \frac{14}{20} \times 100 = 70\%$$

Thus his percentile rank is fourteenth (i.e.70%) out of 20.

For grouped distribution, the percentile rank is simply the solution of the equation (3.19) for π_r:

$$\pi_r = \frac{F_{(r)-1} + f_r\left(\frac{P_r - l_r}{h}\right)}{n} \times 100 \qquad \ldots (3.20)$$

Let us obtain the percentile rank for an age of 35.5 for the data in Table 3.2. Here P_{30}=35.5, falling in the range 34.5–39.5, $F_{(30)-1} = 12$, $f_{30} = 15$, $l_{30.} = 34.5$, h=5 and n =50. Using (3.20)

$$\pi_r = \frac{12 + 15\left(\frac{35.5 - 34.5}{5}\right)}{50} \times 100 = 30\%$$

Hence the percentile rank is 30%. This implies that 30% of the workers are aged 35.5 years or below.

3.5.6 The Deciles

When a distribution is divided into ten equal parts, each division is called a **decile**. Thus, there are 9 deciles in a distribution, which are denoted by D_1, $D_2, \ldots D_9$. Obviously $D_5 = \tilde{m} = P_{50}$.

The method of determining the deciles is similar to that for median, percentiles and quartiles. To compute 6th decile (D_6), for example, for the distribution in Example 3.19, determine $rn/10 = (6n)/10$. For n= 70, this quantity is 42. This being an integer, the average of 42nd and 43rd values will be the 6th decile. By inspection of the distribution, D_6=55.

For grouped data, the formula for the rth decile is

$$D_r = l_r + \frac{h}{f_r}\left(\frac{rn}{10} - F_{(r)-1}\right) \qquad \ldots (3.21)$$

where

l_r = Lower limit of the rth decile class

$F_{(r)-1}$ = Cumulative frequency of the class prior to the pre- rth decile class

f_r = Frequency of the rth decile class

h = Width of the decile class

Example 3.21: Obtain D_4 for the distribution given in Table 3.2 and interpret your result.

Solution: Here n=50, r=4, $4n/10$=20, so that $l_4 = 34.5$ and $F_{(4)-1} = 12$, so that

$$D_4 = 34.5 + \frac{5}{15}\left(\frac{4 \times 50}{10} - 12\right) = 37.17$$

This value for the fourth decile implies that approximately 40% workers are under 37.17 years.

3.6 THE MODE

A third statistical measure, the **mode**, is sometimes used as a measure of central tendency of a distribution. The distribution we refer to here may be composed of both categorical and numeric data. The mode is interpreted as the value that occurs most frequently in a distribution.

***Definition* 3.3**: *Mode is the most frequently occurring value in a set of observations. In other words, mode is the value of a variable, which occurs with the highest frequency.*

For nominal data, such as sex (male, female), or marital status (married, single, widowed, divorced) of an individual, it does not make any sense to ask for the mean (or median) sex or mean (or median) marital status unless these variables can be assigned meaningful numerical values. It does however, make sense to ask, which **category** has the **most** people. The term **most** in this case means the largest number of individuals. This means that we are looking for a category of the variable, which has the highest frequency. Such a category, if exists, is called **modal category** and is used to locate the mode in a set of observations or in a distribution. If a population consists of 87 percent Muslims, 11 percent Hindus and the remaining 2 percent are of other religions; the modal category is the **Muslim**, which has the most people. These examples thus justify why we do sometimes need to introduce the concept of **mode** as a measure of central tendency in addition to arithmetic mean and median.

Example 3.22: The responses of 120 athletes on their preferred color of track suits were as follows:

3.5 Measures of Position

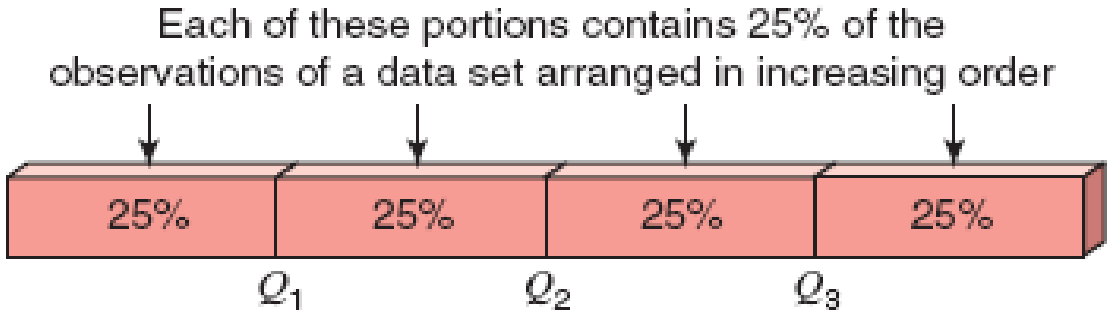
- Quartiles and Interquartile Range
- Percentiles and Percentile Rank

Quartiles and Interquartile Range

Definition

Quartiles are three summery measures that divide a ranked data set into four equal parts. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median.

Figure 3.11 Quartiles.



