Chapter 3

Data Description

1 Introduction

This chapter shows the statistical methods that can be used to summarize data. The most familiar of these methods is the finding of averages. Loosely stated, the average means the *central of the distribution* of the most typical case. Measures of average are also called measures of central tendency and include *mean*, *median, mode* and *midrange.*

In addition to knowing the average, one must know how data values are dispersed. That is, do the data values cluster around the mean? Or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called *measure of variation* or *measure of dispersion*. These measures include the *range, variance* and *standard deviation.*

Finally, another set of measures is necessary to describe data. These measures are called measures of position. The most common position measures are *percentiles, deciles,* and *quartiles.*

2 Measures of Central Tendency

Measures found by using all the data values in the population are called *parameters.* Measures obtained by using the data values of the samples are called *statistics.*

A **statistic** is a characteristic or measure obtained by using the data values from a sample. A **parameter** is a characteristic or measure obtained by using all the data values for a specific population.

2.1 Mean (Arithmetic Mean)

The mean is the sum of the values, divided by the total number of the values. The symbol $\bar{X}$ represents the sample mean.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

or $$\bar{X} = \frac{\sum X}{n}$$

where n represents the total number of the values in the sample.

For a population, the Greek letter μ (mu) is used for the mean

$$\mu = \frac{X_1 + X_2 + \cdots X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

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

Example
The data represent the numbers of days off per year for a sample of individuals selected from nine different counties. Find the mean 20, 26, 40, 36, 23, 42, 35, 24, 30
Solution

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

Hence the mean of the number of days off is 30.7

Example
Use the frequency distribution of the miles that 20 randomly selected runners ran during a given week. (The last example in section 4 of chapter 2). Find the mean.

Class boundaries	Frequency	Cumulative frequency
5.5-10.5	1	1
10.5-15.5	2	3
15.5-20.5	3	6
20.5-25.5	5	11
25.5-30.5	4	15
30.5-35.5	3	18
35.5-40.5	2	20

Solution

Step1: Make a table as shown

A Class	B Frequency (f)	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		

Step2: Find the midpoints of each class and enter them in **C**.
Step3: For each class, multiply the frequency by the midpoint and place the product in **D.**
The completed table is shown here.

A Class	B Frequency (f)	C Midpoint(X_m)	D $f \cdot X_m$
5.5-0.5	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	99
35.5-40.5	2	38	76

Step4: Find the sum in column **D.**

Step5: Divide the sum by n to get the mean.

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

The step for finding the mean for grouped data are summarize as follows

Step1: Make a table as shown

A	B	C	D
Class	**Frequency** (f)	**Midpoint** (X_m)	$f \cdot X_m$

Step2: Find the midpoint of each class and place them in column **C.**

Step3: Multiply the frequency by the midpoint of each class and place the product in column **D.**

Step4: Find the sum of column **D**.

Step5: Divide the sum obtained in column **D** by the sum of frequencies obtained in **B.**

The formula for the mean is

$$\bar{X} = \frac{\sum f \cdot X_m}{n}$$

2.2 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 data set is ordered, it is called *a data array.*

A **median** is the midpoint of the data array. The symbol for the median is *MD*.

Steps in computing the median of a data array:

Step1: Arrange the data in order

Step2: Select the midpoint

Example

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

Solution

Step1: Arrange the data in order

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

Step2: Select the middle value

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

Hence, the median is 401 rooms.

Example

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

Arrange the data in order

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

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

Hence the median is 218 days.

For grouped data, we cannot find the exact median. It can be estimated, however, by (1) locating the class in which the median lie and (2) interpolating within that class to arrive at the median. The formula is

$$\text{Median} = LB + \frac{\frac{n}{2} - CF}{f} \times W$$

where:

LB denotes lower class boundary of the class containing the median
n denotes the total of frequencies
f denotes the frequency of the class containing median
W denotes the class width.
CF denotes cumulative number of frequencies in the classes preceding the class containing median.

Example
Below is a frequency table showing the distribution of grade points of students. Find the median.

