Unit-5
Regression Analysis, Hypothesis Testing

Content: Regression analysis, Linear Regression, Hypothesis testing, Tests

What Is a Regression?

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, ***"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the***

regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Why do we use Regression Analysis?

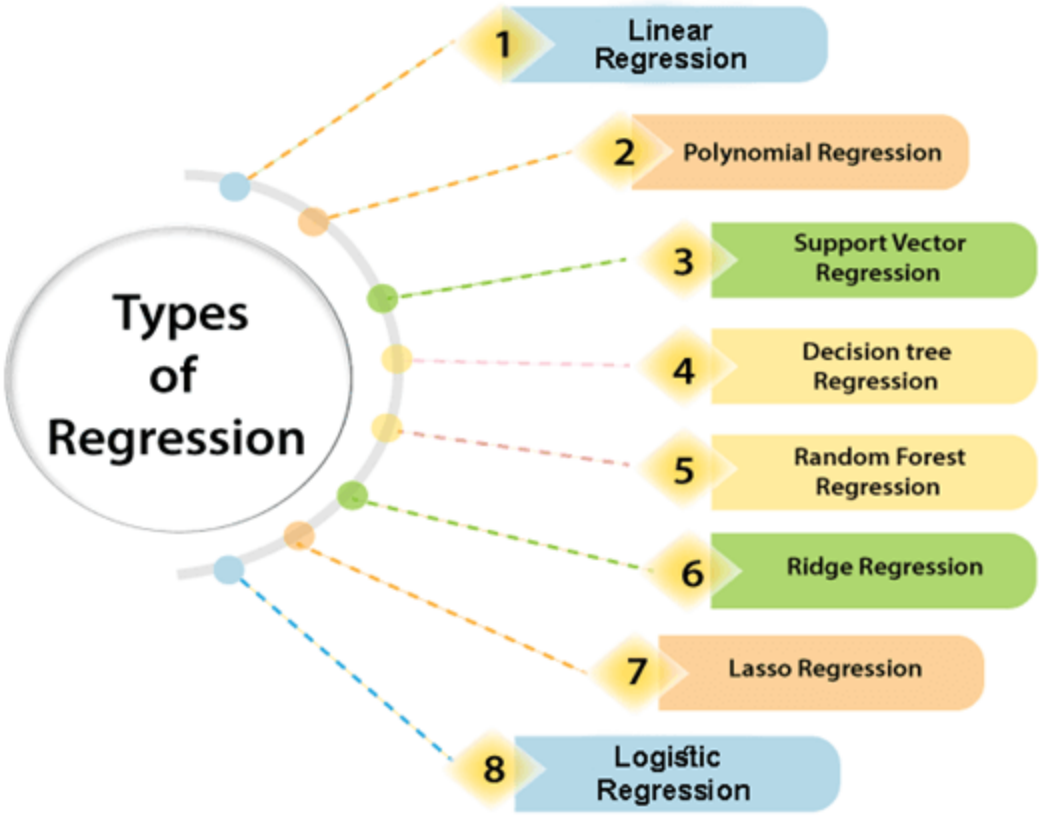
As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

Types of Regression

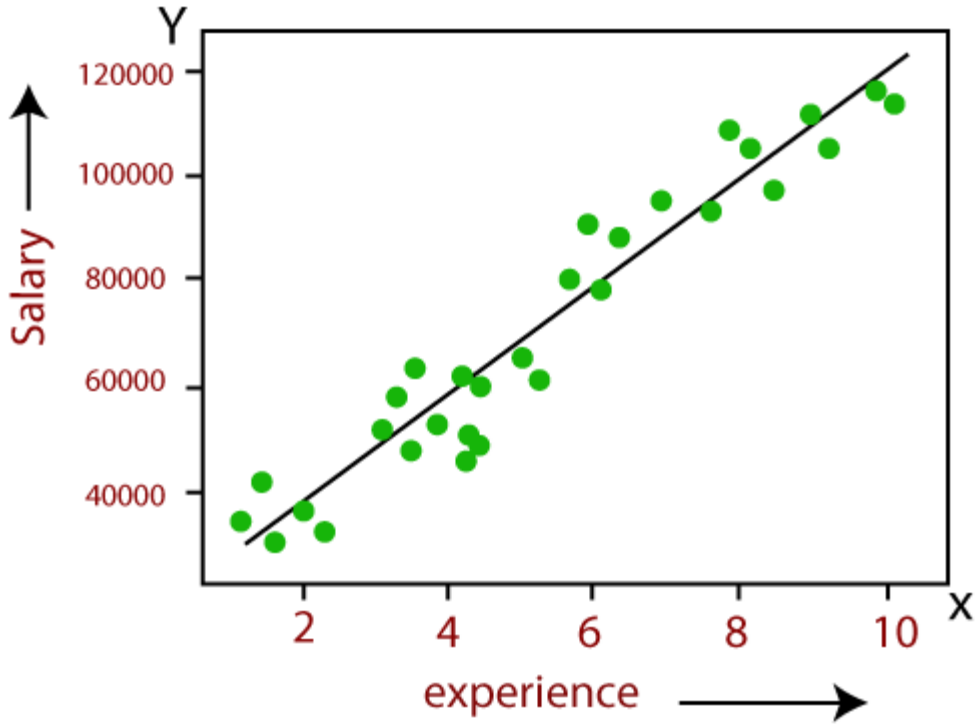
There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