Quartiles and Interquartile Range

Calculating Interquartile Range

The difference between the third and first quartiles gives the ***interquartile range***; that is,

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

Example 3-20

Refer to Table 3.3 in Example 3-5, which gives the 2008 profits (rounded to billions of dollars) of 12 companies selected from all over the world. That table is reproduced below.

a) Find the values of the three quartiles. Where does the 2008 profits of Merck & Co fall in relation to these quartiles?

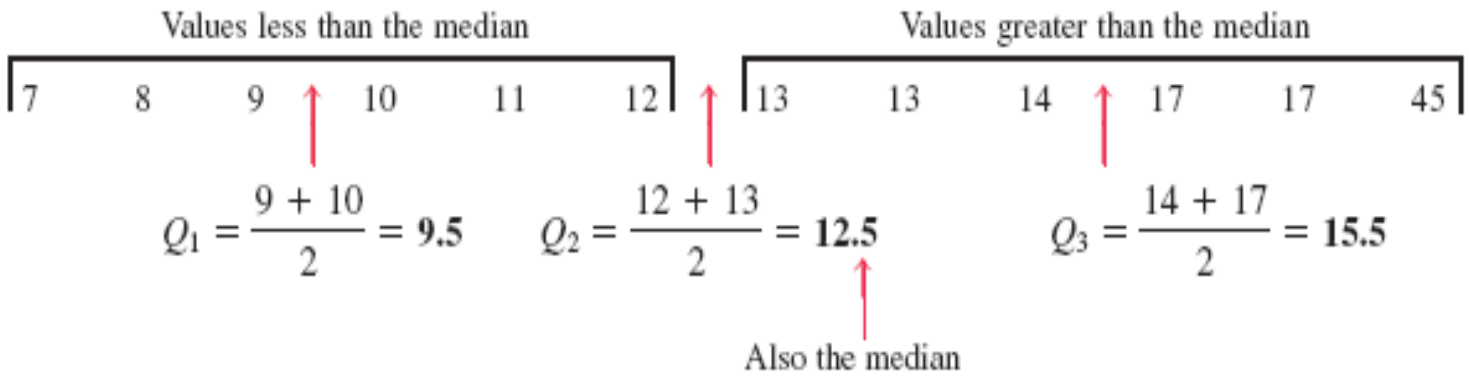
b) Find the interquartile range.

Table 3.3

Company	2008 Profits (billions of dollars)
Merck & Co	8
IBM	12
Unilever	7
Microsoft	17
Petrobras	14
Exxon Mobil	45
Lukoil	10
AT&T	13
Nestlé	17
Vodafone	13
Deutsche Bank	9
China Mobile	11

Example 3-20: Solution

a)



By looking at the position of \$8 billion, which is the 2008 profit of Merck & Co, we can state that this value lies in the **bottom 25%** of the profits for 2008.

Example 3-20: Solution

b)

IQR = Interquartile range = $Q_3 - Q_1$

$= 15.5 - 9.5$

$=$ **$6 billion**

Example 3-21

The following are the ages (in years) of nine employees of an insurance company:

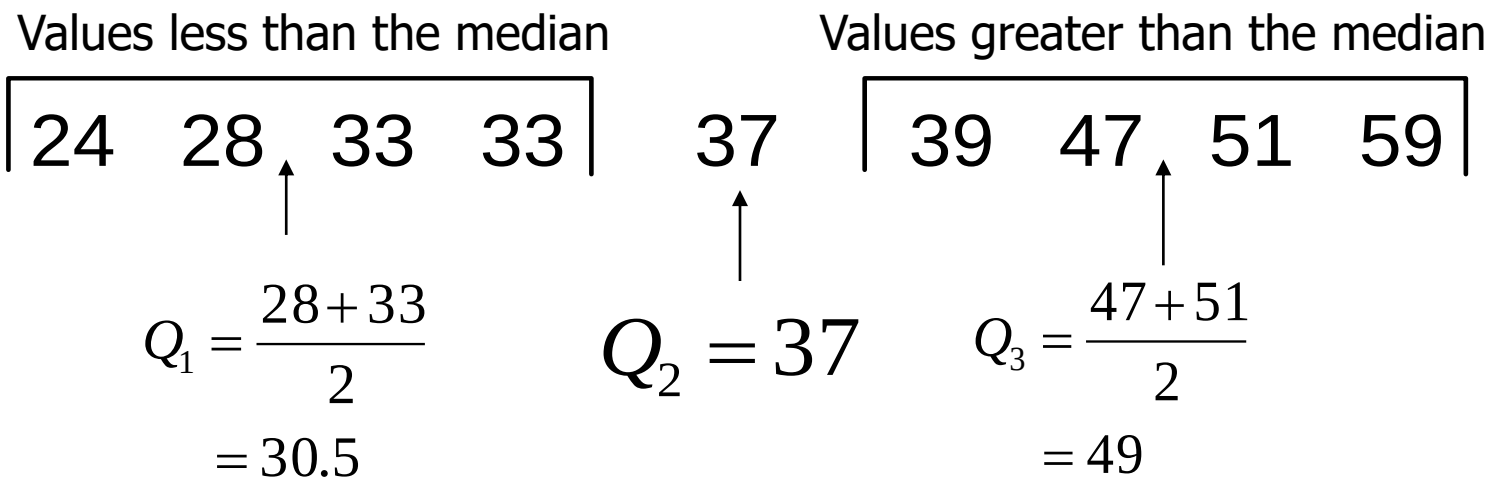
47 28 39 51 33 37 59 24 33

a) Find the values of the three quartiles. Where does the age of 28 fall in relation to the ages of the employees?

b) Find the interquartile range.

