Exp-6: Study and Computation of descripive statistics in Python using Pandas, NumPy and SciPy.

Objectives:

> 1. Understand the basics of statistics.
> 2. Understand, compute and demonstrate the various **statistical measures**.

Important Shortcut Keys

> - A -> To **create** cell **above**
> - B -> To **create** Cell **below**
> - D D -> For **deleting** the cell
> - M -> To **markdown** the Cell
> - Y -> For **code** the cell
> - Z -> To **undo** the deleted cell

1. Basics of Statistics

1.1 Statistics

- Statistics is a science dealing with the **Collection**, **Analysis**, **Interpretation**, and **Presentation** of numerical data.
- Study of Statistics can be done in various ways. One way is to subdivide statistics into two branches namely:
    1. Descriptive Statistics and
    2. Inferential Statistics

1.2 Population versus Sample

- To understand the difference between Descriptive Statistics and Inferential Statistics, definitions of **Population** and **Sample** are helpful.
- **Population:** a collection of persons, objects, or items of interest.

    - Examples:
        - “all workers presently employed by Microsoft"
        - "all automobiles"
        - "all Tata Motor cars produced from 2019 to 2021"
        - “all washing machines produced on February 3, 2019, by the LG Company at Pune plant"
    - Census: When researchers gather data from the whole population for a given measurement of interest, they call it a **census**. Most people are familiar with U.S. census. Every 10 years, the government attempts to count all persons living in this country.

- **Sample:** a portion of whole and, if properly taken, is representative of the whole.
    - For various reasons (to be discussed in experiment 8), researchers often prefer to work with a sample of the population instead of the entire population (census).
    - For example, in conducting quality-control experiments to determine the average life of lightbulbs, a lightbulb manufacturer might randomly sample only 75 lightbulbs during a production run.

1.3 Descriptive Statistics versus Inferential Statistics

- **Descriptive Statistics:** If a researcher *gathers data on a group to describe or reach conclusions about that same group*, the statistics are called *descriptive statistics*.
- **Inferential Statistics:** If a researcher *gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken*, the statistics are *inferential statistics* (*inductive statistics*).

1.4 Parameter versus Statistic

- **Parameter:** *descriptive measure of the population is called as parameter*.
    - Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), population variance ($σ^2$), and population standard deviation (σ).
- **Statistic:** *descriptive measure of a sample is called as statistic*.
    - Statistics are usually denoted by Roman letters. Examples of statistics are sample mean ($\bar{x}$), sample variance (s^2), and sample standard deviation (s).
- Differentiation between the terms parameter and statistic is important only in the use of inferential statistics.

    - An analyst often wants to estimate the value of a parameter or conduct tests about the parameter.
    - However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to take a census.
    - In such cases, the analyst can take a random sample of the population, calculate a statistic on the sample, and infer by estimation the value of the parameter.
    - The basis for inferential statistics, then, is the ability to make decisions about parameters without having to complete a census of the population.
- For example, a manufacturer of washing machines would probably want to determine the average number of loads that a new machine can wash before it needs repairs.
    - The parameter is the population mean or average number of washes per machine before repair.
    - A company researcher takes a sample of machines, computes the number of washes before repair for each machine, averages the numbers, and estimates the population value or parameter by using the statistic, which in this case is the sample average. Figure 1.2 demonstrates the inferential process.

1.5 Need of Probability in Inferential Statistics

- Inferences about parameters are made under uncertainty. Unless parameters are computed directly from the population, the statistician never knows with certainty whether the estimates or inferences made from samples are true. In an effort to estimate the level of confidence in the result of the process, statisticians use probability statements.

1.6 Variable, Measurement, and Data in Statistics

- **Variable:** *variable is a characteristic of any entity being studied that is capable of taking on different values.*
    - Examples: "total sales", "time spent in shopping store", "labor productivity", "customer satisfaction" etc.
- **Measurement:** *a measurement is taken when a standard process is used to assign numbers to particular attributes or characteristics of a variable.*
    - Example: Some measurements are obvious like: "total sales", "time spent in shopping store". But on the other hand some measurements need to be defined carefully like: "labor productivity", "customer satisfaction".
- **Data:** *data are recorded measurements.*