- **Linear Regression**
- **Logistic Regression**
- **Polynomial Regression**
- **Support Vector Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Ridge Regression**
- **Lasso Regression:**



Linear Regression:

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



- Below is the mathematical equation for Linear regression:

1. Y= aX+b

Here, Y = dependent variables (target variables),
X= Independent variables (predictor variables),
a and b are the linear coefficients

Some popular applications of linear regression are:

- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

What is Simple Linear Regression?

Linear regression is called to be a simple linear regression if there is only one independent variable. And mathematically it can be represented as

y=b0+b1x1+E

Where :

y: dependent variable

b0: intercept

b1: coefficient of x1(independent variable)

E: error

What is Multiple Linear Regression?

Linear regression is called multiple linear regression if there is more than one independent variable. And mathematically it can be represented as

y= b0+b1X1+b2X2+…+bnXn+E

Where :

y: dependent variable

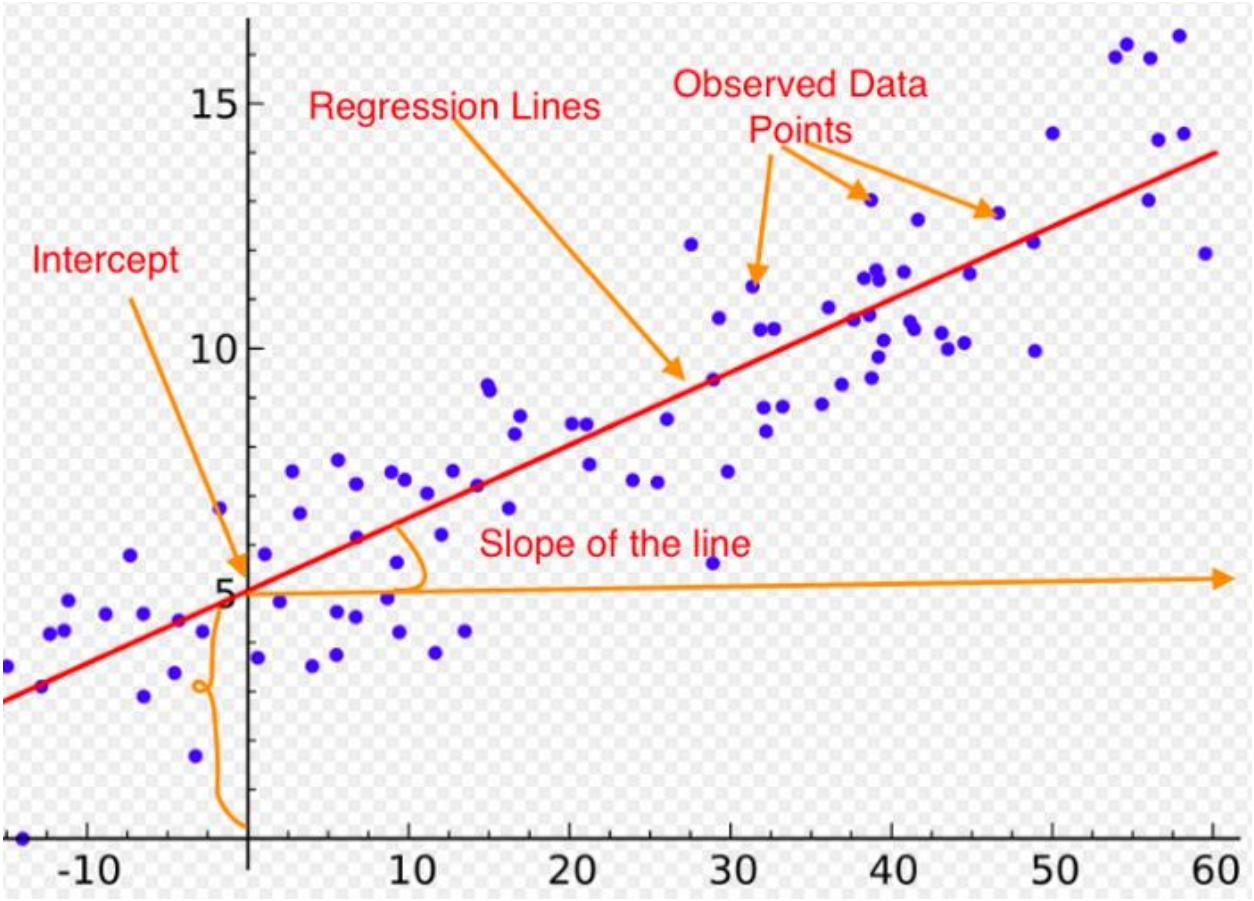
b0: intercept

b1: coefficient of x1(independent variable)

