Modul 3

Basic Statistic

Purwadhika
Startup and Coding School

What is Statistic?

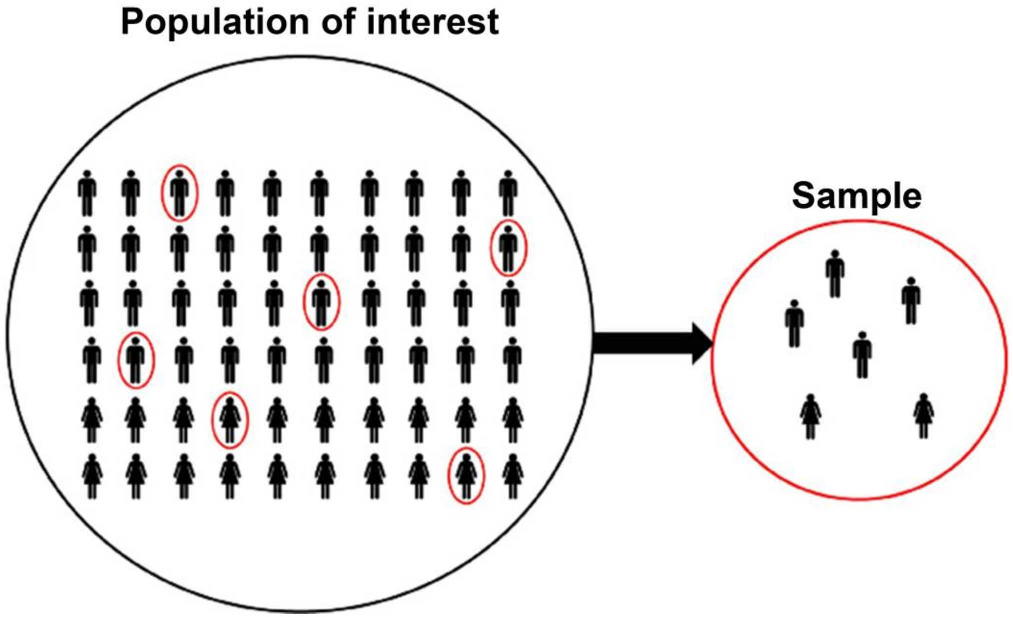
- Statistic could be described as methodology for collecting, analyzing, interpreting, and drawing conclusion from information. In other word, statistic is the science of gaining information from numerical and categorical data.
- Furthermore, statistic is the science of dealing with uncertain phenomenon and events.
- It is a very broad subject, with many application on various field.

Statistic

- Statistic provide methods for:
    - Design: Planning and carrying out research studies
    - Description: Summarizing and exploring data
    - Inference: Making predictions and generalizing about phenomena represented by the data

Population and Sample

- Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during the research.
- Set of individual or object observed as representation of the population is called sample



Population and Sample

- Population always represent the target of an investigation. We learn about the population from the samples
- Finite Population is a population that could be physically listed. Ex:
  - Student at Purwadhika
  - Number of chair at the classroom
- Hypothetical Population is a population that was abstract and arise from the phenomenon under consideration. Ex:
  - Factory producing light bulb, if the factory keep the same equipment, using the same produce method, and raw materials. The bulb produced could be consider as hypothetical population

Population and Sample

- It is always only a certain features of individual person or object that under investigation at the same time. Not all features wanted to be measured from individual at the population. This give Population and sample another definition.
- Population (statistical) is a set of measurements corresponding to the entire collection of units for which inferences are to be made
- A sample from statistical population is the set of measurements that are actually collected during the investigation.

Type of Statistic

- **Descriptive Statistic**

Branch of statistic to summarize and describe the data. It consist of methods for organizing and summarizing information.

- **Inferential Statistic**

Branch of statistic to use the data sample to make an inference about a population. It consist of methods for drawing and measuring the reliability of conclusion based on the sample from the population

Descriptive Statistic

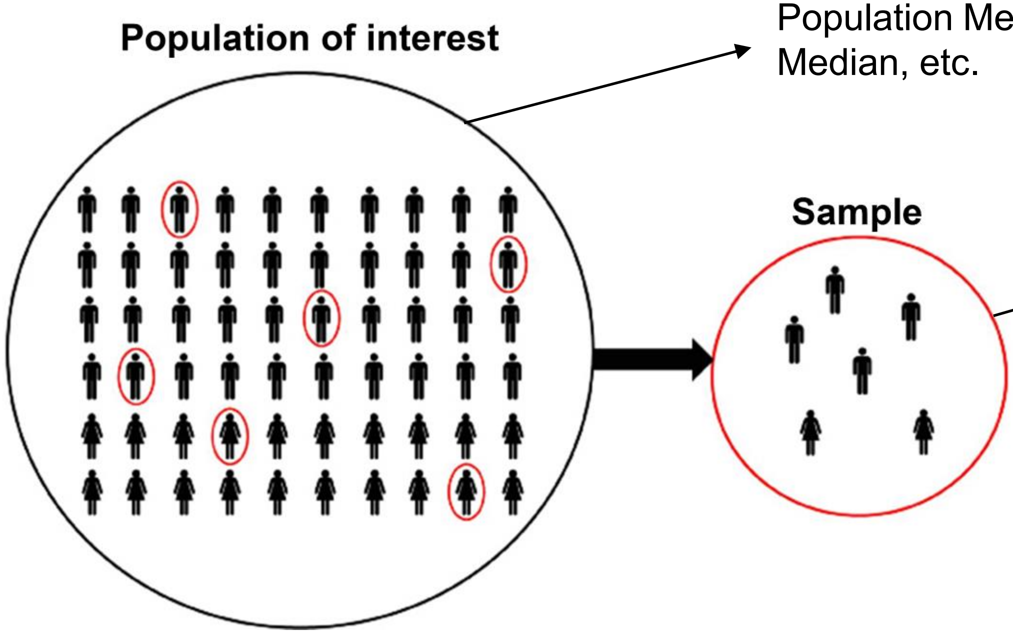
- Descriptive statistic includes the construction of graphs, charts, tables, and calculation of various descriptive measures such as averages, variation, and percentile.
- There are generally 2 ways to describe the data:
  - **Measures of Central Tendency.** It is the way of describing the central position of a frequency distribution for a group of data. We can describe it by using, for example Mean, Median, Mode
  - **Measures of Spread.** It is the way to summarize the group of data by describing how spread the data are. We can describe it by using, for example Range, Quartile, Variance, and Standard Deviation

Inferential Statistic

- Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn. It is, therefore, important that the sample accurately represents the population. The process of achieving this is called sampling.
- Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population.
- Inferential statistics method include point estimation, interval estimation, and hypothesis testing which are all based on probability theory.

Parameter

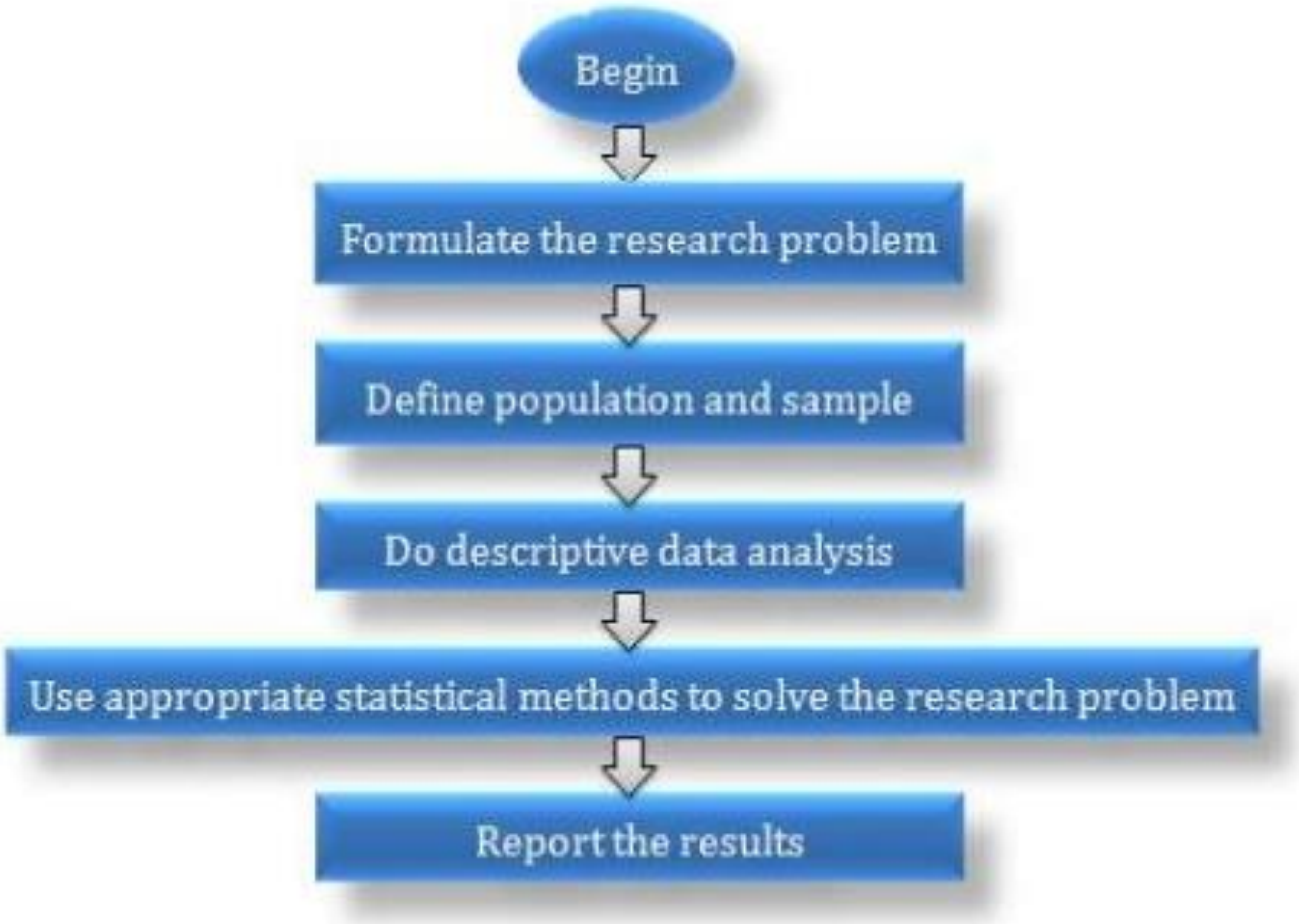
- Often feature of the population under research could be summarized by numerical parameter. Hence the research problem usually becomes as on investigation of the values of parameters.
- Population parameter are unknown and sample statistic used to make inference about it



Population Parameter such as Population Mean, Population Median, etc.