1.7 Data Measurement

- Immense volume of data are gathered by businesses every day, representing myriad items.
- For example, **numbers** represent **dollar costs** of items produced, **geographical locations** of retail outlets, **weights** of shipments, and **rankings** of subordinates at yearly reviews.
- All such data should not be analyzed the same way statistically because the entities represented by the numbers are different.
- For this reason, the analyst needs to know the level of data measurement represented by the numbers being analyzed.
- Lets think about two numbers 40 and 80. They could represent
    - the weights of two objects being shipped,
    - the ratings received on a consumer test by two different products, or
    - Cricket jersey numbers of a wicket-keeper and a fine-leg position player.
    - **Think:** Although 80 kg is twice as much as 40 kg, the wicket-keeper is probably not twice as big as the fine-leg position player!
    - **Think:** Averaging the two weights seems reasonable, but averaging the football jersey numbers makes no sense.
- Thus appropriateness of the data analysis depends on the level of measurement of the data gathered.
- The phenomenon represented by the numbers determines the level of data measurement.

Four common levels of data measurement follow:

1. Nominal (Lowest level of data measurement)
2. Ordinal
3. Interval and
4. Ratio (Highest level of data measurement)

1.7.1 Nominal Level

- Lowest level of data measurement
- used only to **classify** and **categorize**
- Examples:
    - Employee identification numbers
    - Which of the following employment classifications best describes your area of work?
        1. Educator
        2. Construction worker
        3. Manufacturing worker
        4. Lawyer

5. Doctor
6. Other

- sex, religion, ethnicity, geographic, location, place of birth, telephone numbers, adhar card number and Zip codes etc.

- To analyze nominal data, statistical techniques are limited.
- chi-square statistic can be applied.
- **nonmetric data or qualitative data**.

1.7.2 Ordinal Level

- Higher level than nominal level measurement.
- In addition to the nominal-level capabilities, ordinal-level measurement can be used to **rank** or **order** people or objects.
- Examples:
  - using ordinal data, a supervisor can evaluate three employees by ranking their productivity with the numbers 1 through 3.
    - The supervisor could identify one employee as the most productive, one as the least productive, and one as somewhere in between by using ordinal data.
    - However, the supervisor could not use ordinal data to establish that the intervals between the employees ranked 1 and 2 and between the employees ranked 2 and 3 are equal; that is, she could not say that the differences in the amount of productivity between workers ranked 1, 2, and 3 are necessarily the same.
  - Some questionnaire Likert-type scales are considered by many researchers to be ordinal in level. See the following question:

This tutorial is helpful:					
		not at all	somewhat	moderately	very much	extremely	
	1	2	3	4	5

- Certain statistical techniques are specifically suited to ordinal data, but many other techniques are not appropriate for use on ordinal data.
  - For example, it does not make sense to say that the average of “moderately helpful” and “very helpful” is “moderately helpful and a half.”
- **nonmetric data or qualitative data**.

1.7.3 Interval Level

- Next to the highest level of data.

- In addition to the ordinal-level capabilities, **distances between consecutive numbers have meaning and the data are always numerical**.
- The distances represented by the differences between consecutive numbers are equal; that is, interval data have equal intervals.
- Examples:
    - Fahrenheit temperature (Celsius temperature)
        - temperature can be ranked.
        - the amounts of heat between readings, such as 20°F, 21°F and 22°F are same.
        - Here ratio doesn't work: 40° F is not twice as hot as 20° F.
        - Here it will work: an increase of 40° F is twice as much as an increase of 20° F.
- **The zero point is a matter of convention or convenience and not a natural or fixed zero point**.
- **Zero is just another point on the scale and does not mean the absence of the phenomenon**.
    - For example, zero degrees Fahrenheit is not the lowest possible temperature.
- **metric data or quantitative data**.

1.7.4 Ratio Level

- Highest level of data-measurement.
- In addition to the interval-level capabilities, **ratio data have an absolute zero, and the ratio of two numbers is meaningful**.
- The notion of absolute zero means that zero is fixed, and the zero value in the data represents the absence of the characteristic being studied.
- The value of zero cannot be arbitrarily assigned because it represents a fixed point. This definition enables the statistician to create ratios with the data.
- Examples:
    - Kelvin temperature
        - temperature can be ranked.
        - the amounts of heat between readings, such as 20K, 21K and 22K are same.
        - 0 K is the lowest possible temperature. So, ratio will work.
        - 40 kelvin is twice as hot as 20 kelvin.
    - Weight
    - Height

- Time
- Volume

- **metric data or quantitative data**.

