Hypothesis Testing

In our day to day routine we come across many questions such as:

- How much liters of water should be alotted on an average to a household in Delhi *[Nation saw a heated debate on this during Delhi elections few years back]*
- How much stock of apple juice should i order for this for powai location
- By what amount should we increase credit card limits for a certain group of customers
- Should BCCI select Ishant Sharma for Australia tour in spite of his poor performance in last 4 test matches in India
- Should we provide additional tuition for language courses to science students.

To figure out answers to these question at very basic level, you don't really need any hardcore statistics knowledge. For example lets take first and last questions from the ones listed above.

To figure out how much water on an average a Delhi household would need, we can simply ask every house hold regarding the same. But its much harder than said. There are too many house holds to cover. Resources and time required to do this will definitely not justify the utility of outcome. So what do we do then? Instead of connecting with each and every household, we can talk to handful of them. Or in other words , take a small sample. And consider their average consumption as the general answer. But we need to be careful about what all households we are selecting as sample. Water consumption needs might not be same across all regions of Delhi, also there will be certain difference between residential and commercial entities. Our **sample** should contain observations from all these different segments *[strata in population]* in order to truly represent entire Delhi.

Taking up next question of whether science students need additional tuition for language courses. Underlying thought here is that science students perform poorly in language courses in comparison to students coming from other streams. To figure out if that is the case we'll collect student performance data and see the average performance of students from both science and non-science streams. If science students are performing significantly worse than their counterparts, we'll take the decision in favor of providing them with additional tuition.

Common theme in strategy of solving both these problem was to collect data and then use it to verify, refute our claims or estimate the value of parameter. What our methods lacked though, was rigorous statistical framework. We'd build the same as we progress in this module.

We'll be discussing various concepts here on wards which might seem slightly disconnected in the beginning but everything will fall in place as we near the end. Starting with Population and Sample, we have already used these concepts above , its time to formally introduce them

Population & Samples

Population is a huge collection of data points. This can be anything from Age of all graduating engineers of India in last decade or Number of pages in individual books published on topic statistics in last century. If we want to figure out a parameter *[any characteristic]* value for this "population" we can measure in theory each and every observation to find the average value.

As we witnessed earlier, this is rarely feasible. And for all practical purposes to estimate value of this population parameter *[such as average , standard deviation etc]* we work with a sample instead.

These "sample" observations need to be randomly chosen in order to avoid personal bias. Also they should represent all strata present with in the population.