Statistic sample such as Statistic Mean, Statistic Median, etc.

Statistical Data Analysis



Begin

Formulate the research problem

Define population and sample

Do descriptive data analysis

Use appropriate statistical methods to solve the research problem

Report the results

Variables

- A characteristic that varies from one person or thing to another is called a variable. Ex: Height, Weight, Eye Color, etc.
- Variable could be divided as quantitative (numerical) and qualitative (categorical) variable.

Qualitative Variable

- Qualitative variables or discrete variable could be categorized as:
  - **Nominal**: Variables that have two or more categories, but did not have intrinsic order. Ex: Type of fruit (Apple, Banana, Grape)
  - **Dichotomous or Binary**: Nominal variables which only have two categories. Ex: Gender (Male or Female), Own a house (Yes or No), Type of properties(Commercial or Residential)
  - **Ordinal**: Variables that have two or more categories, but the categories could be ranked or ordered. Ex: Satisfaction level (Satisfied, Normal, Not Satisfied)

Qualitative Variable

- Number of observation that fall into particular class of the qualitative variables is called the frequency (count) of the class. A table listing all classes and their frequencies is called a frequency distribution.
- Often, we interested in the relative frequency or the percentages of the class. We could find it by dividing the frequency of the class by total number of observations and multiply it by 100. The percentage would be represented as decimal.

Quantitative Variable

- Quantitative variable or Continuous variable can be further divided into:
  - **Interval**: Variable which their central characteristic could be measured along continuum value and have numerical value. Ex: Temperature, difference between 20C and 30C is the same as difference between 30C and 40C
  - **Ratio**: Interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever.

Quantitative Variable

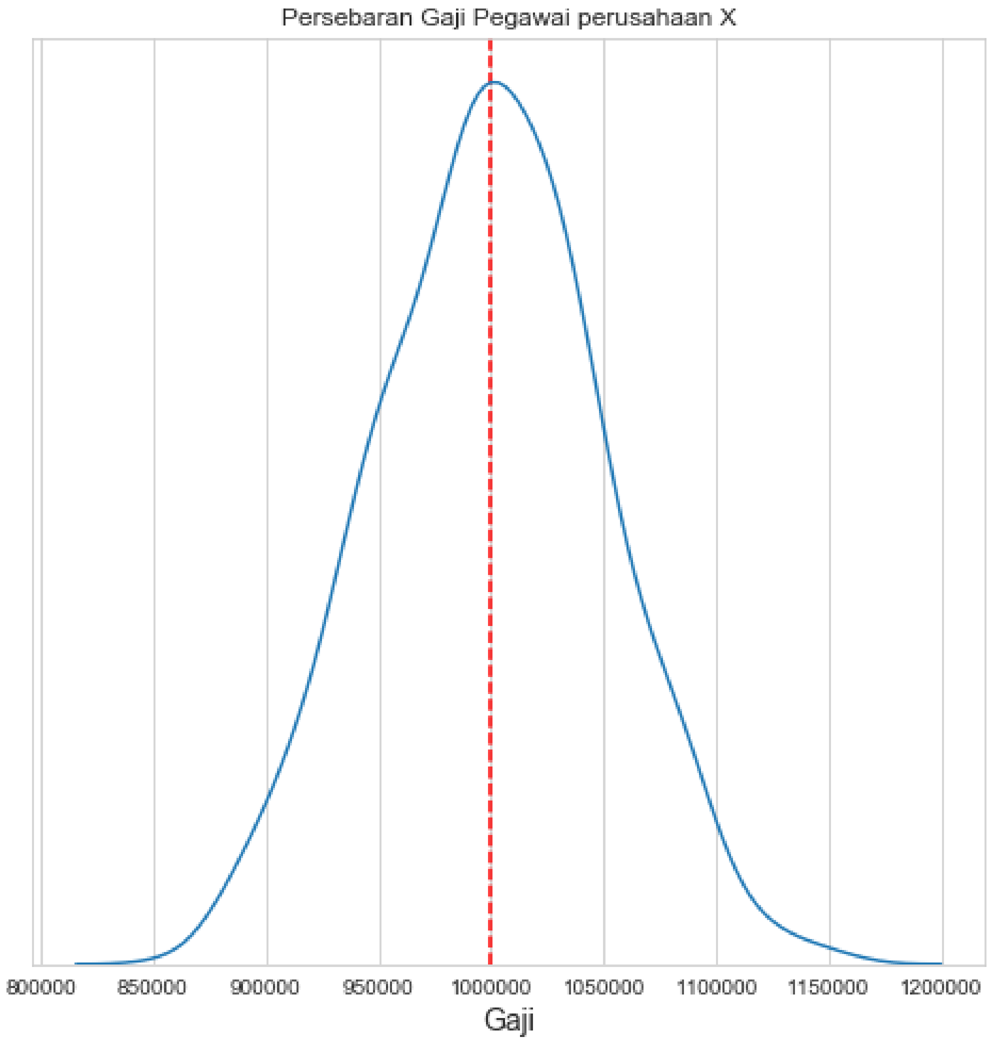
- If the quantitative variable can only obtain few different values, then the data could be summarized as qualitative variables in frequency tables.
- If the quantitative variables have a lot of different values, then the data must be grouped into classes before the table of frequencies could be formed. The main step of grouping quantitative variables into classes are:
  1. Find The minimum and the maximum values variable in the data set
  2. Choose interval of equal length that cover the range between the minimum and the maximum without overlapping. These are called class intervals, the end points called class limits
  3. Count the number of observation in the data that belong to each class interval. The count in each class is the class frequency
  4. Calculate the class frequencies of each class

Probability Distribution

- *Probability Distribution* is a mathematical function which assign a likelihood for a random variable to having a certain value. In other word, likelihood of random variable to having a certain value would depend on the *Probability Distribution*.

Probability Distribution

- *Probability Distribution* is useful for inferring the most probable event to be happen, likelihood of an event to occur, and the likelihood interval for event to occur. For example, company X employ 1000 employees and I want to take a sample of the salary for each employee. If I create the distribution plot of the company X employees salary, it could be visualized in the image beside.
- Picture beside giving us the *Probability Distribution* of the salary likelihood in the company X. We know if we randomly take a salary sample from company X, we probably would get a higher chance to acquire value that close to the red line or the mean. In other hand, the probability for us to get the random sample with salary value less than 900000 or more than 1100000 is small.



Probability Distribution Properties

- In statistic, *probability* could be written as

  p(x) = probability or likelihood of random variable to having a value x

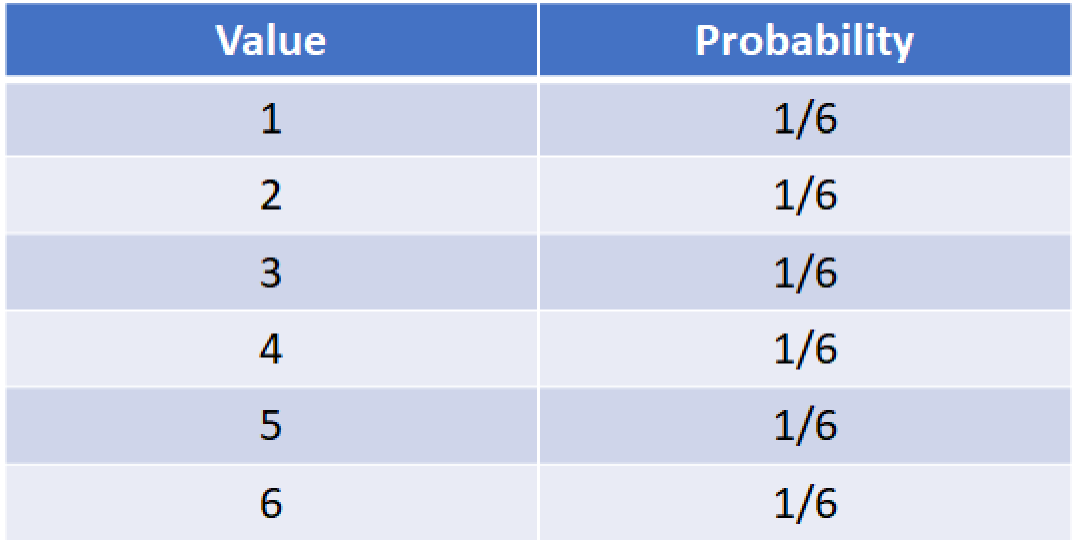
- If we sum all the probability that could happen in the *Probability Distribution*, it would equal to 1. Moreover, the probability would only exist within range value 0 to 1.
- In addition, depend on what variable data type we have, the *Probability Distribution* type would be also different. That is:
  - *Discrete Probability Distribution* or *Probability Mass Function* for discrete or categorical variable
  - *Continuous Probability Distribution* or *Probability Density Function* for continuous or quantitative variable

Probability Mass Function (PMF)

- *Probability Mass Function* is a *Probability Function* for *discrete* or *categorical* variable. For example, die-roll event. This event following the *Probability Mass Function* because there are no existing value in between (6-side die only have value probability 1,2,3,4,5,6, and there are no condition that we could get a value 1.5).
- Moreover, every probability of the value in the PMF would have probability more than 0 and sum of the probability would equal to 1. Why is that? Because, if we roll a die we would certainly getting a value from 6 possibility in the die-roll event.