Example 3-21: Solution

a)



The age of 28 falls in the **lowest 25%** of the ages.

Question 2: Find the first quartile, second quartile and third quartile of the given information of the following sequence 4, 77, 16, 59, 93, 88 ?

Solution:

First, lets arrange of the values in an ascending order :

4, 16, 59, 77, 88, 93

Given n = 6

$\therefore$ Lower quartile = $(n+14)^{th}$ term

$= (6+14)^{th}$ term

$= (74)^{th}$ term

$= 1.7^{th}$ term

Here we can consider the 2^{nd} term (rounding 1.7 to nearest whole integer) from the set of observation.

$\Rightarrow 2^{nd}$ term = 16

Lower quartile = 16

Upper quartile = $(3(n+1)4)^{th}$ term

= $(3(6+1)4)^{th}$ term

= $(214)^{th}$ term

= 5.25^{th}

Here we can consider the 5^{th} term (rounding 5.25 to nearest whole integer) from the set of observation.

$\Rightarrow 5.25^{th} = 88$

Upper quartile = 88

Inter - quartile = Upper quartile - lower quartile

= 88 - 16

= 72

Example 3-21: Solution

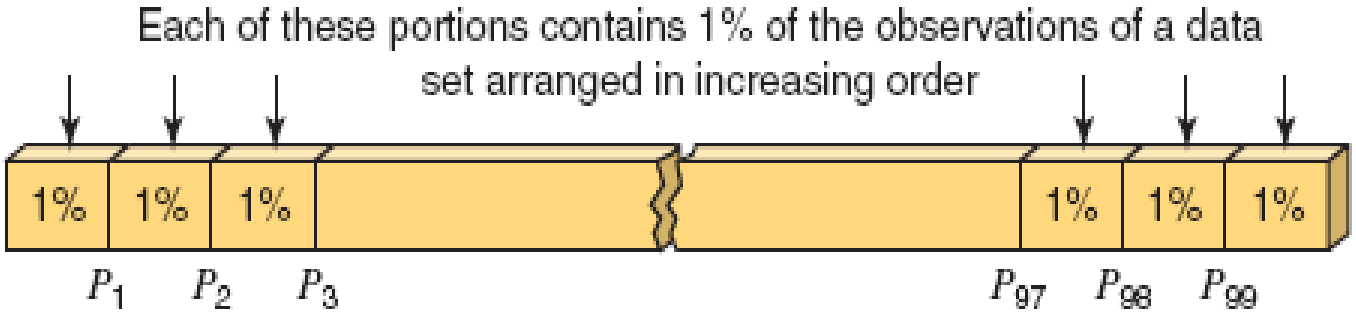
b)

IQR = Interquartile range = $Q_3 - Q_1$

$= 49 - 30.5$

$=$ **18.5 years**

Percentiles and Percentile Rank



Percentiles and Percentile Rank

Calculating Percentiles

The (approximate) value of the kth percentile, denoted by P_k, is

$$P_k = \text{Value of the } \left(\frac{kn}{100}\right)\text{th term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

Example 3-22

Refer to the data on 2008 profits for 12 companies given in Example 3-20. Find the value of the 42nd percentile. Give a brief interpretation of the 42nd percentile.

Example 3-22: Solution

The data arranged in increasing order as follows:

7 8 9 10 11 12 13 13 14 17 17 45

The position of the 42nd percentile is

$$\frac{kn}{100} = \frac{(42)(12)}{100} = 5.04\text{th term}$$

Example 3-22: Solution

The value of the 5.04th term can be approximated by the value of the fifth term in the ranked data. Therefore,

P_k = 42nd percentile = 11 = $11 billion

Thus, approximately 42% of these 12 companies had 2008 profits less than or equal to $11 billion.

Percentiles and Percentile Rank

Finding Percentile Rank of a Value

$$\text{Percentile rank of } x_i = \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100$$

Example 3-23

Refer to the data on 2008 profits for 12 companies given in Example 3-20. Find the percentile rank for $14 billion profit of Petrobras. Give a brief interpretation of this percentile rank.

Example 3-23: Solution

The data on revenues arranged in increasing order is as follows:

7 8 9 10 11 12 13 13 14 17 17 45

In this data set, 8 of the 12 values are less than \$14 billion. Hence,

$$\textbf{Percentile rank of 14} = \frac{\textbf{8}}{\textbf{12}} \times \textbf{100} = \textbf{66.67\%}$$

Example 3-23: Solution

Rounding this answer to the nearest integral value, we can state that about 67% of the companies in these 12 had less than $14 billion profits in 2008. Hence, 33% of these 12 companies had $14 billion or higher profits in 2008.