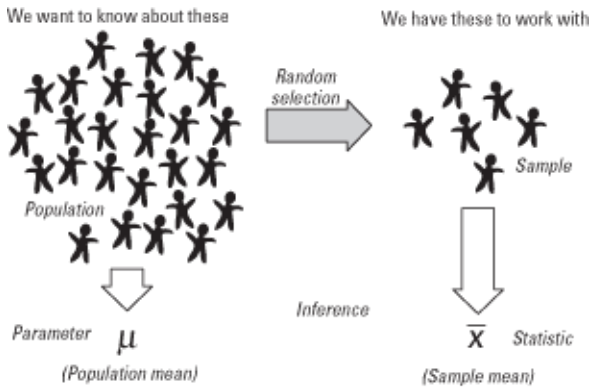


Figure 1: PnS

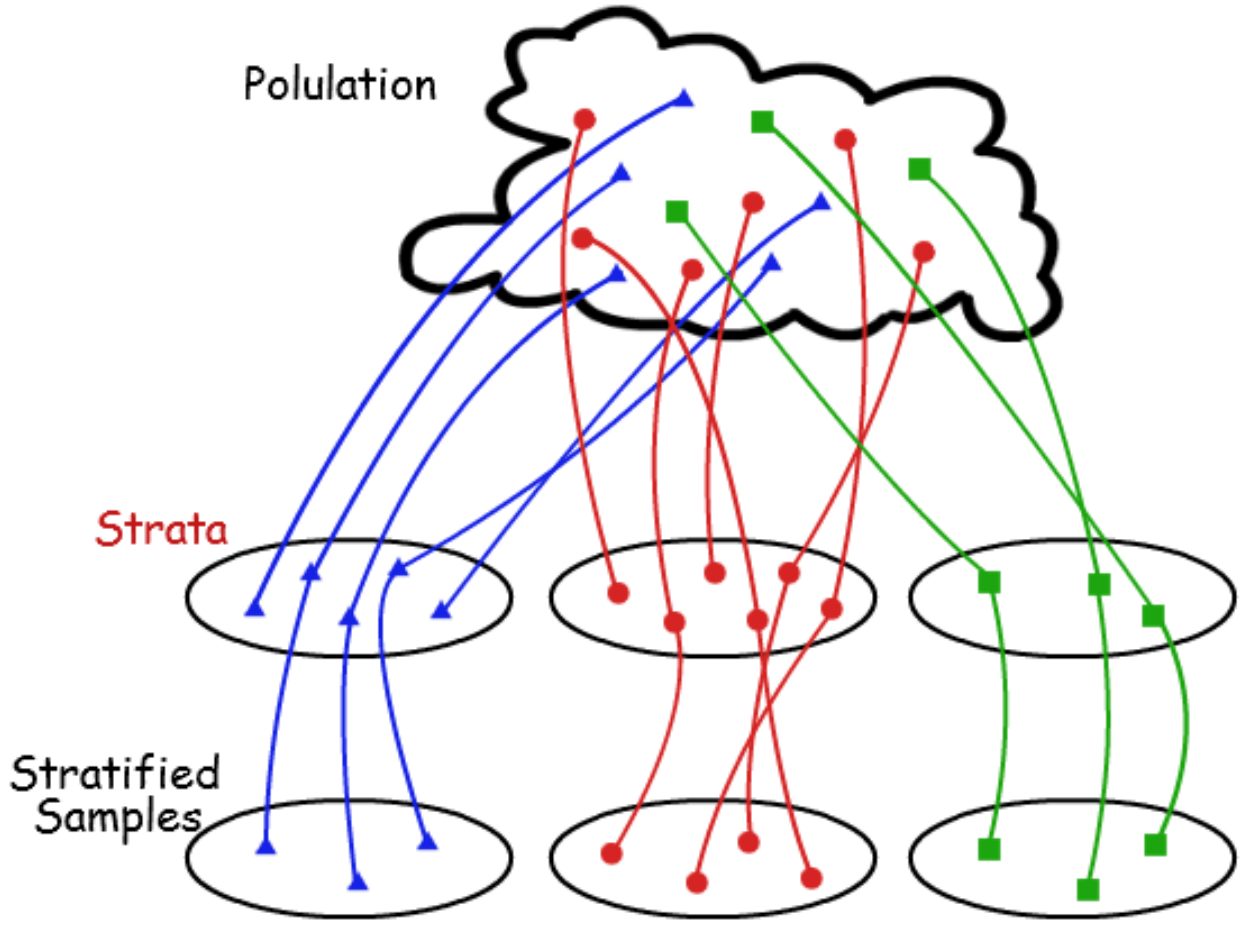


Figure 2: StS

Estimates & Errors

Purpose of the sampling from the population was to *estimate* the population parameter such as average age of graduating engineers or average number of pages in statistics books.

We don't really know what the actual value of population parameter is. Estimates give us an idea what it might be. As your sample size grows bigger and bigger, estimate will be more close to real value of population parameter.

Since these samples contain randomly chosen observations, each new sample will give you a different estimate for population parameter. What this means is that an estimate from sample will always contain error. The term error is not same as *mistake*.

Errors are inherent part of estimates by design. Hypothesis Testing is a frame work to quantify these errors.

Histograms, Probabilities, Cumulative Probabilities and Distributions

We all have seen those frequency bar charts at some point of our time while working with excel sheets.Lets consider scores of 40 students in a quiz. First few data points look like this:

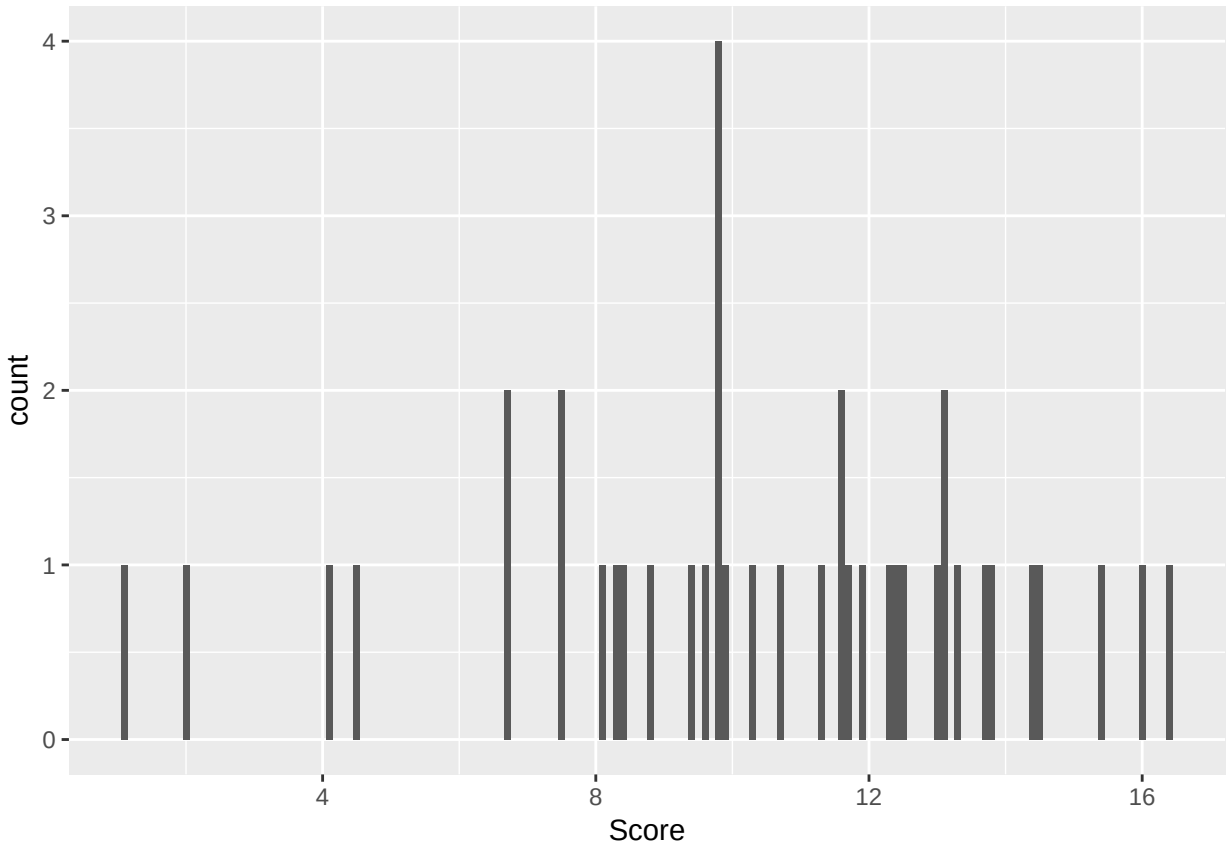
\	Score
1	7.5
2	10.7
3	6.7
4	16.4

... full table truncated

We can convert this to frequency counts, which tells us how many students got a certain score.

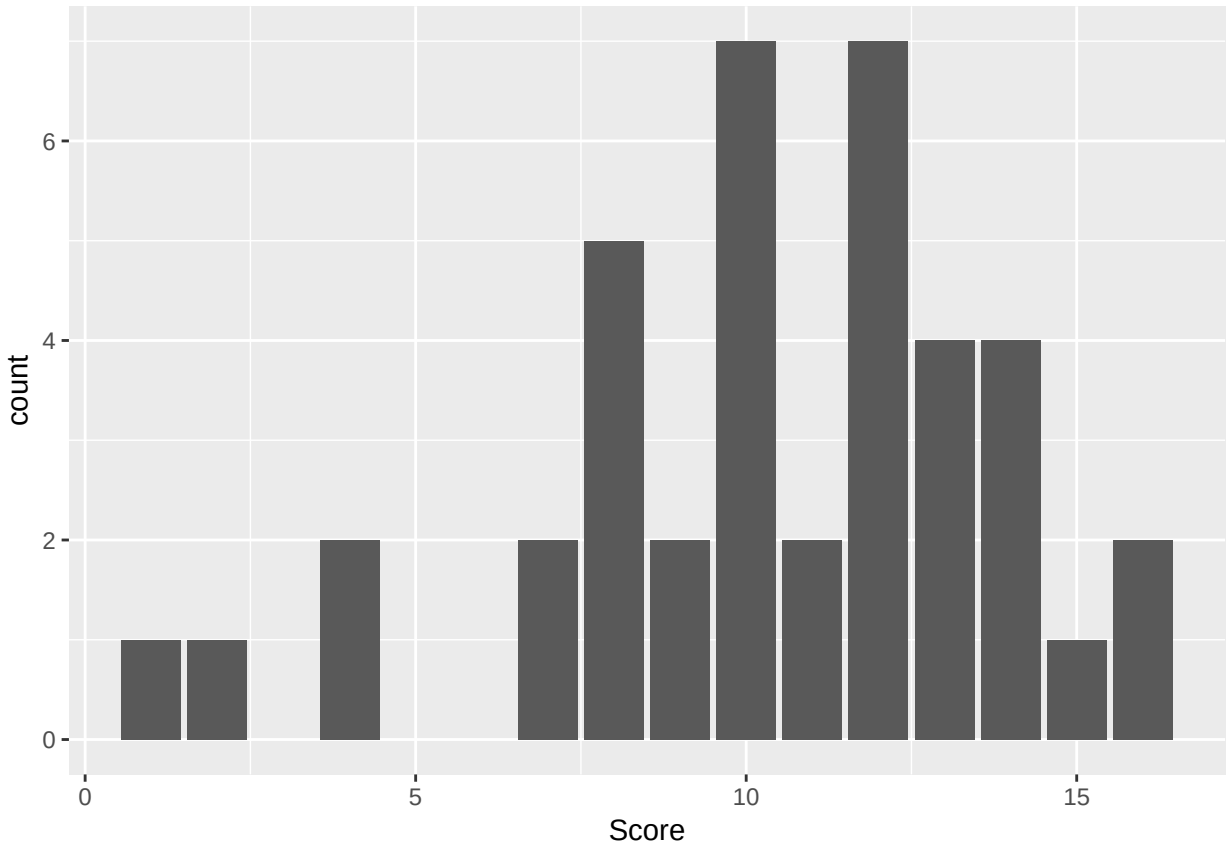
Score	Frequency
1.1	1
2.0	1
4.1	1
4.5	1
6.7	2
7.5	2
... full	table truncated

