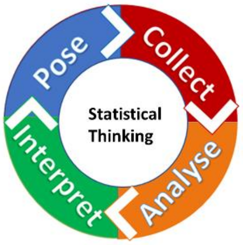
Study Unit 1

Statistical Thinking



Statistical Thinking Outline

- Population data vs sample data
- Measure of central tendency
- Measure of spread
- Measure of partition
- five-number summary in statistics
- Scale of measurement
- Correlation analysis

Study Unit Duration

This Study Unit requires a minimum of 3 hours' formal study time.

You may spend an additional 2-3 hours on revision.

Preamble

Statistical thinking can align one's thoughts with the fundamental principle of statistics to make better decisions under uncertainty. In other words, understanding statistics is important for anyone that wants to make a good decision since it is applicable in every field of human activity. With the understanding of basic statistical methods, you will know when to apply the right tool to a given problem and think statistically. This course will help you understand those statistical concepts and apply them to solve a life problem.

Learning Outcomes of Study Unit 1

Upon completion of this study unit, you should be able to:

1.1 Employ the use of population data, sample data, parameter and statistics analysis of data.

1.2 Use measure of central tendency to compute statistical analysis from an observational study

1.3 Use measure of spread and partition to compute statistical analysis from an observational study

1.4 Categorize data according to a scale of measurement and use Pearson correlation to estimate correlation among variables in the data

Terminologies, Acronyms and their Meaning

μ	Population mean
$\bar{X}$	Sample mean
n	Sample size
N	Population size
$\sum$	sum of
np	NumPy
pd	Pandas

sns	Seaborn
Q1	First quartile
Q3	Third quartile
IQR	Interquartile range
σ^2	Population variance
s^2	Sample variance

Prerequisites

A basic understanding of Python programming in CS14 (Programming in Python) is required.

1.1: Population Data versus Sample Data

1.1.1 Population

A **population** is a complete set. Population data is a collection of **all** elements in the population. For example:

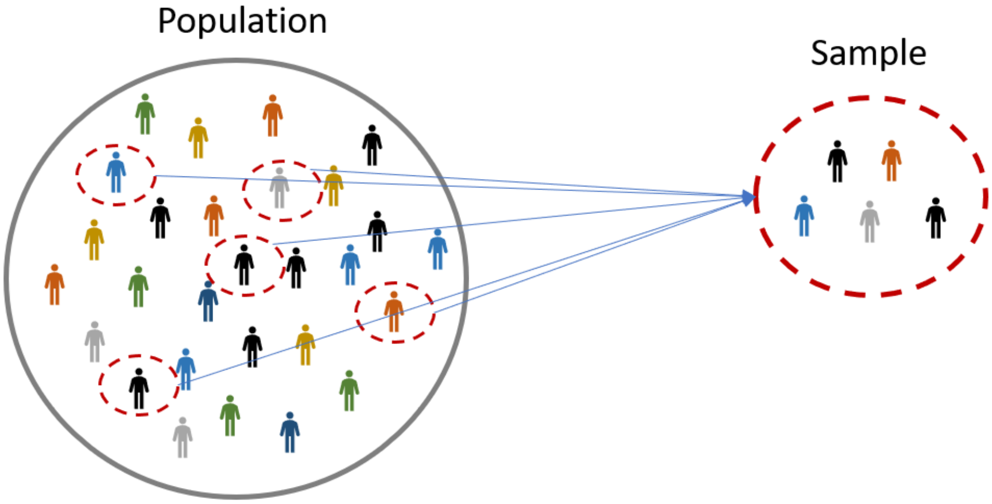
- All the height of graduating students at the Eastern College Somalia
- All the weight of females in Ethiopia
- All the ages of students in South Sudan
- All countries in Africa

If the desired information is available for all items in the population, we have what is referred to as a **census**. In practice, we rarely have a complete set of data. We usually collect data in samples, such as the weight of 10 female students in Ethiopia.

1.1.2 Sample

A **sample** is a subset or fraction of the population. For example:

- 150 fish randomly sampled from Adar River in South Sudan
- 5 best students selected from each University in Sudan
- Taking 10 East African countries out of the 54 African countries



A pictorial difference between population and a sample

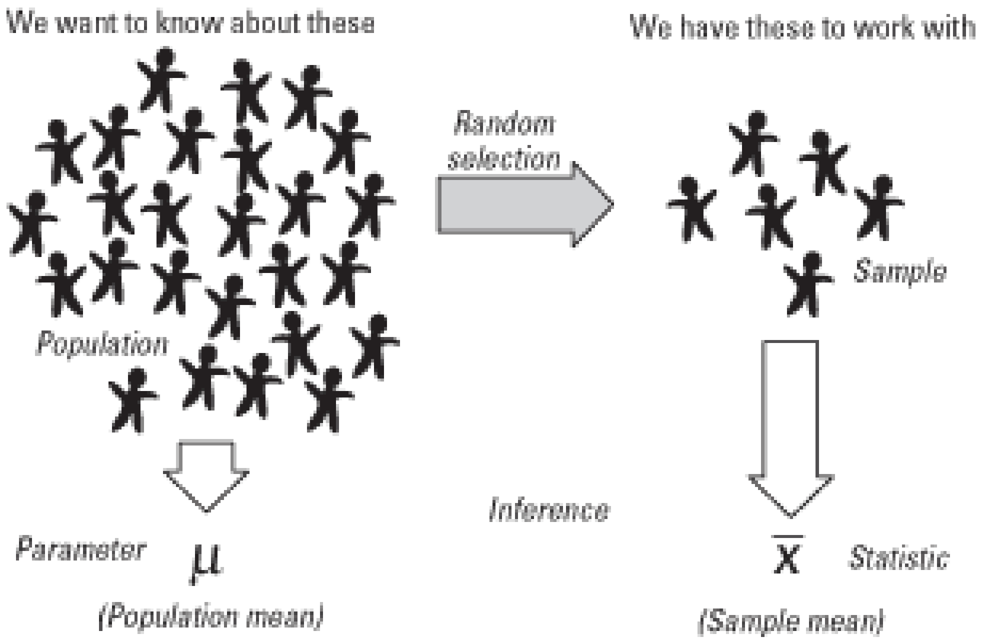
1.1.1 Parameters and Statistics

Parameter

Parameters are the characteristics of a population. It is a descriptive measure for a population and are typically written using Greek letters. The population mean is μ (pronounce as mu). The population variance is σ^2 (sigma squared), population standard deviation is σ (sigma), and population proportion is *P*.

Statistic

Statistic is the characteristics of a sample. It is a descriptive measure for a sample and are typically written using Roman letters. The sample mean is $\bar{X}$(x-bar), the sample variance is s^2, the sample standard deviation is s, and the sample proportion is p. Sample statistics are used to estimate unknown population parameters.



Source

In this section, we will examine descriptive statistics in terms of measures of central tendency, measures of dispersion and measure of partition. These descriptive statistics help us to identify the center and spread of the data. You will get familiar to the formula and how to get the results in Python. Before we start, let's load packages that will be needed in this unit.

Importing packages

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1.2 Measure of central tendency

The most common measures of central tendency are the mean, median, and the mode. Measure of central tendency is also known as measure of location or average.

1.2.1 Mean

The arithmetic mean or simply mean is the sum of all the elements divided by the number of elements.

The sample mean is denoted by $\bar{X}$ while the population mean is denoted as μ.

$$\text{Sample mean} = \bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Population mean} = \mu = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Example 1

The sample ages of 10 students in CS 2 are shown below:

14, 25, 17, 21, 11, 17, 22, 25, 16, and 13

$$\text{Mean} = \frac{14 + 25 + 17 + 21 + 11 + 17 + 22 + 25 + 16 + 13}{10} = \frac{181}{10} = 18.1$$

1.2.2 How to get mean in Python?

We can use **np.mean()** to calculate the mean age. We need to first convert the raw data to Numpy array.

```
import numpy as np

age = np.array([14, 25, 17, 21, 11, 17, 22, 25, 16, 13])

np.mean(age)
```

18.1

or simply use **age.mean()**

```
age.mean()
```

18.1

Example 2

The Professor of Computer Science whose name is Addisu works at Addis Ababa University, Ethiopia. On a weekly basis, he used to give money in Ethiopian Birr (ETB) to the best student in his weekly assignment. The following are the amounts he has given out during CS 2 course:

495, 503, 503, 498, 503, 505, 503, 500, 501, 489, 498, 488, 499, 497, 508, 507, 507, 509, 508, and 503 respectively.