PMF

- If we create a roll-die event probability table, we would acquire table such as below:



Value	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

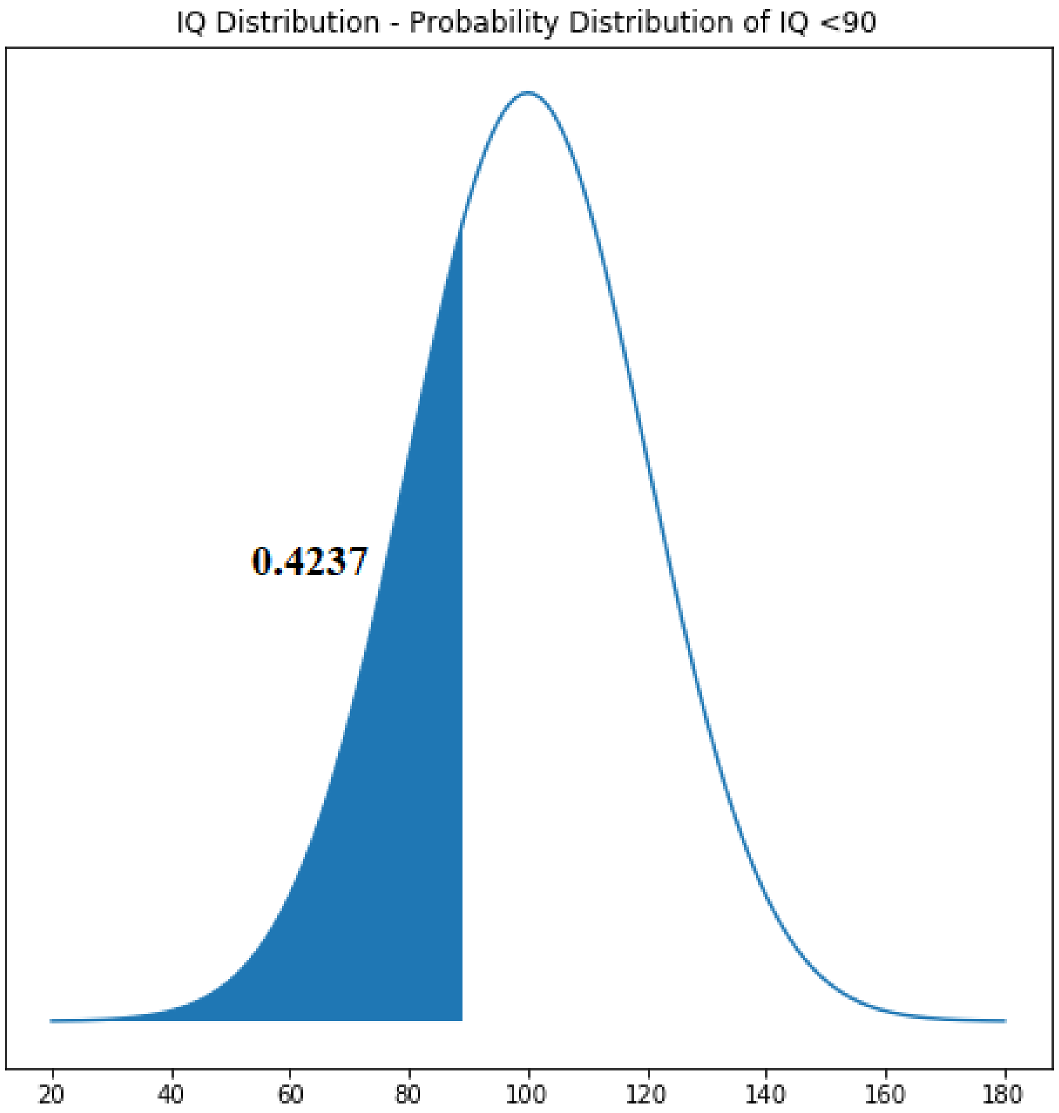
- From the table above, we can see that sum of the probability would equal to 1

Probability Density Function (PDF)

- If a variable contain infinite value between 2 values then this variable following the *Probability Density Function* (PDF). As an example are the body height and employee salaries. Different from PMF, probability in the PDF could be 0. For example, probability to measure someone height exactly at 175 cm (no more or less than few picometer) is 0 or at least close to 0.

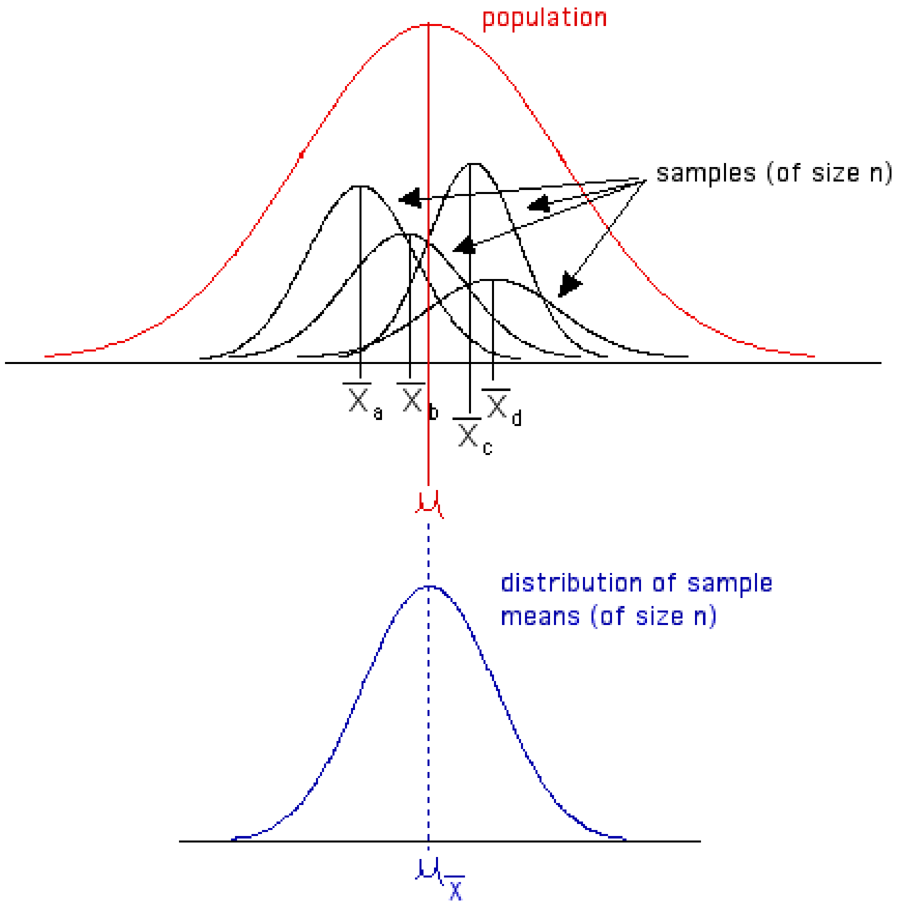
PDF

- Probability on PDF are measured based on the value range, not just one value. For example, we have IQ distribution from value 20 to 180 with sample mean 100 and standard deviation 20.
- In picture beside, we could for example want to know the probability to acquire a sample with IQ less than 90. We can measure the probability by measure the total area of the colored area or *area under the curve*. You could check this article to know more about the measurement. From the data, if we take a random sample, the probability to get a sample with IQ less than 90 is 42.37%.



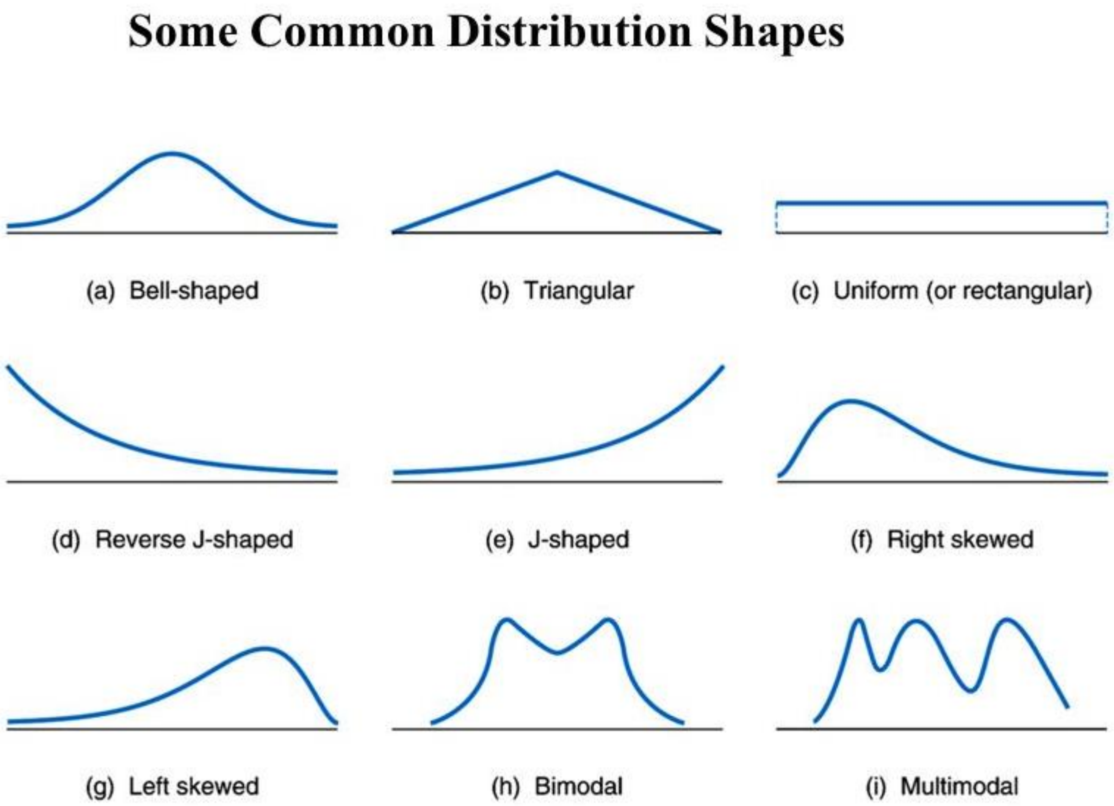
Population and Sample Distribution

- Frequency distribution for a variable apply both to population and to samples from that population.
- As the sample size increases, the sample relative frequency in any class interval gets closer to the true population relative frequency.



Sample Distribution

- One way to summarize a sample distribution is to describe its shape.
- A group for which the distribution is bell-shaped is fundamentally different from a group with uniform-shaped distribution



Some Common Distribution Shapes

(a) Bell-shaped

(b) Triangular

(c) Uniform (or rectangular)

(d) Reverse J-shaped

(e) J-shaped

(f) Right skewed

(g) Left skewed

(h) Bimodal

(i) Multimodal

Type of Measurements

- There are three most common way to describe the central measurement of the frequency distribution:
  - **Mode**: Value of a qualitative or a countable quantitative variable where the frequency is occurring the most.
  - **Median**: The middle value in the ordered list. If the number of observation is odd, then the sample median is the observed value exactly in the middle. If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list. In either case, the sample median position is at n+1/2 when n is the number of observation
  - **Mean**: The sum of observed values in a data divided by the number of observations. The most commonly used measure of center for quantitative variable.

