Descriptive Statistics

Objectives:

After studying this chapter, the student will:

1. Understand how data can be appropriately organized and displayed.
2. Understand how to reduce data sets into a few useful, descriptive measures.
3. Be able to calculate and interpret measures of central tendency, such as the mean, median, and mode.
4. Be able to calculate and interpret measures of dispersion, such as the range, variance, and standard deviation.
5. Be able to calculate and interpret measures of position, such as the percentile, quartiles, and standard score.
6. Be able to calculate and interpret measures of shape, such as the skewness, and kurtosis.

Introduction:

Raw Data:

Data recorded in the sequence in which they are collected and before they are processed or ranked.

Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Status of 50 Students freshman, sophomore, junior, and senior

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

Data in raw form are usually not easy to use for decision making.

The ordered array:

It is just a sorted list of data.

- Shows range (min to max).
- Provides some signals about variability within the range.
- May help identify outliers (unusual observations).
- If the data set is large, the ordered array is less useful.

Organizing and Graphing Qualitative Data:

1. Frequency Distribution (Frequency Table)

A **frequency distribution for qualitative data** lists all categories and the number of elements that belong to each of the categories.

Example:

Status of 50 students.

Categories	Tally	Frequency (f)
		Sum =

Relative Frequencies

It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values.

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

Percentage

To find the percentage, we multiply the relative frequency by 100%.

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100$$

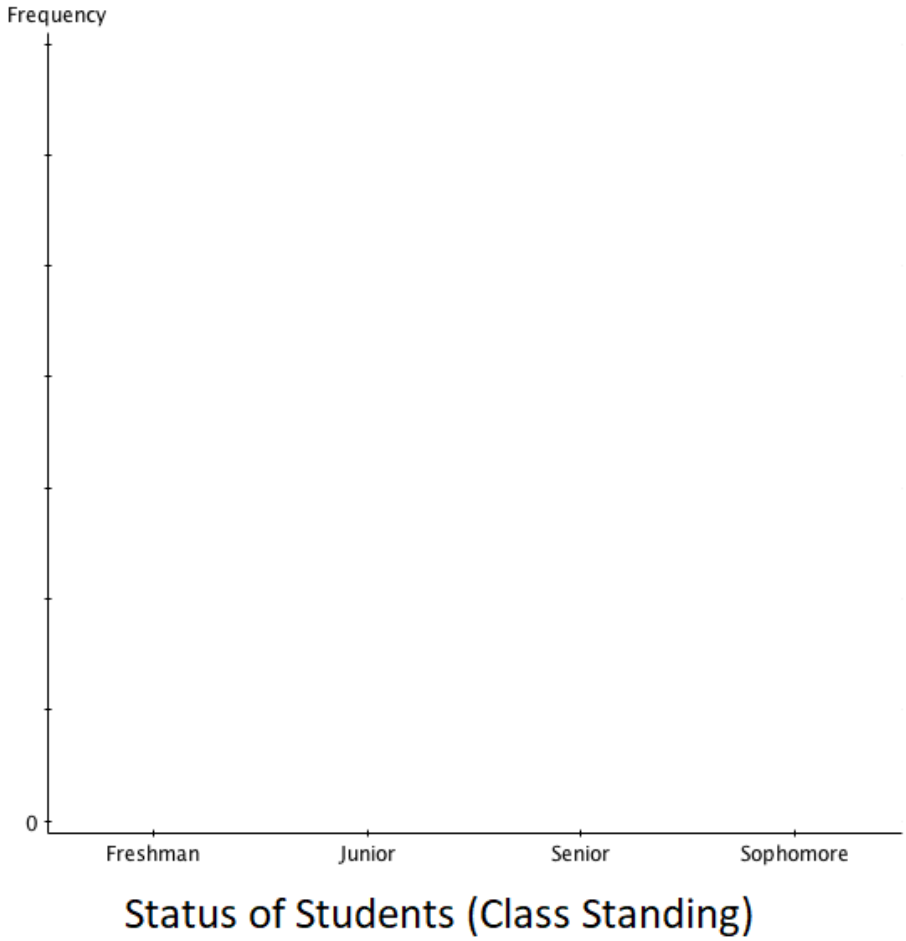
Categories	Frequency (f)	Relative Frequency	Percentage
Freshman			
Sophomore			
Junior			
Senior			
Sum =			

2. Bar Graph

A graph made of bars whose heights represent the frequencies of respective categories.

Example:

Status of 50 students.



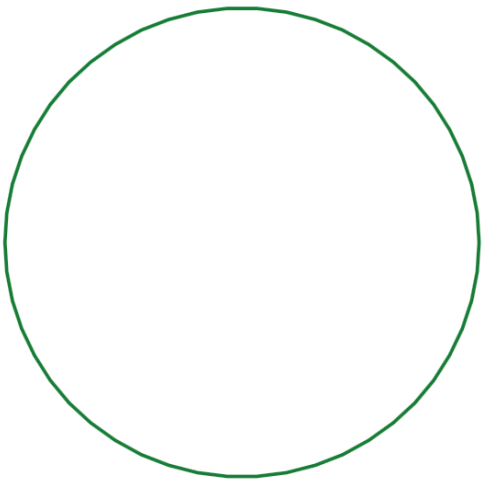
3. Pie Chart

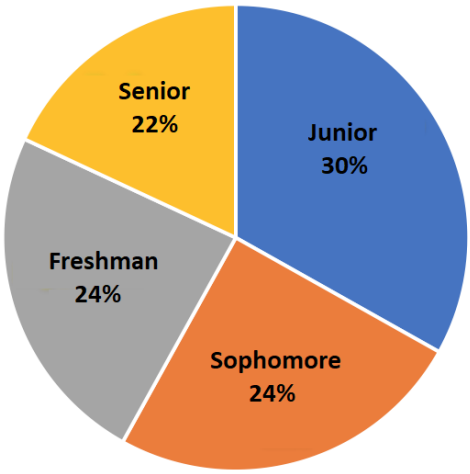
A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories.

Example:

Status of 50 students.

Categories	Relative Frequency	Angle Size
Freshman	12/50	
Sophomore	12/50	
Junior	15/50	
Senior	11/50	
Sum =	50/50	





Organizing and Graphing Quantitative Data:

Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

1. Frequency Distribution (Frequency Table)

A **frequency distribution for quantitative data** lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called grouped data.

Steps for organizing quantitative data into frequency distribution:

Step 1: Determine the number of classes wanted, which depends on the number of observations (sample size) available.

1. Number of classes, c = K such that $2^K \geq n$.
2. $c = \sqrt{n}$, (Take the integer part of the answer).
3. Sturge's formula: $c = 1 + 3.3 \log n$.
4. You may use the following table:

Sample Size	Number of Classes
Less than 16	Not enough data
16 – 31	5
32 – 63	6
64 – 127	7
128 – 255	8
256 – 511	9
512 – 1023	10

1024 – 2047	11
2048 – 4095	12
4096 – 8191	13
…	…

Step 2: Locate the largest observation (Max) and the smallest observation (Min).

Step 3: Find the range: $R = Max - Min$.

Step 4: Find the minimum class width required to cover this range by dividing the range by the number of categories desired.

Step 5: Find the actual class width to be used by rounding the minimum class width up to the same number of decimal places as the data itself.

Step 6: Find the lower boundary for the first class, which lies half unit below the smallest observation (or any number below the smallest value).

Data reported to the nearest	Unit	Half-Unit
Whole number	1	0.5
Tenth (1 decimal place)	0.1	0.05
2 decimal places	0.01	0.005
3 decimal places	0.001	0.0005
…	…	…

Step 7: Find the remaining class boundaries by adding the class width to the preceding boundary value.

Step 8: Count the number of items in each class.

Class Boundaries	Tally	Frequency (f)	Relative Frequency	Percentage	Midpoint
Sum =					

Class Width

The difference between the two boundaries of a class gives the class width. The class width is also called the class size.

$$\text{Class width} = \text{Upper boundary} - \text{Lower boundary}$$

Class Midpoint

The class midpoint or mark is obtained by dividing the sum of the two boundaries of a class by 2.

$$\text{Class midpoint or mark} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

2. Histogram

A **histogram** is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on the vertical axis.