b2: coefficient of x2(independent variable)
bn: coefficient of xn (independent variable)
E: Error

What is Regression Line?

The **Regression line** is a straight line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest. The formula for the best-fitting line (or regression line) is
y = a + bx,
where**:**
"b" is the slope of the *line*
"a" is the y-intercept.
"x" is an explanatory variable.
"y" is a dependent variable



Regression line attempts to define the predicted value of "y" (dependent variable) for a given value of "x" (independent variable). The best-fit regression line attempts to minimise the sum of the squared distances between the observed(actual) data points and the predicted ones.
The intercept of regression lines helps us to estimate the value of "y" (dependent variable), having no effects of "x" (independent variable).

Why is multicollinearity a problem in linear regression?

If independent variables are not purely independent of each other than they are correlated. And as a result, it leads to change in one variable that will induce the shift in associated correlated variables. If they possess a strong correlation, then it is more difficult to keep one variable unchanged with a change to the other variable.
Hence, this causes the problem for linear regression models to estimate the relationship between a dependent variable and independent variables, as correlated independent variables change simultaneously.

Multicollinearity reduces the power of linear regression models to identify significantly important independent variables.

1. Constant variance for different values of the dependent variable will have the same variance in their error: This is also called homoscedasticity in error. Let's understand, Why homoscedasticity in error is important in linear regression?

Heteroscedasticity is the antonym of homoscedasticity. Due to heteroscedasticity, it becomes difficult to determine the coefficients of standard errors. Also, we get an unreliable standard error.

1. Independence of errors: Errors are the deviation of predicted values from the actual values. This assumes that the errors of dependent variables are random and are in no correlation with each other
2. The quantitative data condition: Regression can only be performed on quantitative data. Regression analysis is not a good technique to find the trend in qualitative data.

Implementation in R

In R programming, **lm()** function is used to create linear regression model.
Syntax: lm(formula)
Parameter:
formula: represents the formula on which data has to be fitted To know about more optional parameters, use below command in console: help("lm")
Multiple Regression
Multiple regression is another type of regression analysis technique that is an extension of the linear regression model as it uses more than one predictor variables to create the model.
Mathematically,

To fit a linear regression model in R, we can use the **lm()** command.

To view the output of the regression model, we can then use the **summary()** command.

This tutorial explains how to interpret every value in the regression output in R.
Interpreting Regression Output in R

The following code shows how to fit a multiple linear regression model with the built-in **mtcars** dataset using hp, drat, and wt as predictor variables and mpg as the response variable:

```
#fit regression model using hp, drat, and wt as predictors
model<- lm(mpg ~ hp + drat + wt, data = mtcars)

#view model summary
```

```
summary(model)

Call:
lm(formula = mpg ~ hp + drat + wt, data = mtcars)

Residuals:
   Min     1Q  Median     3Q    Max
-3.3598 -1.8374 -0.5099  0.9681  5.7078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.394934   6.156303   4.775 5.13e-05 ***
hp          -0.032230   0.008925  -3.611 0.001178 **
drat         1.615049   1.226983   1.316 0.198755
wt          -3.227954   0.796398  -4.053 0.000364 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.561 on 28 degrees of freedom
Multiple R-squared:  0.8369,           Adjusted R-squared:  0.8194
F-statistic: 47.88 on 3 and 28 DF,  p-value: 3.768e-11
```

Implementation in R

Multiple regression in R programming uses the same **lm()** function to create the model.
Syntax: lm(formula, data)
Parameters:
formula: represents the formula on which data has to be fitted
data: represents dataframe on which formula has to be applied

Here is how to interpret every value in the output:
Call

```
Call:
lm(formula = mpg ~ hp + drat + wt, data = mtcars)
```

This section reminds us of the formula that we used in our regression model. We can see that we used **mpg** as the response variable and **hp**, **drat**, and **wt** as our predictor variables. Each variable came from the dataset called **mtcars**.
Residuals

```
Residuals:
   Min     1Q  Median     3Q    Max
-3.3598 -1.8374 -0.5099  0.9681  5.7078
```

This section displays a summary of the distribution of residuals from the regression model. Recall that a residual is the difference between the observed value and the predicted value from the regression model.