We can plot these counts as bar charts .

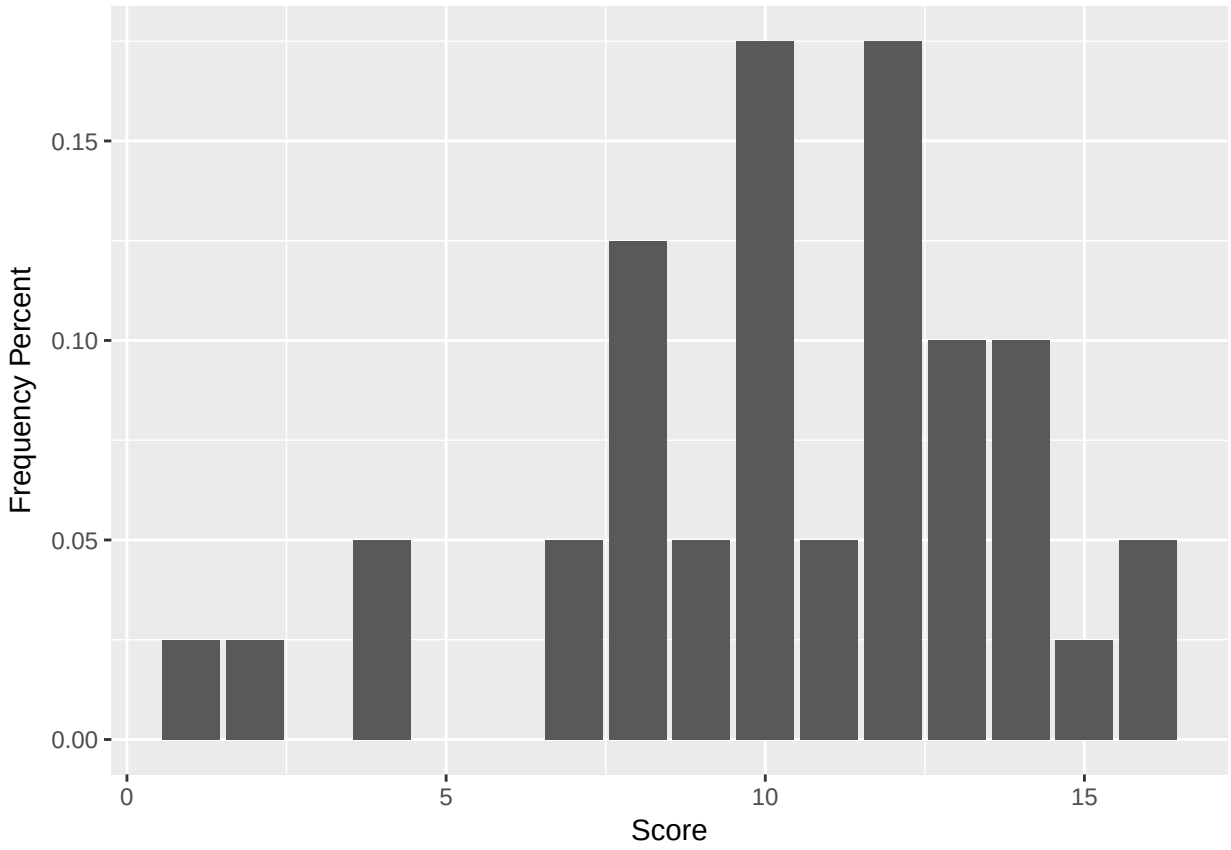


We can see that this bar chart however is not very informative as many frequencies are simply equal and many intervals remain blank.It doesn't give us very good idea as to what kind of values are more frequent and so on. Since we have less number of data points we can club them into classes. For example all scores like 9.2, 9.4 9.6 etc can be clubbed into class 9.