1.8 Comparision of the Four Levels of Data

Contents	Nominal	Ordinal	Interval	Ratio
Classification, categorization	[Yes]	[Yes]	[Yes]	[Yes]
Ranking, ordering	[No]	[Yes]	[Yes]	[Yes]
distance bewteen two numbers has meaning	[No]	[No]	[Yes]	[Yes]
ratio bewteen two numbers has meaning	[No]	[No]	[No]	[Yes]

- Nominal data are the most limited data in terms of the types of statistical analysis that can be used with them.
- Ordinal data allow the researcher to perform any analysis that can be done with nominal data and some additional analyses.
- With ratio data, a statistician can make ratio comparisons and appropriately do any analysis that can be performed on nominal, ordinal, or interval data. Some statistical techniques require ratio data and cannot be used to analyze other levels of data.
- Statistical techniques can be separated into two categories: parametric statistics and nonparametric statistics.
  - **Parametric statistics:** data must be either interval or ratio level. [metric data only]
  - **Nonparametric statistics:** any data levels.

Question #1: Level of Data Measurement

Because of increased competition for patients among providers and the need to determine how providers can better serve their clientele, hospital administrators sometimes administer a quality satisfaction survey to their patients after the patient is released. The following types of questions are sometimes asked on such a survey. These questions will result in what level of data measurement?

1. How long ago were you released from the hospital?
2. Which type of unit were you in for most of your stay?
   _____ Coronary care
   _____ Intensive care
   _____ Maternity care
   _____ Medical unit

_____ Pediatric/children's unit
_____ Surgical unit

3. In choosing a hospital, how important was the hospital's location?
   _____ Very Important
   _____ Somewhat Important
   _____ Not Very Important
   _____ Not at All Important
4. What was your body temperature when you were admitted to the hospital?
5. Rate the skill of your doctor:
   _____ Excellent
   _____ Very Good
   _____ Good
   _____ Fair
   _____ Poor

Answer #1: Write your answer in a Page and insert here as an image....

2. Statistical Measures

Experiment 3 and 4 presented graphical techniques for organizing and visualizing data. Even though such graphical techniques allow the business analyst to make some general observations about the shape and spread of the data, a more complete understanding of the data can be attained by summarizing the data using statistics. Here we will present such statistical measures, including measures of central tendency, measures of position, measures of variability, and measures of shape.

2.1 Measures of Central Tendency

- Measures of central tendency yield information about the center, or middle part, of a group of numbers.
- Measure of central tendency presented here are:
  1. Mode
  2. Arithmetic mean

3. Median
4. Percentiles
5. Quartiles

2.1.1 Mode

The mode is the **most frequently occurring value** in a set of data.

- Bimodal: Datasets with two modes.
- Multimodal: Data sets with more than two modes.
- The concept of mode is often used in determining sizes.
- Example:
    - Manufacturers who produce cheap rubber flip-flops that are sold for as little as Rs 100.00 around the world might produce them in only one size in order to save on machine setup costs.
    - In determining the one size to produce, the manufacturer would most likely produce flip-flops in the modal size.
- The mode is an appropriate measure of central tendency for nominal-level data.

Mode of the Grouped Data

- $mode = L + \frac{d_1}{d_1+d_2}(W)$
- where,
    - $L =$ lower limit of the mode class
    - $d_1 =$ frequency difefence between the mode class and preceding class
    - $d_2 =$ frequency difefence between the next class and mode class
    - $f_{med} =$ frequency of median class
    - $W =$ width of mode class

Example: Mode of the Grouped Data

Let the dataset is as follows:

Class Intervals	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

Find the median of above grouped data.

Solution:

- Step-1: Find the mode class:
  - Mode class: Search for class interval whose frequency is maximum
    - In this case: mode class is **30-under 40** whose frequency is maximum.
    - So $L = 30, d_1 = 18 - 6 = 12, d_2 = 18 - 11$ and $W = 10$
- Step-2: Calculate mode
  - $mode = L + \frac{d_1}{d_1 + d_2}(W) = 30 + \frac{12}{12+19}(10) = 36.315$

2.1.2 Median

The median is the **middle value** in an ordered array of numbers.