Which measurement to choose?

- **Mode** should be used when calculating the measure of center for the qualitative variable
- **Mean** is the proper measure of center if dealing with the quantitative variable with symmetric distribution (often bell shaped)
- **Median** is the good choice if the quantitative variable have a skewed distribution. We do not used mean in this case, because mean could be highly influenced by an observation that falls far from the rest of the data (outlier)
- It should be noted that this measurement assume that the sample measurement is corresponding to the population measures of center, which are unknown. The sample measurement can be used to estimate this unknown parameter.

Measures of Variation

- Another important aspect of descriptive study is numerically measuring the extent of variation around the center. Two dataset of the same variable may possess similar position of center but remarkably different with respect to variability.
- Most frequently used measures of variation; the sample range, the sample interquartile range, and the sample standard deviation

Range

- The sample range is obtained by computing the difference between the largest observed value of the variable and the smallest one

Range = Max – Min

- Range is overly sensitive to the sample size

Standard Deviation

- The sample standard deviation is the most frequently used measure of variability. It can be considered as a kind of average of the absolute deviations of observed values from the mean of the variable in question.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

S = Sample Standard Deviation
X_i = Each value of dataset
$\bar{x}$ = Mean of the dataset

- Since Standard Deviation defined by the sample mean, it is preferred measure of variation if the mean is used as the measure of center(ex: Symmetric Distribution)

Standard Deviation

- The more variation in the observed value, the larger the standard deviation for the variable observed.
- Standard Deviation is greatly affected by a few extreme observation (usually outlier).

Sample Statistics and Population Parameters

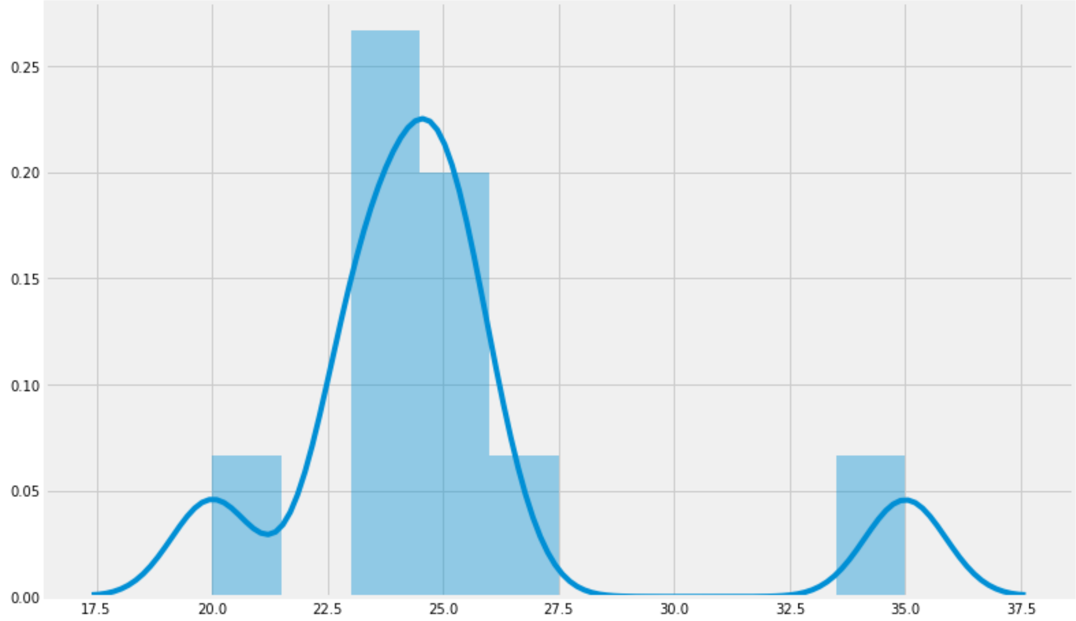
- From the measure of center and variation, sample mean (**$\bar{x}$**) and sample standard deviation (**s**) are the most commonly reported.
- Because the values depend on the samples, it vary from sample to sample. They are called random variable, because their values are unknown before the sample is chosen.
- We should separate between the sample statistic and the corresponding measures for the population. If we define it, **μ** is population mean, **σ** is population standard deviation.
- **μ** and **σ** is always constant, but **$\bar{x}$** and **s** is depending on the sample.

Sampling Distribution

- When the data are produced by random sampling, a statistic is a random variable that obeys the laws of probability theory. The link between probability and data is formed by the sampling distributions of statistic. A sampling distribution shows how a statistic would vary in repeated data production
- Sampling distribution reflect the sampling variability that occurs in collecting data and using sample statistic to estimate parameters.
- In other word, we take our statistic sample (ex: mean) and plot it in the graph.

Sampling Distribution

- Let's take an example, age of the student in the Purwadhika classroom and we get age average 25. We have 10 student in the class with varying age of 23, 23, 24, 22, 35, 25, 25, 25, 24,26. If we plot this into graph, we might get something like this

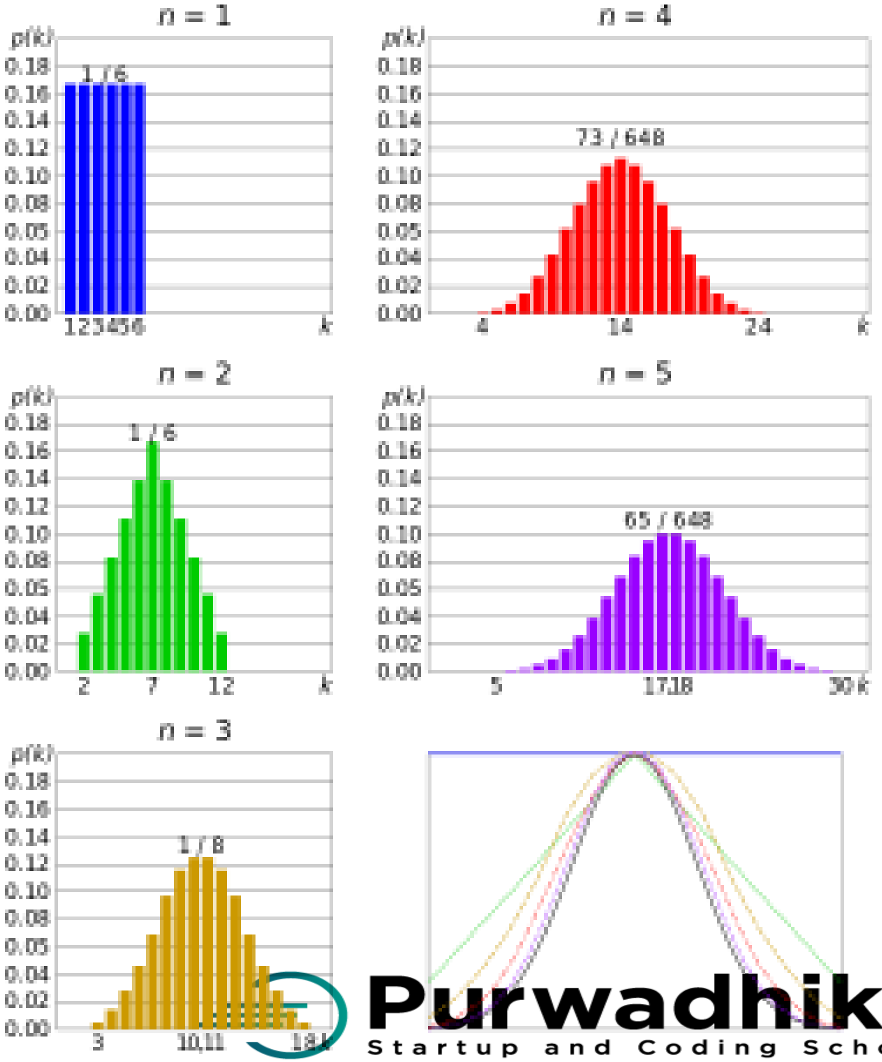


- What the graph look like would be impossible to predict, but according to *Central Limit Theorem* **If we use Ton of data, it would close to the bell curve (Normal Distribution)**.

Central Limit Theorem

- The **Central Limit Theorem** states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30. All this is saying is that as you take more samples, especially large ones, your graph of the sample means will look more like a normal distribution.

Here’s what the Central Limit Theorem is saying, graphically. The picture below shows one of the simplest types of test: rolling a die. The **more times you roll the die**, the more likely the shape of the distribution of the means tends to look like a **normal distribution graph**.

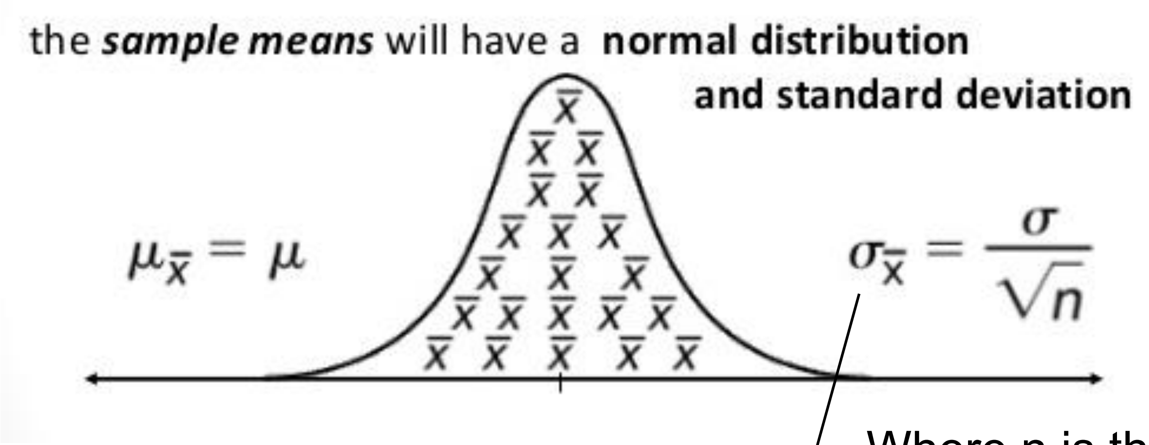


Central Limit Theorem

- An essential component of the Central Limit Theorem is that the **average of your sample means will be the population mean**. In other words, add up the means from all of your samples, find the average and that average will be your actual population mean.
- Similarly, if you find the average of all of the standard deviations in your sample, you'll find the actual standard deviation for your population. It's a pretty useful phenomenon that can help accurately predict characteristics of a population.