Class	Frequency
30-39	4
40-49	6
50-59	8
60-69	12
70-79	9
80-89	7
90-99	4

Solution:

Class	Frequency	CF
30-39	4	4
40-49	6	10
50-59	8	18
60-69	12	30
70-79	9	39
80-89	7	46
90-99	4	50

We have $W = 10$, and $n = 50$ and since $\frac{n}{2} = \frac{50}{2} = 25$, the median falls in the class 60-69, so $LB = 60 - 0.5 = 59.5$. By the formula, we obtain

$$\text{Median} = LB + \frac{\frac{n}{2} - CF}{f} \times W = 59.5 + \frac{25 - 18}{12} \times 10 = 65.33$$

Example
Given the frequency distribution as below. Estimate the median.

Class	Frequency	Class	Frequency
80-89	5	110-119	8
90-99	9	120-129	6
100-109	20	130-139	2

2.3 The Mode

The value that occurs most often in a data set is called the mode.

A data set can have more than one mode or no mode.

Example

The following data represent the duration (in days) of U.S. 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

It is helpful to arrange the data in order, although it is not necessary 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, and 14. Since 8-day voyages occurred 5 times, a frequency larger than any other number, the mode for the data set is 8.

Example

Find the mode for the number of coal employees per county for ten selected counties in southwestern Pennsylvania. 110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752

Solution

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

Example

Eleven different automobiles were listed at a speed of 15 miles per hour for stopping distances. The data are shown below. Find the mode. 15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26

Solution

Since 18 and 24 both occur 3 times, the modes are 18 and 24. This data set is said to be *bimodal.*

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

Example

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

Solution

A Class	B Frequency (f)	
5.5-10.5	1	
10.5-15.5	2	
15.5-20.5	3	
20.5-25.5	5	← Modal class
25.5-30.5	4	
30.5-35.5	3	
35.5-40.5	2	

Solution

The modal class is 20.5-25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather that boundaries. Hence the mode could also be given as 23 miles per week.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are nominal or categorical.

Example

A survey showed this distribution for the number of students enrolled in each field. Find the mode.

Business	1425
Liberal Arts	878
Computer Science	632
Education	471
General Study	95

Solution

Since the category with the highest frequency is business, the most typical case is a business major.

2.4 The Midrange

The midrange is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange

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

Example

Find the midrange of this data set: 2, 3, 6, 8, 4, 1.

Solution

The highest value is 8, and the lowest value is 1, hence, the midrange is

$$\text{MR} = \frac{1+8}{2} = 4.5$$

2.5 The Weighted Mean

Sometimes, one must find the mean of a data set in which not all values are equally represented. This type of mean that consider an additional factor is called the weighted mean and is used when the values are not all equally represented.

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

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_iX_i}{\sum_{i=1}^{n} w_i}$$

where $X_i, i = 1, 2, ..., n$ are the values and $w_i, i = 1, 2, ..., n$ are the weights.

Example below shows how the weighted mean is used to compute a grade point average. Since courses vary in their credit value, the number of credits must be used as weights.

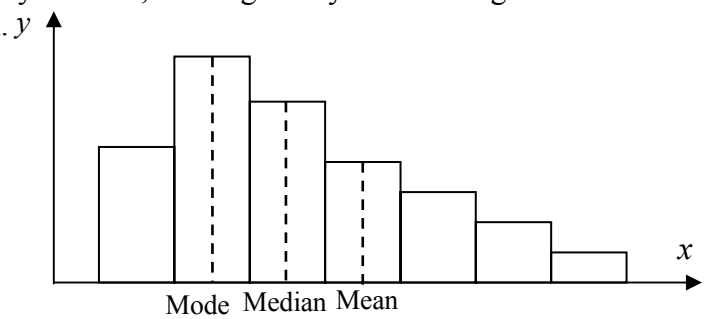
Example
A student received an A in English Composition (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology (4 credits) and a D in Physical Education (2 credits). Assuming A= 4 grade points, B=3 grade points, C=2 grade points, D=1 points and F=0 grade point, find the student grade point average.

Solution

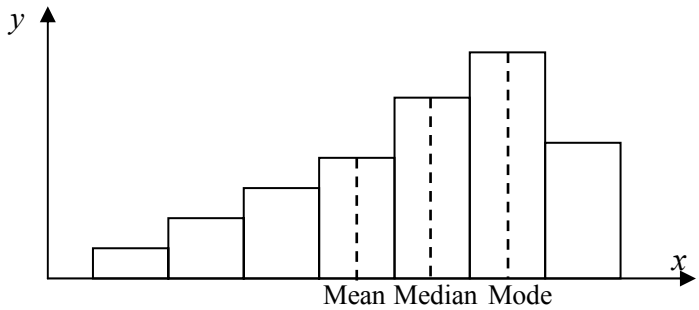
Courses	**Credits(w)**	**Grade X**
English Composition	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology	4	B (3 points)
Physical Education	2	D (1 points)