- For an array with an odd number of terms, the median is the middle number.
- For an array with an even number of terms, the median is the average of the two middle numbers.
- The following steps are used to determine the median:
  - Step 1: Arrange the observations in an ordered data array. (Let the size of array is **n**)
  - Step 2: For an odd number of terms, find the middle term of the ordered array. That is **${(\frac{n+1}{2})}^{th}$** term is the median.
  - Step 3: For an even number of terms, find the average of the middle two terms. That is the average of **${(\frac{n}{2})}^{th}$** and **${(\frac{n+2}{2})}^{th}$** is the median.
- Advantage of median: **The median is unaffected by the magnitude of extreme values**. This characteristic is an advantage, because large and small values do not inordinately influence the median.
- Disadvantage of median: not all the information from the numbers is used to calcualte median.
- The level of data measurement must be at least ordinal for a median to be meaningful.
- Best estimate in least absolute deviation criterian.

Median of the Grouped Data

- $median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}$(W)
- where,
  - $L =$ lower limit of the median class

- $cf_p =$ cumulative frequency of class preceding the median class
- $f_{med} =$ frequency of median class
- $W =$ width of median class
- $N =$ total of frequency

Example: Median of the Grouped Data

Let the dataset is as follows:

Class Intervals	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

Find the median of above grouped data.

Solution:

- Step-1 Find the cumulative frequencies.

Class Intervals	Frequency(f)	Cumulative frequency (cf_p)
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	1	50
---	---	---
Total	50	Not required
---	---	---

- Step-2: Find the median class:
  - $\frac{N}{2} = \frac{50}{2} = 25$
  - Search for class interval in which 25 is less than or equal to cumulative frequency.
    - In this case: class for value 25 is **40-under 50**
  - So median class is **40-under 50**
    - So $L = 40$, $f_{med} = 11$, $cf_p = 24$ and $W = 10$
- Step-3: Calculate median

- $median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W) = 40 + \frac{\frac{50}{2} - 24}{11}(10) = 40.909$

2.1.3 Mean

The arithmetic mean is the **average of a group of numbers** and is computed by summing all numbers and dividing by the number of numbers.

- Formula for Mean Calculation:
  - Population Mean: $\mu = \frac{\sum x_i}{N}$, where N is the population size
  - Sample Mean: $\bar{x} = \frac{\sum x_i}{n}$, where n is the sample size
- Advantage of mean: The mean is affected by each and every value, which is an advantage.
- Disadvantage of mean:The mean uses all the data, and each data item influences the mean. It is also a disadvantage because extremely large or small values (outliers) can cause the mean to be pulled toward the extreme value.
- The level of data measurement must be at least interval for a mean to be meaningful.
- Best estimate in least square criterian.
- Unbiased estimate of population means i.e. $E[\bar{x}] = \mu$

Mean of the Grouped Data

- Let there are N data points grouped into K clasess with frequencies $f_1 + f_2 + \ldots + f_K$.
  - So, $f_1 + f_2 + \ldots + f_K = N$
- Then mean is defiend as a weighted average of class midpoints (M).
- Here, Class frequencies (f) are the weights.
- $\mu = \frac{\sum fM}{\sum f} = \frac{f_1M_1 + f_2M_2 + \ldots + f_KM_K}{f_1 + f_2 + \ldots + f_K} = \frac{f_1M_1 + f_2M_2 + \ldots + f_KM_K}{N}$

Example: Mean of the Grouped Data

Let the dataset is as follows:

Class Intervals	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3

Class Intervals	Frequency
70-under 80	1

Find the mean of above grouped data.

Solution:

Class Intervals	Frequency(f)	Class Midpoints(M)	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
---	---	---	---
Total	50	Not required	2150
---	---	---	---

So the mean is: $\mu = \frac{\sum fM}{} = \frac{2150}{} = 43.0$

Question #2: Mode, Median and Mean

Shown below is a list of the top 13 shopping centers in the United Kingdom by retail size in 1000 square meters (m2). Calculate the mode, median, and mean for these data and write the python code for the same.

Shopping Center	Size (1000 m^2)
MetroCentre	190.0
Trafford Centre	180.9
Westfield Stratford City	175.0
Bluewater	155.7
Liverpool One	154.0
Westfield London	149.5
Merry Hill	148.6
Manchester Arndale	139.4
Meadowhall	139.4
Lakeside	133.8
St. David's	130.1
Bullring	127.1
Eldon Square	125.4

Answer #7(a):

Write your answer in a Page and insert here as an image....

