Chapter 1:
Overview and Descriptive Statistics

Learning Objectives

1. Define Statistics, Population, Sample, and Process
2. Methods in descriptive statistics
3. Measures of location
4. Measures of variability

1.1
Statistics, Population, Sample, and Process

Statistics is the science of problem solving using **DATA**. A statistical investigative process involves with collecting, organizing, summarizing and analyzing information in order to draw conclusions or to make decisions.

Descriptive Statistics deals with organizing, displaying, and summarizing the data collected.

Inferential Statistics consists of methods that use sample results to help make decisions or predictions about a population.

Census is a survey that includes **every member** of the population whose desired information can be obtained.

Sample survey is a technique of collecting information from a **portion** of the population.

Variables are the characteristics of the individuals (sampling units) within the population.

Random Variables are variables that take values **at random**.

Key Point: ***Possible values of variables vary.*** *Consider the variable "height". If all individuals had the same height, then obtaining the height of one individual would be sufficient in knowing the heights of all individuals. Of course, this is not the case. As investigators, we wish to identify the factors that influence variability.*

Scales of Measurement

- Scales of measurement include:
  - Nominal
  - Interval
  - Ordinal
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

Scales of Measurement

- Nominal

- Data are labels or names used to identify an attribute of the element.
- A nonnumeric label or numeric code may be used.

Scales of Measurement

- Ordinal

- The data have the properties of nominal data and the order or rank of the data is meaningful.
- A nonnumeric label or numeric code may be used.

Scales of Measurement

- Interval

The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.

Interval data are always numeric.

Scales of Measurement

- Ratio

The data have all the properties of interval data and the ratio of two values is meaningful.

Variables such as distance, height, weight, and time use the ratio scale.

This scale must contain a zero value that indicates that nothing exists for the variable at the zero point.

The list of observations a variable assumes is called **data.**

Qualitative data are observations corresponding to a qualitative variable.

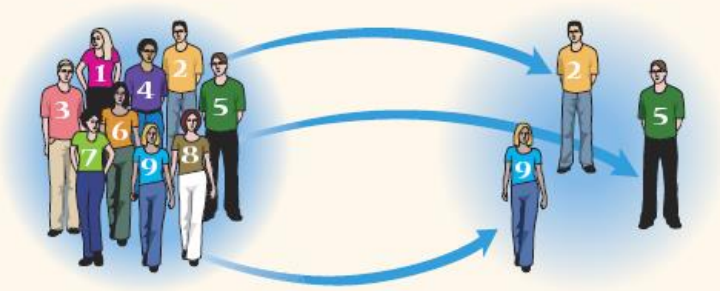
Gender is a variable. The possible values for Gender are male or female. So, the values: Male or Female are data.

Quantitative data are observations corresponding to a quantitative variable.

Weight is a variable. A weight of 154 lbs is an observation.

- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.