The minimum residual was **-3.3598**, the median residual was **-0.5099** and the max residual was **5.7078**.

Coefficients

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.394934   6.156303   4.775 5.13e-05 ***
hp          -0.032230   0.008925  -3.611 0.001178 **
drat         1.615049   1.226983   1.316 0.198755
wt          -3.227954   0.796398  -4.053 0.000364 ***

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This section displays the estimated coefficients of the regression model. We can use these coefficients to form the following estimated regression equation:

mpg = 29.39 – .03*hp + 1.62*drat – 3.23*wt

For each predictor variable, we’re given the following values:

Estimate: The estimated coefficient. This tells us the average increase in the response variable associated with a one unit increase in the predictor variable, assuming all other predictor variables are held constant.

Std. Error: This is the standard error of the coefficient. This is a measure of the uncertainty in our estimate of the coefficient.

t value: This is the t-statistic for the predictor variable, calculated as (Estimate) / (Std. Error).

Pr(>|t|): This is the p-value that corresponds to the t-statistic. If this value is less than some alpha level (e.g. 0.05) than the predictor variable is said to be statistically significant.

If we used an alpha level of α = .05 to determine which predictors were significant in this regression model, we’d say that **hp** and **wt** are statistically significant predictors while **drat** is not.

Assessing Model Fit

```
Residual standard error: 2.561 on 28 degrees of freedom
Multiple R-squared:  0.8369,     Adjusted R-squared:  0.8194
F-statistic: 47.88 on 3 and 28 DF,  p-value: 3.768e-11
```

This last section displays various numbers that help us assess how well the regression model fits our dataset.

Residual standard error: This tells us the average distance that the observed values fall from the regression line. The smaller the value, the better the regression model is able to fit the data.

The degrees of freedom is calculated as n-k-1 where n = total observations and k = number of predictors. In this example, mtcars has 32 observations and we used 3 predictors in the regression model, thus the degrees of freedom is 32 – 3 – 1 = 28.

Multiple R-Squared: This is known as the coefficient of determination. It tells us the proportion of the variance in the response variable that can be explained by the predictor variables.

This value ranges from 0 to 1. The closer it is to 1, the better the predictor variables are able to predict the value of the response variable.

Adjusted R-squared: Ths is a modified version of R-squared that has been adjusted for the number of predictors in the model. It is always lower than the R-squared.

The adjusted R-squared can be useful for comparing the fit of different regression models that use different numbers of predictor variables.

F-statistic: This indicates whether the regression model provides a better fit to the data than a model that contains no independent variables. In essence, it tests if the regression model as a whole is useful.

p-value: This is the p-value that corresponds to the F-statistic. If this value is less than some significance level (e.g. 0.05), then the regression model fits the data better than a model with no predictors.

When building regression models, we hope that this p-value is less than some significance level because it indicates that the predictor variables are actually useful for predicting the value of the response variable.

Assumptions in Regression

Regression is a parametric approach. 'Parametric' means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets which doesn't fulfill its assumptions. Therefore, for a successful regression analysis, it's essential to validate these assumptions.

So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

Let's look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests

that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
5. The error terms must be normally distributed.

What if these assumptions get violated ?

Let's dive into specific assumptions and learn about their outcomes (if violated):

1. Linear and Additive: If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

How to check: Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms (X, X^2, X^3) in your model to capture the non-linear effect.

2. Autocorrelation: The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients. Let's understand narrow prediction intervals with an example:

For example, the least square coefficient of X^1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in presence of autocorrelation, the standard error reduces to 1.20. As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

How to check: Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

3. Multicollinearity: This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you'll end up with an incorrect conclusion that a variable strongly / weakly affects target variable.

Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That's not good!
How to check: You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

4. Heteroskedasticity: The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.
How to check: You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

5. Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.
How to check: You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

Interpretation of Regression Plots
Until here, we've learnt about the important regression assumptions and the methods to undertake, if those assumptions get violated.
But that's not the end. Now, you should know the solutions also to tackle the violation of these assumptions. In this section, I've explained the 4 regression plots along with the methods to overcome limitations on assumptions.