Calculate the mean amount of money he has given out using Python.

Solution

```
money = np.array([495, 503, 503, 498, 503, 505, 503, 500, 501, 489, 498, 488, 499, 497, 508, 507,  507, 509, 508, 503])

money.mean()
```

```
501.2
```

The mean amount of money Professor Addisu has given out is 501.2 ETB.

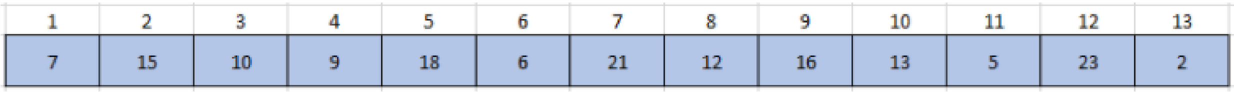
1.2.3 Median

The median is the middle value when the data is arranged in ascending or descending order.

When the number of observations is odd, then the middle number when the observations are arranged is the median.

Consider the following data:

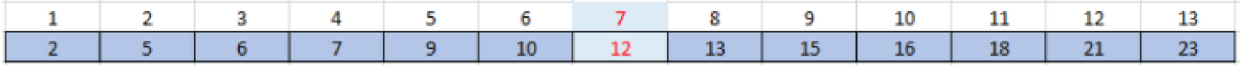
13 students registered for the programming with Python course (CS 2 1) and the performance of the students were shown below:



1	2	3	4	5	6	7	8	9	10	11	12	13
7	15	10	9	18	6	21	12	16	13	5	23	2

You will notice that the data are not sorted (not well arranged). We need to rearrange it in ascending order (from smallest to highest) or descending order (from highest to lowest) and take the middle number as the median because the number of observations is 13 i.e. it is odd.

After arranging in ascending order, we have:



1	2	3	4	5	6	7	8	9	10	11	12	13
2	5	6	7	9	10	12	13	15	16	18	21	23

Therefore, the median is in the 7th position and the value of the median is 12.

1.2.4 How to get median in Python?

We can use np.median() to calculate the median after we first convert the raw data to Numpy array.

```
performance = np.array([7, 15, 10, 9, 18, 6, 21, 12, 16, 13, 5, 23, 2])

np.median(performance)
```

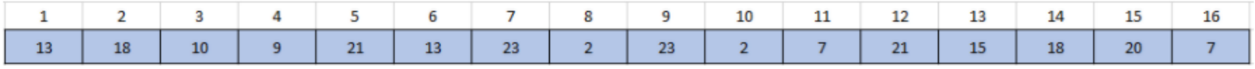
```
12.0
```

Numpy array object has no attribute .median(). So we can never use performance.median()

Example 2

Consider the following data:

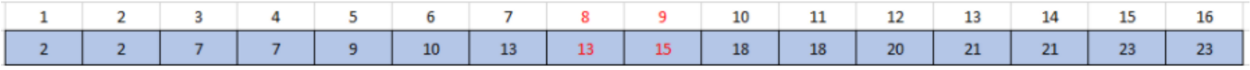
16 students registered for the introduction to data science course (CS 2 2) and the performance of the students were shown below:



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
13	18	10	9	21	13	23	2	23	2	7	21	15	18	20	7

You will notice that the number of observations is 16. So, if the number of observations is even, then the median will lie within the two middle numbers when the numbers are sorted in ascending or descending order of magnitude. In that case, the median is the mean of the two middle numbers.

Let's rearrange in ascending order



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	2	7	7	9	10	13	13	15	18	18	20	21	21	23	23

The median falls between 8th and 9th position which correspond to the value of 13 and 15 respectively. Therefore, the median is the mean of 13 and 15 i.e. $\frac{13+15}{2} = 14$

Let's verify the result in Python

```
performance = np.array([13, 18, 10, 9, 21, 13, 23, 2, 23, 2, 7, 21, 15, 18, 20, 7])
np.median(performance)
```

```
14
```

1.2.5 Mode

The mode is the value that occurs the most frequently in the data set. You can easily get the mode by tabulating your dataset.

Example 1

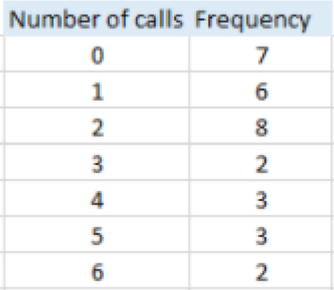
Mr. Jamal a native of Somalia recorded the number of phone calls he received during the month of December 2020. The amount of calls he received for each day were as follows:

0, 2, 6, 2, 2, 0, 0, 1, 1, 5, 3, 1, 0, 2, 3, 1, 2, 1, 4, 4, 5, 0, 5, 1, 2, 2, 2, 0, 4, 0, and 6.

Find the number of calls he received most in December.

Solution

If we count the number of times each call was received, we will have a table like this:



Number of calls	Frequency
0	7
1	6
2	8
3	2
4	3
5	3
6	2

We can read the table as follows:

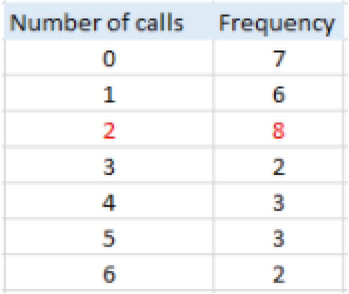
No call (0) appears 7 times

1 call appears 6 times

2 calls appear 8 times

and so on.

Therefore, calls with the highest frequency (number of times it appeared) is the mode.



Number of calls	Frequency
0	7
1	6
2	8
3	2
4	3
5	3
6	2

In this case, 2 is the mode

1.2.6 How to get mode in Python?

We can use scipy module (package) to calculate mode. Let's import it.

```
from scipy import stats
```

We then use **stats.mode()** function to calculate the mode of the raw data when we might have converted it to Numpy array.

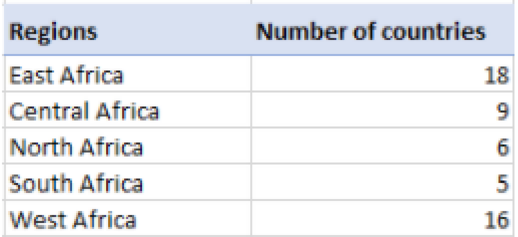
```
calls = np.array([0, 2, 6, 2, 2, 0, 0, 1, 1, 5, 3, 1, 0, 2, 3, 1, 2, 1, 4, 4, 5, 0, 5, 1, 2, 2, 2, 0, 4, 0, 6])

stats.mode(calls)
```

```
ModeResult(mode=array([2]), count=array([8]))
```

As you can see, the mode is 2 and its frequency is 8.

Example 2

According to the United Nations (UN), there are 54 countries in Africa. Each country in Africa is grouped in to regions. So, we have regions like Central Africa, East Africa, North Africa, Southern Africa, and West Africa.



Regions	Number of countries
East Africa	18
Central Africa	9
North Africa	6
South Africa	5
West Africa	16

Which region has the highest number of countries?

Solution

From the table above, we can see that the East Africa region has the highest number of countries (18) is the mode. Therefore, the mode is East Africa.

Class activity 1 (Peer to peer review activity)

Peer to Peer Interaction

Visit the LMS, locate forum activity and participate in the discussion

Thirty farmers were asked how many farm workers they hire during a typical harvest season in Tigray. Their responses were:

4, 5, 6, 5, 1, 2, 8, 0, 4, 6, 7, 8, 4, 6, 7, 9, 8, 6, 7, 5, 5, 4, 2, 1, 9, 3, 3, 4, 6, 4.

Find the mean, median, and the mode of the farm workers hired using Python.

1.3 Measures of spread and Portion

1.3.1 Measurement of Spread

A measure of spread also known as measure of dispersion is used to describe the variability in the data. That is, it tells us how wide the dataset is.

The most common measures of spread are:

- Range

- Standard deviation
- Variance
- Mean absolute deviation

Range

The range is the difference between the highest and lowest values in a data set. It is the simplest measure of spread. The formula for range is:

$Range = maximum\ value - minimum\ value$

The Range tells you how much is in between the lowest value and highest value.

Example

For example, let us consider the following data set:

14, 25, 17, 21, 11, 17, 22, 25, 16, and 13.

Calculate the range is the dataset.

