Statistical Concepts for Predictive Modelling

There are a few statistical concepts, such as hypothesis testing, p-values, normal distribution, correlation, and so on without which grasping the concepts and interpreting the results of predictive models becomes very difficult. Thus, it is very critical to understand these concepts, before we delve into the realm of predictive modelling.

In this lab, we will be going through and learning these statistical concepts so that we can use them in the upcoming labs. This lab will cover the following topics:

- **Random sampling and central limit theorem**: Understanding the concept of random sampling through an example and illustrating the central limit theorem's application through an example. These two concepts form the backbone of hypothesis testing.
- **Hypothesis testing**: Understanding the meaning of the terms, such as null hypothesis, alternate hypothesis, confidence intervals, p-value, significance level, and so on. A step-by-step guide to implement a hypothesis test, followed by an example.
- **Chi-square testing**: Calculation of chi-square statistic. A description of usage of chi-square tests with a couple of examples.
- **Correlation**: The meaning and significance of correlations between two variables, the meaning and significance of correlation coefficients and calculating and visualizing the correlation between variables of a dataset.

Random sampling and the central limit theorem

Let's try to understand these two important statistical concepts using an example. Suppose one wants to find the average age of people in Ireland. Now, the safest and brute-force way of doing this will be to gather age information from each citizen and calculate the average for all these ages. But, going to each citizen and asking their age or asking them to tell their age by some method will take a lot of infrastructure and time. It is such a humongous task that census, which attempts to

do just that, happens once a decade and what will happen if you decided to do so in a non-census year?

Statisticians face such issues all the time. The answer lies in random sampling. Random sampling means that you take a group of 1000 individuals (or 10000, depending on your capacity, obviously the more the merrier) and calculate the average for this group. You call this **A1**. Getting to this is easier as 1000 or 10000 is within your reach. Then you select a second group of 1000 or 10000 people and calculate their average. You call this **A2**. You do this 100 times or 1000 times and call them **A3**, **A4**,…, **A100** or **A3**, **A4**,…, **A1000**.

Then according to the most fundamental theorem in statistics called the central limit theorem:

- The average of **A1**, **A2**,…, **A100** will be a good estimator of the average age of the residents of Ireland. If **Am** is the estimated average age of the residents of Ireland, then it is given by:

$$Am = A1 + A2 + \ldots\ldots + A100 / 100$$

- If the number of such samples is sufficiently large, then the distribution of these averages will roughly follow a normal distribution. In other words, **A1**, **A2**,…, **A100** will be normally distributed.

Now, the thing is that we are no more interested in finding the exact value of the average age, but we are settling for an estimator of the same. In such a case, we will have to make do with defining a range of values in which the actual value might lie. Since we have assumed a normal distribution for the average age values of these groups, we can apply all the properties of a normal distribution to quantify the chances of this average age being greater or lesser than a certain number.

Hypothesis testing

The concept we just discussed in the preceding section is used for a very important technique in statistics, called hypothesis testing. In hypothesis testing, we assume a hypothesis (generally related to the value of the estimator) called null hypothesis and try to see whether it holds true or not by applying the rules of a normal distribution. We have another hypothesis called alternate hypothesis.

Null versus alternate hypothesis

There is a catch in deciding what will be the null hypothesis and what will be the alternate hypothesis. The null hypothesis is the initial premise or something that we assume to be true as yet. The alternate hypothesis is something we aren't sure about and are proposing as an alternate premise (almost often contradictory to the null hypothesis) which might or might not be true.

So, when someone is doing a quantitative research to calibrate the value of an estimator, the known value of the parameter is taken as the null hypothesis while the new found value (from the research) is taken as the alternate hypothesis. In our case of finding the mean age of Ireland, we can say that based on the demographics of Ireland, a researcher can claim that the mean age should be less than 35. This can serve as the null hypothesis. If a new agency claims otherwise (that it is greater than 35), then it can be termed as the alternate hypothesis.

Z-statistic and t-statistic

Assume that the value of the parameter assumed in the null hypothesis is $\mathbf{A_o}$. Take a random sample of 100 or 1000 people or occurrences of the event and calculate the mean of the parameter, such as mean age, mean delivery time for pizza, mean income, and so on. We can call it $\mathbf{A_m}$. According to the central limit theorem, the distribution of population means that random samples will follow a normal distribution.

The Z-statistic is calculated to convert a normally distributed variable (the distribution of population mean of age) to a standard normal distribution. This is because the probability

values for a variable following the standard normal distribution can be obtained from a precalculated table. The Z-statistic is given by the following formula:

$$Z = (Am - Ao) / (\sigma / \sqrt{n})$$

In the preceding formula, the **σ** stands for the standard deviation of the population/occurrences of events and **n** is the number of people in the sample.

Now, there can be two cases that can arise:

- **Z- test (normal distribution)**: The researcher knows the standard deviation for the parameter from his/her past experience. A good example of this is the case of pizza delivery time; you will know the standard deviation from past experiences:

$$Z = (Am - Ao) / (\sigma / \sqrt{n})$$

A_o (from the null hypothesis) and **n** are known. A_m is calculated from the random sample. This kind of test is done when the standard deviation is known and is called the **z-test** because the distribution follows the normal distribution and the standard-normal value obtained from the preceding formula is called the **Z-value**.

- **t-test (Student-t distribution)**: The researcher doesn't know the standard deviation of the population. This might happen because there is no such data present from the historical experience or the number of people/event is very small to assume a normal distribution; hence, the estimation of mean and standard deviation by the formula described earlier. An example of such a case is a student's marks in an exam, age of a population, and so on. In this case, the mean and standard deviation become unknown and the expression assumes a distribution other than normal distribution and is called a **Student-t** distribution. The standard value in this case is called **t-value** and the test is called **t-test**.