Simple Random Sampling

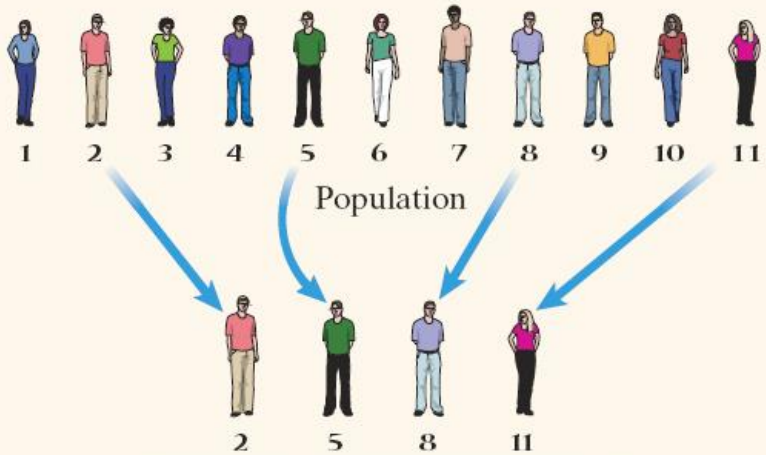


Population Sample

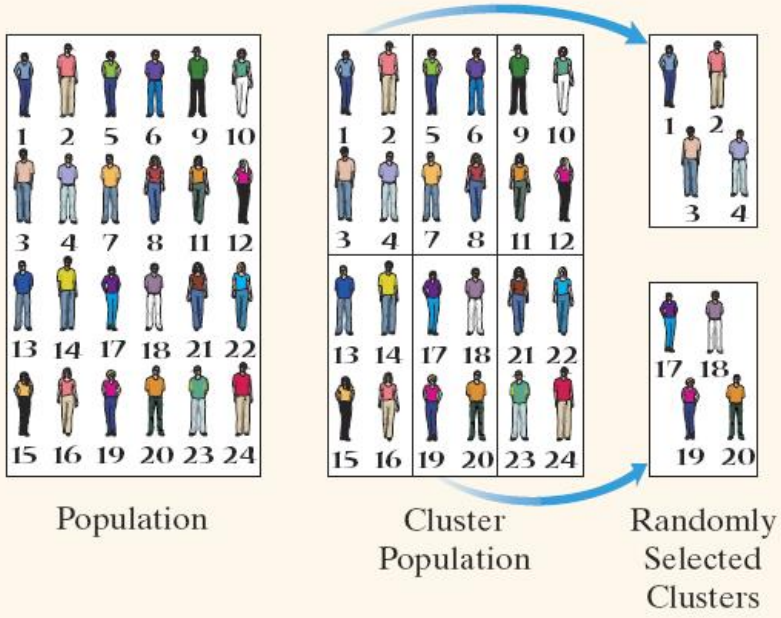
Stratified Sampling

Population Strata Sample

Systematic Sampling



Cluster Sampling



1.2 Methods in descriptive statistics

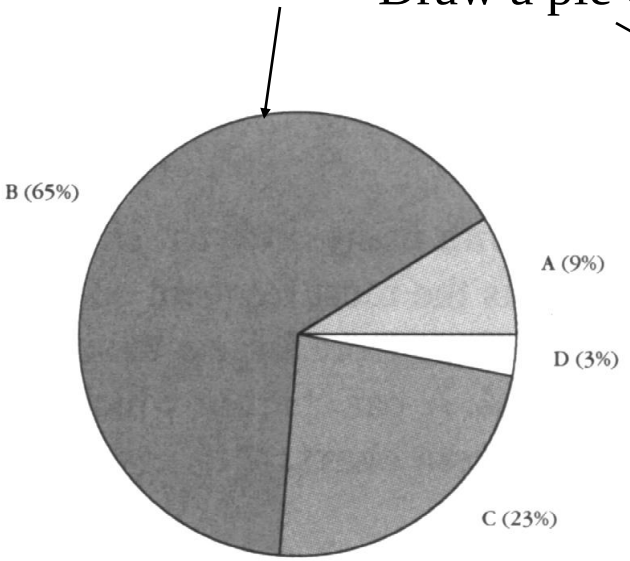
- A statistical table can be used to display data graphically as a data distribution: consists of Class, Class Frequency, Relative Frequency or Percentage.
- For qualitative data, three measurements are available for the list of categories:
  - the frequency, or number of measurements
  - the relative frequency, or proportion = frequency / Total # of observations
  - the percentage
    - A pie chart is the familiar circular graph that shows how the measurements are distributed among the categories.
    - A bar chart shows the same distribution of measurements in categories, with the height of the bar measuring how often a particular category was observed.
    - A bar chart in which the bars are ordered from largest to smallest is called a Pareto chart.

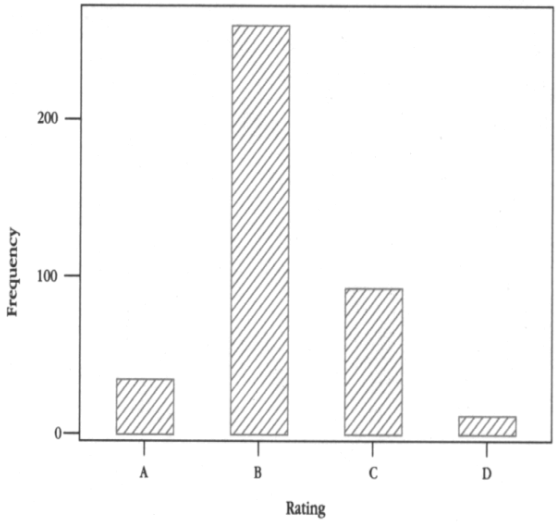
A survey of 400 individuals are made to rate the quality of a school.