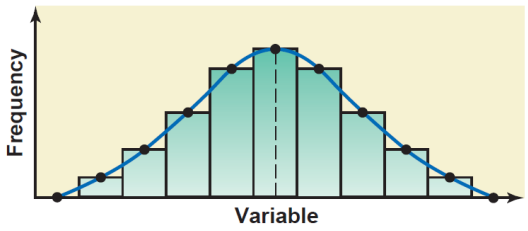
The frequencies, relative frequencies, or percentages are represented by the heights of the bars.

In a histogram, the bars are drawn adjacent to each other.

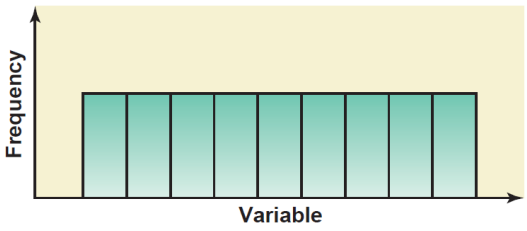
Shapes of Histograms

A histogram can assume any one of a large number of shapes. The most common of these shapes are

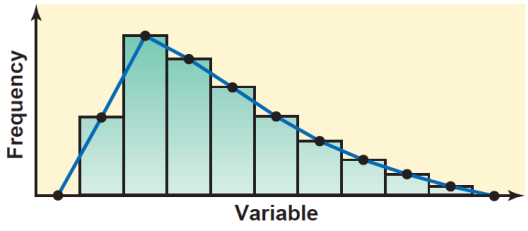
1. Symmetric.
2. Skewed.
3. Uniform or rectangular.



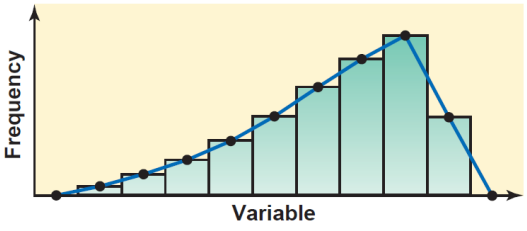
Symmetric histograms.



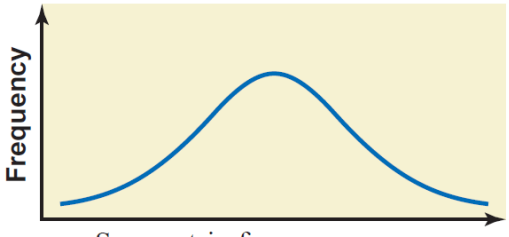
A histogram with uniform distribution.



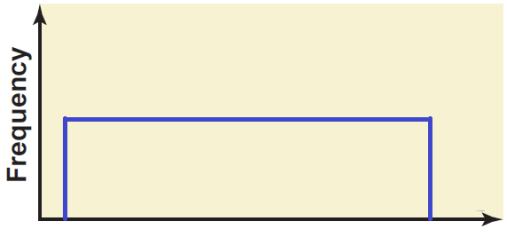
skewed-to-the-right histogram



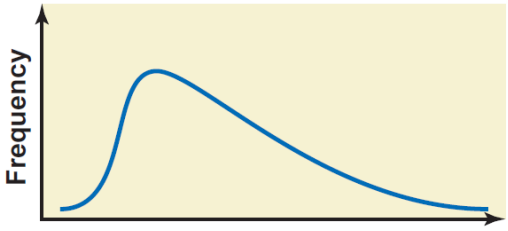
skewed-to-the-left histogram



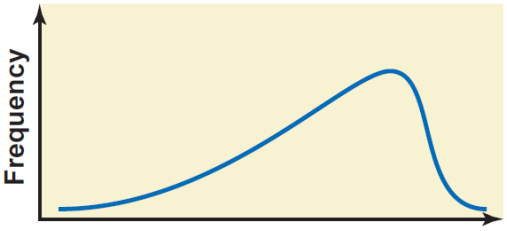
Symmetric frequency curves.



Uniform



Frequency curve skewed to the right.



Frequency curve skewed to the left.

3. Polygon

A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines.

4. Ogive

An **ogive** is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at heights equal to the cumulative frequencies of respective classes.

Cumulative Frequency Distribution

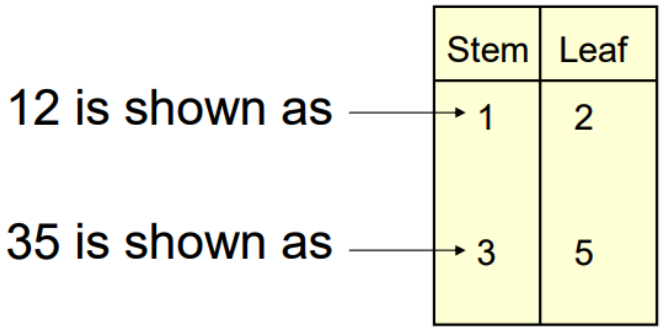
A **cumulative frequency distribution** gives the total number of values that fall below the upper boundary of each class.

Class Boundaries	Frequency (f)	Cumulative Frequency
14.5 - 18.5	3	
18.5 - 22.5	22	
22.5 - 26.5	16	
26.5 - 30.5	4	
30.5 - 34.5	3	
34.5 - 38.5	2	
Sum =		

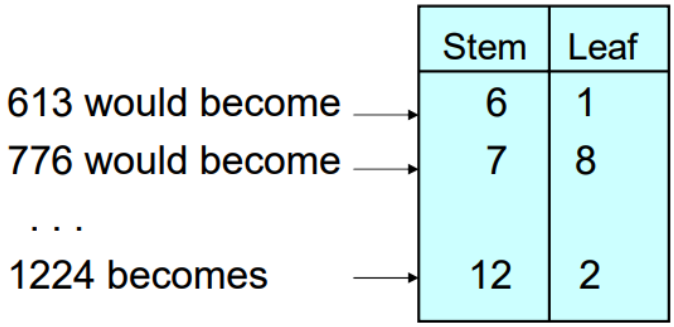
5. Stem-and-Leaf Display

In a **stem-and-leaf display** of quantitative data, each value is divided into two portions—a stem and a leaf. The leaves for each stem are shown separately in a display.

- In case of two-digit number: use the 10's digit for the stem unit.



- In case of three or four-digit number: round off the 10's digit to form the leaves.

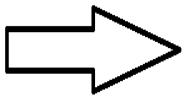


Example:

Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

Stem	Leaves



Stem	Leaves

6. Single-Valued Classes

If the observations in a data set assume only a few distinct (integer) values, it may be appropriate to prepare a frequency distribution table using single-valued classes—that is, classes that are made of single values and not of intervals.

This technique is especially useful in cases of discrete data with only a few possible values.

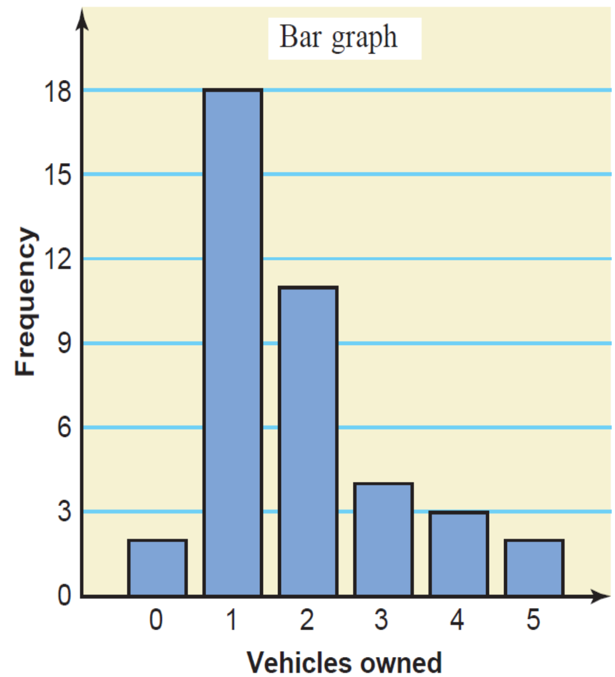
Example:

The administration in a large city wanted to know the distribution of vehicles owned by households in that city. A sample of 40 randomly selected households from this city produced the following data on the number of vehicles owned.