Lets see how does the bar chart looks like once we do this



Lets convert these counts to frequency percentages by dividing them by total counts *[40]*



By looking at this Frequency Percent chart and assuming that this sample represent entire population of students, i can say that 20% of students score in "class 11". Or in other words, if i randomly pick a student's score , probability of it being in class 11 is 0.2/20% .

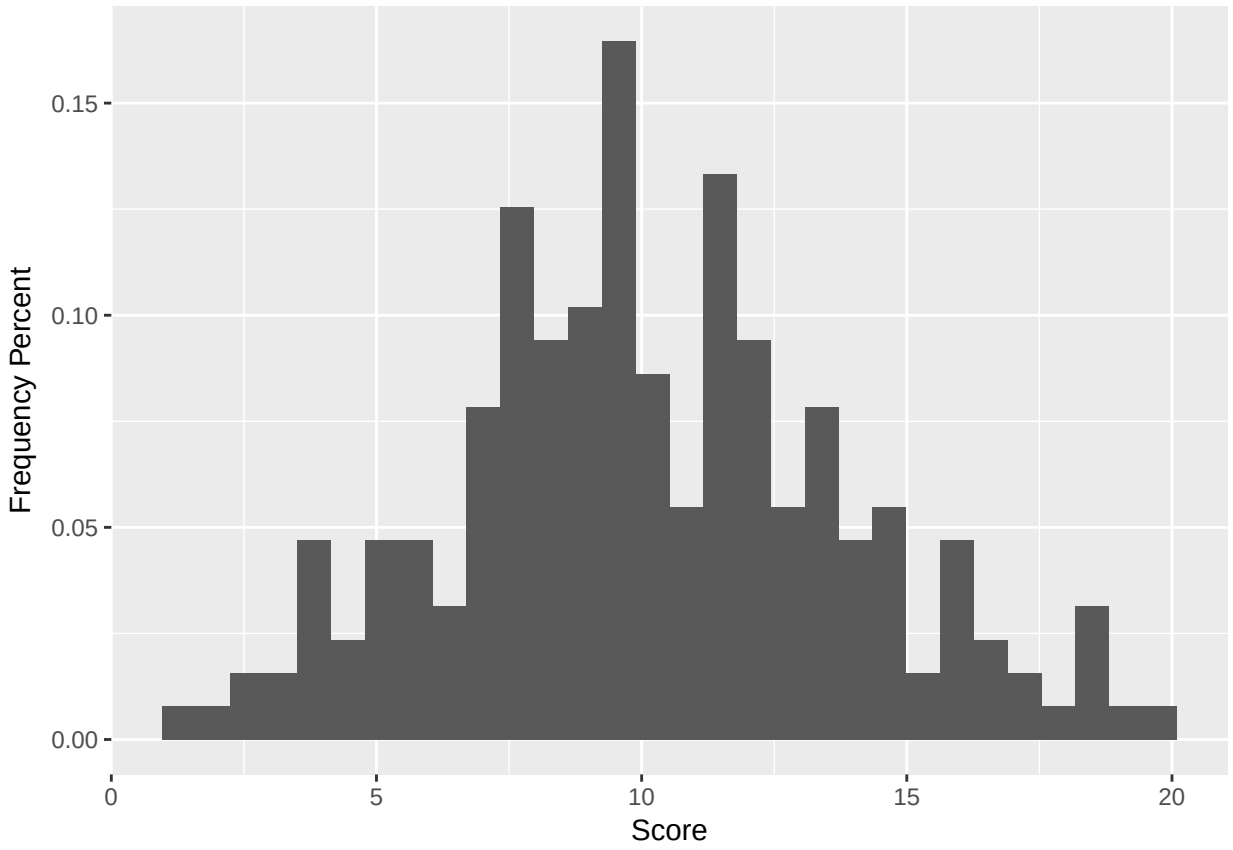
If you ask me what is the probability that a student would score between 5 to 10. I will simply add all the probabilities associated with "class 5" to "class 10".

$$P(5 \leq Score \leq 10) = P(Score = 5) + P(Score = 6) + P(Score = 7) + P(Score = 8) + P(Score = 9) + P(Score = 10)$$