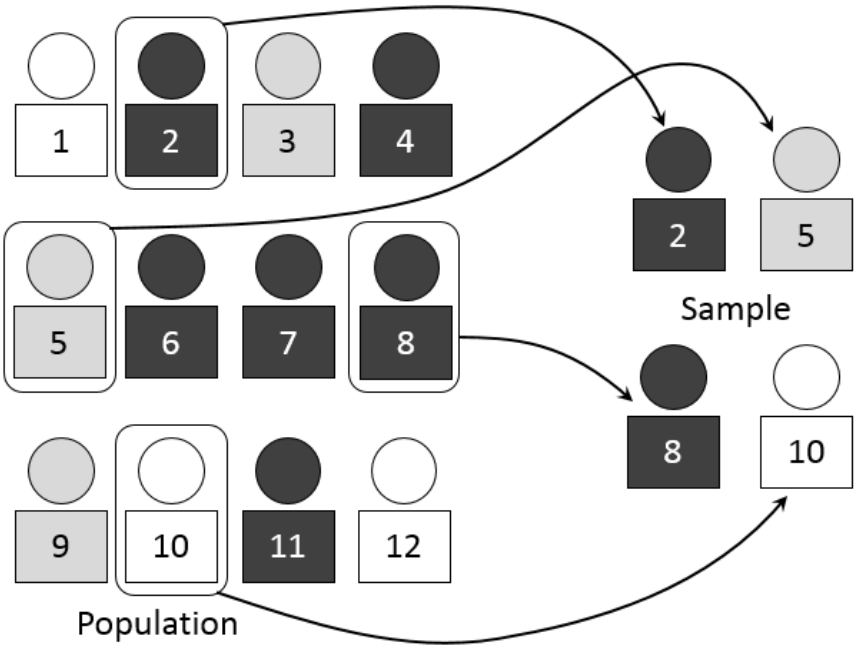
Sampling

2. What are population and sample in statistics?

Two basic but vital concepts in statistics are those of population and sample. We can define them as follows.

- **Population** is the entire group that you wish to draw data from (and subsequently draw conclusions about). While in day-to-day life, the word is often used to describe groups of people (such as the population of a country) in statistics, it can apply to any group from which you will collect information. This is often people, but it could also be cities of the world, animals, objects, plants, colors, and so on.

- **A sample** is a representative group of a larger population. Random sampling from representative groups allows us to draw broad conclusions about an overall population. This approach is commonly used in polling. Pollsters ask a small group of people about their views on certain topics. They can then use this information to make informed judgments about what the larger population thinks. This saves time, hassle, and the expense of extracting data from an entire population (which for all practical purposes is usually impossible).



What is descriptive statistics?

Descriptive statistics are used to describe the characteristics or features of a dataset. The term 'descriptive statistics' can be used to describe both individual quantitative observations (also known as 'summary statistics') as well as the overall process of obtaining insights from these data. We can use descriptive statistics to describe both an entire population or an individual sample. Because they are merely explanatory, descriptive statistics are not heavily concerned with the differences between the two types of data.

What is inferential statistics?

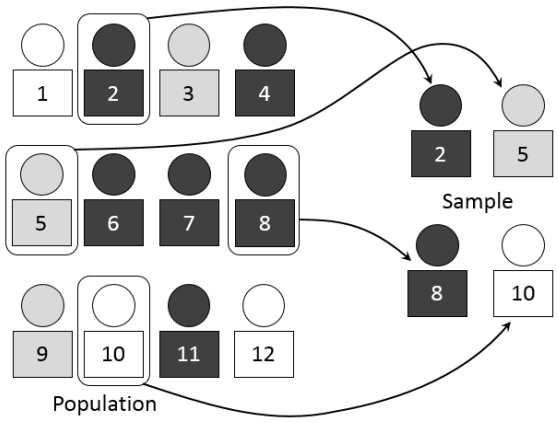
So, we've established that descriptive statistics focus on summarizing the key features of a dataset. Meanwhile, inferential statistics focus on making generalizations about a larger

population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.

Unsurprisingly, the accuracy of inferential statistics relies heavily on the sample data being both accurate and representative of the larger population. To do this involves obtaining a random sample. If you've ever read news coverage of scientific studies, you'll have come across the term before. The implication is always that random sampling means better results. On the flipside, results that are based on biased or non-random samples are usually thrown out. Random sampling is very important for carrying out inferential techniques, but it is not always straightforward!