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{3\times4+3\times2+4\times3+2\times1}{3+3+4+2} = \frac{32}{12} = 2.7$$

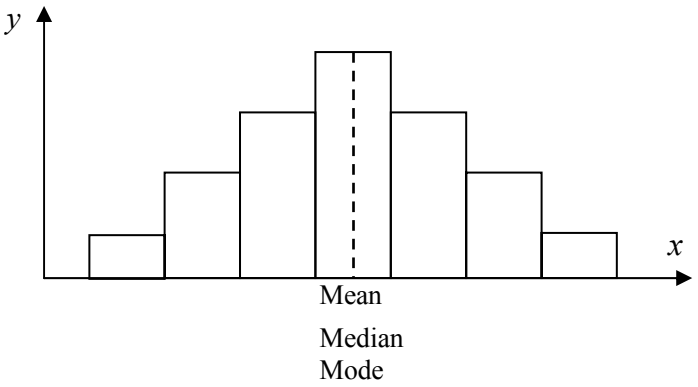
2.6 Distribution Shapes
Frequency distribution can assume many shapes. The most important shapes are positively skewed, symmetric, and negatively skewed. Figures below show histograms of each.



(a) Positively skewed or right-skewed



(b) Negatively skewed or left-skewed



(c) Symmetric

3 Measures of Variation

For the spread or variability of a data set, three measures are commonly used: *range, mean deviation, variance,* and *standard deviation.* Each measure will be discussed in this section.

3.1 Range

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

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

Example

Table below shows the duration in months the outdoor paint will last before fading. Find the range of the paint.

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

Solution

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 \text{ months}$$

3. 2 Mean Deviation

Mean Deviation is the arithmetic mean of the absolute values of the deviations from the arithmetic mean. The formula is

$$\text{Mean Deviation} = \frac{\sum \left| X - \bar{X} \right|}{n}$$

where X is the value of each observation

$\bar{X}$ is the arithmetic mean of the values.

n is the number of observations in the sample

Example

The number of patients seen in the emergency room in a hospital for a sample of 5 days last year was: 103, 97, 101, 106, and 103. Determine the mean deviation and interpret.

Solution

First find the arithmetic mean

$$\text{Mean} = \frac{103+97+101+106+103}{5} = 102$$

then we construct the table

Number of cases	$X-\bar{X}$	$\lvert X-\bar{X}\rvert$
103	103-102= 1	1
97	97-102= -5	5
101	101-102= -1	1
106	106-102= 4	4
103	103-102= 1	1
		Total 12

$$\text{Mean deviation} = \frac{\lvert X-\bar{X}\rvert}{n} = \frac{12}{5} = 2.4$$

Hence the mean deviation is 2.4 patients per day. The number of patients deviates, on average, by 2.4 patients from the mean of 102 patients per day.

Example

The weight of a group of crates being shipped to Ireland is (in pounds)
95, 103, 105, 110, 104, 105, 112, and 90. a) What is the range of the weights?
b) Compute the arithmetic mean weight. c) Compute the mean deviation of the weights. (answer: a) 22, b) 103, c) 5.25 pounds)

To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation*.

3.3 Population Variance and Standard Deviation

The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma)

The formula for the population variance is

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

where

X is individual value.
μ is the population mean.
N is the population size.

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

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

Example
Find the variance and standard deviation for brand B paint data in the above example. The months were 35, 45, 30, 35, 40, 25
Solution
Step1: Find the mean

$$\mu = \frac{\sum X}{N} = \frac{35+45+30+35+40+25}{6} = \frac{210}{6} = 35$$

Step: Subtract the mean from each value, and place the result in column B of the table
Step3: Square each result and place the squares in column C of the table.

A X	**B** $X-\mu$	**C** $(X-\mu)^2$
35	0	0
45	10	100
30	-5	25
35	0	0
40	5	25
25	-10	100

Step4: Find the sum of the square in C

$$\sum (X-\mu)^2 = 0+100+25+0+25+100 = 250$$

Setp5: Divide the sum by N to get the variance

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} = \frac{250}{6} = 41.7$$

Step6: Take the square root to get the standard deviation

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5

3.4 Sample Variance and Standard Deviation

The formula for the sample variance, denoted by s^2, is

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

where X is the individual value
$\bar{X}$ is the sample mean
n is the sample size.