Central Limit Theorem

- Standard deviation that formed by the sample means would be called **Standard Error**.
- Standard Error shows the inference population variability, and would decrease as the sample size increase


the *sample means* will have a **normal distribution**
and standard deviation
$\mu_{\bar{x}} = \mu$
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
$\bar{x}$

Where n is the sample size for each mean

Standard Error

Estimation

- We already know that we could use the sample data to estimate the unknown population parameter. Statistical inference use sample data to forms two types of estimators of parameters:
  - **Point Estimate** (Single Number)
  - **Confidence Interval** (Interval)

Point Estimate

- Point estimate consist of a single number, calculated from the data, that is the best single guess for the unknown parameter.
- The available information is assumed to be in the form of random sample X1,X2,….Xn of size n taken from the population.
- For example, to estimate the **μ** (Population Mean) we could use estimator **x̄**.
- A good point estimator of a parameter is one with sampling distribution that is centered around parameter, and smallest Standard Error.

Confidence Interval

- It is more desirable to produce an interval values that is likely contain the true value of the unknown parameter, rather than point estimation which could be bias.
- A confidence interval estimate consist of interval numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident the parameter lies in the interval. The percentage is called the **confidence level**.
- The confidence level is a number from 0 to 1, which represent the percentage (Often it is 0.90 (90%), 0.95(95%), or 0.99(99%))

Confidence Interval

- Confidence Interval is divided whether the population variance is known or not. We apply what we called Z distribution if the population variance is known, and the T distribution of the population variance is unknown. Although if the sample size is more than 50, we use the Z distribution. Confidence Interval of Population Mean could be denoted as below

$$\bar{x} \pm \left(t_{n-1,\alpha/2}\right) \cdot \frac{S}{\sqrt{n}}$$

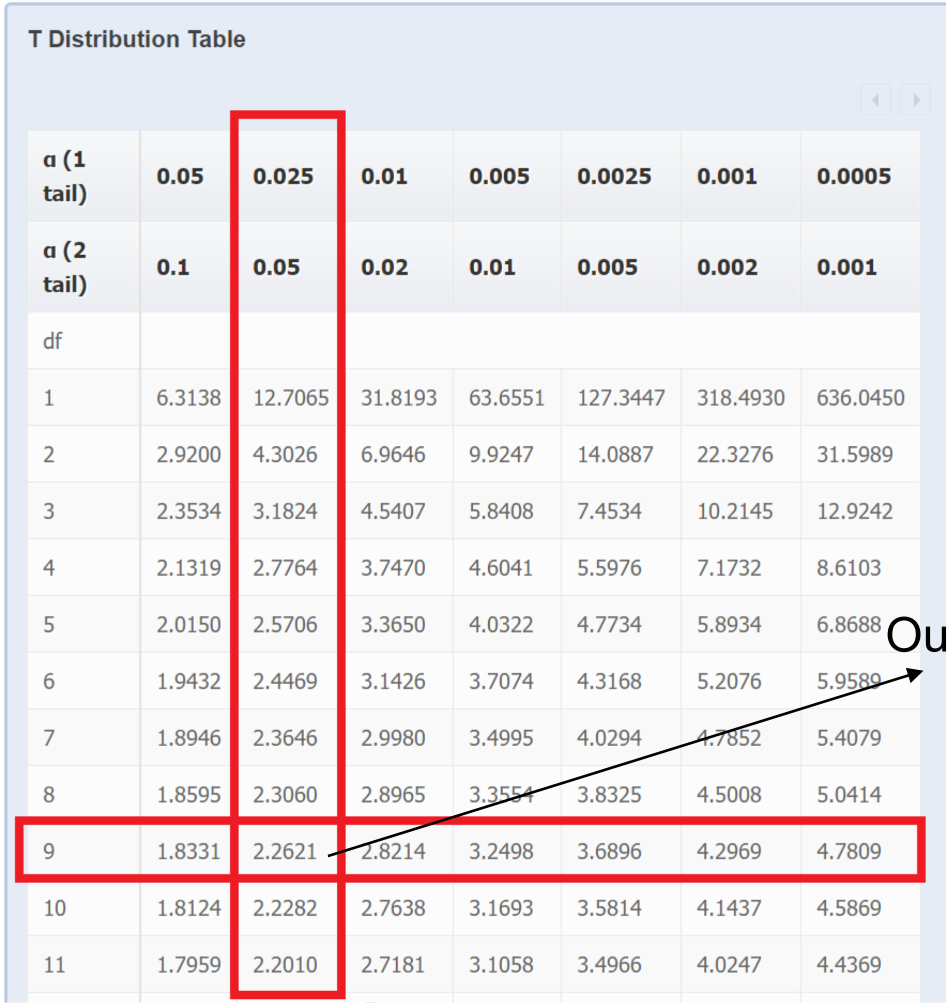
This value is called T-score. We get it from the T-distribution table. Check this link for the table https://www.easycalculation.com/statistics/t-distribution-critical-value-table.php

n-1 also called degree of freedom (df)

Legend
$\bar{x}$ = sample mean
S = sample standard variance
n = Number of Samples
α = 1 – Confidence level

Confidence Interval

- Let’s take a further example from our previous Purwadhika student age. We already have the sample mean 25, the number of sample is 10, and if we calculate the Standard Error it would be 1.16.
- Often, we take the 95% confidence level, it means the **α** is 0.05 and we divide it by 2 and which would result 0.025 (this is for one-tail test). As the sample size is 10, we would have df equal to 9. Now, lets calculate the T-score from the table.
- If we put everything in the Confidence Interval of the mean, it would be 25 ± 2.2621*1.16. This give our confidence interval of Purwadhika Student age between 22.38 and 27.62, with 95% confidence and 5% chance the population parameter is outside of the range.
- This interval would be more accurate with more data



T Distribution Table

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369

Our T-scor

Hypothesis

- A common aim in many studies it to check whether the data agree with certain predictions. These predictions are hypothesis about variables measured in the study. Ex: Would this new medicine work? Would richer people buy bigger apartment? etc.
- If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this: "If I…(do this to an independent variable)….then (this will happen to the dependent variable)." Ex: If the (Purwadhika Student wearing hat during study) then (it would affect their exam score)

Hypothesis Testing

- There are few steps in the data-driven decision making:
  1. Formulate a Hypotheses
  2. Find the right test
  3. Execute the test
  4. Make a Decision
- Hypothesis testing in statistics is a way to test the results of a survey or experiment to see if we have meaningful results. We are basically testing whether the results are valid by figuring out the odds that the results have happened by chance. If the results may have happened by chance, the experiment won't be repeatable and so has little use.
- We called the hypothesis testing a **Significance Test**. All significance test have five elements: Assumptions, Hypotheses, Test Statistic, P-value, and Conclusion.

Significance Test

- Before we go to test the hypothesis, we need to rephrase our Hypothesis for the significance test into two different hypothesis; The Null Hypothesis (**H_0**) and the Alternative Hypothesis (**H_1** or **H_A**)
- The null hypothesis is the hypothesis that is directly tested, the alternative hypothesis is a hypothesis that contradicts the null hypothesis.
- The significance test is analyzing the strength of the sample evidence against the null hypothesis (whether the data contradict the null hypothesis) hence suggesting that the alternative hypothesis is true.
- Often time, we set the null hypothesis as hypothesis that we want to reject so we could accept the alternative hypothesis

Significance Test

- All significance test require certain assumption for the tests to be valid. These assumption refer to:
  - Type of Data
  - The form of population distribution
  - Method of Sampling
  - Sample size
  - etc.
- Because of this assumption, there are many kind of Significance Test to be used, depend on the data we have

T-Test

- The t-test tells you how significant the differences between groups are; In other words it lets you know if those differences (measured in means/averages) could have happened by chance.
- The t score is a ratio between the **difference between two groups and the difference within the groups**. The larger the t score, the more difference there is between groups. The smaller the t score, the more similarity there is between groups. A t score of 3 means that the groups are three times as different *from* each other as they are within each other. When you run a t-test, the bigger the t-value, the more likely it is that the results are repeatable.
  - A large t-score tells you that the groups are different.
  - A small t-score tells you that the groups are similar.

T-Test

- There are **three main types of t-test:**
  - An Independent Samples t-test compares the means for two groups.
  - A Paired sample t-test compares means from the same group at different times (say, one year apart).
  - A One sample t-test tests the mean of a single group against a known mean.
- The test depend on the hypothesis we have or create

Hypothesis Testing

- Let's use our previous example, If the (Purwadhika Student wearing hat during study) then (it would affect their exam score) We could rephrase the hypothesis as:

  H_0: There are no significant difference between wearing the hat during study and exam score

  H_A: There are significant difference between wearing the hat during study and exam score

- From the hypothesis, it seems we could use Independent sample T-Test for hypothesis testing

Independent Sample T-Test

- It is the most common type of the T-Test, because it help compare the means of the two different dataset.
- This test is specifically used when:
  - We do not know the population mean or standard deviation.
  - We have two independent, separate samples.
- Let’s go back to our hypothesis, we state in the $\mathbf{H_0}$ that There are no significant difference between wearing the hat during study and exam score. This could be stated in the mathematical notation as:

$$\mu_{hat} = \mu_{no\ hat}$$

- The Alternative could be stated as:

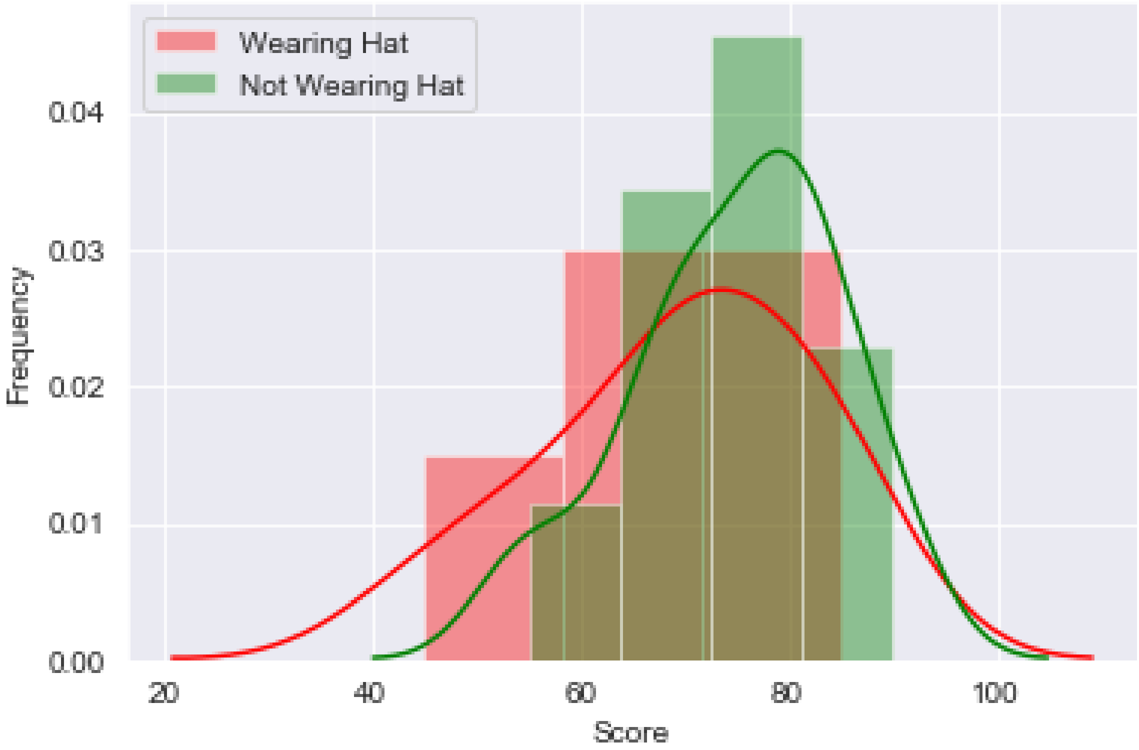
$$\mu_{hat} \neq \mu_{no\ hat}$$

Assumption of Independent T-Test

- **Assumption of Independence:** you need two independent, categorical groups that represent your independent variable. In our case, it is the ‘Wearing Hat’ feature
- **Assumption of normality**: the dependent variable should be approximately normally distributed. The dependent variable should also be measured on a continuous scale. In our example, the “test score” would be the dependent variable.
- **Assumption of Homogeneity of Variance:** The variances of the dependent variable should be equal.

Hypothesis Testing

Let's try the hypothesis testing with sample data of 20 students. We have Wearing Hat Feature as Independent Variable, and Exam Score as the Dependent Variable. The distribution of the data could be seen on image below. Its approximately to the normal distributed, but the sample is too small; but for the meantime, let’s work with this data.



Wearing Hat	Exam Score
Yes	90
No	85
Yes	100
Yes	45
No	80
No	70
No	90
Yes	85
Yes	55
No	80
Yes	75
Yes	60
No	70
Yes	85
No	90
Yes	85
No	55
No	75
No	65
Yes	90

Independent T-Test

- We need to put our data to this formula below:

$$t = \frac{\mu_A - \mu_B}{\sqrt{\left[\frac{\left(\Sigma A^2 - \frac{(\Sigma A)^2}{n_A}\right) + \left(\Sigma B^2 - \frac{(\Sigma B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

Legend:
μ_A: Mean of data set A
μ_B: Mean of data set B
ΣA^2: Sum of the squares of data set A
ΣB^2: Sum of the squares of data set B
n^A: Number of items in data set A
n^B: Number of items in data set B

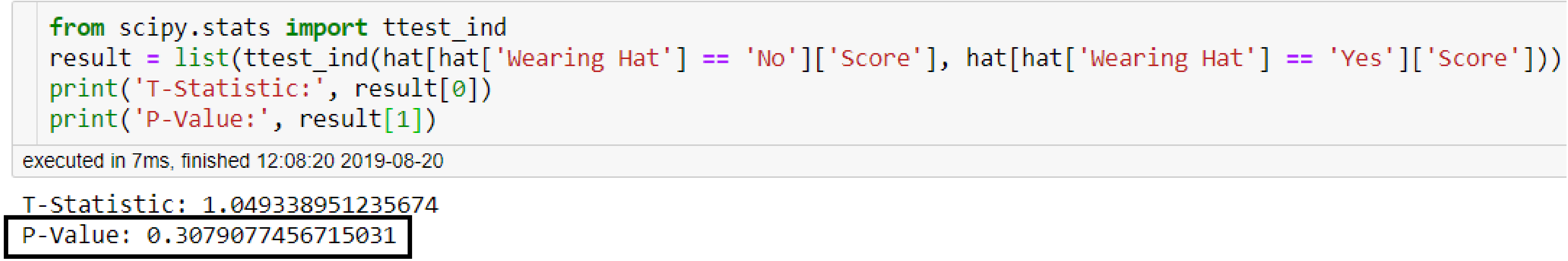
- Let’s assume the first dataset (A) is the Sample Not Wearing Hat, and the second dataset (B) is the sample Wearing Hat.
- The t-statistic is 1.05 after we calculate it
- We also need to calculate the degree of freedom (df), which have the formula as (n_A-1 + n_B-1). Our df is 18.

Independent T-Test

- Like before, we need to set our confidence level (set it as 95% or **α** = 0.05 , as it is the most common level)
- With our confidence level and degree of freedom, we need to check the t-table for our t-score comparison. Notice on the table there is 1-tail, and 2-tail on the **α**. This choice of 1 or 2 tail is depend on our hypothesis, in our case it is 2-tail as our alternative hypothesis could be either more or less rather than in one direction
- From the t-table, we get t-score of 2.10
- The calculated value of 1.0493 is less than the cutoff of t-score 2.10 from the table. Therefore p-value > .05. As the p-value is greater than the alpha level, we cannot conclude that there is a difference between means. It means we accept our Null Hypothesis.

Independent T-Test

- What is this P-value that suddenly shown up? P-value or probability value is the smallest value where we could still reject our null hypothesis. In statistical hypothesis testing, **for a given statistical model, the probability that, when the null hypothesis is true**, the statistical summary (such as the sample mean difference between two groups) would be equal to, or more extreme than, the actual observed results.
- To acquire the P-value, we need to come back to our t-table but often the best we could get is the P-value interval if we use the T-test. For this case, lets use python to get our P-value.



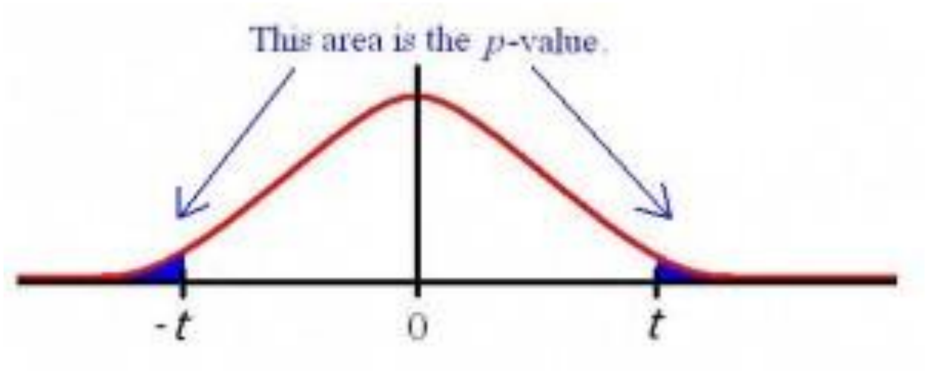
```python
from scipy.stats import ttest_ind
result = list(ttest_ind(hat[hat['Wearing Hat'] == 'No']['Score'], hat[hat['Wearing Hat'] == 'Yes']['Score']))
print('T-Statistic:', result[0])
print('P-Value:', result[1])
```

executed in 7ms, finished 12:08:20 2019-08-20

```
T-Statistic: 1.049338951235674
P-Value: 0.3079077456715031
```

Rejection Region

- The **rejection region** (also called a **critical region**, the place where you reject the null hypothesis) is a part of the hypothesis testing process. Specifically, it is an area of probability that tells if the hypothesis is probably true. The area of the rejection region is depending on the significance level we decide (often it is 95% confidence, so the rejection region is 5%). In image below is two-tailed test rejection region (2.5% on each side)



Every rejection region can be drawn on a probability distribution. The image shows a t-distribution with a two-tailed rejection region. It's also possible to have a rejection region in one tail only.

- We reject the hypothesis if the P-value is in the rejection region
- The type of the test is depending on the hypothesis we have. For more information check this link: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/

Independent T-Test

- As our P-value is around 0.3, the value is already bigger than our significance level which is 0.05. This mean, we get a statistical evidence to accept our Null hypothesis.
- Of course, P-value is only a probability value when **the observed data is sufficiently inconsistent with the Null hypothesis**. It was not mean that the P-value is the probability of the Null hypothesis is true. It is only one of the evidence.

Analysis of Variance (ANOVA)

- If we have test to examine the difference between the two population, we could also test if there is a difference between more than two population. This is where we use ANOVA. T-test is preferred when only two mean are present, but if want to compare more than two mean, we use ANOVA.
- Using ANOVA would mean we test the null hypothesis as stated below (assume we have population):

$$\mathbf{H_0}:\mu_1 = \mu_2 = \mu_3 = \mu_4$$

- With the alternative hypothesis ($\mathbf{H_A}$) that there are difference between the population mean. The weakness of ANOVA is that we would not know which population group precisely is different.

ANOVA

- There are many kind of ANOVA, but the most common to be used is either one-way ANOVA or two-way ANOVA. The differences is stated below.
- The main point difference is the number of Independent Variable is used

BASIS FOR COMPARISON	ONE-WAY ANOVA	TWO-WAY ANOVA
Meaning	One-way ANOVA is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	Two-way ANOVA is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple level of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles needs to be satisfied.