Standard distribution can also be estimated once the mean is estimated, if the number of samples is large enough. Let us call the estimated standard distribution **S**; then the **S** is estimated as follows:

$$S = \sum (Ai - Ao)^2 / (n - 1)$$

The t-statistic is calculated as follows:

$$t = (Am - Ao) / (S / \sqrt{n})$$

The difference between the two cases, as you can see, is the distribution they follow. The first one follows a normal distribution and calculates a Z-value. The second one follows a Student-t distribution and calculates a t-value. These statistics that is Z-statistics and t-statistics are the parameters that help us test our hypothesis.

Confidence intervals, significance levels, and p-values

Let us go back a little in the last lab and remind ourselves about the cumulative probability distribution.

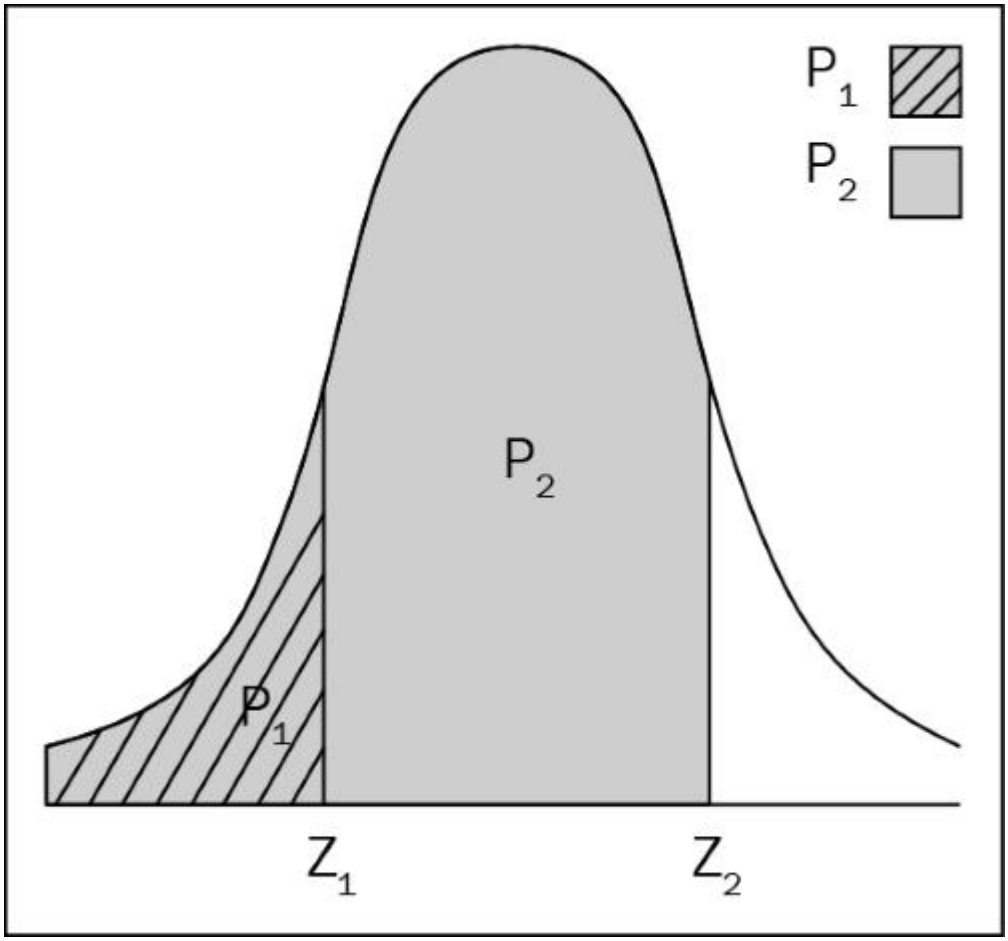

Fig. 4.1: A typical normal distribution with p-values

Let us have a look to the preceding figure, it shows a standard normal distribution. Suppose, **Z**1 and Z2 are two Z-statistics corresponding to two values of random variable and **p1** and **p2**are areas enclosed by the distribution curve to the right of those values. In other

words, **p1** is the probability that the random variable will take a value lesser than or equal to **Z1** and **p2** is the probability that the random variable will take a value greater than **Z2**.

If we represent the random variable by **X**, then we can write:

$$P(X < Z1) = p1$$
$$P(X < Z2) = p2$$

Also, since the sum of all the exclusive probabilities is always **1**, we can write:

$$P(X > Z1) = 1 - p1$$
$$P(X > Z2) = 1 - p2$$

For well-defined distributions, such as the normal distribution, one can define an interval in which the value of the random variable will lie with a confidence level (read probability). This interval is called the confidence interval. For example, for a normal distribution with mean **μ** and standard deviation **σ**, the value of the random variable will lie in the interval **[μ-3σ, μ+3σ]**with 99% probability. For any estimator (essentially a random variable) that follows a normal distribution, one can define a confidence interval if we decide on the confidence (or probability) level. One can think of confidence intervals as thresholds of the accepted values to hold a null hypothesis as true. If the value of the estimator (random variable) lies in this range, it will be statistically correct to say that the null hypothesis is correct.

To define a confidence interval, one needs to define a confidence (or probability level). This probability needs to be defined by the researcher depending on the context. Lets call this **p**. Instead of defining this probability **p**, one generally defines (**1-p**) that is called level of significance. Let us represent it by **ß**. This represents the probability that the null hypothesis won't be true. This is defined by the user for each test and is usually of the order of 0.01-0.1.