The standard deviation of a sample s is

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

where X is the individual value
$\bar{X}$ is the sample mean
n is the sample size.

Computation Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance	**Standard deviation**
$s^2 = \dfrac{\sum X^2 - \left[\left(\sum X\right)^2 / n\right]}{n-1}$	$s = \sqrt{\dfrac{\sum X^2 - \left[\left(\sum X\right)^2 / n\right]}{n-1}}$

Example

Find the sample variance and standard deviation for the amount of European auto sales for a sample of 6 years shown. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Solution

Step1: Find the sum of the values.

$$\sum X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$$

Step2: Square each value and find the sum

$$\sum X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$$

Step3: Substitute in the formula and solve

$$s^2 = \frac{\sum X^2 - \left[\left(\sum X\right)^2 / n\right]}{n-1} = \frac{958.94 - \left[\left(75.6^2\right)/6\right]}{5} = 1.28$$

The variance of the sample is 1.28 and hence the sample standard deviation is 1.13

3.5 Variance and Standard Deviation for Grouped Data

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

Example

Find the variance and the standard deviation for the frequency distribution of the data representing the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoints
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

Step 1: Make a table as shown, and find the midpoint of each class.

A Class	B Frequency(f)	C Midpoints (X_m)	D $f \cdot x_m$	E $f \cdot X_m^2$
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		

Step 2: Multiply the frequency by the midpoint for each class, and place the products in column **D.**

$$1 \cdot 8 = 8, 2 \cdot 13 = 26, ..., 2 \cdot 38 = 76$$

Step 3: Multiply the frequency by the square of the midpoint, and place the products in column **E.**

$$1 \cdot 8^2 = 64, 2 \cdot 13^2 = 338, ..., 2 \cdot 38^2 = 2888$$

Step 4: Find the sums of columns **B, D,** and **E.** The sum of column **B** is *n*, the sum of column **D** is $\sum f \cdot X_m$ and the sum of column **E** is $\sum f \cdot X_m^2$

A Class	B Frequency(f)	C Midpoints (X_m)	D $f \cdot x_m$	E $f \cdot X_m^2$
5.5-10.5	1	8	8	64
10.5-15.5	2	13	26	338
15.5-20.5	3	18	54	972
20.5-25.5	5	23	115	2645
25.5-30.5	4	28	112	3136
30.5-35.5	3	33	99	3267
35.5-40.5	2	38	76	2888
			490	13310

Step 5: Substitute in the formula and solve for s^2 to get the variance

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

$$= \frac{13310 - \left[(490)^2 / 20\right]}{20-1}$$

$$= 68.7$$

Step 6: Take the square root to get the standard deviation

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

Procedure for Finding the Sample Variance and Standard Deviation of Grouped Data

Step 1: Make a table as shown and find the midpoint of each class

A Class	B Frequency(f)	C Midpoints	D $f \cdot x_m$	E $f \cdot X_m^2$

$$(X_m)$$

Step 2: Multiply the frequency by the midpoint for each class and place the products in column **D.**
Step 3: Multiply the frequency by the square of the midpoint and place the products in column **E**.
Step 4: Find the sum of column **B, D,** and **E.**
Step 5: Substitute in the formula and solve to get the variance

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

Step 6: Take the square root to get the standard deviation.

3.6 Coefficient of Variation

A statistic that allows one to compare standard deviations when the units are different (to compare the standard deviations of two different variables), is called *the coefficient of variation.*

The **coefficient of variation** is the standard deviation divided by the mean. The result is expressed as a percentage.

For Samples	**For Populations**
$\text{CVar} = \frac{s}{\bar{X}} \cdot 100\%$	$\text{CVar} = \frac{\sigma}{\mu} \cdot 100\%$

Example
The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

Solution
The coefficients of variation are

$$\text{CVar} = \frac{s}{\bar{X}} = \frac{5}{87} \cdot 100\% = 5.7\% \text{ sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100\% = 14.8\% \text{ commissions}$$

Since the coefficient of variation is larger for commission, the commissions are more variable than the sales.

Example
The mean for the number of pages of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variances.
(answer: 3.6% pages, 4.3% advertisments)

3.7 Chebyshev's theorem
Chebyshev's theorem, developed by the Russian mathematician Chebyshev (1821-1894), specifies the proportions of the spread in terms of the standard deviation.