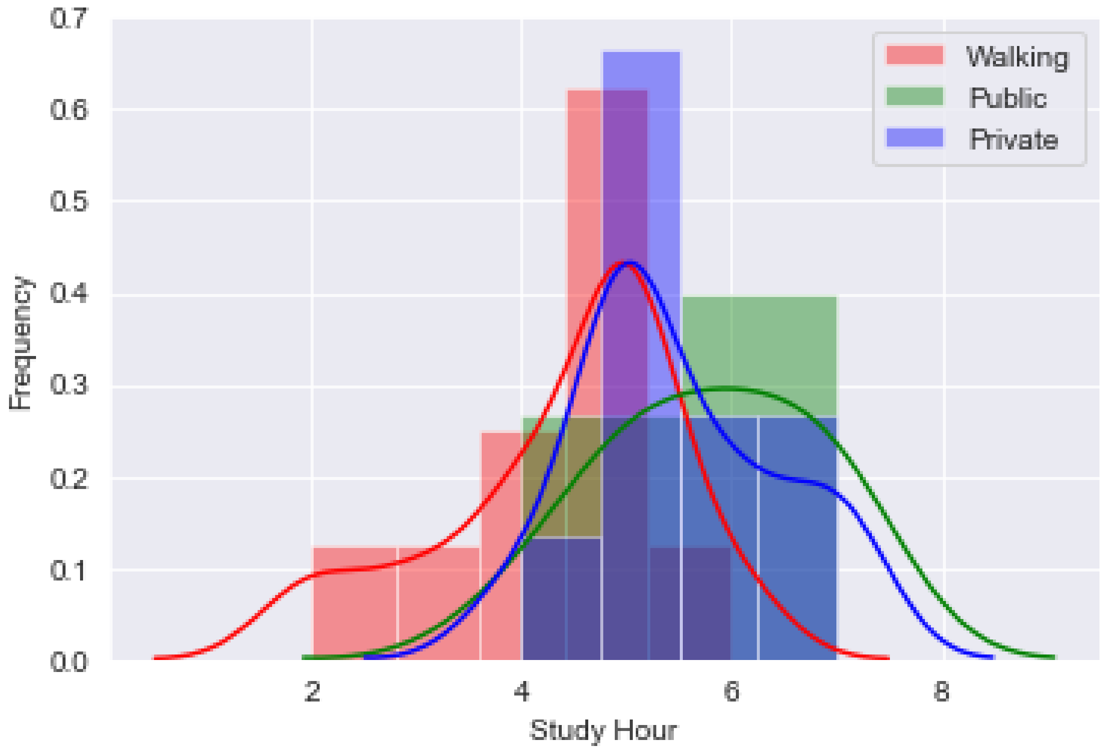
Purwadhika
Startup and Coding School

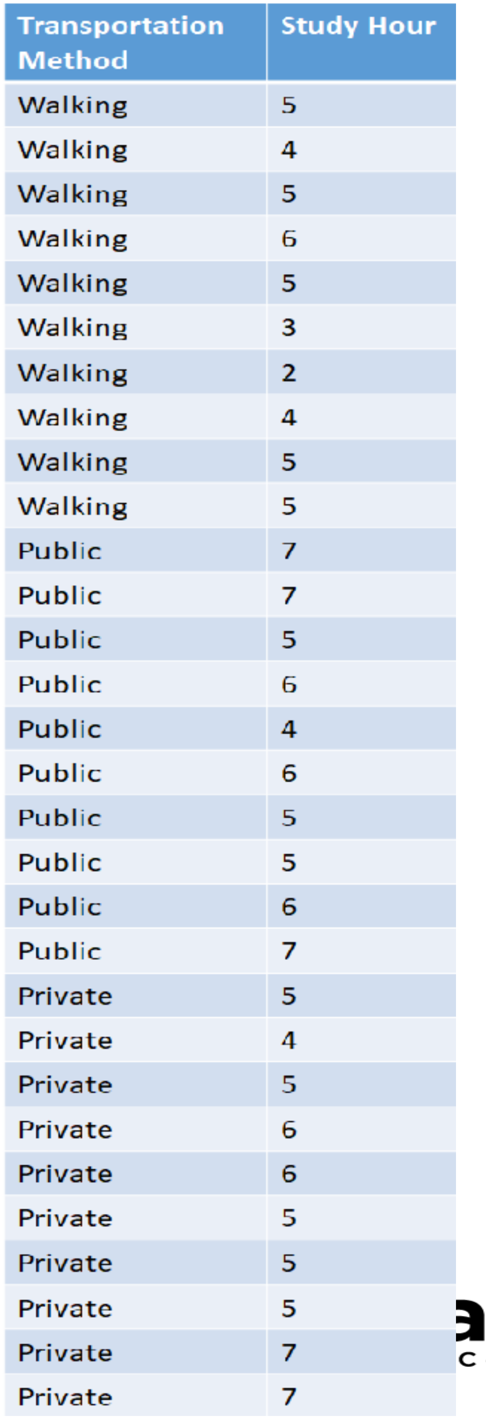
Assumption of One-Way ANOVA

- Normality – That each sample is taken from a normally distributed population
- Sample independence – that each sample has been drawn independently of the other samples
- Variance Equality – That the variance of data in the different groups should be the same
- Your dependent variable – here, “weight”, should be continuous – that is, measured on a scale which can be subdivided using increments (i.e. grams, milligrams)

One-Way ANOVA

Let's try the One-Way ANOVA with an example. Let's say right now we have sample data of 30 students at Purwadhika with Independent Variable of Transportation Method to Purwadhika and Dependent Variable hour of study. Its approximately to the normal distributed, lets try to work with this data.





Transportation Method	Study Hour
Walking	5
Walking	4
Walking	5
Walking	6
Walking	5
Walking	3
Walking	2
Walking	4
Walking	5
Walking	5
Public	7
Public	7
Public	5
Public	6
Public	4
Public	6
Public	5
Public	5
Public	6
Public	7
Private	5
Private	4
Private	5
Private	6
Private	6
Private	5
Private	5
Private	5
Private	7
Private	7

adhika
Coding School

One-Way ANOVA

- ANOVA basically is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. It test the hypothesis equality between population.
- How likely are the population means to be equal? This depends on 3 pieces of information from our samples:
  - the variance between sample means ($MS_{Between}$);
  - the variance within our samples (MS_{Within}) and
  - the sample sizes.
- We basically combine all this information into a single number: **test statistic F**

MSBetween

- We could get information of MS$_{\text{Between}}$ from two information, Sums of Squares Between Samples Mean (SS$_{\text{Between}}$) and Degree of Freedom in Mean Square Between(df$_{\text{Between}}$)
- **SS$_{\text{Between}}$ expresses the total amount of dispersion among the sample means.** Larger SS$_{\text{Between}}$ indicates that the sample means differ more. And the more different our sample means, the more likely that our population means differ as well. The formula is stated below:

$$SS_{between} = \Sigma\, n_j\, (\overline{X}_j - \overline{X})^2$$

Legend

$\overline{X}_j$ denotes a group mean;

$\overline{X}$ is the overall mean;

n_j is the sample size per group.

- When comparing k means, the degrees of freedom (df) is:

k - 1

Purwadhika
Startup and Coding School

MSBetween

- **MSBetween is basically the variance among sample means.** The formula could be stated as below: $$MS_{between} = \frac{SS_{between}}{df_{between}}$$

- Let’s try to calculate the $MS_{Between}$ from our data. First, we need to calculate the $SS_{Between.}$ Refer to the notebook for the values.

  $=(10*(4.4-5.23)^2 + 10*(5.8 -5.23)^2 + 10*(5.5 -5.23)^2)$

  $=6.889 + 3.249 + 0.729$

  $=10.867$

- $df_{between}$ is 3- 1 = 2
- $MS_{Between}$ is 10.867/2 = 5.4335

MSWithin

- If our population means are really equal, then what difference between sample means -$\text{MS}_{\text{Between}}$- can we reasonably expect? Well, this depends on the variance *within* subpopulations. $\text{MS}_{\text{Within}}$ basically is the variance within groups.
- Just like $\text{MS}_{\text{Between}}$ we get the information from **sums of squares within** ($\text{SS}_{\text{within}}$) indicates the total amount of dispersion within groups and **degrees of freedom within** ($\text{df}_{\text{Within}}$) is (n – k) for n observations and k groups.

- $\text{SS}_{\text{Within}}$ could be stated as below:

$$SS_{within} = \Sigma\,(X_i - \overline{X}_j)^2$$

Legend

$\overline{X}_j$ denotes a group mean;

X_i denotes an individual observation (“data point”).

MSWithin

- Just like $\text{MS}_{\text{Between,}}$ the calculation is the same. It stated as below:

$$MS_{within} = \frac{SS_{within}}{df_{within}}$$

- Let's try to calculate the $\text{MS}_{\text{Within}}$, we need to calculate the $\text{SS}_{\text{Within}}$ and $\text{df}_{\text{Within}}$ first. Refer to the notebook for the values.

$$=((4.4 – 5)^2 + (4.4 – 4)^2+...+(5.5-7)^2)$$

$$=30.5$$

$\text{df}_{\text{Within}} = 30 -3 = 27$

$\text{MS}_{\text{Within}} = 30.5/27$

$=1.1296$

F-Statistic

- Now, we ready to calculate our F-Statistic:

$$F = \frac{MS_{between}}{MS_{within}}$$

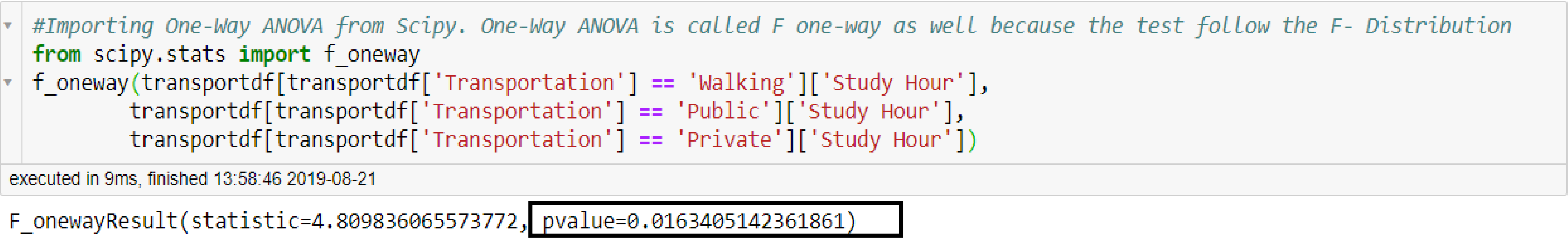
F = 5.4335/ 1.1296

F = 4.8101

Now we try to find the F-critical value from the F-distribution table using df$_{\text{Within}}$ and df$_{\text{Between}}$. Check this link for the F-Distribution Table: http://homepages.wmich.edu/~hillenbr/619/AnovaTable.pdf

F-Statistic

- Use the df_{Within} for the column, and $df_{Between}$ for the row. Using the table we get F(2,27)-critical value 3.35
- As our critical value is smaller than our F-statistic (3.35 < 4.8101), we could assume that our P-value would be less than our significant level 0.05. If we want to get the exact number of P-value, we need to use the power of computation.



```
#Importing One-Way ANOVA from Scipy. One-Way ANOVA is called F one-way as well because the test follow the F- Distribution
from scipy.stats import f_oneway
f_oneway(transportdf[transportdf['Transportation'] == 'Walking']['Study Hour'],
        transportdf[transportdf['Transportation'] == 'Public']['Study Hour'],
        transportdf[transportdf['Transportation'] == 'Private']['Study Hour'])
```

executed in 9ms, finished 13:58:46 2019-08-21

```
F_onewayResult(statistic=4.809836065573772, pvalue=0.0163405142361861)
```

- We could see that the P-value is 0.016 which is less than 0.05. It mean we have evidence to reject our Null hypothesis and accept the Alternative hypothesis that there are difference between the population mean.