Answer #7(b):

Write your Python code below:

```
# Write your python code to calculate mode, median and mean.Then press Shift+Enter to execute
arr = [190, 180.9, 175, 155.7, 154, 149.5, 148.6, 139.4, 139.4, 133.8, 130.1, 127.1, 125.4]

import statistics
import numpy as np
from scipy import stats
mode = statistics.mode(arr)
median = statistics.median(arr)
mean = statistics.mean(arr)

print('Mode: ', mode)
print('Median: ', median)
print('Mean: ', mean)
```

```
Mode:  139.4
Median:  148.6
Mean:  149.91538461538462
```

▶ Click here for the solution

2.2 Measures of Position

- While measures of central tendency are important, they do not tell the whole story.
    - For example, suppose the mean score on a statistics exam is 80%. From this information, can we determine a range in which most people scored? The answer is no. There are two other types of measures, measures of position and variability, that help paint a more concise picture of what is going on in the data. In this section, we will consider the measures of position and discuss measures of variability in the next one.
- A measure of position determines the position of a single value in relation to other values in a sample or a population data set.
- Measures of position give a range where a certain percentage of the data fall.
- These are not sensitive to the influence of a few extreme observations.
- Measure of Position presented here are:
    1. Percentiles
    2. Quartiles

2.2.1 Percentiles

Percentiles are measures that **divide a group of data** into **100 parts**. 90th percentiles indicates that at least 90% of the data lie below it, and at most 10% of the data are above it.

- The nth percentile is the value such that at least n percent of the data are below that value and at most (100 − n) percent are above that value.
- The level of data measurement must be at least oridnal for a percentile to be meaningful.
- median and 50th percentile are the same value.
- The following steps are used to determine the percentile:
    - Step 1: Arrange the observations in an ordered data array. (Let the size of array is **n**)
    - Step 2: Calculate the pth percentile location i by: $i=\frac{p}{100}(n-1)$. (indexing starts from 0)
    - Step 3:
        - (a) if i is a **whole number**, pth percentile is the value at the i position.
        - (b) if i is **not a whole number**, pth percentile is given by: $V_i + (V_j - V_i) \times fraction$
            - where, V_i = value at $\lfloor{i}\rfloor$ position, V_j = value at $\lceil{i}\rceil$ position, and fraction is the fractional part of i.

2.2.2 Quartiles

Quartiles are measures that **divide a group of data** into **4 parts**.

- These three measures are denoted first quartile (denoted by Q1), second quartile (denoted by Q2), and third quartile (denoted by Q3).
- Q2 is the same as the median of a data set.
    - Q2 = 50th percentile (median)
- Q1 is the value of the middle term among the observations that are less than the median
    - Q1 = 25th percentile (lower quartile)
- Q3 is the value of the middle term among the observations that are greater than the median
    - Q3 = 75th percentile (upper quartile)

Question #3: Percentile and Quartile

Find the 25^{th}, 30^{th}, 50^{th} and 75^{th} percentile of the following array and verify by writing the python code.

```
arr = [5, 12, 13, 14, 17, 19]
```

Answer #3(a):

Write your answer in a Page and insert here as an image....

Answer #3(b):

Write your Python code below....

```
# Write your python code to calculate required percentile values.Then press Shift+Enter to ex
arr = [5, 12, 13, 14, 17, 19]

import numpy as np

print("arr", arr)

print("25th percentile: ", np.percentile(arr, 25))

print("30th percentile: ", np.percentile(arr, 30))

print("50th percentile: ", np.percentile(arr, 50))

print("75th percentile: ", np.percentile(arr, 75))
```

```
arr [5, 12, 13, 14, 17, 19]
25th percentile:  12.25
30th percentile:  12.5
50th percentile:  13.5
75th percentile:  16.25
```

▶ Click here for the solution

2.3 Measures of Variability

- measures of variability are used to describe the spread or the dispersion of a set of data.
- Using measures of variability in conjunction with measures of central tendency and position makes possible a more complete numerical description of the data.
- To introduce the idea of variability, consider this example. Two vending machines A and B drop candies when a coin is inserted. The number of pieces of candy one gets is random. The following data are recorded for six trials at each vending machine:

    - **Vending machine A:**
        - Pieces of candy from this machine: 1, 2, 3, 3, 5, 4
        - one can calculate: **mean** = 3, **median** = 3, **mode** = 3
    - **Vending machine B:**
        - Pieces of candy from this machine: 2, 3, 3, 3, 3, 4
        - one can calculate: **mean** = 3, **median** = 3, **mode** = 3
    - They have the same centers but what about the spreads?
- There are many ways to describe variability or spread including:
    1. Range
    2. Interquartile range (IQR)
    3. Variance and Standard Deviation
    4. z- scores
    5. Coefficient of variation

2.3.1 Range

The range is the difference between the largest value of a data set and the smallest value of a set.

$$Range = maximum - minimum$$

- Although it is usually a single numeric value, some business analysts define the range of data as the ordered pair of smallest and largest numbers (smallest, largest).
- It is a crude measure of variability.
- The range is easy to compute but it is affected by extreme values.
- One important use of the range is in quality assurance, where the range is used to construct control charts.

2.3.2 Interquartile Range

The interquartile range is the difference between upper and lower quartiles and denoted as IQR.

$$IQR = Q3 - Q1$$

- It is the range of the middle 50% of the data.
- It is not affected by extreme values.
- Example: In describing a real estate housing market, realtors might use the interquartile range as a measure of housing prices when describing the middle half of the market for buyers who are interested in houses in the midrange.
- It is also used in the construction of box-and-whisker plots.

2.3.3 Variance and Standard Deviation

The variance is the **average of the squared deviations about the arithmetic mean** for a set of numbers.

The Standard Deviation is the **square root of the variance**.

- Formula for Variance Calculation:
    - Population Variance: $\sigma^2 = \frac{\sum{{x_i - \mu}^2}}{N}$, where N is the population size
    - Population Standard Deviation: $\sigma = \sqrt{\frac{\sum{{x_i - \mu}^2}}{N}}$, where N is the population size
    - Sample Variance: $s^2= \frac{\sum{{x_i - \bar{x}}^2}}{n-1}$, where n is the sample size
    - Sample Standard Deviation: $s= \sqrt{\frac{\sum{{x_i - \bar{x}}^2}}{n-1}}$, where n is the sample size
        - Why do we divide by $n-1$ instead of by n?: The main use for sample variances and standard deviations is as estimators of population variances and standard deviations. Dividing by (n-1) rather than n gives it a special property that we call an "unbiased estimator". Therefore s^2 is an **unbiased estimator** for the population variance.
- These measures capture the spreadness. More the variance, more is the spreadness and vice-versa.
- These measures are also used as a part of other analyses, such as computing confidence intervals and in hypothesis testing.

Meaning of Standard Deviation

What is a standard deviation? What does it do, and what does it mean? Insight into the concept of standard deviation can be gleaned by viewing the manner in which it is applied.

Two ways of applying the standard deviation are:

1. **Empirical Rule:** The empirical rule is an important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are **normally distributed**. The empirical rule is used for only three numbers of standard deviations: 1σ, 2σ, and 3σ. **Distance from the Mean | Values within Distance** ----------|---------- $\mu \pm 1\sigma$ | 68 $\mu \pm 2\sigma$ | 95 $\mu \pm 3\sigma$ | 99.7

2. **Chebyshev's Theorem:** It states that at least $1 - 1/k^2$ proportion of values will fall within $\pm k$ standard deviations of the mean regardless of the shape of the distribution as long as k is greater than 1.
    - The empirical rule applies only when data are known to be approximately normally distributed.
    - Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is nonnormal.
    - Even though Chebyshev's theorem can in theory be applied to data that are normally distributed, the empirical rule is more widely known and is preferred whenever appropriate.
    - Chebyshev's theorem is not a rule of thumb, as is the empirical rule, but rather it is presented in formula format and therefore can be more widely applied.

2.3.4 z- Scores

A z score represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed. Using z scores allows translation of a value's raw distance from the mean into units of standard deviations.