Solution

Range = Maximum value - Minimum value

Maximum value = 25 (The highest value of the dataset)

Minimum value = 11 (The lowest value in the dataset)

Range = 25 - 11 = 14

Therefore, the range of the data is 15.

The limitation of range is that it doesn’t provide a very accurate picture of the variability.

Variance

The variance is the average of the squared differences from the mean. It is a measure of how far each value in the data set is from the mean. A small variance indicates that the data points tend to be very close to the mean and to each other. A high variance indicates that the data points are very spread out from the mean and from each other.

The sample variance is denoted by s^2 while the population variance is denoted by σ^2.

The population variance is:

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

The sample variance is:

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

where,

$\sum$ = sum of …

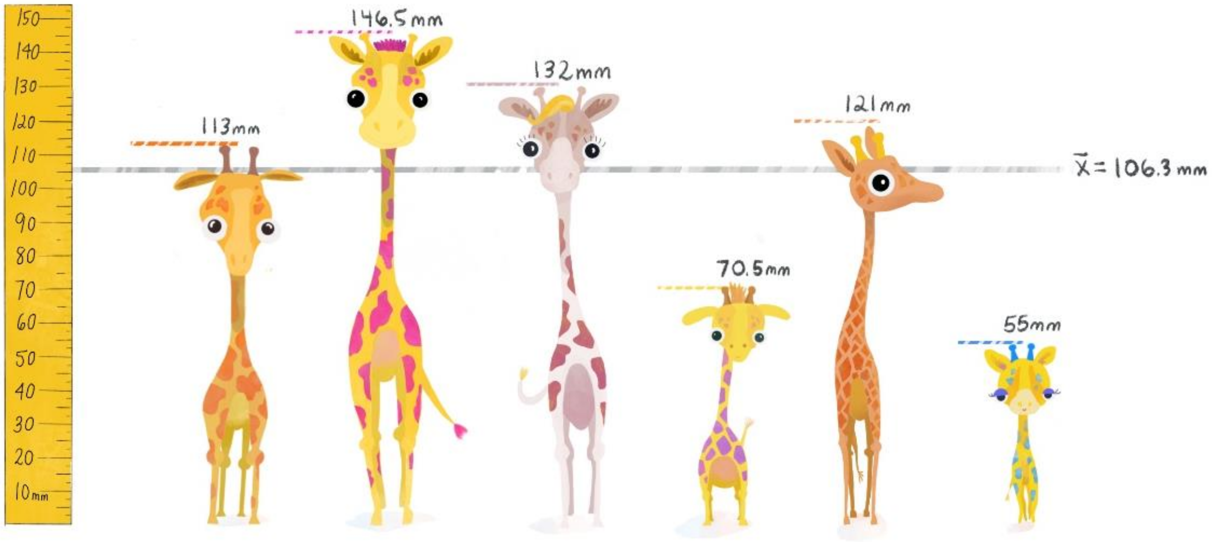
μ = population mean

$\bar{X}$ = sample mean

n = sample size

N = population size

Example 1



Source: https://tinystats.github.io/teacups-giraffes-and-statistics/

The height of 5 sample Giraffes in South Sudan zoo are shown above and their heights are as follows:

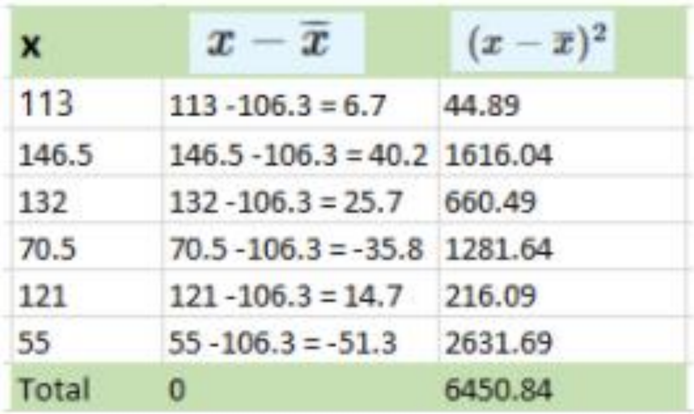
113, 146.5, 132, 70.5, 121, and 55.

Calculate the variance of the height of Giraffes.

Solution

Variable: x = height of a Giraffe

The mean for this data set is $\bar{X} = \frac{113+146.5+132+70.5+121+55.}{6} = 106.3$ mm



x	$x - \bar{x}$	$(x - \bar{x})^2$
113	113 -106.3 = 6.7	44.89
146.5	146.5 -106.3 = 40.2	1616.04
132	132 -106.3 = 25.7	660.49
70.5	70.5 -106.3 = -35.8	1281.64
121	121 -106.3 = 14.7	216.09
55	55 -106.3 = -51.3	2631.69
Total	0	6450.84

The sample variance is

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

$$\sum(x_i - \bar{x})^2 = 6450.84$$

$$n - 1 = 6 - 1 = 5$$

$$s^2 = \frac{6450.84}{5} = 1290.1$$

Therefore, the variance of the height of Giraffes is 1290.1

How to get variance in Python?

With numpy.var() we can get the variance of an array.

```
heights = np.array([113, 146.5, 132, 70.5, 121, 55])

np.var(heights)
```

```
1075.1389
```

We will notice that **np.var()** gave us a value that is difference from what we calculated in the formula. It is so because **np.var()** can only calculate the population variance.

However, we see that it does not have the same units of heights. It is strange. Therefore, we introducing concept of standard deviation

---

One problem with the variance is that it does not have the same unit of measure as the original data. For example, original data containing heights measured in millimeter (mm) has a variance measured in square millimeter (mm^2).

Standard deviation

The **standard deviation** is square root of variance

The population standard deviation is

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

The sample standard deviation is

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

Example

In example 1, we got the variance of the height of Giraffes to be 1290.1

Therefore, the standard deviation is $\sqrt{1290.1}$ = 35.92

The standard deviation of the sample height of Giraffes in South Sudan is 35.92 mm

How to get standard deviation in Python?

With **numpy.std( )** we can get the variance of an array.

```
heights = np.array([113, 146.5, 132, 70.5, 121, 55])

np.std(heights)
```

```
32.7893
```

1.3.2 Measure of partition

Measures of partition are measures that divide a distribution into specified number of parts.

The following are the most common measure of partition:

- Quartiles
- Percentile

Quartile

This measure divides frequency distribution into four equal parts, Q1 (first/lower quartile), Q2 (second quartile), Q3 (third/upper quartile) and Q4.

If the data set consist of n items and we arranged then in ascending order then

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{ item,}$$

$$Q_2 = \left(\frac{n+1}{2}\right)^{th} \text{ item and}$$

$$Q_3 = 3\left(\frac{n+1}{4}\right)^{th} \text{ item}$$

Important note:

The second quartile is also called the median.

Q_4 is not always useful!

Interquartile range

Interquartile range (IQR) is the difference between the third quartile (Q_3) and the first quartile (Q_1)

Example 1

Compute Q1, Q3 and the inter quartile range (IQR) for the data relating to the marks of 13 students in the Introduction to Statistical Thinking quiz given below:

10, 15, 10, 9, 18, 16, 14, 12, 16, 13, 15, 20, 17.

Solution

n = 13

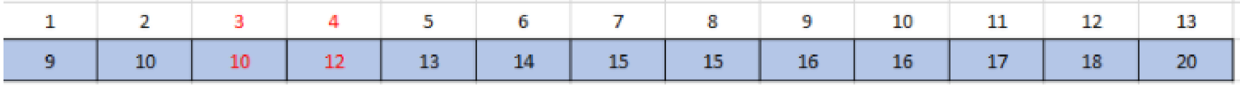
Arrange the values in ascending order

9, 10, 10, 12, 13, 14, 15, 15, 16, 16, 17, 18, 20

$Q_1 = (\frac{n+1}{4})^{th}$ position

$Q_1 = (\frac{13+1}{4})^{th}$ = 3.5 position

So, the first quartile is within third and forth position. We therefore, find the average of the values.