5	1	1	2	0	1	1	2	1	1	1	3	3	0	2	5	1	2	3	4
2	1	2	2	1	2	2	1	1	1	4	2	1	1	2	1	1	4	1	3

Frequency Distribution of Vehicles Owned

Vehicles Owned	Number of Households (f)
	$\Sigma f =$



Exercises:

Q1. The following are the number of babies born during a year in 60 community hospitals.

30	55	27	45	56	48	45	49	32	57	47	56
37	55	52	34	54	42	32	59	35	46	24	57
32	26	40	28	53	54	29	42	42	54	53	59
39	56	59	58	49	53	30	53	21	34	28	50
52	57	43	46	54	31	22	31	24	24	57	29

(a) Use these data to prepare:

1. A frequency distribution
2. A relative frequency distribution
3. A cumulative frequency distribution
4. A cumulative relative frequency distribution

5. A histogram
6. A frequency polygon
7. An ogive
8. Stem-and-Leaf display.

(b) Describe these data relative to symmetry and skewness.

Q2. The following table summarizes the scores obtained by 500 students on a questionnaire designed to measure managerial ability. (Scores are integer values that range from 0 to 20. A high score indicates a high level of ability).

Measurement Class	Relative Frequency
1.5 - 3.5	0.05
3.5 - 5.5	0.15
5.5 - 7.5	0.20
7.5 - 9.5	0.25
9.5 - 11.5	0.15
11.5 - 13.5	0.10
13.5 - 15.5	0.05
15.5 - 17.5	0.05

1. What proportion of the scores lie between 3.5 and 5.5?
2. What proportion of the scores are higher than 11.5?
3. How many students scored less than 5.5?
4. Construct a frequency histogram.
5. Describe the shape of the distribution.

Q3. The following data give the number of turnovers (fumbles and interceptions) by a college football team for each game in the past two seasons.

3	2	1	4	0	2	2	1	0	3	2	3
0	2	3	1	4	1	3	2	4	0	1	2

1. Prepare a frequency distribution table for these data using single-valued classes.
2. Calculate the relative frequencies and percentages for all classes.
3. In how many games did the team commit two or more turnovers?
4. Draw a bar graph for the frequency distribution of part a.

Q4. The FBI gathers data on violent crimes. For 20,000 murders committed over the past few years, the following fictitious data set represents the classification of the weapon used to commit the crime.

12,500 committed with guns

2,000 with a knife

5000 with hands

500 with explosives

1. Construct a pie chart to describe this data.
2. Construct a bar graph to describe this data.

Measures of Central Tendency:

A **measure of central tendency** is a value that represents a typical, or central, entry of a data set.

The three most commonly used measures of central tendency are the:

1. Arithmetic mean (Average),
2. Median,
3. Mode.

1. Arithmetic Mean:

The **mean** of a data set is the sum of the data entries divided by the number of entries.

It is the most popular and best understood measure of central tendency for quantitative data.

To find the mean of a data set, use one of the following formulae.

Population Mean:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

where,

μ: is the population mean (μ is a Greek letter, read as mu).

$\sum_{i=1}^{N}$: is the summation notation.

X_i: is the value of element i in the population.

N: is the population size.

Note that μ is a descriptive measures of the entire population, so we call it a parameter of the population.

Sample Mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where,

$\bar{X}$: is the sample mean (read as X bar).

X_i: is the value of element i in the sample.

n: is the sample size.

Note that $\bar{X}$ is a descriptive measures of a sample, so we call it a statistic.

We will often use the sample mean $\bar{X}$ to estimate (make an inference about) the population mean μ.

Example:

All 57 residents in a nursing home were surveyed to see how many times a day they eat meals.

1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,

3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4,

4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5.

What is the mean for the number of meals eaten per day?

Solution:

1 meal (2 people)

2 meals (7 people)

3 meals (28 people)

4 meals (12 people)

5 meals (8 people)

Example:

The following cholesterol levels of 10 people were measured in mg/dl:

260, 150, 165, 201, 212, 243, 219, 227, 210, 240.

Calculate the mean.

Solution:

Properties of the mean:

- Most common measure of central tendency.
- The mean is unique: A data set has only one mean (generally not part of the data set).
- All the values are included in computing the mean (and hence is a good representative of the data).
- The mean is affected by outliers (unusually large or small data values).
- The mean is not an appropriate measure for ordinal or nominal variables.
- The mean is affected accordingly if the data values are given mathematical treatment by any constant item.
- The sum of the deviations of each value from the mean will always be zero.
- Expressed symbolically:

• $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$	OR	$\sum_{i=1}^{N}(X_i - \mu) = 0.$

- μ: fixed but usually unkown.
- $\bar{X}$: random but known.

Example:

260, 150, 165, 201, 212, 243, 219, 227, 210, 240.

Example:

In the previous example, replace the value 260 by 931 and recalculate the mean.

Solution:

Example:

Find the mean of 10, 20, 30, 40, and 50.

Solution:

1. Add 10 to each value and find the mean.

Solution:

2. Subtract 10 from each value and find the mean.

Solution:

c. Multiply each value by 10 and find the mean.

Solution:

d. Divide each value by 10 and find the mean.

Solution:

e. Make a general statement about each situation.

Solution:

2. Median:

The **median** of a data set is that value which divides the set into two equal parts such that the number of values equal to or greater than the median is equal to the number of values equal to or less than the median.

- If the data set has an odd number of entries, the median is the middle data entry.
- If the data set has an even number of entries, the median is the mean of the two middle data entries.

Example:

The following cholesterol levels of 10 people were measured in mg/dl:

260, 150, 165, 201, 212, 243, 219, 227, 210, 240.

Calculate the median.

Solution:

Example:

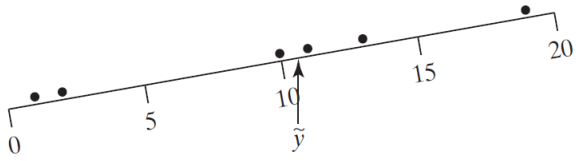
In the previous example, add the value 120 and recalculate the median.

Solution:

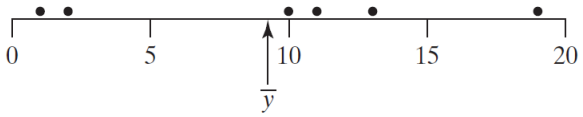
Properties of the median:

- There is only one median for a set of data (may be part of the data set).
- The median uses the position rather than the specific value of each data entry. So, it is not affected by outliers.
- It is useful as a descriptive measure for skewed distributions.

Mean versus Median



The median divides the data into two equal pieces



The mean is the "point of balance" of the data

3. Mode:

The **mode** of a data set is the data entry that occurs most frequently.

- If no entry is repeated, the data set has no mode.
- If two entries occur with the same most frequency, each entry is a mode and the data set is called bimodal.

Properties of the mode:

- The mode may not exist.
- The mode may not be unique.
- The mode may or may not equal the mean and median.
- The mode is not affected by extreme values.
- The mode always corresponds to one of the actual observations (unlike the mean and median).
- The mode can be used for either qualitative or quantitative data.

Example:

The data set: 1, 2, 3, 4, 6, 8, 9 has no mode.

The data set: 1, 2, 3, 3, 4, 5 has one mode (Unimodal).

The data set: 1, 1, 2, 3, 4, 4, 5 has two modes (Bimodal).

The data set: 1, 1, 2, 3, 3, 4, 5, 5 has three modes (Trimodal).

Example:

Six strains of bacteria were tested to see how long they could remain alive outside their normal environment. The time, in minutes, 2, 3, 5, 7, 8, 10. Find the mode.