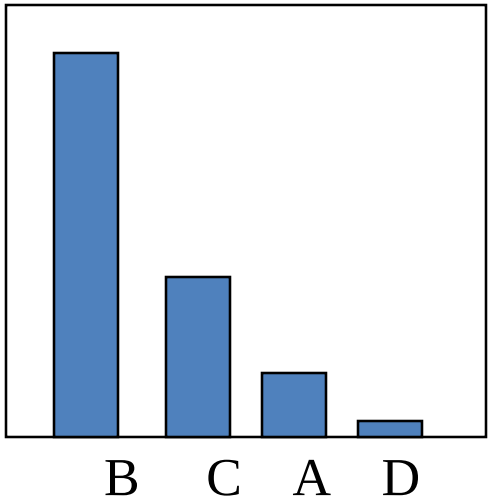
The data is summarized:

Rating	Frequency	Relative Frequency	Percentage
A	35	.0875	8.75%
B	260	.65	65%
C	93	.2325	23.25%
D	12	0.03	3%
Total	400	1	100%

Draw a pie chart, a Bar Chart and a Pareto chart







Stem and leaf plots:

- This plot presents a graphical display of the data using the actual numerical values of each data point.

Constructing a Stem and Leaf Plot:

1. Divide each measurement into two parts: the stem and the leaf.
2. List the stems in a column, with a vertical line to their right.
3. For each measurement, record the leaf portion in the same row as its matching stem.
4. Order the leaves from lowest to highest in each stem.
5. Provide a key to your stem and leaf coding so that the reader can recreate the actual measurements if necessary.

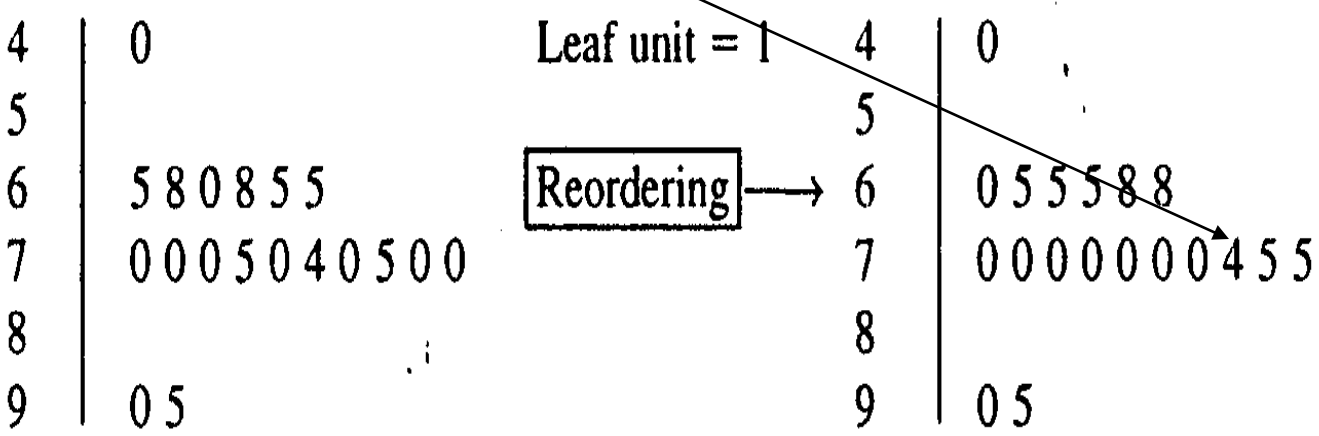
Example

The following Table lists the prices (in dollars) of 19 different brands of walking shoes. Construct a tem and leaf plot to display the distribution of the data.

90	70	70	70	75	70
65	68	60	74	70	95
75	70	68	65	40	65
70					

The price 74 is represented by the stem 7 and leaf 4. The price obtained by: 74 x (Leaf Unit) = 74x(1) = 74.

Solution



Histograms

What is it?