1	2	3	4	5	6	7	8	9	10	11	12	13
9	10	10	12	13	14	15	15	16	16	17	18	20

$$Q_1 = \frac{10 + 14}{2} = 12$$

$Q_3 = 3(\frac{n+1}{4})^{th}$ position

$Q_3 = 3(\frac{13+1}{4})^{th} = 3 \times \frac{14}{4} = \frac{42}{4}$ = 10.5th position

The third quartile is within 10th and 11th positions. We therefore, find the average of the values in the above figure.

$$Q_3 = \frac{16 + 17}{2} = 16.5$$

Interquartile range (IQR) $= Q_3 - Q_1 = 16.5 - 12 = 4.5$

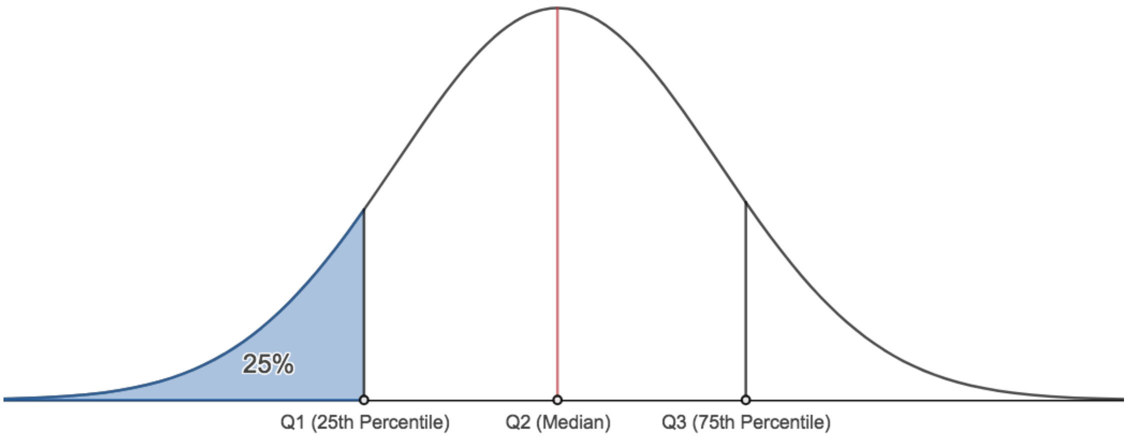
Important note:

The Quartiles are divisions of data by 25%, therefore,

Quartile 1 (Q1) can be called the **25th** percentile

Quartile 2 (Q2) can be called the **50th** percentile

Quartile 3 (Q3) can be called the **75th** percentile



How to get quartile in Python?

With **numpy.quantile()** method we can get the quartile values of an array.

numpy.quantile(a, q)

Where:

- a is the array
- q is sequence of quantiles to compute, which must be between 0 and 1 inclusive.

Therefore, for Q1 (q= 0.25), for Q2 (q = 0.5), and for Q3 (q= 0.75)

Example

From the example 1 above, use NumPy to compute Q1, Q3 and the inter quartile range

Solution

```
marks = np.array([10, 15, 10, 9, 18, 16, 14, 12, 16, 13, 15, 20, 17])

np.quantile(a= marks, q = [0.25, 0.75])
```

```
array([12., 16.])
```

Q1 = 12

Q3 = 16

Interquartile range (IQR) = Q3 - Q1 = 16 -12 = 4

Percentiles

This measure divides frequency distribution into 100 equal parts, P1 (first percentile), P2 (second percentile), ⋯ P50 (median), ⋯, P75 (75th percentile), ⋯, and P100.

Percentiles indicate the percentage of scores that fall below a particular value. They tell you where a score stands relative to other scores. For example, you are the fourth tallest person in a group of 20 for the high jump training in Kenya.

80% of people are shorter than you:

That means you are at the 80th percentile. If your height is 1.60m then "1.60m" is the 80th percentile height in that group.

Important note:

Quartile 1 (Q1) can be called the **25th** percentile

Quartile 2 (Q2) can be called the **50th** percentile

Quartile 3 (Q3) can be called the **75th** percentile

How to get Percentile in Python?

With numpy.percentile() method we can get the percentile values of an array.

numpy.percentile(a, q)

Where:

- a is the array
- q is sequence of percentiles to compute, which must be between 0 and 100 inclusive.

Therefore, for P25 (q = 25), for P50 (q = 50), and for P75 (q = 75)

Example 2

Mr Afwerki, a Tigrinya, has a hen that lays ten eggs. Each egg was weighed and recorded in gramms (g) as follows:

59, 56, 61, 68, 52, 53, 69, 54, 57, 51.

Calculate the 75th percentile of the weight of the egg using Python.

Solution

In this case, q will be 75. Since we are looking for the 75th percentile.

```
eggs = np.array([59, 56, 61, 68, 52, 53, 69, 54, 57, 51])

np.percentile(a = eggs, q = 75)
```

```
60.5
```

Therefore, the 75th percentile is 60.5. This means that that 75 percent of the eggs are lower than 60.5g.

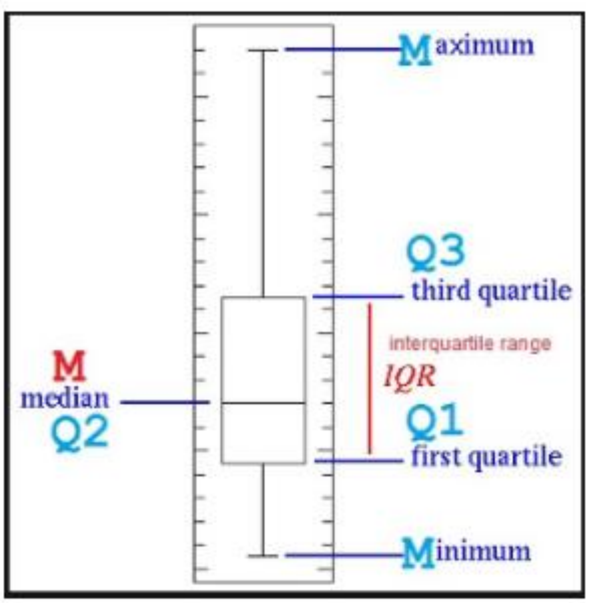
1.3.3 Introduction to five number summary

The five-number summary in statistics is a set of descriptive statistics that provides information about a dataset. It consists of the five most important basic statistics:

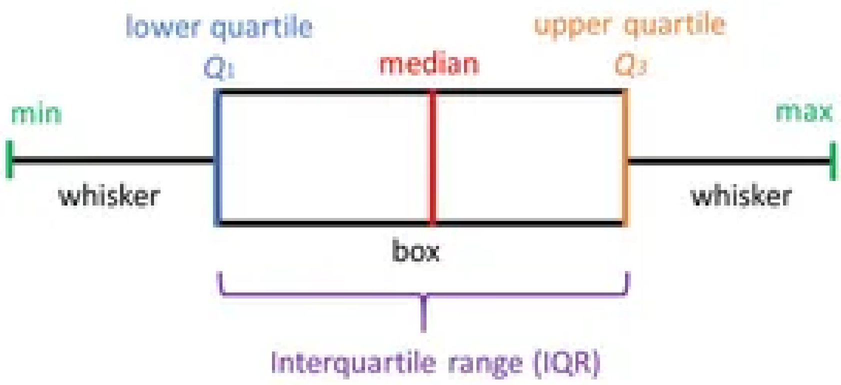
- the minimum (smallest observation)
- the lower quartile or first quartile
- the median (the middle value)
- the upper quartile or third quartile
- the maximum (largest observation)

A box plots

A boxplot can be used to show or represent the five-number summary



Vertical position



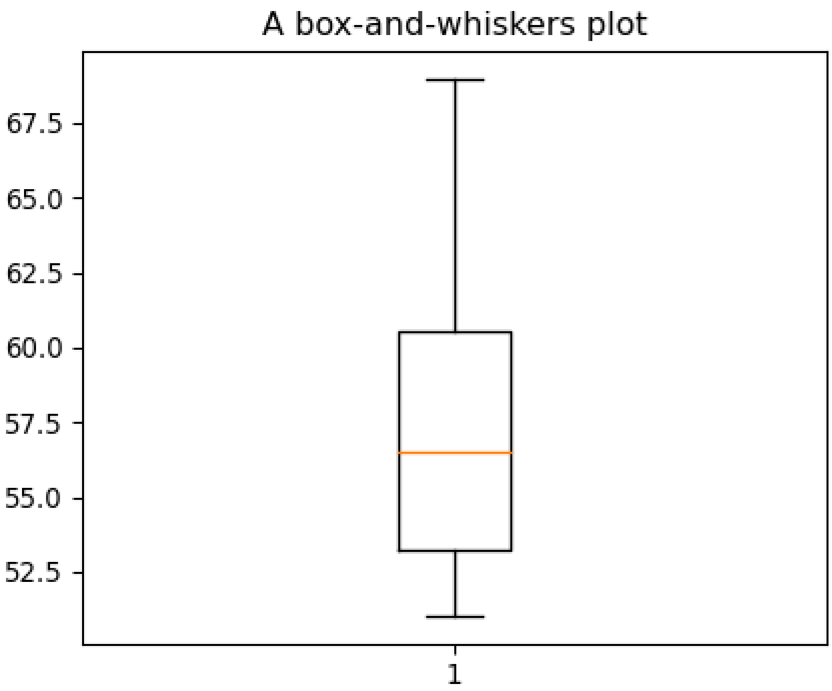
Horizontal position

Example

```
eggs = np.array([59, 56, 61, 68, 52, 53, 69, 54, 57, 51])
```

We will use boxplot in Matplotlib library to plot this

```
plt.boxplot(eggs)

plt.title("A box-and-whiskers plot");
```