Solution:

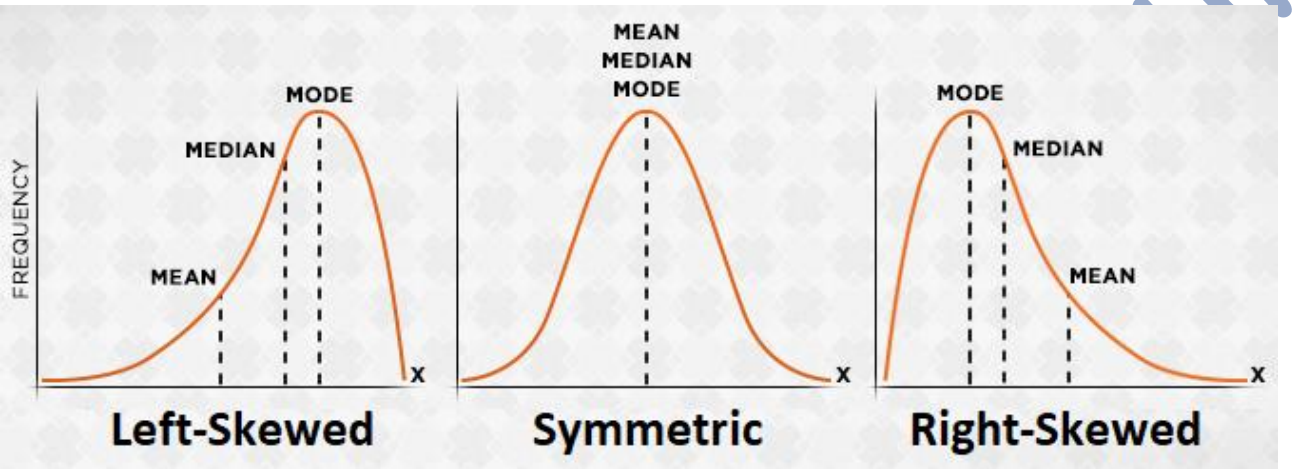
Example:

If for some reason you were categorizing patients by hair color for the hypertension study, and you found that there were 15 individuals with brown hair, 20 individuals with blond hair, 5 with black hair, 2 with purple hair, and 12 with red hair. Which hair color is the mode of this data set?

Solution:

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

1. Symmetric: Mean = Median = Mode.
2. Left-Skewed: Mean < Median < Mode.
3. Right-Skewed: Mean > Median > Mode.



Other Measures of Central Tendency:

1. Trimmed Mean:

It is the mean calculated by discarding a certain percentage of the lowest and the highest scores from the sample.

- For a set of sample data containing n measurements we calculate the percent 100α trimmed mean as follows:
  1. Order the measurements.
  2. Discard the smallest 100α percent and the largest 100α percent of the measurements (The recommended value of α is something between 0.1 and 0.2).
  3. Compute the arithmetic mean of the remaining measurements.
- Note that the median may be regarded as a 50 percent trimmed mean.
- It is most appropriate when the data set contains outliers.

Example:

The caloric content of French fries depends on how they are prepared. A random sample of 10 different 3 ounce servings of French fries from different fast-food restaurants had the calories shown below.

222 255 254 230 249 222 237 287 315 245

Find the 10% trimmed mean.

Solution:

2. Geometric Mean

It is a kind of average of a set of numbers. It can be calculated by multiplying all the numbers and taking the n-th root of the total.

$$GM = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n}$$

- GM is an appropriate measure when:
  - ✓ values change exponentially, and
  - ✓ in case of skewed distribution.
- GM is more commonly used in microbiological and serological research, in which distributions are often right-skewed.
- One important disadvantage of GM is that it cannot be used if any of the values are zero or negative.

- The geometric mean is used for variables whose effect is multiplicative.
- For example, if a tree increases its height by 60% one year, 8% the next year, and 4% the third year, its final height would be the initial height multiplied by 1.60×1.08×1.04=1.80. Taking the geometric mean of these numbers (1.216) and multiplying that by itself three times also gives the correct final height (1.80), while taking the arithmetic mean (1.24) times itself three times does not give the correct final height.

Example:

If a strain of bacteria increases its population by 20% in the first hour, 30% in the next hour and 50% in the next hour. Find the mean rate of growth of the bacteria over the period of 3 hours.

Solution:

3. Harmonic Mean:

It is the reciprocal of the arithmetic mean of the observations. That is,

$$HM = \frac{1}{\frac{\sum_{i=1}^{n}\frac{1}{X_i}}{n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}.$$

- This mean is useful for finding the average speed.
- It is most appropriate when the data set contains outliers.
- It cannot be used if any of the values are zero.

Example:

Suppose a person drove 100 miles at 40 miles per hour and returned driving 50 miles per hour. What is the average speed?

Solution:

Example:

If six birds set up their first nest 1.0, 1.4, 1.7, 2.1, 2.8, and 47 km from the nest they were born in, then

- The arithmetic mean dispersal distance would be 9.33 km,
- The geometric mean would be 2.95 km,
- The harmonic mean would be 1.90 km.

4. Weighted Mean:

It is calculated by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights. That is,

$$WM = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

- It is useful when certain values in a data set are considered more important than others.

  For example, to determine students' grades in a course, an instructor may assign a weight to the final exam that is twice as much as that to each of the other exams.

Example:

Find the average number of grams of fat per ounce of meat or fish that a person would consume over a 5-day period if he ate these:

Meat or fish	Fat (g/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

Solution:

5. Combined Mean:

One property of the mean is that if we know the means and sample sizes of two (or more) data sets, we can calculate the combined mean of both (or all) data sets. The combined mean for two data sets is calculated by using the formula:

$$\bar{X}_c = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

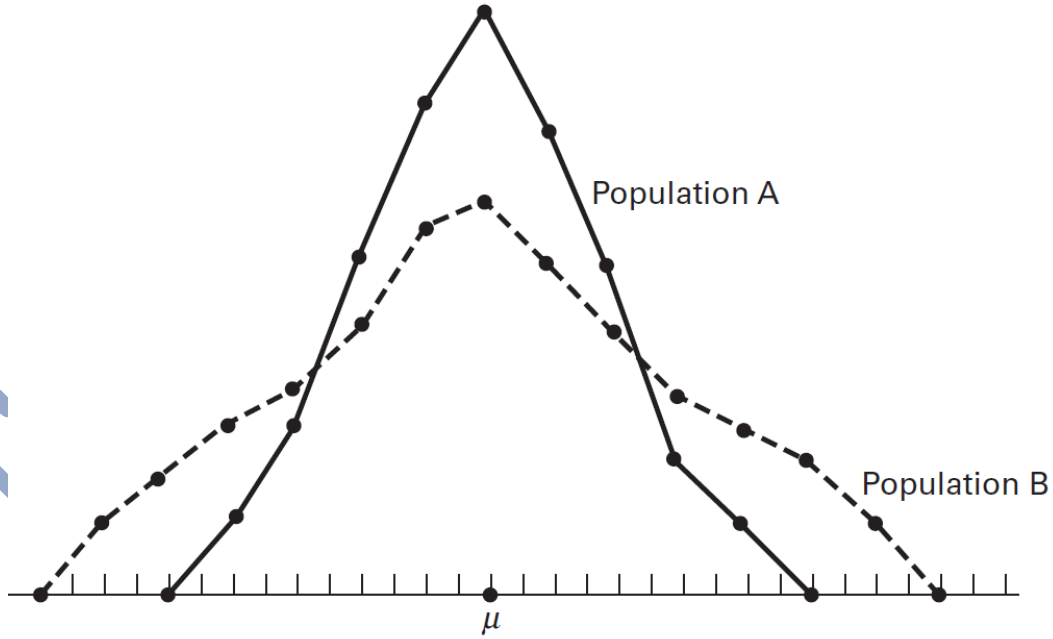
Example:

Suppose a sample of 10 statistics books gave a mean price of $140 and a sample of 8 mathematics books gave a mean price of $160. Find the combined mean.

Solution:

Measures of Dispersion:

- A measure of dispersion conveys information regarding the amount of variability present in a set of data.
- If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data.
- The amount of dispersion may be small when the values, though different, are close together.
- The following Figure shows the frequency polygons for two populations that have equal means but different amounts of variability. Population B, which is more variable than population A, is more spread out.



Two frequency distributions with equal means but different amounts of dispersion.

1. Range:

The **range** is the difference between the largest and smallest value in a set of observations. That is,

$$R = Max - Min.$$

- It is not a good measure of dispersion to use for a data set that contains outliers.
- It is not a very satisfactory measure of dispersion. 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.

2. Variance:

The **variance** of a data set is the average of the squares of the distance each value is from the mean.

- It measures how far a set of numbers are spread out from their average value.
- It is nonnegative and is zero only if all observations are the same.

It is the most popular and best understood measure of central tendency for quantitative data.