An important concept to learn here is the probability value or just a p-value of a statistic. It is the probability that the random variable assumes, it's a value greater than the Z-value or t-value:

$$p - value = P(X > Z)$$

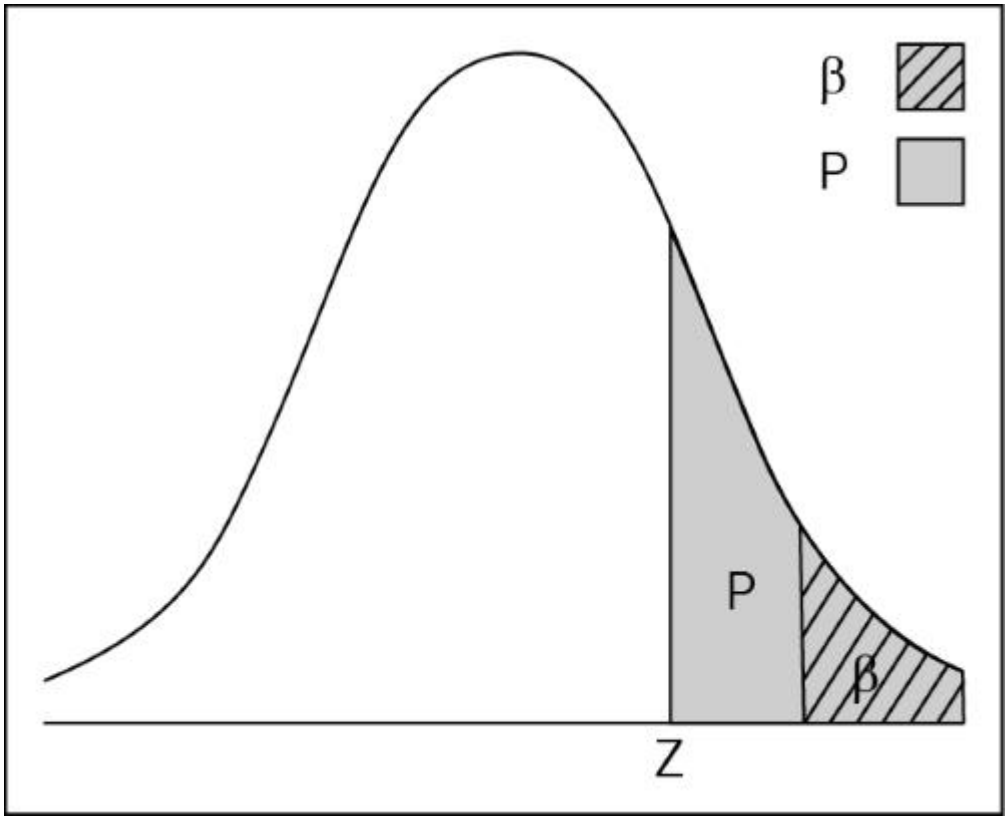

Fig. 4.2: A typical normal distribution with p-values and significance level

Now, this Z-value and the p-value has been obtained assuming that the null hypothesis is true. So, for the null hypothesis to be accepted, the Z-value has to lie outside the area enclosed by **ß**. In other words, for the null hypothesis to be true, the p-value has to be greater than the significance level, as shown in the preceding figure.

To summarize:

- Accept the null hypothesis and reject the alternate hypothesis if **p-value>ß**
- Accept the alternate hypothesis and reject the null hypothesis if **p-value<ß**

Different kinds of hypothesis test

Due to the symmetry and nature of the normal distribution, there are three kinds of possible hypothesis tests:

- Left-tailed
- Right-tailed

- Two-tailed

Left-tailed: This is the case when the alternate hypothesis is a "less-than" type.

The hypothesis testing is done on the left tail of the distribution and hence the name. In this case, for:

- Accepting a null hypothesis and rejecting an alternate hypothesis the **p-value>ß** or **Z>Zß**
- Accepting an alternate hypothesis and rejecting a null hypothesis the **p-value<ß** or **Z<Zß**

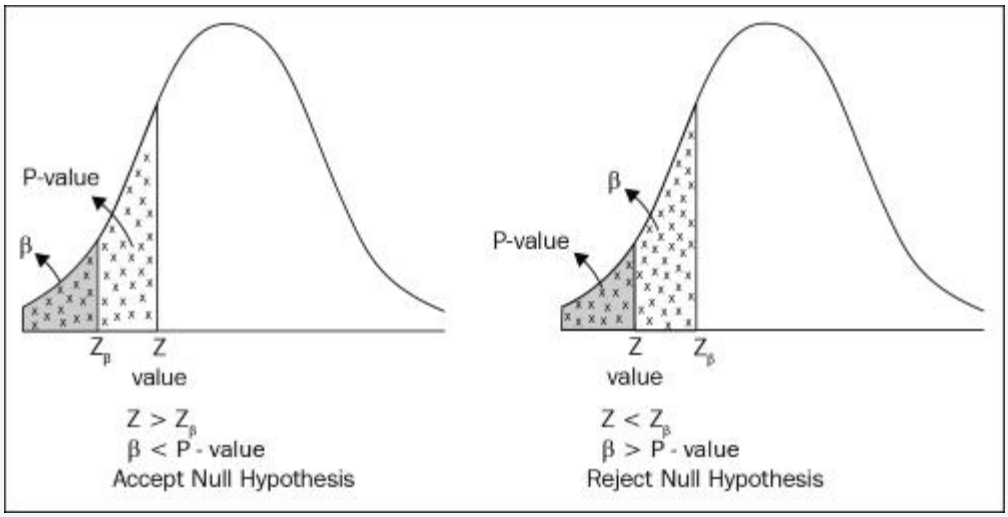

Fig. 4.3: Left-tailed hypothesis testing

Right-tailed: This is the case when the alternate hypothesis is of greater than type. The hypothesis testing is done on the right tail of the distribution, hence the name. In this case, for:

- Accepting a null hypothesis and rejecting an alternate hypothesis the **p-value>ß** or **Z<Zß**
- Accepting an alternate hypothesis and rejecting a null hypothesis the **p-value<ß** or **Z>Zß**