Parametric vs Non-Parametric

- Parametric methods are often those for which we know that the population is approximately normal and non-parametric methods techniques for which we do not have to make any assumption of parameters for the population we are studying.

	Parametric	**Non-parametric**
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

Goodness of fit test

- The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution). In other words, it tells you if your sample data represents the data you would expect to find in the actual population. Goodness of fit tests commonly used in statistics are:
  - The chi-square.
  - Anderson-Darling.
  - Shipiro-Wilk.
- Goodness of fit test is also a hypothesis testing although we often aim in this test for the test for fail to reject the null hypothesis rather than rejecting the null hypothesis. Formally, the hypothesis are stated as:
  - H0: The sample is following the designated distribution
  - H1: The sample is not following the designated distribution

Correlation

- Correlation (to be exact Correlation in Statistic) is a measure of a mutual relationship between two variables whether they are causal or not. This degree of measurement could be measured on any kind of data type (Continuous and Continuous, Categorical and Categorical, Continuous and Categorical).
- Although correlation stated how it measured the mutual relationship, the presence of correlation measurement does not provide strong evidence toward causation.

Covariance

- In statistics, covariance is a measure of the association between variable X and Y. To be exact, it measures the linear relationship tendency of the variables.

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- Covariance is calculated by subtracting each member of the variable by its mean (Centering the data). These centered scores are multiplied to measure whether the increase or decrease in one variable is associated with one another. Finally, the expected value (*E*) of these centered scores is calculated as a summary of the association. The expected value itself in another term is the average or mean (*μ)*.

Pearson Correlation Coefficient

- Pearson Correlation is one of the most used correlations during the data analysis process. Pearson correlation measures the linear relationship between variable continuous X and variable continuous Y and has a value between 1 and -1. In other words, the Pearson Correlation Coefficient measures the relationship between 2 variables via a line.

$$\rho_{X,Y} = \frac{E[(X-E[X])(Y-E[Y])]}{\sigma_X \sigma_Y}$$

- If you notice, the top side of the fraction equation (numerator) is similar to the covariance equation we previously just discussed. This means we could also state the Pearson Correlation Coefficient as below.

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Pearson Correlation Coefficient

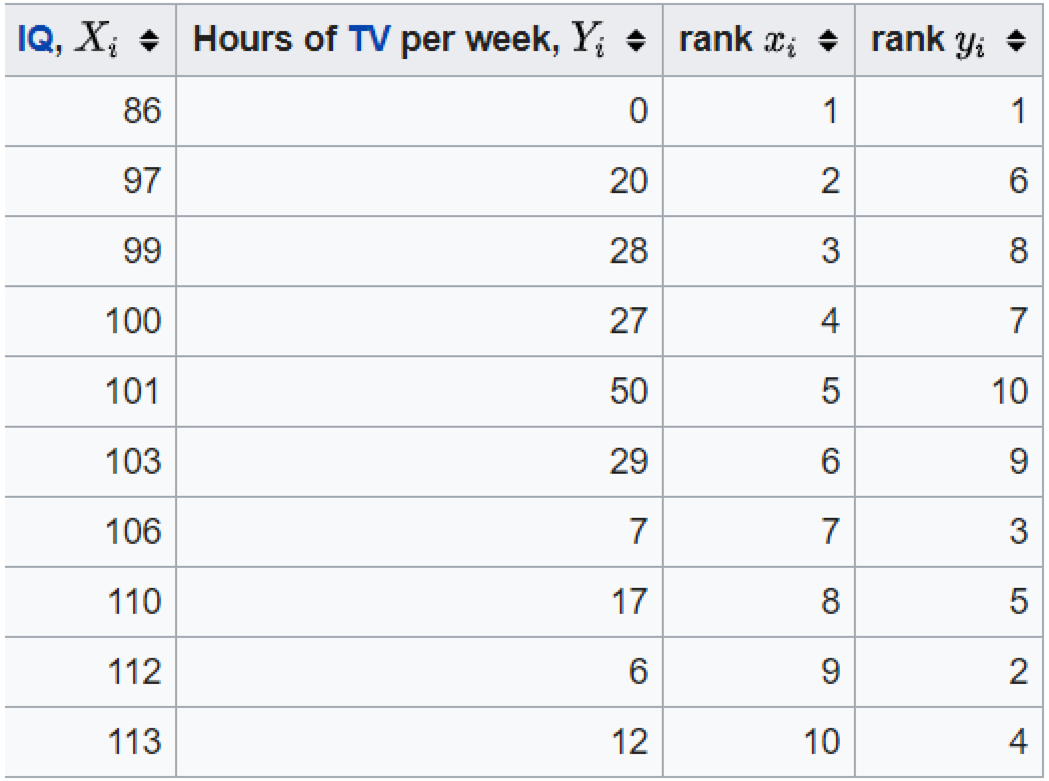
- When the correlation coefficient is **closer to value 1**, it means there is a **positive relationship** between variable X and Y. A positive relationship indicates an increase in one variable associated with an increase in the other. On the other hand, **the closer correlation coefficient is to -1** would mean there is a **negative relationship** which is the increase in one variable would result in a decrease in the other. **If *X* and *Y* are independent**, then the **correlation coefficient is close to 0** although the Pearson correlation can be small even if there is a strong relationship between two variables.

Spearman Rank Correlation

- Unlike the Pearson Correlation Coefficient, Spearman Rank Correlation measures the monotonic relationship (Strictly increase or decrease, not both) between two variables and measured by the rank order of the values. The correlation still measured between continuous variable X and continuous variable Y, although the Spearman Rank Correlation method still relevant to the discrete ordinal variable.

$$\rho_{rank_x}, \rho_{rank_y} = \frac{cov(rank_x, rank_y)}{\sigma_{rank_x}, \sigma_{rank_y}}$$

Spearman Rank Correlation



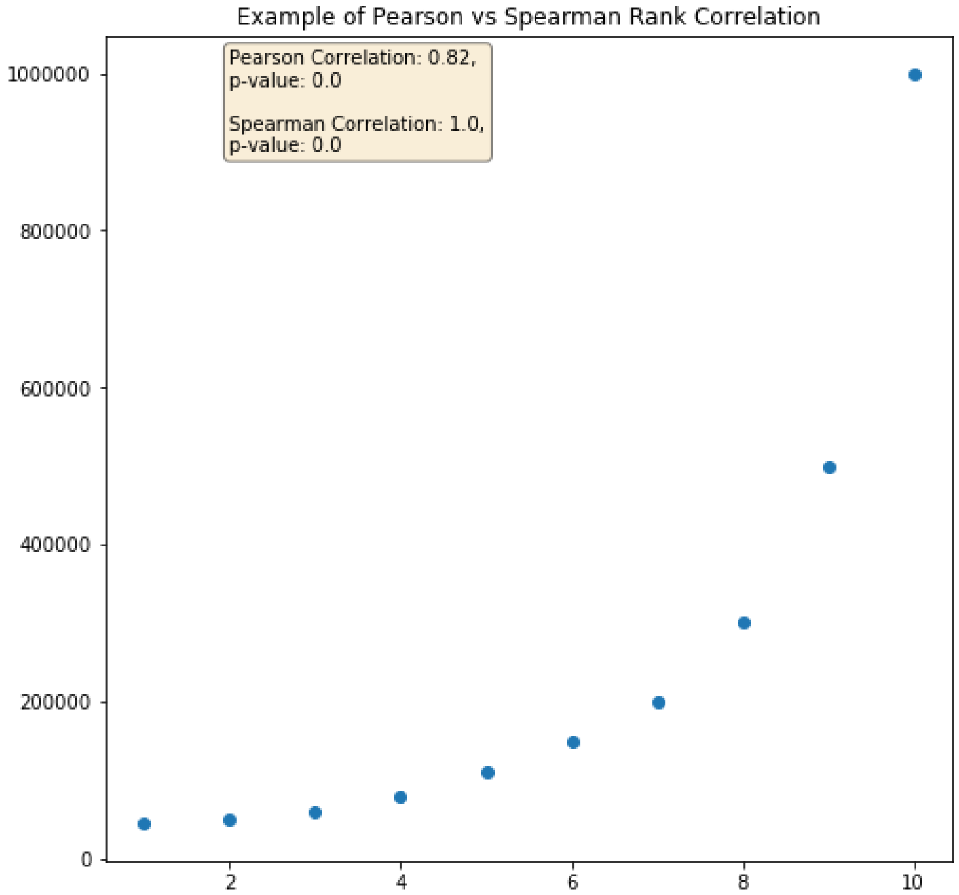
IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i
86	0	1	1
97	20	2	6
99	28	3	8
100	27	4	7
101	50	5	10
103	29	6	9
106	7	7	3
110	17	8	5
112	6	9	2
113	12	10	4

- The easier way to understand the ranking is that we order the data from the smallest to the largest and we assigning the ranking depending on the data order. 1 is the smallest ranking, it means that rank 1 is assigned to the smallest value of the respective column. Why the respective column? we could see from the table that we rank the data based on their respective columns and because we want to know the covariance between the ranking of column X and column Y; we assign the rank w.r.t. each column. To be precise, what we want is the ranking difference between each row.

Spearman Rank Correlation

- Spearman rank correlation could be interpreted similarly as the Pearson correlation coefficient as their value falls between -1 to 1. **The closer the score to 1** means that there is a **positive monotonic relationship** between the variable (the data keep increasing) and vice versa. If **variable X and variable Y independent**, the value would be **equal to 0**.

Difference between Pearson and Spearman Correlation



- Since our data showing the perfect positive monotonic relationship (the data is always increasing) and non-linear relationships; our Spearman correlation is equal to 1. In this case, the Pearson relationship is weaker but still shown a strong association as there is a partial linearity relationship between the data