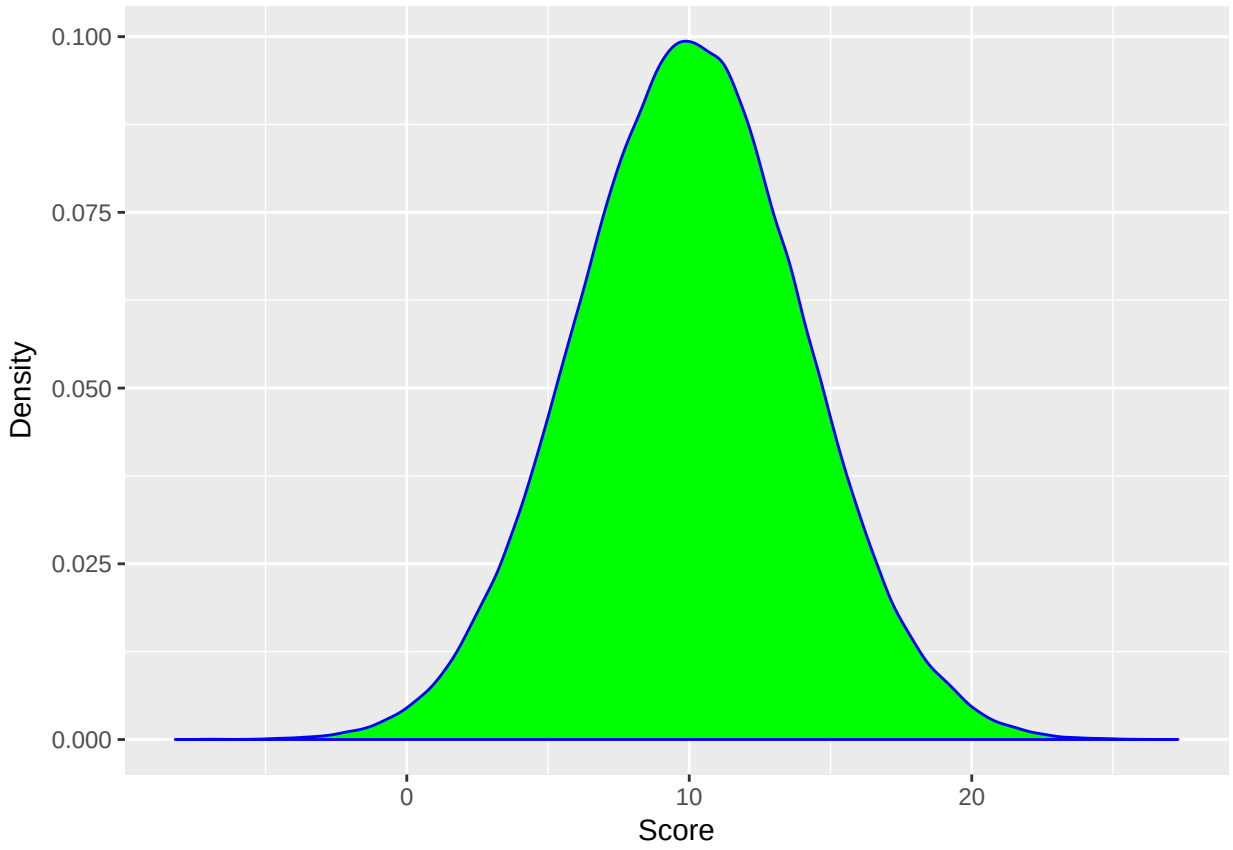
$$P(5 \leq Score \leq 10) = 22.5\%$$

That teaches us , we can comment on probabilities of occurrence in an interval with cumulative probabilities *[addition of probabilities]* . Also by collecting more data we can make our "classes" finer and finer. Which gives us idea about frequency of occurrence of values at much finer level. Lets see how this looks with 200 data points instead of just 40.

This contains much more information in comparison to



We can keep on collecting this data until lets we have practically infinite data points. Imagine that we drew a curve which joins top of all these fine bars now. We could still comment probabilities by looking at point on the curve associated with those values of x.



Lets say this curve is represented by y=f(x), if you pass a value of x, it gives probability of occurrence of that value. Now instead of adding the probabilities of occurrence to get interval probabilities we can use this:

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

[Integration is nothing but summation of infitiely small elements.No, i'm not going to ask you to learn integration now, You just need to know what goes on at the back end. Eventually all this will be done by a software for you.]