Class activity 2 (Tutor guided question)

Peer to Peer Interaction

Visit the LMS, locate forum activity and participate in the discussion

The life expectancy for a person living in one of the 54 countries in Africa in the year 2020 is in the dataset name "Africa life expectancy". The data is from

https://www.indexmundi.com/map/?v=30&r=af&l=en.

1. Import the data as life_expectancy. Note: The data is in the datasets directory.

2. What is the life expectancy for your own country?

3. Find the five-number summary for Life expectancy at birth and the IQR in the dataset

4. Draw a box-and-whiskers plot.

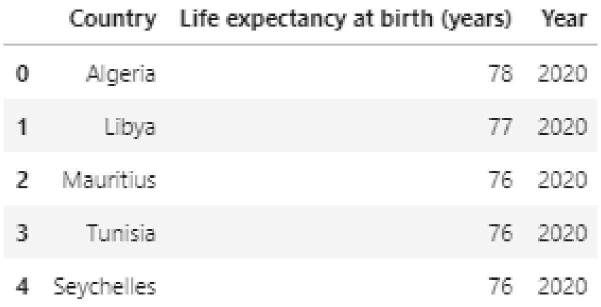
Solution

We need to import various packages that will be needed for this assignment.

```
import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import seaborn as sns
```

Solution 1

```
life_expectancy = pd.read_csv("datasets/Africa life expectancy.csv")

life_expectancy.head()
```

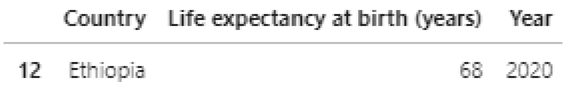


	Country	Life expectancy at birth (years)	Year
0	Algeria	78	2020
1	Libya	77	2020
2	Mauritius	76	2020
3	Tunisia	76	2020
4	Seychelles	76	2020

Solution 2

I am from Ethiopia. Therefore, I will check for my country.

```
life_expectancy[life_expectancy["Country"] = = "Ethiopia"]
```



	Country	Life expectancy at birth (years)	Year
12	Ethiopia	68	2020

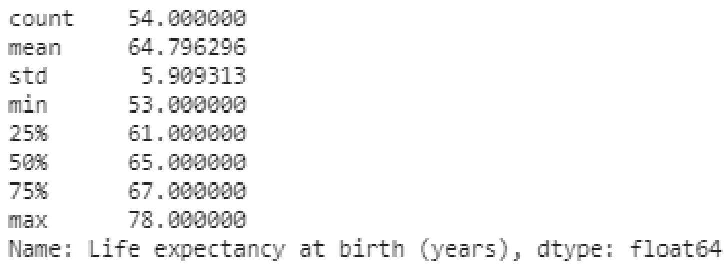
For Ethiopia, the life expectancy is 68 years.

Solution 3

The five number summary are minimum (min), first quartile, median, third quartile, and maximum (max). In this case, we will use **.describe()** attribute.

We will select Life expectancy at birth (years) column and then apply **.describe()** attribute on it.

```
life_expectancy["Life expectancy at birth (years)"].describe()
```



```
count    54.000000
mean     64.796296
std       5.909313
min      53.000000
25%      61.000000
50%      65.000000
75%      67.000000
max      78.000000
Name: Life expectancy at birth (years), dtype: float64
```

The five number summary for the life expectancy are:

Min = 53

First quartile (25%) = 61

Median (50%) = 65

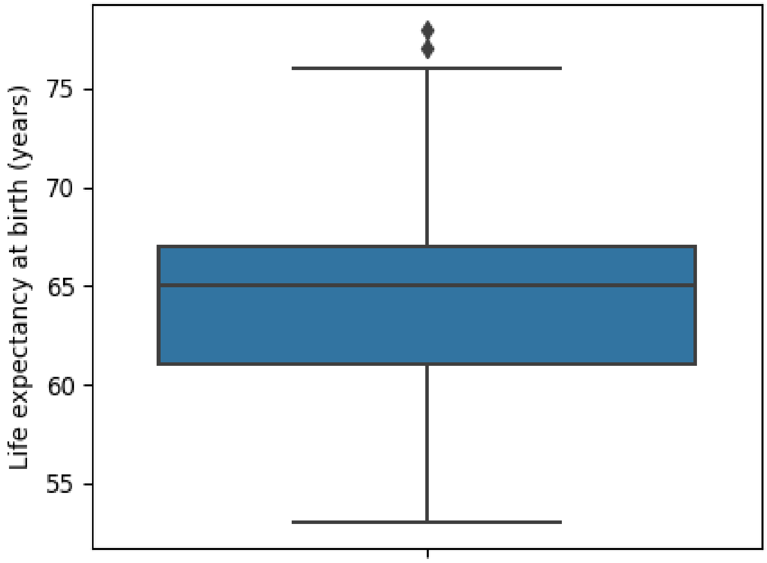
Third quartile (75%) = 67

Max = 78

Solution 4

We can use seaborn package to draw boxplot

```
sns.boxplot(y = "Life expectancy at birth (years)", data = life_expectancy);
```



1.4 Scale of measurement and Correlation Analysis

1.4.1 Scale of measurement

In Introduction to Data Science 1 (CS 2 2), we learnt that data can be qualitative or quantitative. A qualitative data is observed and recorded, for example gender or ethnicity while quantitative data is the type of data that arise as a result of numerical estimation or measurement, for example weight or height of a person. Both qualitative and quantitative have their scale of measurement. In the other hand, each **level of measurement** scale has specific properties that determine the various use of statistical analysis or approach. In this section, we will learn four types of scales such as nominal, ordinal, interval and ratio scale.

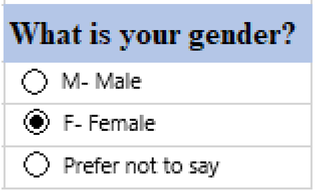
There are four different scales of measurement. The data can be defined as being one of the four scales. The four types of scales are:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale

Nominal Scale

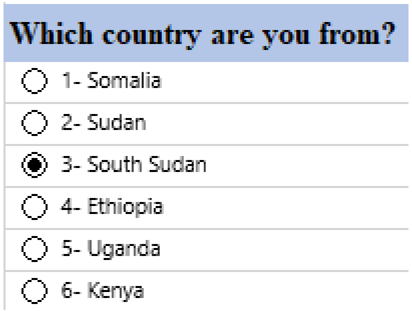
A nominal scale is the 1st level of measurement scale that is used for identification purposes. Sometimes known as categorical variable scale, it is used for labeling variables into distinct classifications. The numbers associated with variables of nominal scale are only tags or labels for categorization or identification purpose. The only statistical analysis that can be performed on a nominal scale is to tabulate it using frequency and percentage. It can also be analyzed graphically using a bar chart or pie chart. Nominal data is known as qualitative data or categorical data. The most common measure of central tendency for the nominal variable is the mode. On the other hand, the median makes no sense for the nominal scale since ranking is meaningless for the nominal data type.

Examples of a nominal scale measurement are shown below:



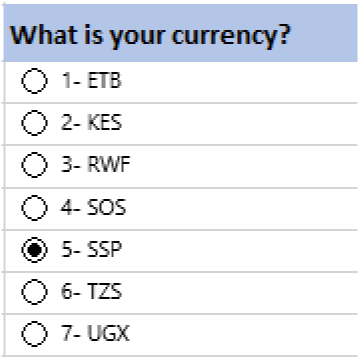
What is your gender?