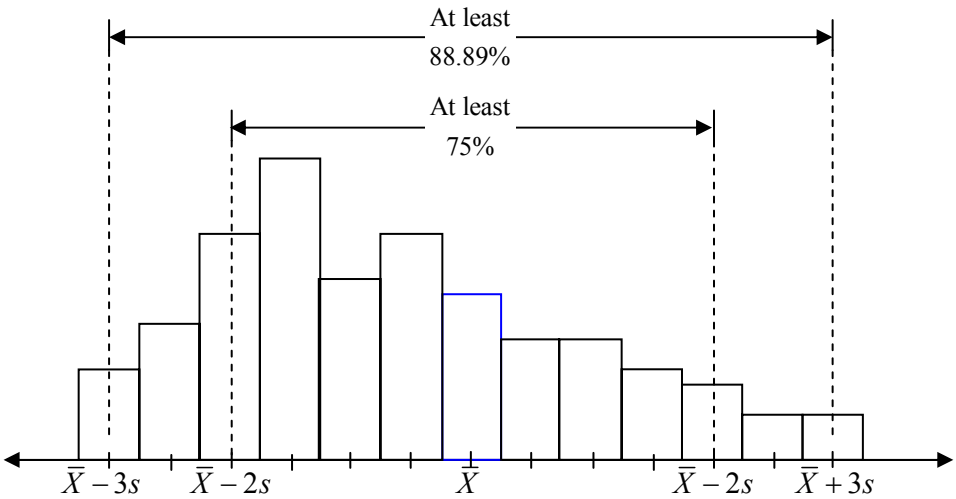
The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $1-1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

This theorem states that at least three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean of the data set. This result is found by substituting $k = 2$ in the expression.

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

Furthermore, the theorem states that at least eight-ninths, or 88.89%, of the data values will fall within 3 standard deviations of the mean.

This theorem can be applied to any distribution regardless of its shape.



Example

The mean price of houses in a certain neighborhood is \$50,000, and the standard deviation is \$10, 000. Find the price range for which at least 75% of the houses will sell.

Solution

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$50,000+2(\$10,000)=\$70,000$$

and

$$\$50,000-2(\$10,000)=\$30,000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30,000 to \$70,000.

Chebyshev's theorem can be used to find the minimum percentage of data values that will fall between any two given values. The procedure is shown in example below.

Example

A survey of local companies found that the mean amount of travel allowance for executives was \$0.25 per mile. The standard deviation was \$ 0.02. Using Chebyshev's

theorems find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

Solution

Step 1: Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

Step 2: Divide the difference by the standard deviation to get *k*.

$$k = \frac{0.05}{0.02} = 2.5$$

Step 3: Use Chebyshev's theorem to find the percentage

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \text{ or } 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

3.8 The Empirical (Normal) Rule

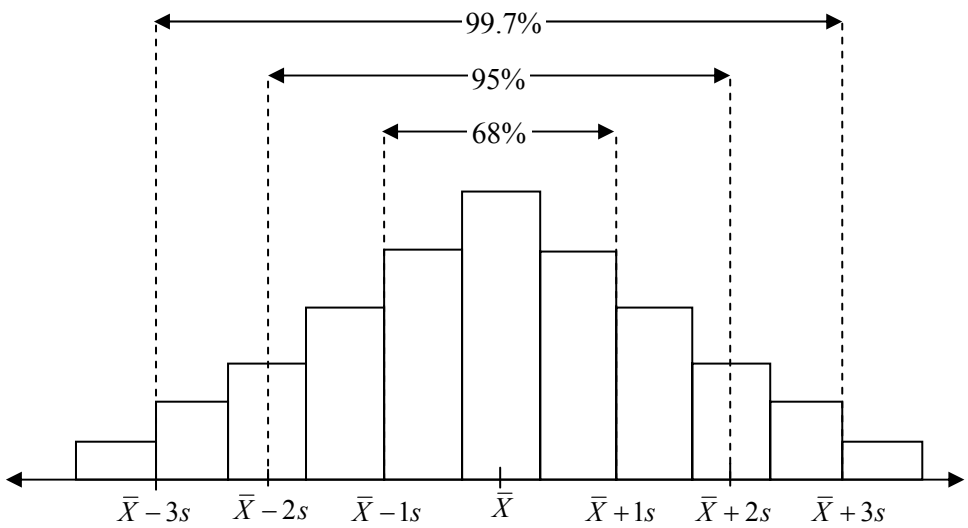
Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule,** are true.