Other properties that it'll follow :

$$0 \leq f(x) \leq 1$$

if you sum probabilities for all possible cases it'll be equal to 1

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

This f(x) here is nothing but distribution curve of the population.By looking at this curve you can get an idea about probability of occurrence of the values from your population.

Normal distribution is such a curve. it has following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

Where μ and σ are population mean and population standard deviations respectively.They are parameters of this functions . For different values of μ and σ , you'll have a different Normal distribution. This is similar to general equation of lines. y=mx+c . For different values of slope(m) and intercept(c) you'll have different lines.

Don't let this weird equation intimidate you.Its just an equation associated with a particular distribution. It also satisfies conditions given above for such a function.

Standardization

We are jumping on to another topic , have patience , things will connect evenutally.

Consider data points for variable X : $x_1, x_2, x_3....x_n$. For these data points we can calculate average and standard deviation as follows.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n}}$$

Lets say now we create another variable Y which takes values : $y_1, y_2, y_3....y_n$ such that:

$$y_i = \frac{x_i - \bar{X}}{S_x}$$

Lets calculate mean and standard deviation for this new variable Y.

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Putting in y_i in terms of x_i

$$\bar{Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})}{S_x n}$$

but

$$\sum_{i=1}^{n} (x_i - \bar{X}) = 0$$

hence

$$\bar{Y} = 0$$

Now lets calculate standard deviation of Y

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}$$

We have already found that $\bar{Y}$ is zero. putting y_i in terms of x_i in the equation above:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{S_x^2 * n}}$$

$$S_y = \frac{1}{S_x} * \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}}$$

We know that factor on the right most is same as formula for S_x. Putting in that gets us:

$$S_y = \frac{S_x}{S_x} = 1$$

These two results $\bar{Y} = 0$ and $S_y = 1$ are very important. Observe that, we did not use any hard coded numbers here. X can be any variable , and if its standardaized in the same manner *[subtract mean and divide by standard deviation]* then the resulting variable will have its mean zero and standard deviation as 1. Also if we are given standardized variable values y_i and mean μ & S.D. σ of non-standardize variable X. We can get values of x_i by using this:

$$x_i = \sigma * y_i + \bar{X}$$

Standard Normal Distribution

We have seen general equation of normal distribution already with parameters μ and σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

Standard Normal Distribution is one specific distribution with $\mu = 0$ and $\sigma = 1$.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp \frac{-x^2}{2}$$

Few results from standard normal distribution

In literature else where you'll find standard normal varibale being represented as Z. We'll follow the same notation.

Following result is self explanatory given that mean of standard normal distribution is 0 and normal distributions in general are symmetric.

$$P(Z \geq 0) = P(Z \leq 0) = 0.50$$

Few other results which also can be broken in symmertical halves as above have been derived as follows:

$$P(-1 \leq Z \leq 1) = 0.682$$

$$P(-2 \leq Z \leq 2) = 0.954$$

$$P(-3 \leq Z \leq 3) = 0.996$$

Since the distribution is symmetric as mentioned before, one sided probabilities as well as remainder probabilites can be easily calculated . For example, using one of the above results you can calculate:

$$P(Z \leq 1) = \frac{0.682}{2} = 0.341$$

$$P(Z \geq 2) = 1 - P(Z \leq 0) - P(0 \leq Z \leq 2) = 1 - 0.5 - \frac{0.954}{2} = 0.023$$

Confidence Intervals

I'm listing few more results from standard normal distribution, similar to as we listed above.

$$P(-1.645 \leq Z \leq 1.645) = 0.90$$
$$P(-1.96 \leq Z \leq 1.96) = 0.95$$
$$P(-2.576 \leq Z \leq 2.576) = 0.99$$

These probability results are stratight forward. One way to interpret any of them is, if you randomly pick an observation from a population which follows standard normal distribution , there is 90% chance/probability that it'll fall in the interval [-1.645,1.645]

Formaly , [-1.645,1.645] is called 90% confidence interval for standard normal distribution. Of course there can be other arbitrary intervals [a,b] such that :

$$P(a \leq Z \leq b) = 0.90$$

But there is only one possible 90% interval which is symmetric about mean of distribution. These symmetric intervals are called Confidence Intervals . These are also written as CI in short notation.