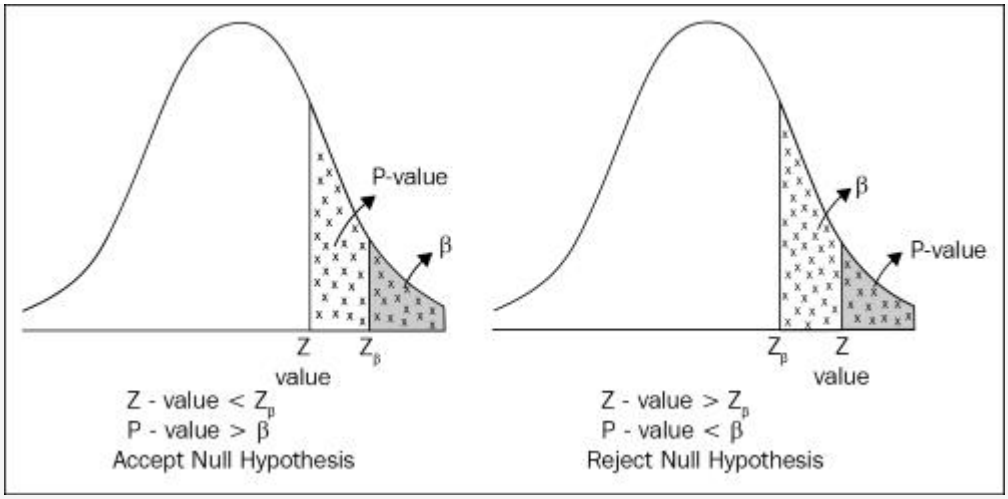

Fig. 4.4: Right-tailed hypothesis testing

Two-tailed: This is the case when the alternate hypothesis has an inequality—less than or more than is not mentioned. It is just an `OR` operation over both kind of tests. If either of the left- or right-tailed tests reject the null hypothesis, then it is rejected. The hypothesis testing is done on both the tails of the distribution; hence, the name.

A step-by-step guide to do a hypothesis test

So how does one accept one hypothesis and reject the other? There has to be a logical way to do this. Let us summarize and put to use whatever we have learned till now in this section, to make a step-by-step plan to do a hypothesis test. Here is a step-by-step guide to do a hypothesis test:

1. Define your null and alternate hypotheses. The null hypothesis is something that is already stated and is assumed to be true, call it **H_o**. Also, assume that the value of the parameter in the null hypothesis is **A_o**.
2. Take a random sample of 100 or 1000 people/occurrences of events and calculate the value of estimator (for example, mean of the parameter that is mean age, mean delivery time for pizza, mean income, and so on). You can call it **Am**.
3. Calculate the standard normal value or Z-value as it is called using this formula:

$$Z = (Am - Ao) / (\sigma / \sqrt{n})$$

In the preceding formula, **σ** is the standard deviation of the population or occurrences of events and **n** is the number of people in the sample.

The probability associated with the Z-value calculated in step 3 is compared with the significance level of the test to determine whether null hypothesis will be accepted or rejected.

An example of a hypothesis test

Let us see an example of hypothesis testing now. A famous pizza place claims that their mean delivery time is 20 minutes with a standard deviation of 3 minutes. An independent market researcher claims that they are deflating the numbers for market gains and the mean delivery time is actually more. For this, he selected a random sample of 64 deliveries over a week and found that the mean is 21.2 minutes. Is his claim justified or the pizza place is correct in their claim? Assume a significance level of 5%.

First things first, let us define a null and alternate hypothesis:

$$Ho: Do = 20(What\ the\ pizza\ guy\ claims)$$
$$Ha: Do > 20(what\ researcher\ claims)$$
$$\sigma = 3, n = 64\ and\ Dm = 24, \beta = 0.05$$

Let us calculate the Z-value:

$$Z = (21.2 - 20) / \left(3 / \sqrt{64}\right) = 3.2$$

When we see the standard normal table for this Z-value, we find out that this value has an area of .9993 to the left of it; hence, the area to the right is **1-.99931**, which is less than **0.05**.

Hence, **p-value<ß**. Thus, the null hypothesis is rejected. This can be summarized in the following figure:

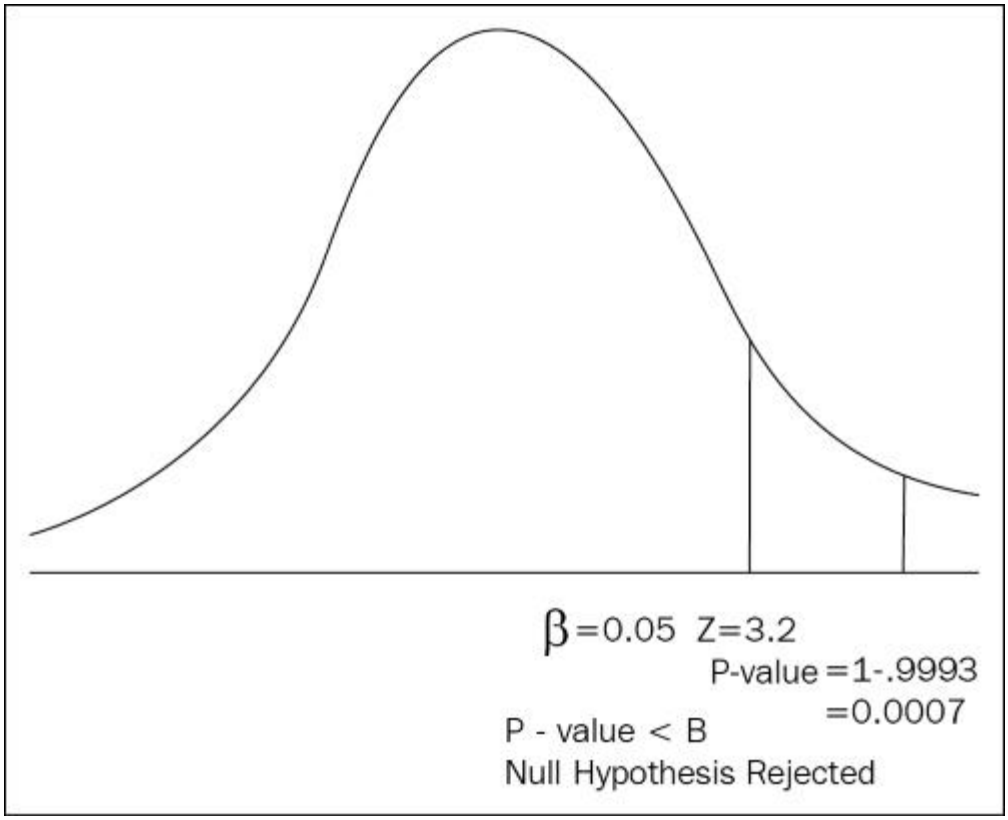

Fig. 4.5: Null Hypothesis is rejected because p-value<significance level

Hence, the researcher's claim that the mean delivery time is more than 20 minutes is statistically correct.

Chi-square tests

The chi-square test is a statistical test commonly used to compare observed data with the expected data assuming that the data follows a certain hypothesis. In a sense, this is also a hypothesis test. You assume one hypothesis, which your data will follow and calculate the expected data according to that hypothesis. You already have the observed data. You calculate the deviation between the observed and expected data using the statistics defined in the following formula:

$$chi-square value(g) = \sum (O-E)^2 / E$$

Where **O** is the observed value and **E** is the expected value while the summation is over all the data points.

The chi-square test can be used to do the following things:

- Show a causal relationship or independence between one input and output variable. We assume that they are independent and calculate the expected values. Then we calculate the chi-square value. If the null hypothesis is rejected, it suggests a relationship between the two variables. The relationship is not just by chance but statistically proven.
- Check whether the observed data is coming from a fair/unbiased source. If the observed data is more skewed towards one extreme, compared to the expected data, then it is not coming from a fair source. But, if it is very close to the expected value then it is.
- Check whether a data is too good to be true. As, it is a random experiment and we don't expect the values to toe the assumed hypothesis. If they do toe the assumed hypothesis, then the data has probably been tampered to make it look good and is too good to be true.

Let us create a hypothetical experiment where a coin is tossed 10 times. How many times do you expect it to turn heads or tails? Five, right? Now, what if we do this experiment 1000 times and record the scores (number of heads and tails). Suppose we observed heads 553 times and a tails in the rest of the trials:

$$Ho: The\ proportion\ of\ head\ and\ tail\ is\ 0.5$$
$$Ha: The\ proportion\ is\ not\ 0.5$$

	Head	Tail
Observed	553	447
Expected	1000*0.5=500	1000*0.5=500

Let us calculate the chi-square value:

$$g = \left[(553-500)^2 + (447-500)^2\right] / 500 = 11.236$$

This chi-square value is compared to the value on a chi-square distribution for a given degree of freedom and a given significance level. The degrees of freedom is the number of categories -1. In this case, it is **2-1=1**. Let us assume a significance level of 0.05.

The chi-square distribution looks a little different than the normal distribution. It also has a peak but has a much longer tail than the normal distribution and is only on one side. As the degree of freedom increases, they start looking similar to a normal distribution:

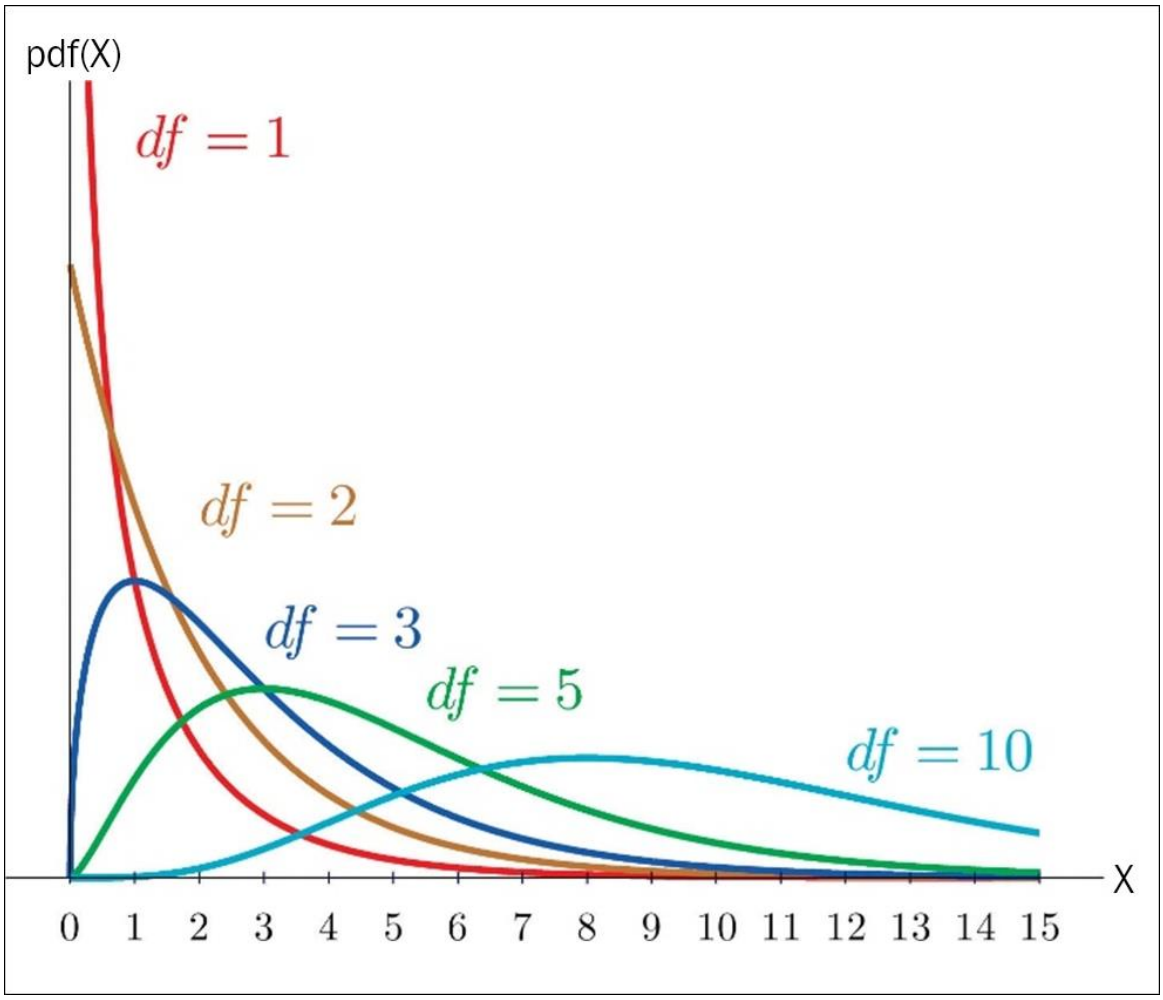

Fig. 4.6: Chi-square distribution with different degrees of freedom

When we look at the chi-square distribution table for a degree of freedom 1 and a significance level of 0.05, we get a value of 3.841. At a significance level of 0.01, we get 6.635. In both the cases, the chi-square statistic is greater than the value from the chi-square distribution, meaning that the chi-square statistic lies on the right of the value from the distribution table.

Hence, the null hypothesis is rejected. That means that the coin is not fair.

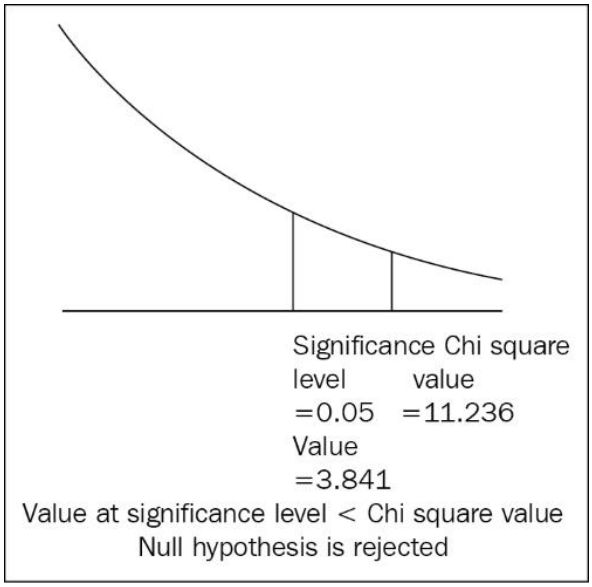

Fig. 4.7: Null hypothesis is rejected because the value of the chi-square statistic at the significance level is less than the value of the chi-square statistic

Let us look at another example where we want to prove that the gender of a student and the subjects they choose are independent.

Suppose, in a group of students, the following table represents the number of boys and girls who have taken Maths, Arts, and Commerce, as their main subjects.

The observed number of boys and girls in each subject is as shown in the following table:

	Maths	Arts	Commerce	Total
Boys	68	52	90	200
Girls	28	37	35	100
Total	96	89	125	300

If the choice of the subjects is irrespective of the gender, then the expected number of boys and girls taking different subjects is, as follows:

	Maths	Arts	Commerce	Total
Boys	(200/300)*96=64	(200/300)*89=59.3	(200/300)*125=83.3	200

	Maths	Arts	Commerce	Total
Girls	(100/300)*96=32	(100/300)*89=29.7	(100/300)*125=41.7	100
Total	96	89	125	300

The deviation element is calculated for each cell using the **$(O-E)^2/E$** formula:

	Maths	Arts	Commerce
Boys	$(68-64)^2/64$	$(52-59.3)^2/59.3$	$(90-83.3)^2/83.3$
Girls	$(28-32)^2/32$	$(37-29.7)^2/29.7$	$(35-41.7)^2/41.7$

On calculating and summing up all the values, the chi-square value comes out to be 5.05. The degree of freedom is the number of categories-1, which amounts to **[(3x2)-1=5]**. Let us assume a significance level of 0.05.

Looking at the chi-square distribution, one can find out that for a 5-degree freedom chi-square distribution, the value of the chi-square statistic at a significance level of 0.05 is 11.07.

The calculated chi-square statistic < chi-square statistic (at significance level=0.05).

Since, the chi-square statistic lies on the left of the value at the significance level, the null hypothesis can't be rejected. Hence, the choice of subjects is independent of the gender.

Correlation

Another statistical idea which is very basic and important while finding a relation between two variables is called correlation. In a way, one can say that the concept of correlation is the premise of predictive modelling, in the sense that the correlation is the factor relying on which we say that we can predict outcomes.

A good correlation between two variables suggests that there is a sort of dependence between them. If one is changed, the change will be reflected in the other as well. One can say that a good correlation certifies a mathematical relation between two variables and due to this mathematical relationship, we might be able to predict outcomes. This mathematical relation can be anything. If **x** and **y** are two variables, which are correlated, then one can write:

$$Y = f(x)$$

If **f** is a linear function, then **a** and **b** are linearly correlated. If **f** is an exponential function, then **a** and **b** are exponentially correlated:

$$Linear\ correlation: y = ax + b$$
$$Exponential\ correlation: y = \exp(a) + b$$

The degree of correlation between the two variables **x** and **y** is quantified by the following equation:

$$correlation\ coefficient(h) = \frac{\sum((x - xm) * (y - ym))}{\sqrt{\sum(x - xm)^{2*} \sum(y - ym)^2}}$$

Where **xm** and **ym** are mean values of **x** and **y**

A few points to note about the correlation coefficient are as follows:

- The value of the correlation coefficient can range from -1 to 1, that is -1<h<1.
- A positive correlation coefficient means that there is a direct relationship between the two variables; if one variable increases, the other variable will also increase and if one decreases the other will decrease as well.
- A positive correlation coefficient means that there is an inverse relationship between the two variables; if one variable increases, the other variable will decrease and if one decreases the other will increase.
- The more the value of the correlation coefficient, the stronger the relation between the two variables.