A **histogram** for a **quantitative data set** is a graph that describes the relative frequency (or frequency) of the variable in which the possible values of the variable are divided into a few groups (classes, or intervals), the relative frequency (or frequency) is represented by a rectangle with **the height** representing **the proportion or relative frequency of occurrence for a particular class** (or group) of the variable being measured. **And these rectangles should touch each other.**

- On the X axis: The class, (or group) of the variable are plotted along the *x* axis.
- On the Y-axis: The relative frequency or frequency of observations within the class is the height on the Y axis.

How to construct a histogram?

Constructing a relative frequency histogram for continuous variables:

1. Choose the number of classes, usually between 5 and 15.
2. Calculate the approximate class width by dividing the difference between the largest and smallest values

   (Range = largest – smallest) by the number of classes.
3. Round the approximate class width up to a convenient number.
4 Locate the class boundaries.

   If discrete, assign one or more integers to a class.

   If continuous, use **Method of left inclusion**: Include the left class boundary point but not the right boundary point in the class.

   – NOTE: Different methods may be used in different software. Some may use right inclusion. Some may add an additional decimal place for the class boundary.
5. Construct a statistical table containing the classes, their boundaries, and their relative frequencies.
6. Construct the histogram like a bar graph with

   the rectangle height = r.f. of the class/ class width

Example

The following Table lists the prices (in dollars) of 19 different brands of walking shoes. Construct a relative histogram to display the distribution of the data.

90	70	70	70	75	70	65	68	60	74	70	95
75	70	68	65	40	65	70					

Solution:

1. Determine # of classes: for example, use k=6 classes
2. Range = 95 -40 = 55,
3. Class width = 55/6 ~ 9.17 ~ 10

   (Round the width up (not round off nor truncate) to a 'convenient number.)
4. Use left-inclusion to determine class boundaries:

   [40,50),[50,60), [60, 70), [70,80),[80,90),[90,100)
5. Construct a Relative Frequency Table – first count # of observations in each class. This is the frequency, call it f_i. Relative frequency (rf_i) = f_i/n , where n is the total # of data points.
6. Draw a two-dimensional graph with

   X-axis: the class boundaries of the variable, and Y-axis: the relative frequency for each class, and a rectangle with the relative frequency as the height for each class.

Relative Frequency Table

Group	Frequency	Relative Frequency
[40,50)	1	1/19
[50,60)	0	0
[60,70)	6	6/19
[70,80)	10	10/19
[80.90)	0	0
[90,100)	2	2/19

Histogram

1.3 Measures of Location

(Mean, Median, Mode, Percenti and Quartiles)

A **parameter** is a descriptive measure of a population.

In most real world cases, the population parameter is not known. For example, the average gas price in the whole nation.

A **statistic** is a descriptive measure of a sample. We use statistic to estimate the corresponding parameter. For example, Average gas price of the nation is not known. However, we can take a random sample of 100 stations and compute the sample average gas price, then use the sample average to estimate the unknown population average.

The **population mean**, is computed using **all** the individuals in a population, the total # of all individuals is N.

The population mean is a ***parameter***.

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N}$$

The **sample mean**, is computed using sample data.

The sample mean is a statistic that is an unbiased estimator of the population mean.

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

NOTE: In real world applications, population mean μ is usually not known, and is estimated by using sample mean $\bar{x}$

EXAMPLE *Computing a Population Mean and a Sample Mean*

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

(a) Compute the population mean of this data.

(b) Then take a simple random sample of $n = 3$ employees. Compute the sample mean. Obtain a second simple random sample of $n = 3$ employees. Again compute the sample mean.

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data is below the median and half the data is above the median. We use *m* to represent the median.

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data is below the median and half the data is above the median. We use *m* to represent the median.

Steps in Computing the Median of a Data Set

1. Arrange the data in ascending order.
2. Determine the number of observations (n).
3. Determine the observation in the middle of the data set. The position is **(n+1)/2**
   (a) If ***(n+1)/2*** *is an integer (i.e. n is* ***ODD****),* locate the data value at the (n+1)/2 position.

   (b) If ***(n+1)/2*** is NOT an integer (i.e. n is **EVEN**), the median is the average of the two data values on either side of the observations that lies in the (n+1)/2 position.