- Formula for z- Scores calculation:
    - Population z- Score: $z = \frac{x_i - \mu}{\sigma}$
    - Sampel z- Score: $z = \frac{x_i - \bar{x}}{s}$
- Physical significance: If a z score is negative, the raw value (x) is below the mean. If the z score is positive, the raw value (x) is above the mean.
- Emperical Rule in terms of z- Score: **z-score value| Values within Distance** ----------|---------- $-1\leq z\leq 1$| 68% $-2\leq z\leq 2$| 95% $-3\leq z\leq 3$| 99.7%

2.3.5 Coefficient of Variation

The coefficient of variation is a statistic that is the ratio of the standard deviation to the mean expressed in percentage and is denoted CV.

- Formula for CV calculation: $CV = \frac{\sigma}{\mu} \times 100$
- Physical significance: Financial investors use the coefficient of variation or the standard deviation or both as **measures of risk**.
    - Example:
        - Imagine a stock with a price that never changes. An investor bears no risk of losing money from the price going down because no variability occurs in the price.

    - Suppose, in contrast, that the price of the stock fluctuates wildly. An investor who buys at a low price and sells for a high price can make a nice profit. However, if the price drops below what the investor buys it for, the stock owner is subject to a potential loss. The greater the variability is, the more the potential for loss.
    - Here, investors use measures of variability such as standard deviation or coefficient of variation to determine the risk of a stock.
- In many cases, what does the coefficient of variation tell us about the risk of a stock that the standard deviation does not?

Question #4: Range & Interquartile Range

Find the range and interquartile range of the following array and verify by writing the python code.

```
arr = [5, 12, 13, 14, 17, 19]
```

Answer #4(a):

Write your answer in a Page and insert here as an image....

Answer #4(b):

Write your Python code below....

```
# Write your python code here.Then press Shift+Enter to execute
arr = [5, 12, 13, 14, 17, 19]

import numpy as np

max_val = max(arr)
min_val = min(arr)

range_of = max_val - min_val

q1 = np.percentile(arr, 25)
q3 = np.percentile(arr, 75)

iqr = q3 - q1

print("Interquartile range: ", iqr)
```

```
Interquartile range:  4.0
```

▶ Click here for the solution

Question #5: Population and Sample Standard Deviation

Find the Population standard devaiation and Sample standard deviation of the following array and verify by writing the python code.

```
arr = [5, 12, 13, 14, 17, 19]
```

Answer #5(a):

Write your answer in a Page and insert here as an image....

Answer #5(b):

Write your Python code below....

```
# Write your python code here.Then press Shift+Enter to execute
import statistics
import numpy as np

arr1 = [5,12,13,14,17,19]

a = statistics.pstdev(arr1)
b = statistics.stdev(arr1)

print('Population standard deviation: ', a)
print('Sample standard deviation: ', b)
```

```
Population standard deviation:  4.422166387140534
Sample standard deviation:  4.844240566555987
```

▶ Click here for the solution

Question #6: Coefficient of Variation

Suppose for the stock "A" and stock "B", five weeks of average prices are given as follows:

```
Stock "A" Prices: 57, 68, 64, 71, 62

Stock "B" Prices: 12, 17, 08, 15, 13
```

Comment on: which one of these two stocks is more riskier?

Answer #6(a):

Write your answer in a Page and insert here as an image....

Answer #6(b):

Write your Python code below....

```
# Write your python code here.Then press Shift+Enter to execute
import statistics
import numpy as np

a1 = [57,68,64,71,62]
a2 = [12, 17, 8, 15, 13]

a = statistics.pstdev(a1)  # or pop_st_dev = np.stdev(arr)
b = statistics.stdev(a2)

print('Population standard deviation: ', a)
print('Sample standard deviation: ', b)
```

```
Population standard deviation:  4.841487374764082
Sample standard deviation:  3.391164991562634
```

Stock A is riskier as compared to stock B as it has greter value of standard deviation.

▶ Click here for the solution

2.4 Measures of Shapes

Measures of shape are tools that can be used to describe the shape of a distribution of data.

In this section, we examine only one measure of shape, skewness.

In addition, we look at box-and-whisker plots.

2.4.1 Skewness

- **Symmetrical Distribution:** A distribution of data in which the right half is a mirror image of the left half is said to be symmetrical. One example of a symmetrical distribution is the normal distribution, or bell curve, shown in Figure.
- **Skewness:** is when a distribution is **asymmetrical** or lacks symmetry.
- Skewness can be:
    1. Left skewed (negatively skewed)
    2. Right skewed (positively skewed) symmtrical-2.jpg

Skewness and the Relationship of the Mean, Median, and Mode

The concept of skewness helps with an understanding of the relationship of the mean, median, and mode.