Although, a strong correlation suggests that there is some kind of a relationship that can be leveraged to predict one based on the other; it doesn't imply that its relation with the other variable is the only factor explaining this, there can be several others. Hence, the most often used quote related to correlation is, "**Correlation doesn't imply causation.**"

Let us try to understand this concept better by looking at a dataset and trying to find the correlation between the variables. The dataset that we will be looking at is a very popular dataset about various costs incurred on advertising by different mediums and the sales for a particular product. We will be using it later to explore the concepts of linear regression. Let us import the dataset and calculate the correlation coefficients:

```
import pandas as pd
advert=pd.read_csv('C:/FILE_PATH/Linear Regression/Advertising.csv')
advert.head()
```

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Fig. 4.8: Dummy dataset

Let us try to find out the correlation between the advertisement costs on TV and the resultant sales. The following code will do the job:

```
import numpy as np
advert['corrn']=(advert['TV']-np.mean(advert['TV']))*(advert['Sales']-
np.mean(advert['Sales']))
advert['corrd1']=(advert['TV']-np.mean(advert['TV']))**2
advert['corrd2']=(advert['Sales']-np.mean(advert['Sales']))**2
corrcoeffn=advert.sum()['corrn']
corrcoeffd1=advert.sum()['corrd1']
corrcoeffd2=advert.sum()['corrd2']
corrcoeffd=np.sqrt(corrcoeffd1*corrcoeffd2)
corrcoeff=corrcoeffn/corrcoeffd
corrcoeff
```

In this code snippet, the formula written above has been converted to code. The value of the correlation coefficient comes out to be 0.78 indicating that there is a descent in positive correlation between TV-advertisement costs and sales; it implies that if the TV-advertisement cost is increased, as a result sales will increase.

Let us convert the preceding calculation to a function, so that we can calculate all the pairs of correlation coefficients very fast just by replacing the variable names. One can do that using the following snippet wherein a function is defined to parameterize the name of the data frame and the column names for which the correlation coefficient is to be calculated:

Copy

```
def corrcoeff(df,var1,var2):
    df['corrn']=(df[var1]-np.mean(df[var1]))*(df[var2]-np.mean(df[var2]))
    df['corrd1']=(df[var1]-np.mean(df[var1]))**2
    df['corrd2']=(df[var2]-np.mean(df[var2]))**2
    corrcoeffn=df.sum()['corrn']
    corrcoeffd1=df.sum()['corrd1']
    corrcoeffd2=df.sum()['corrd2']
    corrcoeffd=np.sqrt(corrcoeffd1*corrcoeffd2)
    corrcoeff=corrcoeffn/corrcoeffd
    return corrcoeff
```

This function can be used to calculate correlation coefficient for any two variables of any data frame.

For example, to calculate the correlation between **TV** and **Sales**columns of the `advert` data frame, we can write it as follows:

$$TV \ \& \ Sales \qquad corrcoeff(advert,'TV','Sales') \quad 0.78$$
$$Radio \ \& \ Sales \qquad corrcoeff(advert,'Radio','Sales') \quad 0.57$$

We can summarize the pair-wise correlation coefficients between the variables in the following table:

	TV	Radio	Newspaper	Sales
TV	1	0.05	0.06	0.78

	TV	Radio	Newspaper	Sales
Radio	0.05	1	0.35	0.57
Newspaper	0.06	0.35	1	0.23
Sales	0.78	0.57	0.23	1

This table is called **Correlation Matrix**. As you can see, it is a symmetric matrix because the correlation between **TV** and **Sales**will be the same as that between **Sales** and **TV**. Along the diagonal, all the entries are **1** because, by definition, the correlation of a variable with itself will always be **1**. As can be seen, the strongest correlation can be found between TV advertisement cost and sales.

Let us see the nature of this correlation by plotting `TV` and `Sales` variables of the advert data frame. We can do this using the following code snippet:

Copy

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.plot(advert['TV'],advert['Sales'],'ro')
plt.title('TV vs Sales')
```

The result is similar to the following plot:

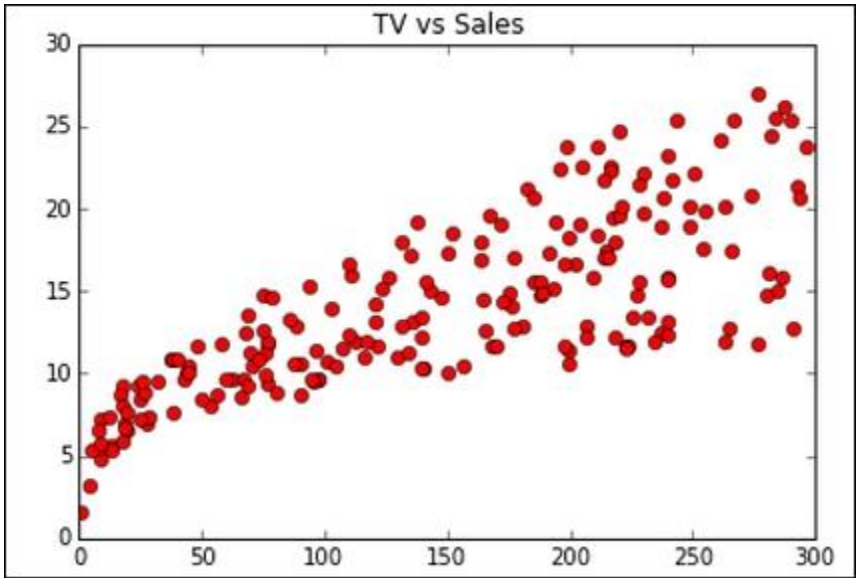

Fig. 4.9: Scatter plot of TV vs Sales

Looking at this plot, we can see that the points are more or less compact and not scattered far away and as the TV advertisement cost increases, the sales also increase. This is the characteristic of two variables that are positively correlated. This is supported by a strong correlation coefficient of 0.78.

Let us plot the variables and see how they are distributed to corroborate their correlation coefficient. For `Radio` and `Sales`, this can be plotted as follows:

Copy

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.plot(advert['Radio '],advert['Sales'],'ro')
plt.title('Radio vs Sales')
```