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

If there are two values that occur with the most frequency, we say the data has is **bimodal**.

EX: Find the mode of the following pulse rate data

80, 76, 65, 68, 72, 73, 65, 80,100

Modes are: 65 and 80

- Qualitative data
  - Cannot add or order them … the mean and median do not exist
  - The mode is the only one of these three measurements that exists
- Find the mode of

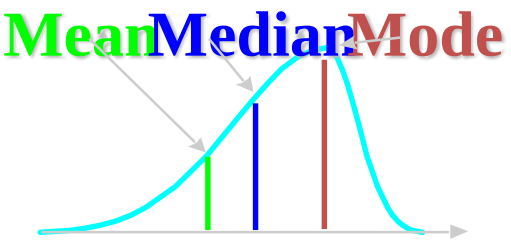
blue, blue, blue, red, green

- The mode is "blue" because it is the value that occurs the most often

Comparison of Mean, Median, and Mode for different shapes of distributions

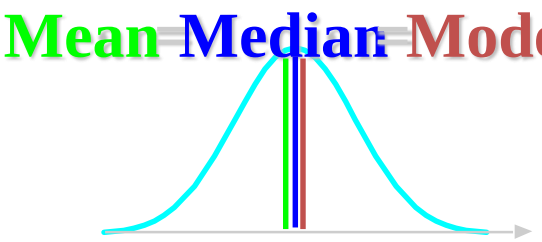
Mean<Median

Left-Skewed



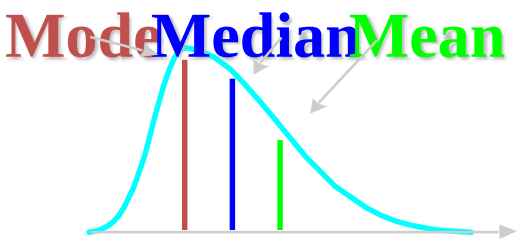
Mean~Median

Symmetric



Mean>Median

Right-Skewed



Percentiles

- **The *p*th percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items take on this value or more.**

Percentiles

- Arrange the data in ascending order.
- Compute index i, the position of the pth percentile.

$$i = (p/100)n$$

- If i is not an integer, round up. The pth percentile is the value in the ith position.
- If i is an integer, the pth percentile is the average of the values in positions i and $i+1$.

Quartiles

- Quartiles are specific percentiles.
- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

1.4 Measures of Variability

Five different measures of variability:

Range, Variance, Standard Deviation,

Interquartile Range(IQR)), Coefficient of Variation

Measures of variability measure the degree that the data values spread. The larger the data values spread, the larger the variation of the data values.

How to measure the variation?

- Range = R = Largest Data Value – Smallest Data Value

- **The sample variance** is :

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

- The sample standard deviation is: $\mathbf{s = \sqrt{s^2}}$

NOTE: the divider: (n-1) is called the Degrees of Freedom.

The **population variance** is symbolically represented by lower case Greek sigma squared.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}$$

The **population standard deviation** is: $\sigma = \sqrt{\sigma^2}$

What is the meaning of variation and how is it used in solving real world problems?