- In a unimodal distribution (distribution with a single peak or mode) that is skewed,
    - the mode is the apex (high point) of the curve
    - the median is the middle value and
    - The mean tends to be located toward the tail of the distribution

Coefficient of Skewness

Formula to calculate skewness: $skew = \frac{\mu - median}{\sigma}$

- If $skew < 0$, the distribution is **negatively skewed**.
- If $skew = 0$, the distribution is **symmetric**.
- If $skew < 0$, the distribution is **positively skewed**.

2.4.2 Box-and-Whisker Plots and Five-Number Summary

- A **box-and-whisker plot**, sometimes called a box plot, is a diagram that utilizes the **upper** and **lower quartiles** along with the **median** and the **two most extreme values** to depict a distribution graphically.
- The box-and-whisker plot is determined from **five specific numbers** sometimes referred to as the **five-number summary**.
    1. Median (Q2)
    2. Lower Quartile (Q1)
    3. Upper Quartile (Q3)
    4. Smallest Value

5. Largest Value

- A Typical Box Plot is shown in Figure below. (In general, its vertically alligned). box.png
- Box: It encolses Q1, median and Q3
- Hinges: Box endpoints (Q1 and Q3)
- Whisker: Lines extended out from the lower (Q1) and upper quartile (Q3) are whiskers.
- Inner fences: Inner fences are established as follows:

  $(Q1 - 1.5 \times IQR)$ to $Q1$

  $Q3$ to $(Q3 + 1.5 \times IQR)$
- Outer fences: Outer fences are established as follows:

  $(Q1 - 3 \times IQR)$ to $Q1$

  $Q3$ to $(Q3 + 3 \times IQR)$
- Outliers: Data values outside the mainstream of values in a distribution are viewed as outliers.
  - Outliers can be merely the more extreme values of a data set.
  - However, sometimes outliers occur due to measurement or recording errors. Other times they are values so unlike the other values that they should not be considered in the same analysis as the rest of the distribution.
  - Mild Outliers: Values in the data distribution that are outside the inner fences but within the outer fences are referred to as mild outliers. Mild outliers are marked with a circle (O) on the boxplot.
  - Extreme Outliers: Values that are outside the outer fences are called extreme outliers. Extreme outliers are marked with an asterisk (*) on the boxplot.
- Thus one of the main uses of a box-and-whisker plot is to identify outliers.
- The other use of box-and-whisker plot is to determine whether a distribution is skewed.
  - The location of the median in the box can relate information about the skewness of the middle 50% of the data.
    - If the median is located on the right side of the box, then the middle 50% are skewed to the left.
    - If the median is located on the left side of the box, then the middle 50% are skewed to the right.
  - By examining the length of the whiskers on each side of the box, a business analyst can make a judgment about the skewness of the outer values.

- If the longest whisker is to the right of the box, then the outer data are skewed to the right.
- If the longest whisker is to the left of the box, then the outer data are skewed to the left.

Question #7: Skewness

Shown below is a list of the top 13 shopping centers in the United Kingdom by retail size in 1000 square meters (m2). Calculate the coefficient of skewness for these data and write the python code for the same.

Shopping Center	Size (1000 m^2)
MetroCentre	190.0
Trafford Centre	180.9
Westfield Stratford City	175.0
Bluewater	155.7
Liverpool One	154.0
Westfield London	149.5
Merry Hill	148.6
Manchester Arndale	139.4
Meadowhall	139.4
Lakeside	133.8
St. David's	130.1
Bullring	127.1
Eldon Square	125.4

Answer #7(a):

Write your answer in a Page and insert here as an image....

Answer #7(b):

Write your Python code below:

```
# Write your python code here.Then press Shift+Enter to execute
from scipy.stats import skew
import numpy as np

arr = np.array([190,189.9,175,155.7,154.0,149.5,148.6,139.4,139.4,133.8,130.1,127.1,125.4])
```

```
ans = skew(arr)
print("Skewness: ",ans)
```

```
Skewness:  0.7306430902445546
```

▶ Click here for the solution

1.2 Import Data

First, we assign the URL of the dataset to "filename".

Note: This file does not have column headers, which we need to assign.

✓ 0s completed at 6:21 PM