○ M- Male
● F- Female
○ Prefer not to say



Which country are you from?

○ 1- Somalia
○ 2- Sudan
● 3- South Sudan
○ 4- Ethiopia
○ 5- Uganda
○ 6- Kenya



What is your currency?

○ 1- ETB
○ 2- KES
○ 3- RWF
○ 4- SOS
● 5- SSP
○ 6- TZS
○ 7- UGX

Other examples include ethnicity, religion, and languages spoken.

When a nominal variable has two categories or levels it is called a binary variable. **Binary** variables are categorical variables that have two possible levels (e.g., yes/no; present/absent). For a binary variable, we may assign a value of one (1) for present or value of zero (0) for absent of a disease. As a rule of thumb, the desired outcome is assigned the value of 1.

Ordinal Scale

The ordinal scale is the 2nd level of measurement that reports the ordering and ranking of data without establishing the degree of variation or difference between them. Ordinal represents the “order.” The attributes on an ordinal scale are usually arranged in ascending or descending order. The ordinal scale allows for rank order (1st, 2nd, 3rd, 4th, etc.) by which data can be sorted, but still does not allow for relative *degree of difference* between them.

Ordinal data is known as qualitative data or categorical data. The most common measure of central tendency for the ordinal scale data is the median. However, the mean (or average) is not allowed while the mode is allowed.

Examples:

Ranking of students' performance in CS 3

- 1st
- 2nd
- 3rd

How would you rate this lecture?

- Excellent
- Very Good
- Good
- Bad
- Poor

Assessing the degree of agreement

- Totally agree
- Agree
- Neutral
- Disagree
- Totally disagree

Interval Scale

The interval scale is the 3rd level of measurement scale. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. Interval scale indicates distance between two entities. The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 70 and 60 degrees is a measurable 10 degrees, as is the difference between 90 and 80 degrees. The interval type allows for the *degree of difference* between items, but not the ratio between them.

We can use mode, median, and arithmetic mean as a measure of central tendency in the interval scale while range and standard deviation can be used as measure dispersion.

Ratio scale

The ratio scale is the 4th level of measurement scale, which is quantitative. It is an extension of the interval scale which possesses a meaningful absolute zero value and because of this, it doesn't have negative value. The ratio scale is compatible with all statistical analysis methods like the measures of central tendency (mean, median, mode, etc.) and measures of dispersion (range, standard deviation, etc.). Examples of ratio scale include height, age, weight, and length.

Pictorial summary of scale of measurement

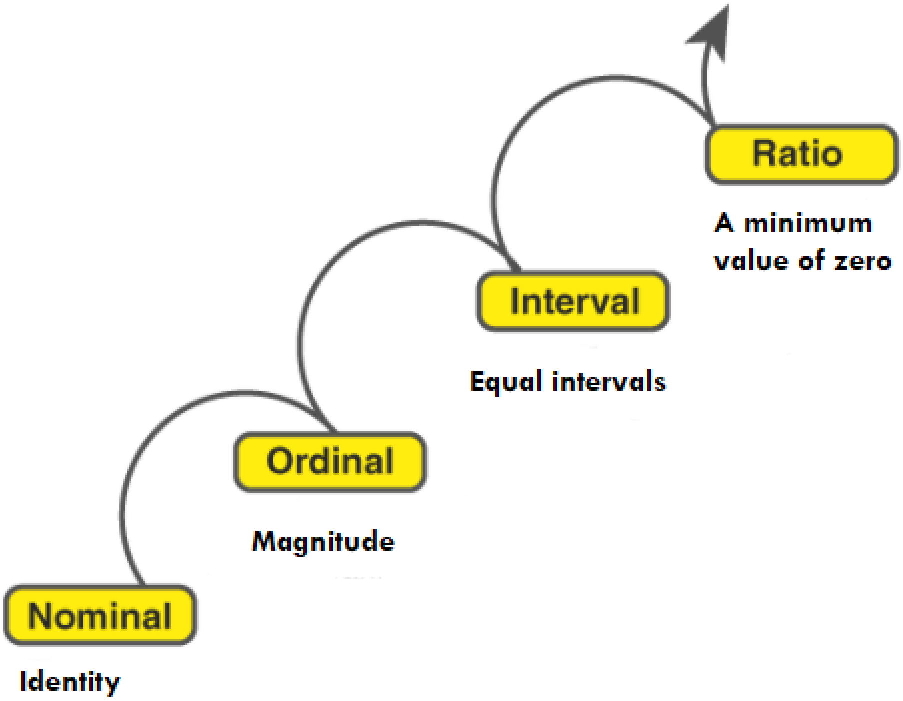
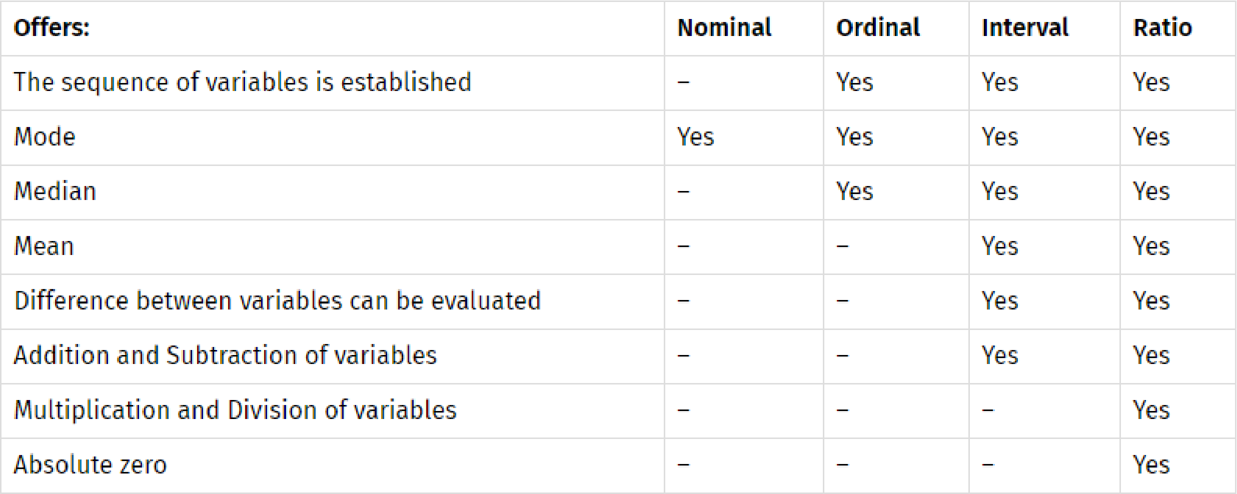


Table of scale of measurement and measure of central tendency



Offers:	Nominal	Ordinal	Interval	Ratio
The sequence of variables is established	–	Yes	Yes	Yes
Mode	Yes	Yes	Yes	Yes
Median	–	Yes	Yes	Yes
Mean	–	–	Yes	Yes
Difference between variables can be evaluated	–	–	Yes	Yes
Addition and Subtraction of variables	–	–	Yes	Yes
Multiplication and Division of variables	–	–	–	Yes
Absolute zero	–	–	–	Yes

1.4.2 Correlation analysis

Correlation measures the strength and direction of the statistical linear relationship between two or more variables in the data. For example, you may want to measure the correlation between height and weight of women in Uganda. We can measure the degree of correlation by using Pearson correlation coefficient. For example, if X and Y are continuous variables, then the correlation between X and Y is given by the formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

where:

r = Pearson r correlation coefficient between x and y

n = number of observations

x = the values of the x-variable in a sample

y = the values of the y-variable in a sample

Just sum up x, y, x^2, y^2 and xy then use the formula above.