Empirical Rule

For Symmetric mound-shaped data

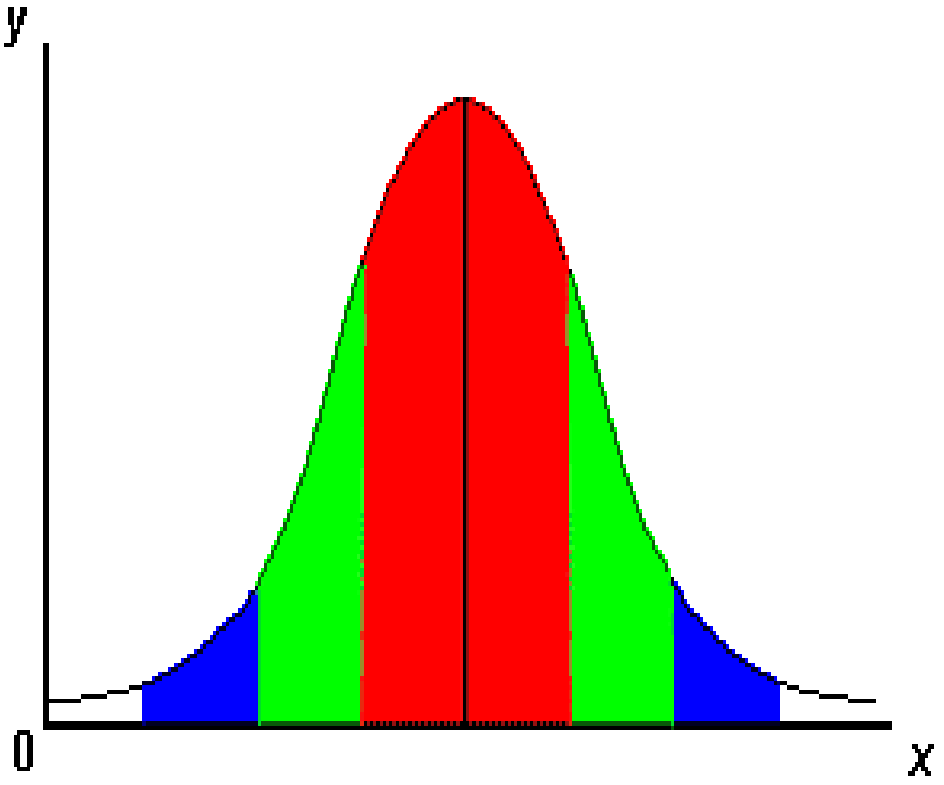
Approximately

68% of the data is between ± 1 s

95% of the data is between ± 2 s

99.7% of the data is between ± 3 s

of the mean



Estimation of s

If the data is mound-shaped, approximately

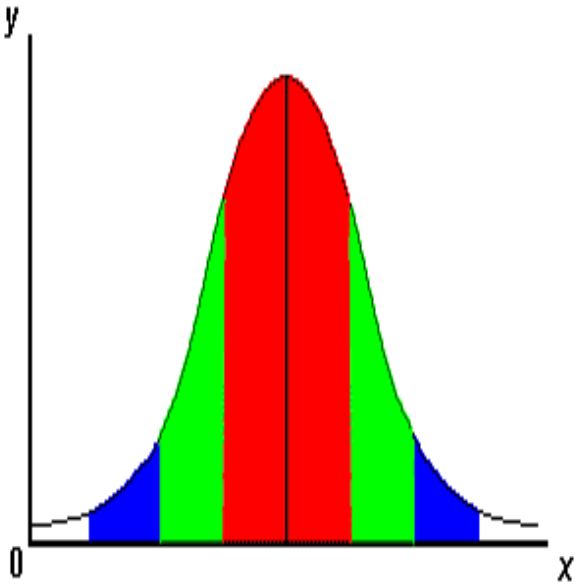
95% of the data lies within 4 s of the mean

100% within 6s of the mean

S ~ range / 4 (sample)

S ~ range / 6 (census)

Therefore, if we know the approximate range of the data, one can estimate S using Range/4 or Range/6 (more conservative)



Population *Z* – score: $z = \frac{x - \mu}{\sigma}$

Sample *Z* – score: $z = \frac{x - \bar{x}}{s}$

Empirical Rule can be expressed based on z-score:

If the *distribution is mounded-shaped, approximately 95% of the data having z-score between –2 and 2., and almost 100% of data having z-scores between –3 and 3.*

One can use z-score to identify potential rare events similar to Empirical Rule:

If the distribution is mounded-shaped, there is only 5% of chance to have z-score outside [-2,2], Therefore, it is a possible outlier.

If the distribution is mounded-shaped, there is less than 1% of chance to have z-score outside [-3,3], Therefore, it is probably an outlier.

Interquartile Range

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Interquartile Range (IQR)

The Five-Number Summary

MINIMUM Q_1 *Median* Q_3 MAXIMUM

IQR (Inter-quartile Range) = $Q_3 - Q_1$

Coefficient of Variation

The coefficient of variation indicates how large the standard deviation is in relation to the mean.

The coefficient of variation is computed as follows:

$$\left(\frac{s}{\bar{x}} \times 100\right)\%$$

for a sample

$$\left(\frac{\sigma}{\mu} \times 100\right)\%$$

for a population