Approximately 68% of the data values will fall within 1 standard deviation of the mean.

Approximately 95% of the data values will fall within 2 standard deviation of the mean.

Approximately 99.7% of the data values will fall within 3 standard deviation of the mean.

For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 (480+90=570 and 480 – 90 =390). Approximately 95% of the scores will fall between 300 and 660 ($480 + 2 \cdot 90 = 660$ and $480 - 2 \cdot 90 = 300$). Approximately 99.7% will fall between 210 and 750 ($480 + 3 \cdot 90 = 750$ and $480 - 3 \cdot 90 = 210$).



4 Measures of Position

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it.

4.1 Standard Scores

A **z score** or **standard 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*. The formula is

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

For samples, the formula is

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

For populations, the formula is

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

The *z score represents the number of standard deviations that a data value falls above or below the mean.*

Example

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

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

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.

Note that if the *z* score is positive, the score is above the mean. If the *z* score is 0, the score is the same as the mean. And if the *z* score is negative, the score is below the mean.

Example

Find the *z* score for each test and state which is higher.

Test A	$X = 38$	$\bar{X} = 40$	$s = 5$
Test B	$X = 94$	$\bar{X} = 100$	$s = 10$

Solution

For test A,

$$z = \frac{X - \bar{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.

When all data for a variable are transformed into z scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A z score, then, is actually the number of standard deviations each value is from the mean for a specific distribution.

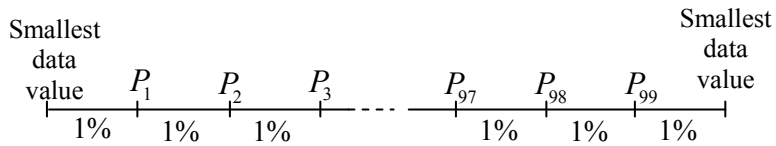
4.2 Percentiles

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 not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Percentiles are symbolized by $P_1, P_2, P_3, \ldots, P_{99}$ and divide the distribution into 100 groups.



There are several mathematical methods for computing percentiles for data. These methods can be used to find approximate percentile rank of a data value or to find a data value corresponding to a given percentile.

Percentile formula

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

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

Example

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile 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

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

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

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

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

Finding a Data Value Corresponding to a Given Percentile

Step1: Arrange the data in order from lowest to highest

Step2: Substitute into the formula

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

where n is total number of values, p is the percentile

Step3A: If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step3B: If c is a whole number, use the value halfway between the cth and $(c+1)^{th}$ values when counting up from the lowest value.

Example

Using the scores in the above example, find the value corresponding to the 25th percentile.

Solution

Step1: Arrange the data in order from lowest to highest

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

Step2: Compute

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

where n is total number of values, p is the percentile

Thus,

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

We round it up to get $c = 3$. Start at the lowest values and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

Example

Using the data set in the above example, find the value that corresponds to the 60th percentile.

Solution

Step1: Arrange the data in order from smallest to largest

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

Step2: Substitute in the formula

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

Step3: Since $c = 6$ is a whole number, we use the value halfway between the 6th and 7th values when counting up form the lowest value

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

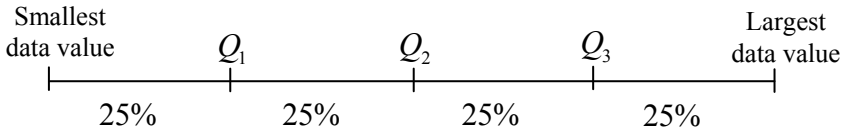
6th value *7th* value

The value halfway between 10 and 12 is 11.

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

4.3 Quartiles and Deciles

Quartiles divide the distribution into 4 groups, separated by Q_1, Q_2, Q_3.
Note that Q_1 is the same as 25th percentile; Q_2 is the same as 50th percentile or the median; Q_3 corresponds to the 75th percentile.



Quartiles can be computed using the formula for computing percentiles. For Q_1 use $P = 25$. For Q_2 use $P = 50$. For Q_3 use $P = 75$.