To find the mean of a data set, use one of the following formulae.

Population Variance:

The formula for the population variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

where,

σ^2: is the population variance (σ is a Greek letter, read as sigm).
μ: is the population mean.
X_i: is the value of element i in the population.
N: is the population size.

Short-Cut Formula for the Population Variance is

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2.$$

Sample Variance:

The formula for the sample variance is

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

where,

$\bar{X}$: is the sample mean.

X_i: is the value of element i in the sample.

n: is the sample size.

Short-Cut Formula for the Sample Variance is

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$$

Degrees of freedom

In computing the variance there are $n - 1$ degrees of freedom because if $n - 1$ values are known, the n-th one is determined automatically. This is because all of the values of $(X_i - \bar{X})$ must add to zero.

3. Standard Deviation:

The variance represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To obtain a measure of dispersion in original units, we merely take the square root of the variance. The result is called the **standard deviation**.

Population Standard Deviation:

The formula for the population standard deviation is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Sample Standard Deviation:

The formula for the sample standard deviation is

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}}$$

- It is the most commonly used measure of variation.
- It shows variation about the mean.
- It has the same units as the original data.

Example:

The following cholesterol levels of 10 people were measured in mg/dl:

260, 150, 165, 201, 212, 243, 219, 227, 210, 240.

1. Calculate the range.
2. Calculate the variance and the standard deviation.

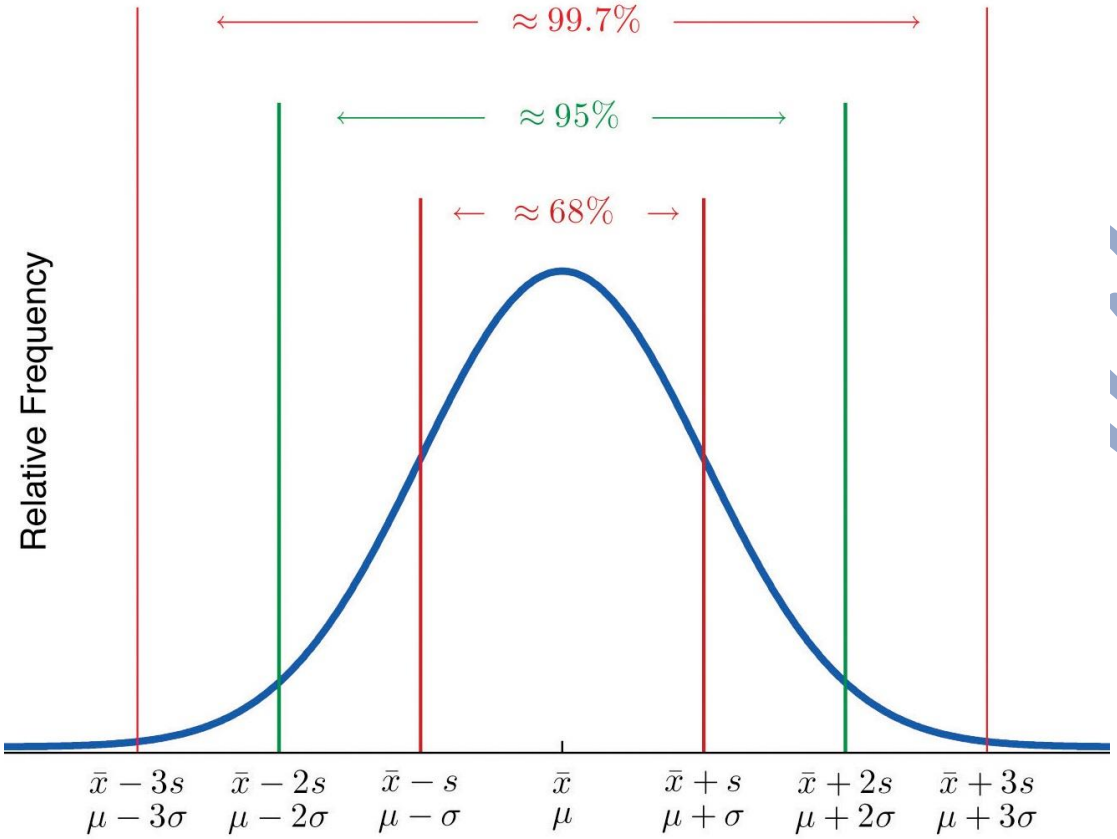
Solution:

Interpreting Standard Deviation

When interpreting the standard deviation, remember that it is a measure of the typical amount an entry deviates from the mean. The more the entries are spread out, the greater the standard deviation.

Empirical Rule

For data with a (symmetric) bell-shaped distribution, the standard deviation has the following characteristics.



1. About 68% of the data lie within one standard deviation of the mean.
2. About 95% of the data lie within two standard deviations of the mean.
3. About 99.7% of the data lie within three standard deviations of the mean.

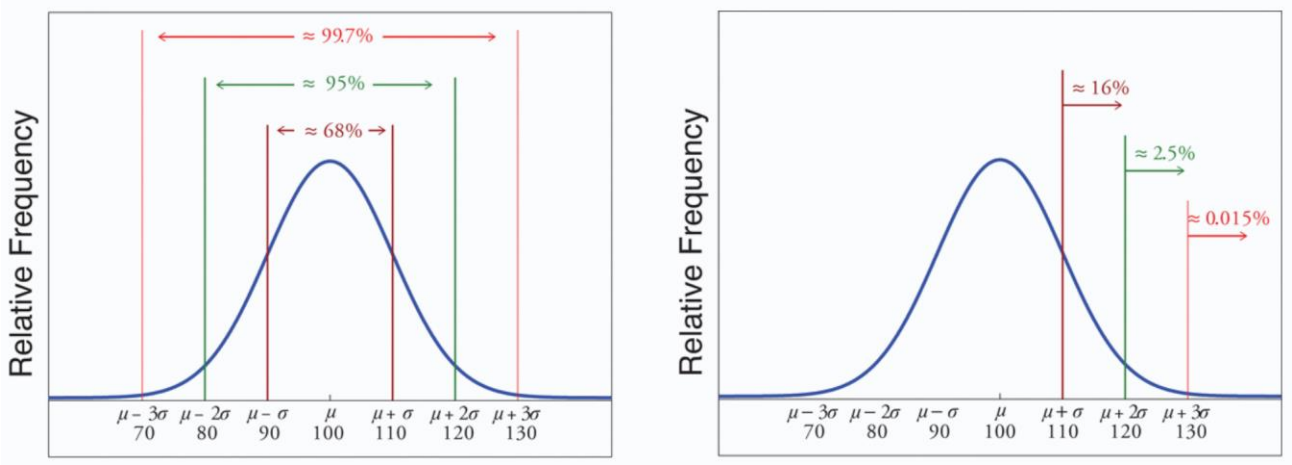
Example:

Scores on IQ tests have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Discuss what the Empirical Rule implies concerning individuals with IQ scores of 110, 120, and 130.

Solution:

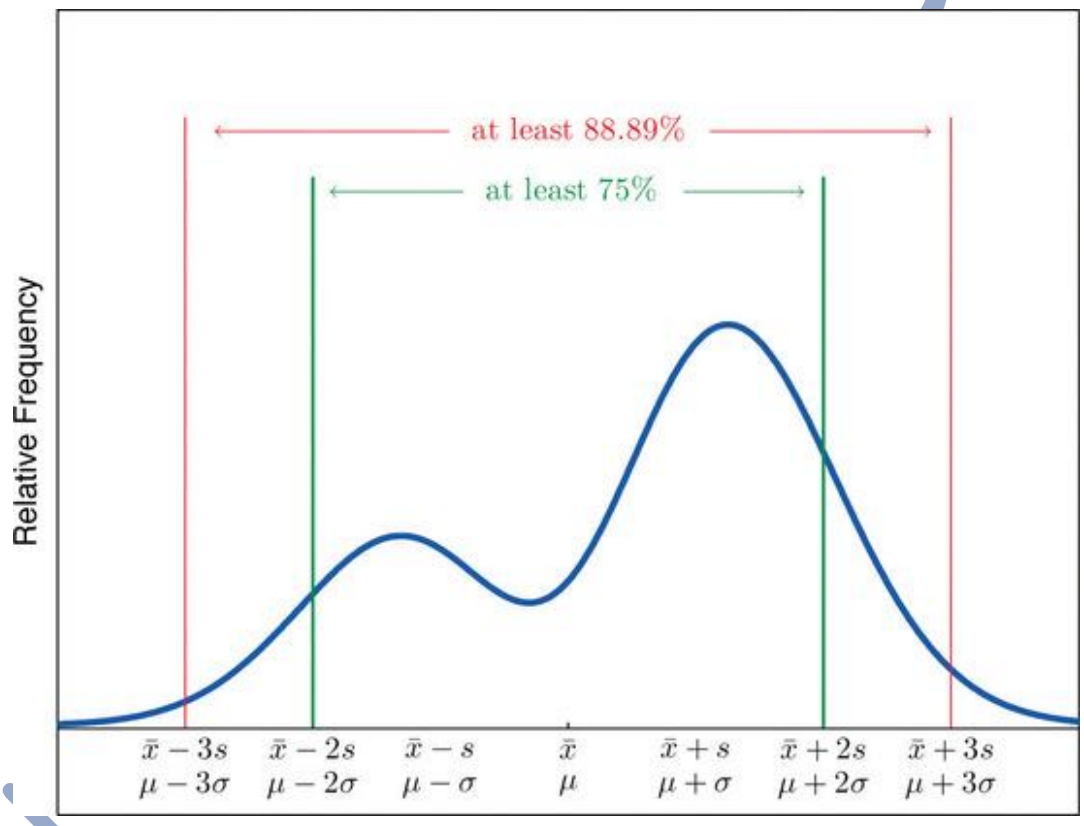
The Empirical Rule states that

1. approximately 68% of the IQ scores in the population lie between 90 and 110,
2. approximately 95% of the IQ scores in the population lie between 80 and 120, and
3. approximately 99.7% of the IQ scores in the population lie between 70 and 130.



Chebyshev's Rule

It applies to any data set, regardless of the shape of the distribution.



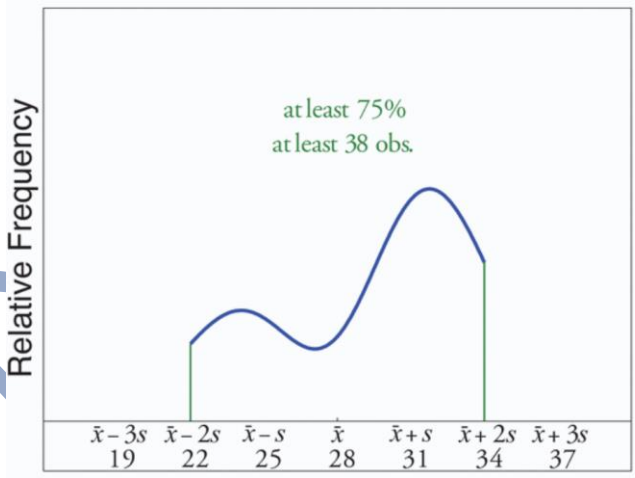
1. No useful information is provided on the fraction of measurements that fall within one standard deviation of the mean.
2. At least 75% of the data will fall within two standard deviations of the mean.
3. At least 89% of the data will fall within three standard deviations of the mean.

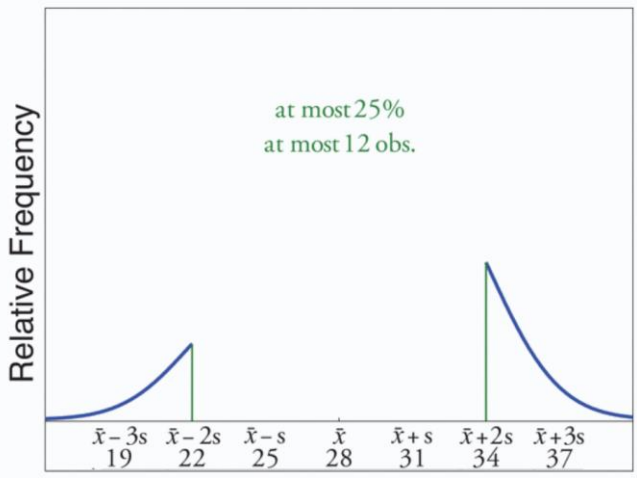
Generally, for any $k > 1$, at least $1 - \frac{1}{k^2}$ of the data will fall within k standard deviations of the mean.

Example:

A sample of size $n = 50$ has mean $\bar{x} = 28$ and standard deviation $s = 3$. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval $(22,34)$? What can be said about the number of observations that lie outside that interval?

Solution:





Measures of Position:

Statisticians often talk about the position of a value, relative to other values in a set of data. The most common measures of position are percentiles, quartiles, and standard scores (z-scores).

1. Percentiles:

The ***k*-th percentile** P of a sample of n observations (denoted by P_k) is that value of the variable such that k percent or less of the observations are less than P_k and $(100 - k)$ percent or less of the observations are greater than P_k.

For example, the 25-th percentile is that value of a variable (P_{25}) such that 25% of the observations are less than that value and 75% of the observations are greater.

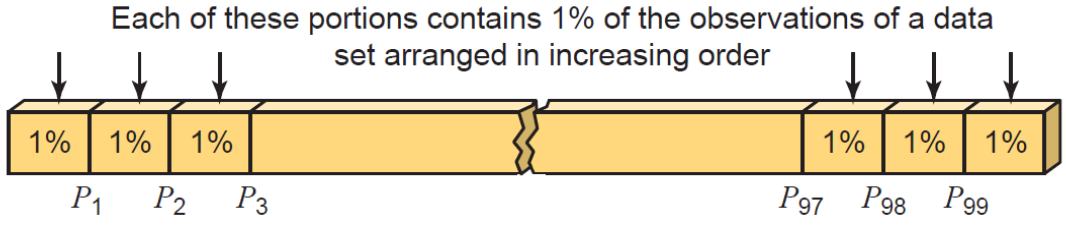
Calculating the k-th percentile (P_k):

1. Order the data from smallest to largest.
2. Determine the location of the percentile (index):

$$\text{index} = \frac{n * k}{100}.$$

   a. If index is not an integer, round it up to the next integer and find the corresponding ordered value.

   b. If index is an integer, say m, calculate the average of the m-th and $(m + 1)$-th ordered values.



Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

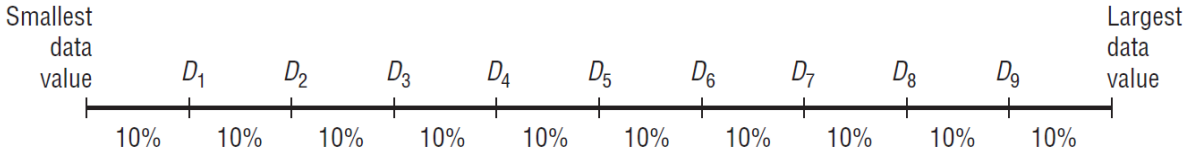
5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

1. Calculate the approximate value of the 90-th percentile.
2. Calculate the approximate value of the 50-th percentile.

Solution:

2. Deciles:

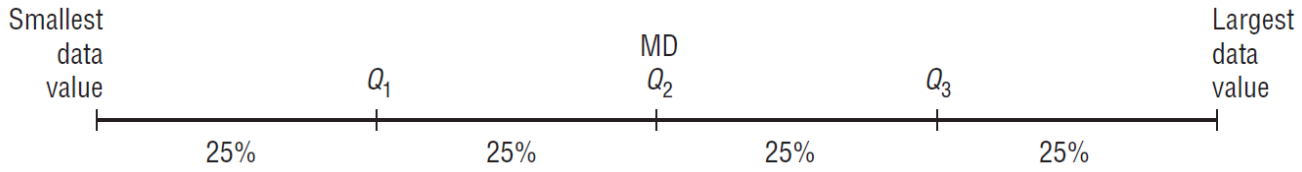
The **deciles** are special cases of percentiles; $D_1 = P_{10}$, $D_2 = P_{20}$, ..., $D_9 = P_{90}$. Thus, deciles divide the data set into 10 equal parts.



3. Quartiles:

Quartiles divide the data set into four equal parts, separated by Q_1, Q_2, Q_3.