Pearson r correlation coefficient is also known as product moment correlation

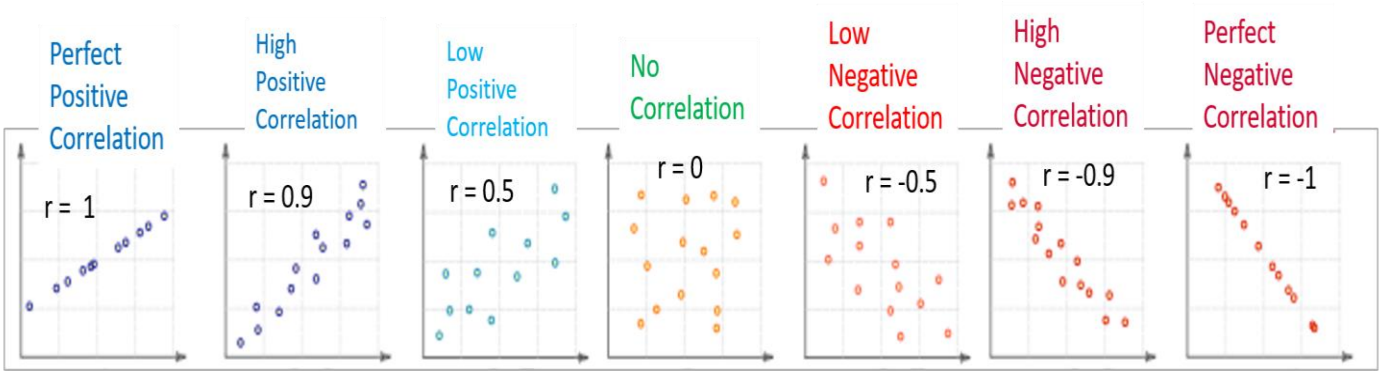
The correlation coefficient (r) can take any values from -1 to 1. The interpretations of the values are:

−1: Perfect negative correlation. The variables tend to move in opposite directions (i.e., when one variable increases, the other variable decreases).

0: No correlation. The variables do not have a relationship with each other.

1: Perfect positive correlation. The variables tend to move in the same direction (i.e., when one variable increases, the other variable also increases).

Positive correlation is a relationship between two variables in which both variables move in the same direction. This is when one variable increase while the other increases and vice versa.



Correlation measures the strength and direction of the linear relationship between two variables. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

How to get correlation in Python?

Python uses **df.corr()** pandas attribute to estimate the correlation between or among the pairs of continuous variables in the DataFrame.

Scatter diagrams

To examine whether there is a correlation between two quantitative variables, you will need to first plot a scatter diagram. The scatter diagram assesses the relationship between the variables and determine whether they are correlated or not.

Example 1

Import **Uganda_women.csv** dataset which contains the height and weight of 300 Ugandan women. The data is in the **datasets** directory.

- Plot a scatter diagram of height against weight and examine the correlation
- Calculate Pearson correlation coefficient between height and weight of Ugandan women

Solution 1

Let's import all the necessary packages:

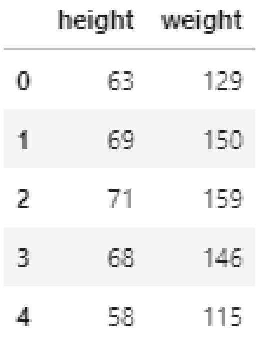
```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```

We need to import Ugandan women dataset and then use Seaborn package to plot the scatter diagram of height and weight.

```
uganda_women = pd.read_csv("datasets/Ugandan_women.csv")
```

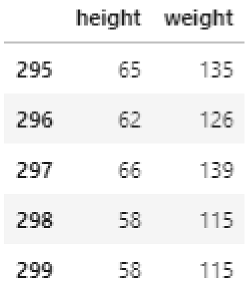
Let us examine the dataset by using **.head()**, **.tail()**, and **.shape** attributes

```
uganda_women.head()
```



	height	weight
0	63	129
1	69	150
2	71	159
3	68	146
4	58	115

```
uganda_women.tail()
```



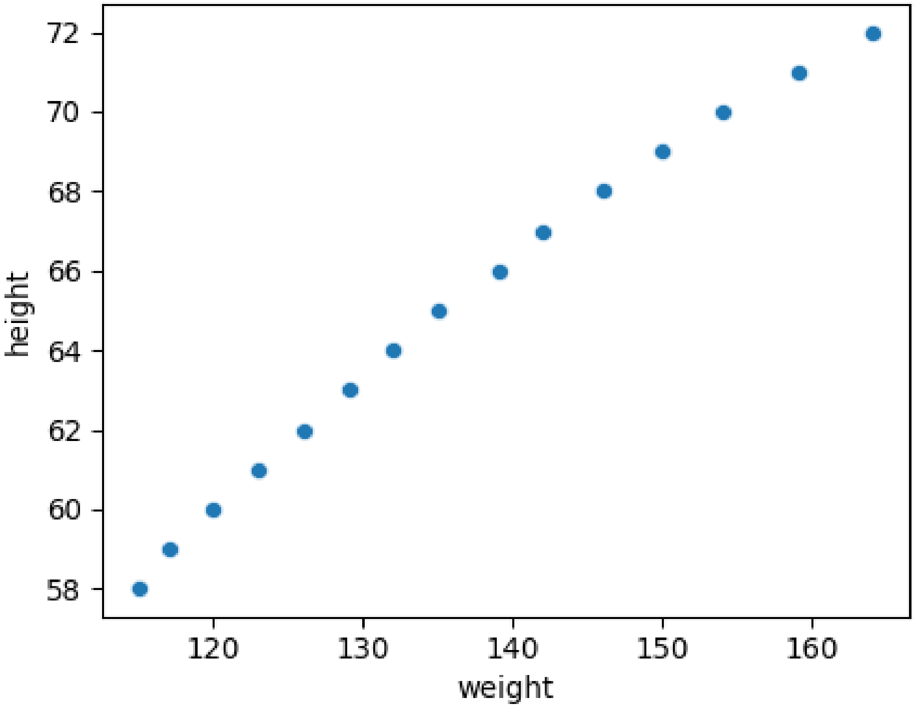
	height	weight
295	65	135
296	62	126
297	66	139
298	58	115
299	58	115

```
uganda_women.shape
```

```
(300, 2)
```

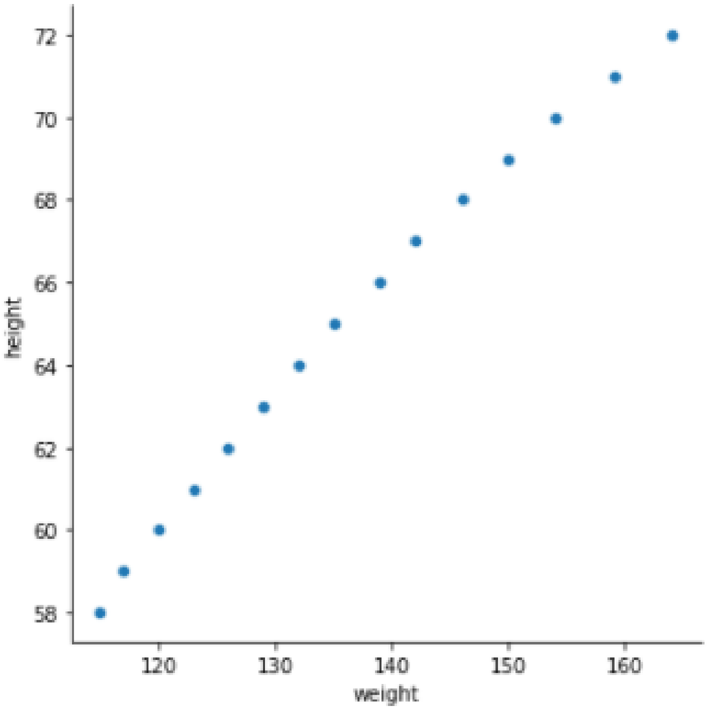
We now plot the scatter diagram using Seaborn **sns.scatterplot()** function.

```
sns.scatterplot(y = "height", x = "weight", data = uganda_women);
```



We can also use **sns.relplot()** function from Seaborn to plot the scatter diagram.

```
sns.relplot(y = "height", x = "weight", data = uganda_women);
```

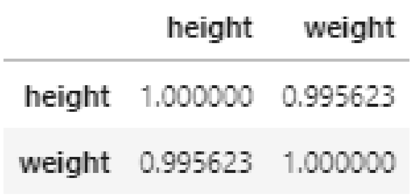


As you can see, there seem to be a very strong positive correlation between height and weight of Ugandan women

Solution 2

We can get estimate of Pearson correlation coefficient by using **uganda_women.corr()**.