However, an easier method for finding quartiles is found in the procedure as follows:

Step1 Arrange the data in order from lowest to highest.
Step2 Find the median of the data value. This is the value for Q_2.
Step3 Find the median of the data values that fall below Q_2. This is the value for Q_1.
Step4 Find the median of the data values that fall above Q_2. This is the value for Q_3.

Example
Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18

Solution:
Step1 Arrange the data in order:
5, 6, 12, 13, 15, 18, 22, 50
Step2 Find the median Q_2
5, 6, 12, 13, 15, 18, 22, 50

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

Step3 Find the median of the data values less than 14

5, 6, 12, 13

$$Q_1 = \frac{6+12}{2} = 9$$

Step4 Find the median of the data values greater than 14

15, 18, 22, 50

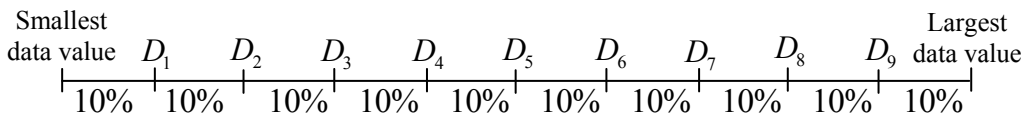
$$Q_3 = \frac{18+22}{2} = 20$$

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

In addition to dividing the data set into four groups, quartiles can bee used as a rough measurement of variability.

The **interquartile range (IQR)** is defined as the difference between Q_1 and Q_3 and it is the range of the middle 50% of the data.

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



Note that D_1 corresponds to P_{10}; D_2 corresponds to P_{20}; etc. Deciles can be found by using the formulas given for percentiles.

These are the relationships among percentiles, deciles and quartiles.

Deciles are denoted by $D_1, D_2, ..., D_9$ and they correspond to $P_{10}, P_{20}, ..., P_{90}$.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}

The median is the same as P_{50} or Q_2 or D_5.

4.4 Outliers

A data set should be checked for extremely high or extremely low values. These values are called outliers.

An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values. An outlier can strongly affect the mean and standard deviation of a variable.

There are several ways to check a data set for outliers. One of which is shown as follows:

Step1 Arrange the data in order and find Q_1 and Q_3

Step2 Find the inter-quartile range $IQR = Q_3 - Q_1$

Step3 Multiply the *IQR* by 1.5.

Step4 Substitute the value obtained in *step3* from Q_1 and add the value to Q_3

Step5 Check the data set for any data value which is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$.

Example: Check the following data set for outliers.

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

Solution:

Find Q_1 and Q_3. $Q_1 = 9, Q_3 = 20$

Find inter-quartile range $IQR = Q_3 - Q_1 = 20 - 9 = 11$

Multiply this value by 1.5 to get $1.5(11) = 16.5$

Subtract the this value from Q_1 and add the same value to Q_3, we obtain

$9 - 16.5 = -7.5$ and $20 + 16.5 = 36.5$

Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence it can be consider an outlier.

5 Exploratory Data Analysis

In **exploratory data analysis** (EDA), data can be organized using a stem and leaf plot. The measure of central tendency used in EDA is the *median*. The measure of variation is the inter-quartile $(Q_3 - Q_1)$. In EDA the data are presented graphically using a **box-plot** (sometimes called a box-and-whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data such as the center and the spread. EDA was developed by *John Tukey*.

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

(i). The lowest value of the data set (i.e., minimum)

(ii). Q_1

(iii). The median

(iv). Q_3

(v). The highest value of the data set (i.e., maximum)

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

A **box-plot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.

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

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

Solution:

Step1 Arrange the data in order

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

Step2 Find the median

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

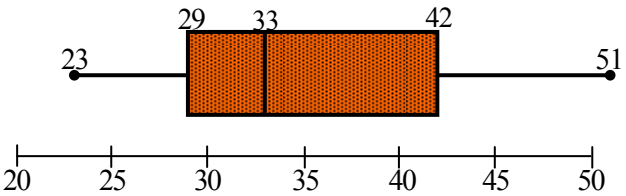
Median

Step3 Find Q_1

23, 27, 29, 30, 31

Step4 Find Q_3

38, 40, 42, 43, 51



Step5 Draw a scale for the data on the *x*-axis.

Step6 Locate the lowest value, Q_1, median, Q_3, and the highest value on the scale.

Step7 Draw a box around Q_1 and Q_3, draw a vertical line through the median, and connect the upper and lower values.