Note that Q_1 is the first quartile (same as the 25-th percentile); Q_2 is the second quartile (same as the 50th percentile, or the median, or D_5); and Q_3 is the third quartile (corresponds to the 75th percentile).



Calculating Q_1, Q_2, and Q_3:

1. Arrange the 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:

Prior to the start of class, resting heart rates were recorded for all females who had signed up for a low-impact aerobics program. The rates for the fifteen females, ages 20 to 25, were

73, 77, 80, 83, 73, 82, 75, 73, 77, 84, 76, 81, 75, 79, 70.

Locate the first quartile.

Solution:

Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

1. Find the values of Q_1, Q_2, and Q_3.
2. Calculate the approximate value of D_1.

Solution:

Percentile Rank

A percentile rank of a given score X indicates the proportion or percentage of data values that fall below X.

Finding percentile rank of a value:

$$Percentile\ rank\ of\ X = \frac{Number\ of\ observations\ below\ X + 0.5}{n} * 100\%$$

Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

Find the percentile rank of 8.

Solution:

Five Number Summary

The combination of the five numbers ($Min,\ Q_1,\ MD,\ Q_3,\ Max$) is called the **five number summary**, and provides a quick numerical description of both the center (location) and spread of a distribution. Each of the values represents a measure of position in the dataset. The min and max providing the boundaries and the quartiles and median providing information about the 25-th, 50-th, and 75-th percentiles.

4. Standard score (z-score):

Sometimes we want to compare data which come from different samples or populations. This is sometimes done using standard units or scores.

The **standard score,** or ***z*-score,** represents the distance between a given measurement X and the mean, expressed in standard deviations.

The **sample** z**-score** for a measurement X is

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

The **population** z**-score** for a measurement X is

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

- The number of standard deviations that a particular measurement deviates from the mean is called z-score.
- A z-score can be negative, positive, or zero.
- If z is negative, the corresponding x-value is below the mean.
- If z is positive, the corresponding x-value is above the mean.
- If $z = 0$, the corresponding x-value is equal to the mean.
- Given values for μ (or $\bar{X}$) and σ (or S), we can go from the "X scale" to the "z scale," and vice versa. Algebraically, we can solve for X and get $X = \sigma z + \mu$. This is also the procedure that is used to get from degrees Celsius (◦C) to degrees Fahrenheit (◦F). The relationship is

  $$C = \frac{F - 32}{1.8}.$$

  Similarly,

  $$F = 1.8 * C + 32.$$

Example:

Student A got a score of 85 in a test whose scores had mean 79 and standard deviation 8. Student B got a score of 74 in a test whose scores had a mean of 70 and a standard deviation 5. Which student got a higher score?

Solution:

Other Measures of Variations:

1. Interquartile Range:

The **interquartile range** (IQR) is the difference between the third and first quartiles. That is,

$$IQR = Q_3 - Q_1$$

2. Coefficient of Variation:

The **coefficient of variation** is a measure of relative variability that expresses standard deviation as a percentage of the mean. It is computed as follows:

For population data:

$$CV = \frac{\sigma}{\mu} * 100\%$$

For sample data:

$$CV = \frac{S}{\bar{X}} * 100\%$$

- It summarizes the amount of variation as a percentage or proportion of the total.
- It is useful when comparing the amount of variation for one variable among groups with different means, or among different measurement variables.

Example:

Suppose two samples of human males yield the following results:

	Sample 1	**Sample 2**
Age	25 years	11 years
Mean Weight	145 Pounds	80 Pounds
Standard Deviation	10 Pounds	10 Pounds

We wish to know which is more variable, the weights of the 25-year-olds or the weights of the 11-year-olds.

Solution:

Example:

The yearly salaries of all employees who work for a company have a mean of $62,350 and a standard deviation of $6820. The years of experience for the same employees have a mean of 15 years and a standard deviation of 2 years. Is the relative variation in the salaries larger or smaller than that in years of experience for these employees?

Solution:

Identifying Outliers:

An **outlier** is a value that is very small or very large relative to the majority of the values in a data set.

Steps for identifying outliers:

1. Arrange data in order.
2. Calculate the first quartile (Q_1), third quartile (Q_3), and the interquartile range (IQR).
3. Compute $Q_1 - 1.5 * IQR$ and $Q_3 + 1.5 * IQR$.
4. Anything outside the interval $(Q_1 - 1.5 * IQR,\ Q_3 + 1.5 * IQR)$ is an outlier.

Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

Check this data set for outliers.

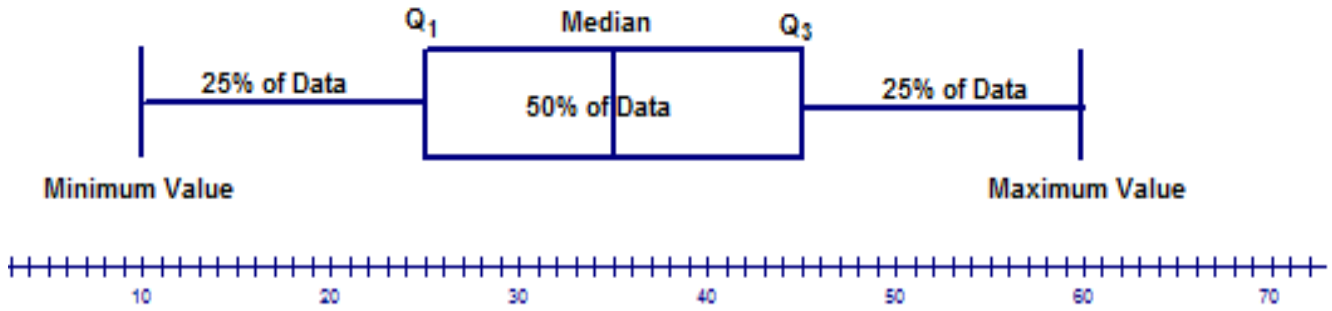
Solution:

Box-and-Whisker Plots:

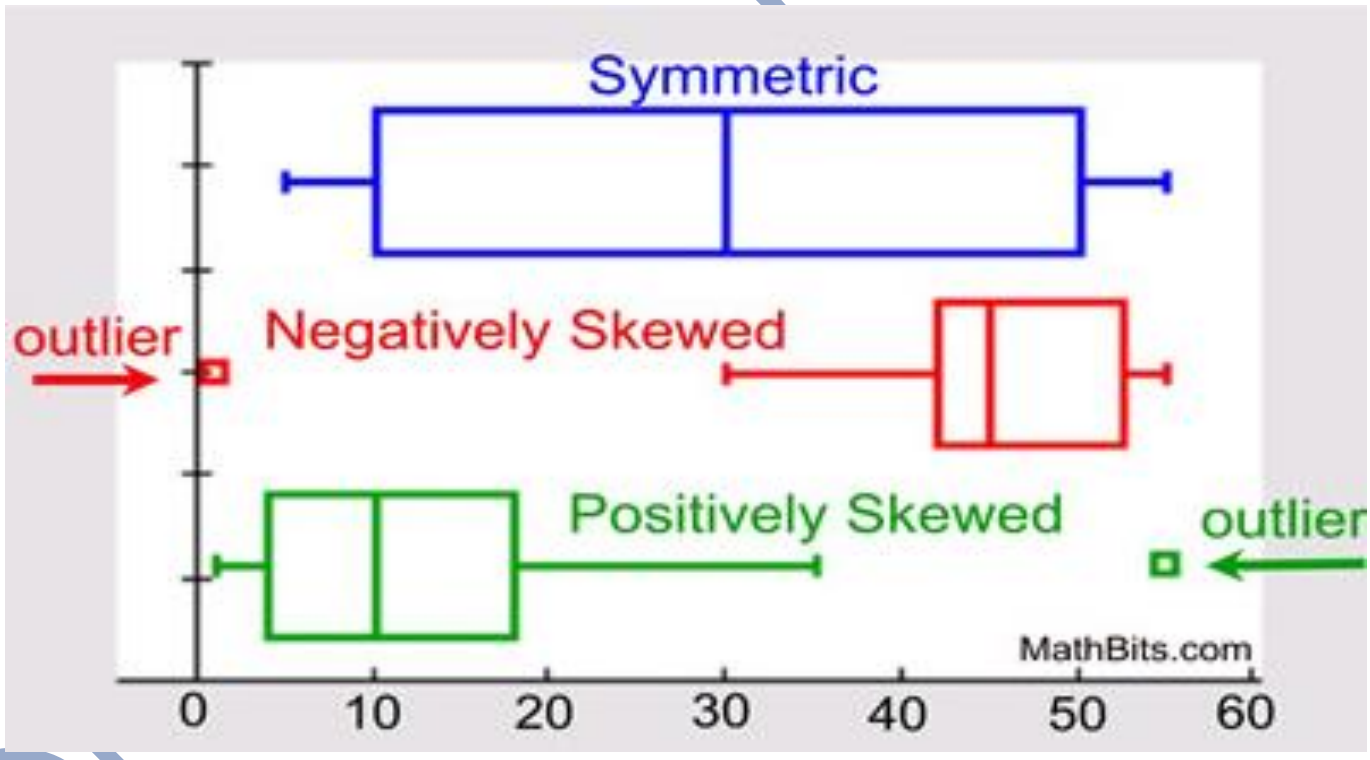
A useful visual device for communicating the information contained in a data set is the **box-and-whisker plot**. The construction of a box-and-whisker plot (sometimes called, simply, a **boxplot**) makes use of the quartiles of a data set and may be accomplished by following these five steps:

- Represent the variable of interest on the horizontal axis.
- Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile and the right end of the box aligns with the third quartile
- Divide the box into two parts by a vertical line that aligns with the median.
- Draw a horizontal line called a whisker from the left end of the box to a point that aligns with the smallest measurement in the data set.

- Draw another horizontal line, or whisker, from the right end of the box to a point that aligns with the largest measurement in the data set.



- Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.



Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

Construct a boxplot for the data and comment on their shape.

Solution:

Measures of Shape:

Measures of shape describe the distribution (or pattern) of the data within a dataset.

1. Coefficient of Skewness:

Skewness is a measure of asymmetry, or more precisely, the lack of symmetry of the data. It is defined as:

$$Skewness = \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}}$$

- The direction of skewness is given by the sign.
- A value of zero means no skewness at all (Symmetric).
- A large negative value means the distribution is negatively skewed (Left-skewed).
- A large positive value means the distribution is positively skewed (Right-Skewed).

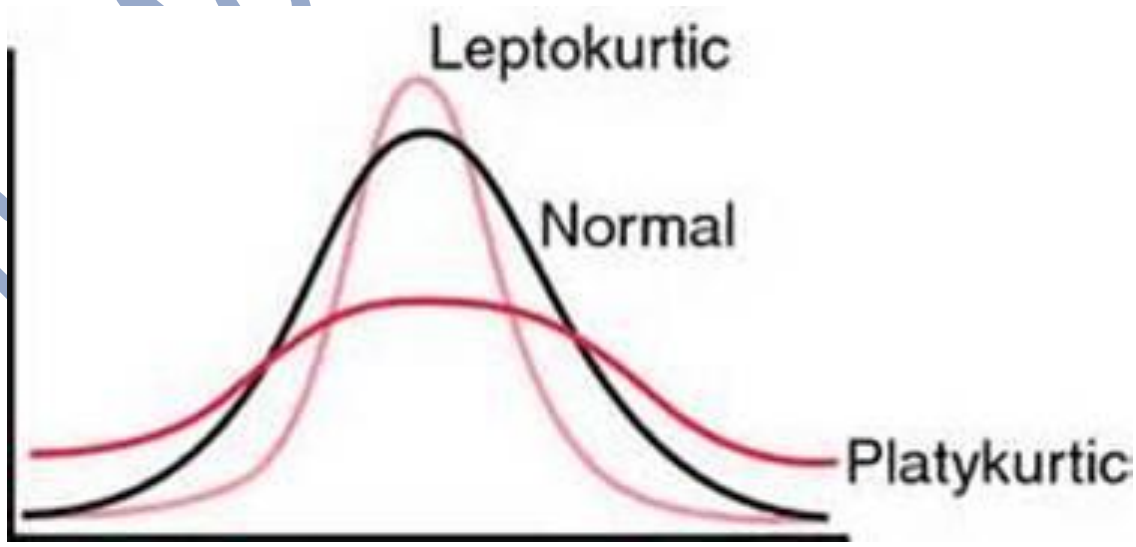
2. Kurtosis:

Kurtosis is a measure of the degree to which a distribution is "peaked" or flat in comparison to a normal distribution whose graph is characterized by a bell-shaped appearance. It is defined as:

$$Kurtosis = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} - 3$$

Types of Kurtosis:

1. $Kurtosis = 0$ ⇒ **Mesokurtic**: the value for a bell-shaped distribution (Normal).
2. $\text{Kurtosis} < 0$ ⇒ **Leptokurtic**: thin or peaked shape (or "light tails").
3. $\text{Kurtosis} > 0$ ⇒ **Platykurtic**: flat shape (or "heavy tails").



Example:

Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10

1. Calculate the coefficient of skewness.
2. Calculate the coefficient of kurtosis.

Solution:

Exercises:

Q1. A student has gotten the following grades on his tests: 87, 95, 76, and 88. He wants an 85 or better overall. What is the minimum grade he must get on the last test in order to achieve that average?

Q2. Find the mean and the median for the LDL cholesterol level in a sample of ten heart patients.

132 139 162 147 133 160 145 150 148 153

Q3. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 500 days, four mice have died but the fifth one survives. The recorded survival times for the five mice are

493 421 222 378 500*

where 500* indicates that the fifth mouse survived for at least 500 days but the survival time (i.e., the exact value of the observation) is unknown.

1. Can you find the sample mean for the data set? If so, find it. If not, why not?
2. Can you find the sample median for the data set? If so, find it. If not, why not?

Q4. Find the mode of the following set of marks.

Marks	1	2	3	4	5
Frequency	6	7	7	5	3

Q5. An 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.

Q6. If the mean of five values is 64, find the sum of the values.

Q7. If the mean of five values is 8.2 and four of the values are 6, 10, 7, and 12, find the fifth value.

Q8. Suppose the average amount of money spent on shopping by 10 persons during a given week is $105.50. Find the total amount of money spent on shopping by these 10 persons.

Q9. Twenty business majors and 18 economics majors go bowling. Each student bowls one game. The scorekeeper announces that the mean score for the 18 economics majors is 144 and the mean score for the entire group of 38 students is 150. Find the mean score for the 20 business majors.

Q10. Seven airline passengers in economy class on the same flight paid an average of $361 per ticket. Because the tickets were purchased at different times and from different sources, the prices varied. The first five passengers paid $420, $210, $333, $695, and $485. The sixth and seventh tickets were purchased by a couple who paid identical fares. What price did each of them pay?

Q11. Suppose that the inflation rates for the last five years are 4%, 3%, 5%, 6%, and 8%, respectively. Thus at the end of the first year, the price index will be 1.04 times the price index at the beginning of the year, and so on. Find the mean rate of inflation over the 5-year period.

Q12. An index is to be computed using the following weights: resting diastolic pressure, 3; resting heart rate, 2; serum cholesterol, 1. Readings are taken weekly for six weeks and then averaged for the period. A patient's average readings are as follows: diastolic pressure, 82; heart rate, 62; cholesterol, 164. What is the weighted mean?

Q13. In a sample of size 1000 the 80^{th} percentile is equal to 30. what is the number of measurements that are less than 30?

Q14. Consider the following sample of five measurements:

2 1 1 0 3.

Calculate the range, sample variance, and sample standard deviation.

1. Add 3 to each measurement and repeat Part (1).
2. Subtract 4 from each measurement and repeat Part (1).
3. Multiply each measurement by 10 and repeat Part (1).
4. Considering your answers to Parts (1, 2, 3, 4), what do you conclude?

Q15. Suppose that 40 and 90 are two elements of a population data set and that their standard scores are -2 and 3, respectively. Using only this information, is it possible to determine the population's mean and standard deviation? If so, find them.

Q16. In a study of water pollution, a sample of mussels was taken and lead concentration (milligrams per gram dry weight) was measured from each one. The following data were obtained:

113.0 140.5 163.3 185.7 202.5 207.2

Calculate the mean, variance, and standard deviation.