The plot we get is as shown in the following figure:

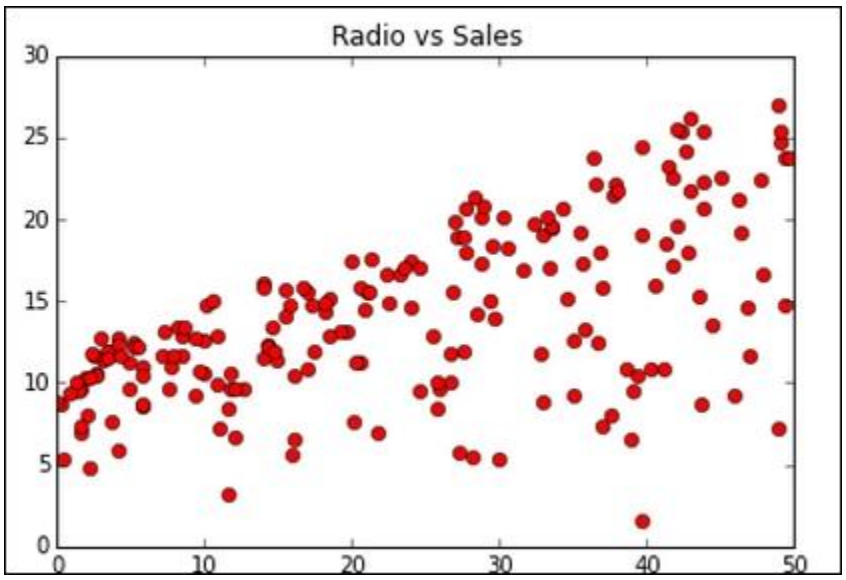

Fig. 4.10: Scatter plot of Radio vs Sales

For Radio and Sales, the points are a little more scattered than TV versus Sales and this is corroborated by the fact that the correlation coefficient for this pair (0.57) is less than that for TV and Sales (0.78).

For plotting `Newspaper vs Sales` data, we can write something similar to the following code:

Copy

```
import matplotlib.pyplot as plt
%matplotlib inline
```

```
plt.plot(advert['Newspaper'],advert['Sales'],'ro')
plt.title('Newspaper vs Sales')
```

The output plot looks similar to the following figure:

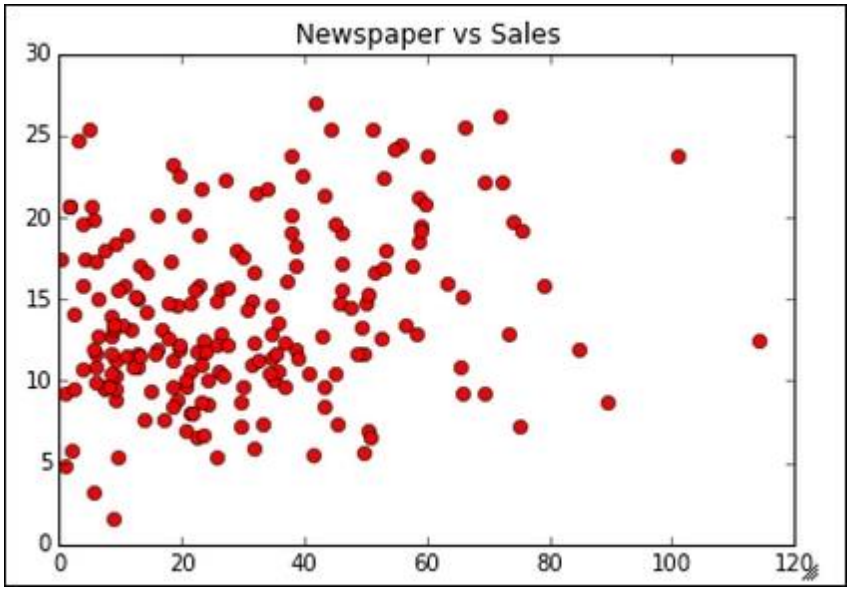

Fig. 4.11: Scatter plot of Newspaper vs Sales

For Newspaper and Sales, the points are way more scattered than in the case of TV and Sales and Radio and Sales. This is further strengthened by a small correlation coefficient of 0.23 between Newspaper and Sales, compared to 0.78 between TV and Sales, and 0.57 between Radio and Sales.

Summary

In this lab, we skimmed through the basic concepts of statistics. Here is a brief summary of the concepts we learned:

- Hypothesis testing is used to test the statistical significance of a hypothesis. The one which already exists or is assumed to be true is a null hypothesis, the one which someone is not sure about or is being proposed as an alternate premise is an alternate hypothesis.
- One needs to calculate a statistic and the associated p-value to conduct the test.
- Hypothesis testing (p-values) is used to test the significance of the estimates of the coefficients calculated by the model.

- The chi-square test is used to test the causal relationship between a predictor and an input variable. It can also be used to check whether the data is fair or fake.
- The correlation coefficient can range from -1 to 1. The closer it is to the extremes, the stronger is the relationship between the two variables.

Linear regression is part of the family of algorithms called supervised algorithms as the dataset on which they are built has an output variable. In a sense, one can say that this output variable governs or supervises the development of the model and hence the name. More on this is covered in the next lab.