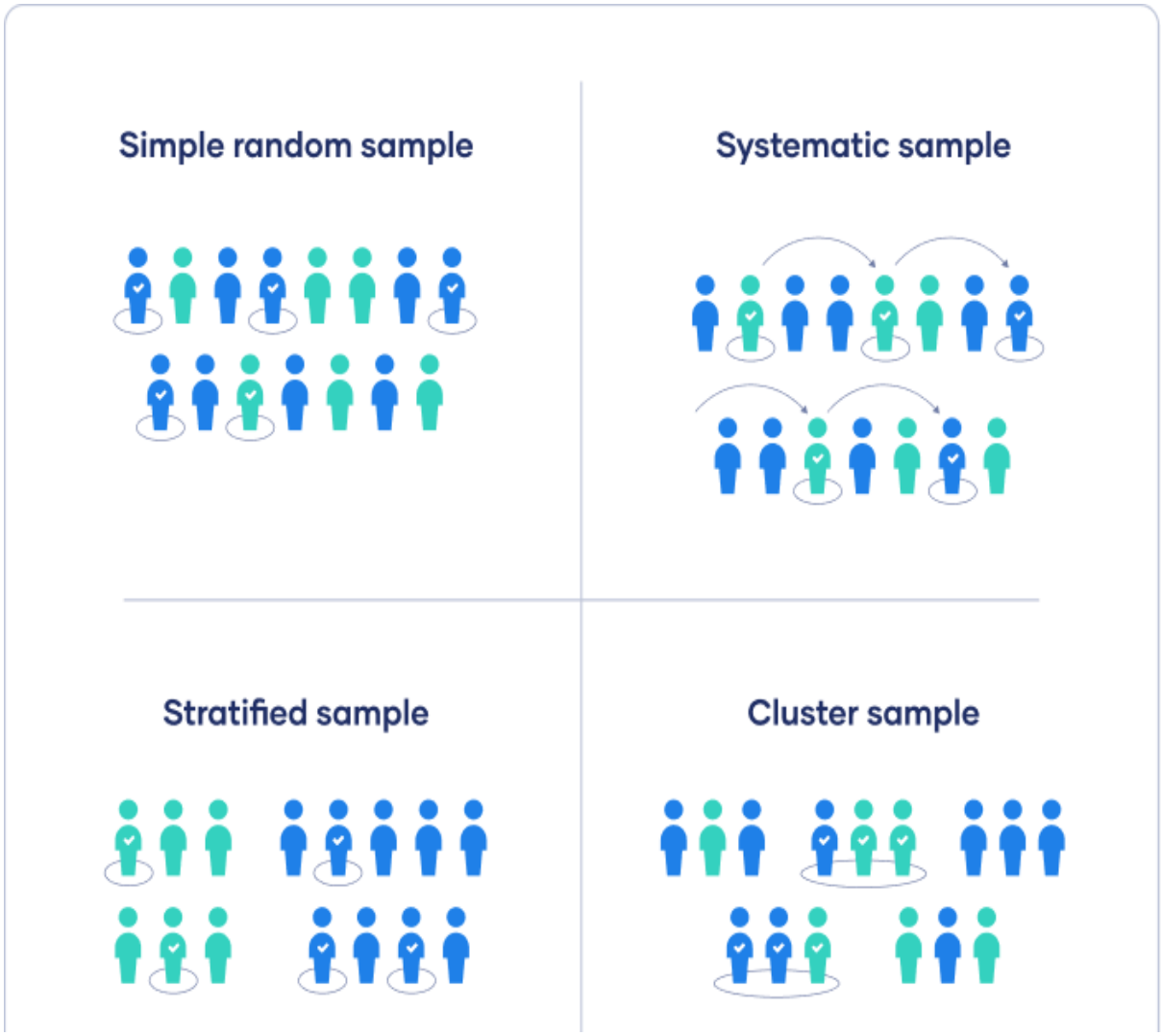
Two basic but vital concepts in statistics are those of population and sample. We can define them as follows.

- **Population** is the entire group that you wish to draw data from (and subsequently draw conclusions about). While in day-to-day life, the word is often used to describe groups of people (such as the population of a country) in statistics, it can apply to any group from which you will collect information. This is often people, but it could also be cities of the world, animals, objects, plants, colors, and so on.
- **A sample** is a representative group of a larger population. Random sampling from representative groups allows us to draw broad conclusions about an overall population. This approach is commonly used in polling. Pollsters ask a small group of people about their views on certain topics. They can then use this information to make informed judgments about what the larger population thinks. This saves time, hassle, and the expense of extracting data from an entire population (which for all practical purposes is usually impossible).



Types of Sampling

- **Probability sampling** involves random selection, allowing you to make strong statistical inferences about the whole group.
- **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect data



The image illustrates the concept of population and sample. Using random sample measurements from a representative group, we can estimate, predict, or infer characteristics about the larger population. While there are many technical variations on this technique, they all follow the same underlying principles.

What is inferential statistics?

So, we've established that descriptive statistics focus on summarizing the key features of a dataset. Meanwhile, inferential statistics focus on making generalizations about a larger population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.

Unsurprisingly, the accuracy of inferential statistics relies heavily on the sample data being both accurate and representative of the larger population. To do this involves obtaining a random sample. If you've ever read news coverage of scientific studies, you'll have come across the term before. The implication is always that random sampling means better results. On the flipside, results that are based on biased or non-random samples are usually thrown out. Random sampling is very important for carrying out inferential techniques, but it is not always straightforward!

What Is Hypothesis Testing in Statistics?

Hypothesis Testing is a type of statistical analysis in which you put your assumptions about a population parameter to the test. It is used to estimate the relationship between 2 statistical variables.