```
uganda_women.corr()
```



	height	weight
height	1.000000	0.995623
weight	0.995623	1.000000

This is a result of the Pearson correlation. We will see that correlation on the same variable will always be 1. Our concern is to look at the correlation between height and weight. Therefore, the correlation between height and weight is 0.9956.

Example 2

This dataset is from a restaurant in Nairobi, Kenya, where many come to eat food and after eating based on a total bill they hate paid some tips. The following are the variables in the tips dataset:

total_bill: Cost of the meal in Kenya Shilling

tip: Gratuity in Kenya Shilling

gender: Sex of person paying for the meal

smoker: Whether they smoke in the party or not

day: Day of the week for the party

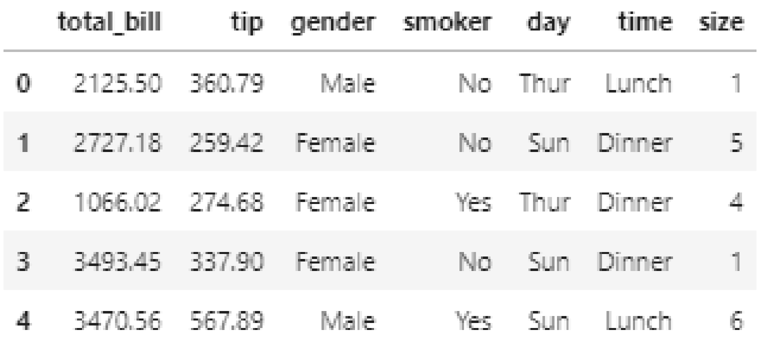
time: Time of the day whether for lunch or dinner

size: Size of the party

- Import the tip dataset as **tips** (this is in the datasets folder)
- Calculate the pairwise correlation among the continuous variables in the data
- Visualize by using scatter diagram

Solution 1

```
tips = pd.read_csv("datasets/tips.csv")

tips.head()
```

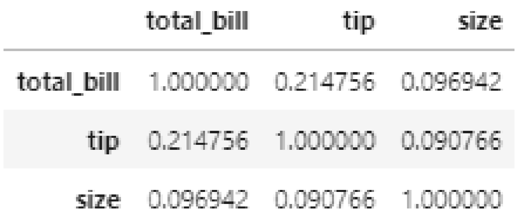


	total_bill	tip	gender	smoker	day	time	size
0	2125.50	360.79	Male	No	Thur	Lunch	1
1	2727.18	259.42	Female	No	Sun	Dinner	5
2	1066.02	274.68	Female	Yes	Thur	Dinner	4
3	3493.45	337.90	Female	No	Sun	Dinner	1
4	3470.56	567.89	Male	Yes	Sun	Lunch	6

Solution 2

To get the pairwise correlation among the continuous variables in the data, we use **tips.corr()**

```
tips.corr()
```



	total_bill	tip	size
total_bill	1.000000	0.214756	0.096942
tip	0.214756	1.000000	0.090766
size	0.096942	0.090766	1.000000

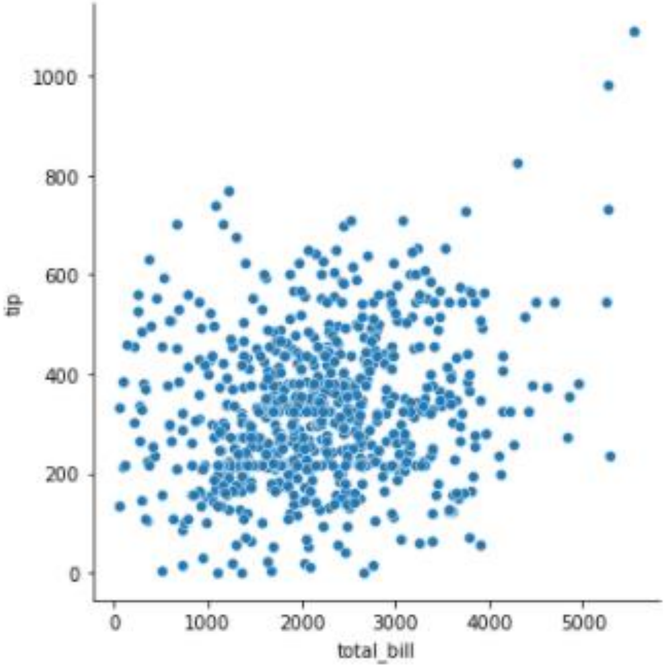
Solution 3

We can plot the scatter diagram between:

- **tip** and **total_bill**
- **tip** and **size**

tip and total_bill scatter diagram

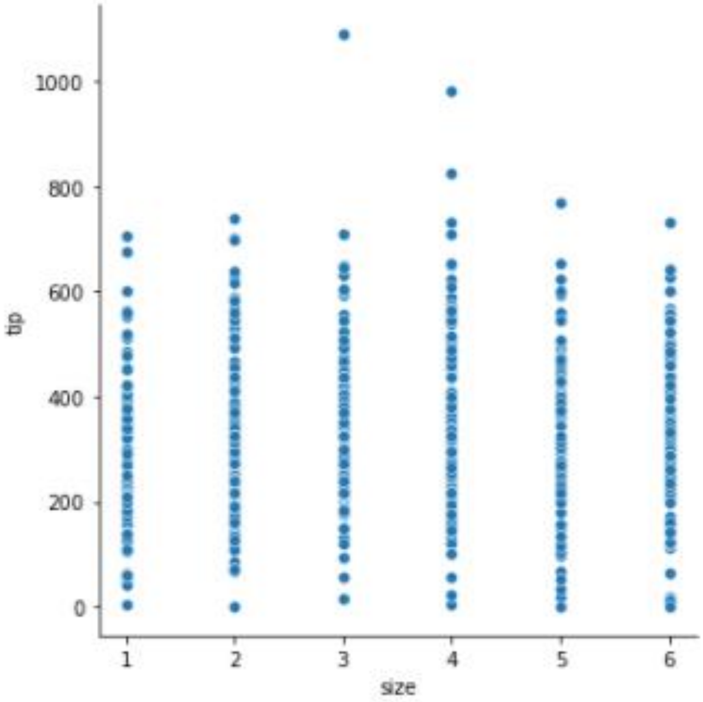
```
sns.relplot(y = "tip", x = "total_bill", data = tips)
```



There is a low positive ($r = 0.214756$) linear correlation between tip and total bill.

tip and size scatter diagram

```
sns.relplot(y = "tip", x = "size", data = tips)
```



As you can see, there is no linear relationship between tip and size of the party. This diagram looks like that because size of the party is a categorical variable i.e. it has a limited value.

Class activity 2 (Peer to peer review activity)

Peer to Peer Interaction

Visit the LMS, locate forum activity and participate in the discussion

The palmer penguins data contains size measurements for three penguin species observed on three islands in the Palmer Archipelago, Antarctica.

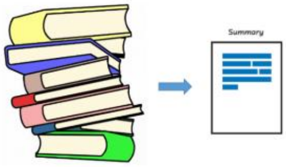
1. Import the necessary libraries
2. Import penguins dataset. Note that this dataset is in the **activity_datasets** directory.
3. Create a pairwise correlation coefficient for all the continuous variables in the dataset
4. Plot a scatter diagram with the x axis as bill_length_mm and y axis as flipper_length_mm by using sns.scatter() function

Correlation and Causation

There is a difference between correlation and causality. The famous expression “correlation does not mean causation” is crucial to understand the two statistical concepts.

If two variables are correlated, it does not imply that one variable causes the changes in another variable. The correlation only assesses relationships between variables, and there may be different factors that lead to relationships. Causation may be a reason for the correlation, but it is not the only possible explanation.

Summary of Study Unit 1



In this study unit, you have learnt that:

1. A population is a complete set while a sample is a subset or fraction of the population.
2. Parameters are descriptive measure for a population. They include the population mean (μ), the population variance (σ^2), population standard deviation (σ), and population proportion (P).
3. Statistic is a descriptive measure for a sample. They include the sample mean ($\bar{X}$), the sample variance (s^2) and the sample standard deviation (s), and sample proportion (p). Sample statistics are used to estimate unknown population parameters.
4. The most common measures of central tendency are the mean, median, and the mode.
5. The measure of spread or dispersion is used to describe the variability in the data, and they include range, standard deviation, variance, and mean absolute deviation.
6. The four scale of measurement in statistics are nominal scale, ordinal scale, interval scale and ratio scale
7. Correlation measures the strength and direction of the statistical linear relationship between two or more variables in the data

Additional resources

For more resources in this section please consider the following:

- https://bit.ly/population-versus-sample
- https://bit.ly/scale-of-measurement
- https://www.mathsisfun.com/data/correlation.html