Let's discuss few examples of statistical hypothesis from real-life -

- A teacher assumes that 60% of his college's students come from lower-middle-class families.
- A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.

Now that you know about hypothesis testing, look at the two types of hypothesis testing in statistics.

Null Hypothesis and Alternate Hypothesis

The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.
H0 is the symbol for it, and it is pronounced H-naught.
The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. H1 is the symbol for it.
Let's understand this with an example.
A sanitizer manufacturer claims that its product kills 95 percent of germs on average.
To put this company's claim to the test, create a null and alternate hypothesis.
H0 (Null Hypothesis): Average = 95%.
Alternative Hypothesis (H1): The average is less than 95%.
Another straightforward example to understand this concept is determining whether or not a coin is fair and balanced. The null hypothesis states that the probability of a show of heads is equal to the likelihood of a show of tails. In contrast, the alternate theory states that the probability of a show of heads and tails would be very different.

Example:

A company is claiming that their average sales for this quarter are 1000 units. This is an example of a simple hypothesis.
Suppose the company claims that the sales are in the range of 900 to 1000 units. Then this is a case of a composite hypothesis.

One-Tailed and Two-Tailed Hypothesis Testing

The One-Tailed test, also called a directional test, considers a critical region of data that would result in the null hypothesis being rejected if the test sample falls into it, inevitably meaning the acceptance of the alternate hypothesis.
In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.
In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.
If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

Example:

Suppose H0: mean = 50 and H1: mean not equal to 50
According to the H1, the mean can be greater than or less than 50. This is an example of a Two-tailed test.
In a similar manner, if H0: mean >=50, then H1: mean <50
Here the mean is less than 50. It is called a One-tailed test.

Type 1 and Type 2 Error

A hypothesis test can result in two types of errors.
Type 1 Error: A Type-I error occurs when sample results reject the null hypothesis despite being true.

Type 2 Error: A Type-II error occurs when the null hypothesis is not rejected when it is false, unlike a Type-I error.

Example:

Suppose a teacher evaluates the examination paper to decide whether a student passes or fails.

H0: Student has passed

H1: Student has failed

Type I error will be the teacher failing the student [rejects H0] although the student scored the passing marks [H0 was true].

Type II error will be the case where the teacher passes the student [do not reject H0] although the student did not score the passing

Level of Significance

The alpha value is a criterion for determining whether a test statistic is statistically significant. In a statistical test, Alpha represents an acceptable probability of a Type I error. Because alpha is a probability, it can be anywhere between 0 and 1. In practice, the most commonly used alpha values are 0.01, 0.05, and 0.1, which represent a 1%, 5%, and 10% chance of a Type I error, respectively (i.e. rejecting the null hypothesis when it is in fact correct).

P-Value

A p-value is a metric that expresses the likelihood that an observed difference could have occurred by chance. As the p-value decreases the statistical significance of the observed difference increases. If the p-value is too low, you reject the null hypothesis.

Here you have taken an example in which you are trying to test whether the new advertising campaign has increased the product's sales. The p-value is the likelihood that the null hypothesis, which states that there is no change in the sales due to the new advertising campaign, is true. If the p-value is .30, then there is a 30% chance that there is no increase or decrease in the product's sales. If the p-value is 0.03, then there is a 3% probability that there is no increase or decrease in the sales value due to the new advertising campaign. As you can see, the lower the p-value, the chances of the alternate hypothesis being true increases, which means that the new advertising campaign causes an increase or decrease in sales.

Implementing Hypothesis testing

Implementing Hypothesis testing is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis (H_o) and (Ha or H1).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

Though the specific details might vary, the procedure you will use when testing a hypothesis will always follow some version of these steps.

Step 1: State your null and alternate hypothesis

After developing your initial research hypothesis (the prediction that you want to investigate), it is important to restate it as a null (H_o) and alternate (H_a) hypothesis so that you can test it mathematically.

The **alternate hypothesis** is usually your initial hypothesis that predicts a relationship between variables. The **null hypothesis** is a prediction of no relationship between the variables you are interested in.

Hypothesis testing exampleYou want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

- H_0: Men are, on average, not taller than women.
  H_a: Men are, on average, taller than women.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

Hypothesis testing example To test differences in average height between men and women, your sample should have an equal proportion of men and women, and cover a variety of socio-economic classes and any other control variables that might influence average height. You should also consider your scope (Worldwide? For one country?) A potential data source in this case might be census data, since it includes data from a variety of regions and social classes and is available for many countries around the world.

Step 3: Perform a statistical test

There are a variety of statistical tests available, but they are all based on the comparison of **within-group variance** (how spread out the data is within a category) versus **between-group variance** (how different the categories are from one another).

If the between-group variance is large enough that there is little or no overlap between groups, then your statistical test will reflect that by showing a low *p*-value. This means it is unlikely that the differences between these groups came about by chance.

Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a high *p*-value. This means it is likely that any difference you measure between groups is due to chance.

Your choice of statistical test will be based on the type of variables and the level of measurement of your collected data.

Hypothesis testing example Based on the type of data you collected, you perform a one-tailed *t*-test to test whether men are in fact taller than women. This test gives you:

- an estimate of the difference in average height between the two groups.
- a *p*-value showing how likely you are to see this difference if the null hypothesis of no difference is true.

Your *t*-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2 cm to infinity. The *p*-value is 0.002.

Step 4: Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the *p*-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

Hypothesis testing example In your analysis of the difference in average height between men and women, you find that the *p*-value of 0.002 is below your cut off of 0.05, so you decide to reject your null hypothesis of no difference.

Step 5: Present your findings

The results of hypothesis testing will be presented in the results and discussion sections of your research paper, dissertation or thesis.

In the results section you should give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated *p*-value). In the discussion, you can discuss whether your initial hypothesis was supported by your results or not.

In the formal language of hypothesis testing, we talk about rejecting or failing to reject the null hypothesis. You will probably be asked to do this in your statistics assignments.

Stating results in a statistics assignment In our comparison of mean height between men and women we found an average difference of 13.7 cm and a *p*-value of 0.002; therefore, we can reject the null hypothesis that men are not taller than women and conclude that there is likely a difference in height between men and women.

However, when presenting research results in academic papers we rarely talk this way. Instead, we go back to our alternate hypothesis (in this case, the hypothesis that men are on average taller than women) and state whether the result of our test did or did not support the alternate hypothesis.

If your null hypothesis was rejected, this result is interpreted as "supported the alternate hypothesis."

Stating results in a research paper We found a difference in average height between men and women of 14.3cm, with a *p*-value of 0.002, consistent with our hypothesis that there is a difference in height between men and women.

These are superficial differences; you can see that they mean the same thing.

You might notice that **we don't say that we reject or fail to reject the alternate hypothesis**. This is because hypothesis testing is not designed to prove or disprove anything. It is only designed to test whether a pattern we measure could have arisen spuriously, or by chance.

If we reject the null hypothesis based on our research (i.e., we find that it is unlikely that the pattern arose by chance), then we can say our test **lends support to our hypothesis**. But if the pattern does not pass our decision rule, meaning that it could have arisen by chance, then we say the test is **inconsistent with our hypothesis**.

T-Test

When to use a *t* test

A *t* test can only be used when comparing the means of two groups (a.k.a. pairwise comparison). If you want to compare more than two groups, or if you want to do multiple pairwise comparisons, use an ANOVA test or a post-hoc test.

The *t* test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The *t* test assumes your data:

1. are independent
2. are (approximately) normally distributed
3. have a similar amount of variance within each group being compared (a.k.a. homogeneity of variance)

If your data do not fit these assumptions, you can try a nonparametric alternative to the *t* test, such as the Wilcoxon Signed-Rank test for data with unequal variances.

One-sample, two-sample, or paired *t* test?

- If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a **paired** *t* **test**. This is a within-subjects design.
- If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a **two-sample** *t* **test** (a.k.a. **independent** *t* **test**). This is a between-subjects design.
- If there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample** *t* **test**.

One-tailed or two-tailed *t* test?

- If you only care whether the two populations are different from one another, perform a **two-tailed** *t* **test**.
- If you want to know whether one population mean is greater than or less than the other, perform a **one-tailed** *t* **test.**

Interpretation from Outputs of t-test

```
 Welch Two Sample t-test

data:  Petal.Length by Species
t = -33.719, df = 30.196, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.331287 -3.836713
sample estimates:
   mean in group setosa mean in group virginica
                  1.456                   5.540
```

- The output provides:
- An explanation of what is being compared, called data in the output table.
- The t value: -33.719. Note that it's negative; this is fine! In most cases, we only care about the absolute value of the difference, or the distance from 0. It doesn't matter which direction.
- The degrees of freedom: 30.196. Degrees of freedom is related to your sample size, and shows how many 'free' data points are available in your test for making comparisons. The greater the degrees of freedom, the better your statistical test will work.
- The p value: 2.2e-16 (i.e. 2.2 with 15 zeros in front). This describes the probability that you would see a t value as large as this one by chance.
- A statement of the alternative hypothesis (Ha). In this test, the Ha is that the difference is not 0.
- The 95% confidence interval. This is the range of numbers within which the true difference in means will be 95% of the time. This can be changed from 95% if you want a larger or smaller interval, but 95% is very commonly used.
- The mean petal length for each group.