Extrapolating Results for general normal distribution

From our lessions of standardization we can say that if we have a general normal distribution following variable X such that

$$X \sim N(\mu, \sigma^2)$$

We can standardize this as follows

$$Z = \frac{X - \mu}{\sigma}$$

Where Z will follow normal distribution but with mean 0 and standard deviation 1. which is standard normal distribution. So given any of the above probability results we can do following *[Lets for X $\mu = 10$ and $\sigma = 2$]:

$$P(-1.645 \leq Z \leq 1.645) = 0.90$$

$$P(-1.645 \leq \frac{X-\mu}{\sigma} \leq 1.645) = 0.90$$

$$P(-1.645 * \sigma + \mu \leq X \leq \mu + 1.645 * \sigma) = 0.90$$

putting in values of μ and σ for X in above , we get

$$P(5.065 \leq X \leq 14.935) = 0.90$$

Which tells us that 90% confidence interval for $X \sim N(10, 9)$ is [5.065,14.935]. If you are wondering that this CI doesn't look symmetric. It is , about the mean 10.

It doesn't make sense to claculate these CI results for infinitely many possible normal distributions . Howeve using the technique shown above we can calculate this for any general normal distribution , with having pre-calculated results for just standard normal distribution.

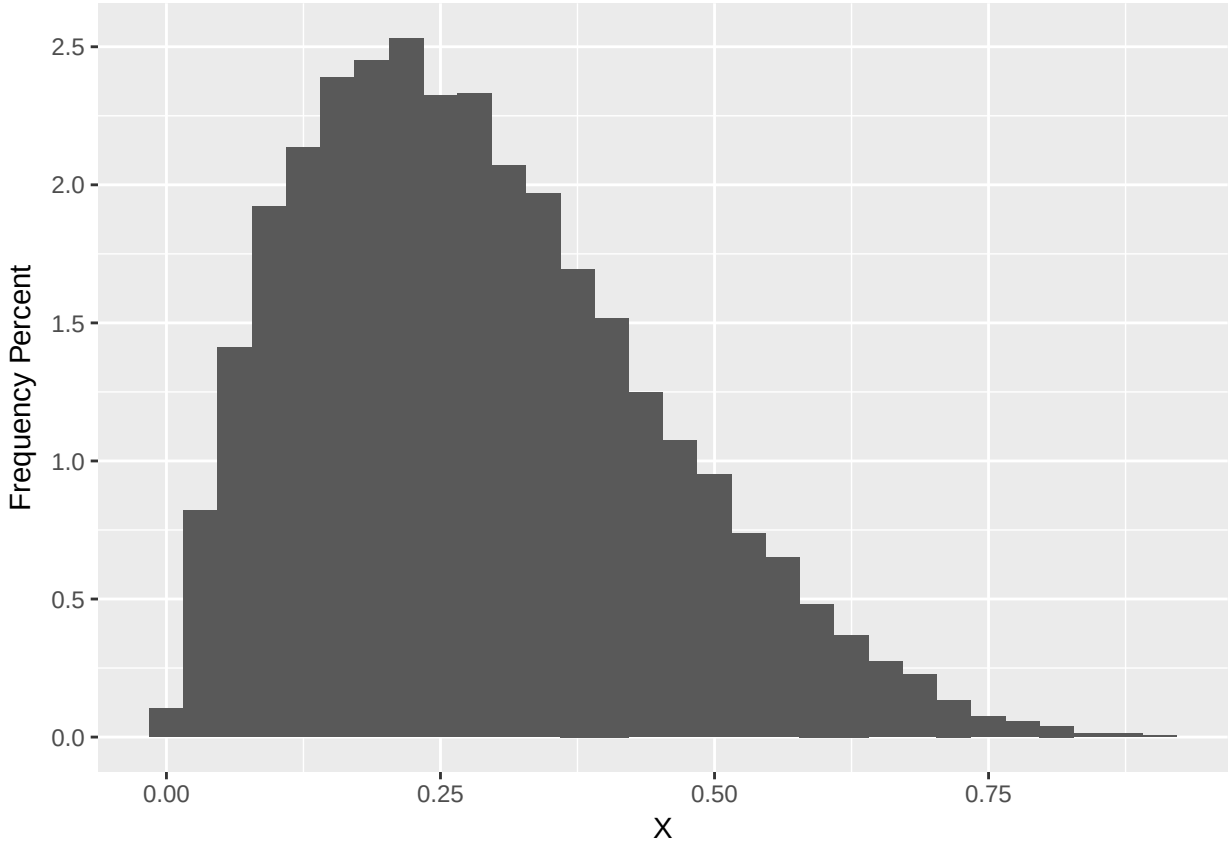
But there are many other possible distributions out there, why are we so keen about exploring Normal distribution? Next segment will give you that answer.

Central Limit Theorem

Averages of samples [of sufficient large size] follow normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ where (μ , σ^2) are mean and variance of the population and n is the sample size. If sample size is small, normal distribution is replaced by t-distribution. [This is irrespective of distribution being followed by the variable values]

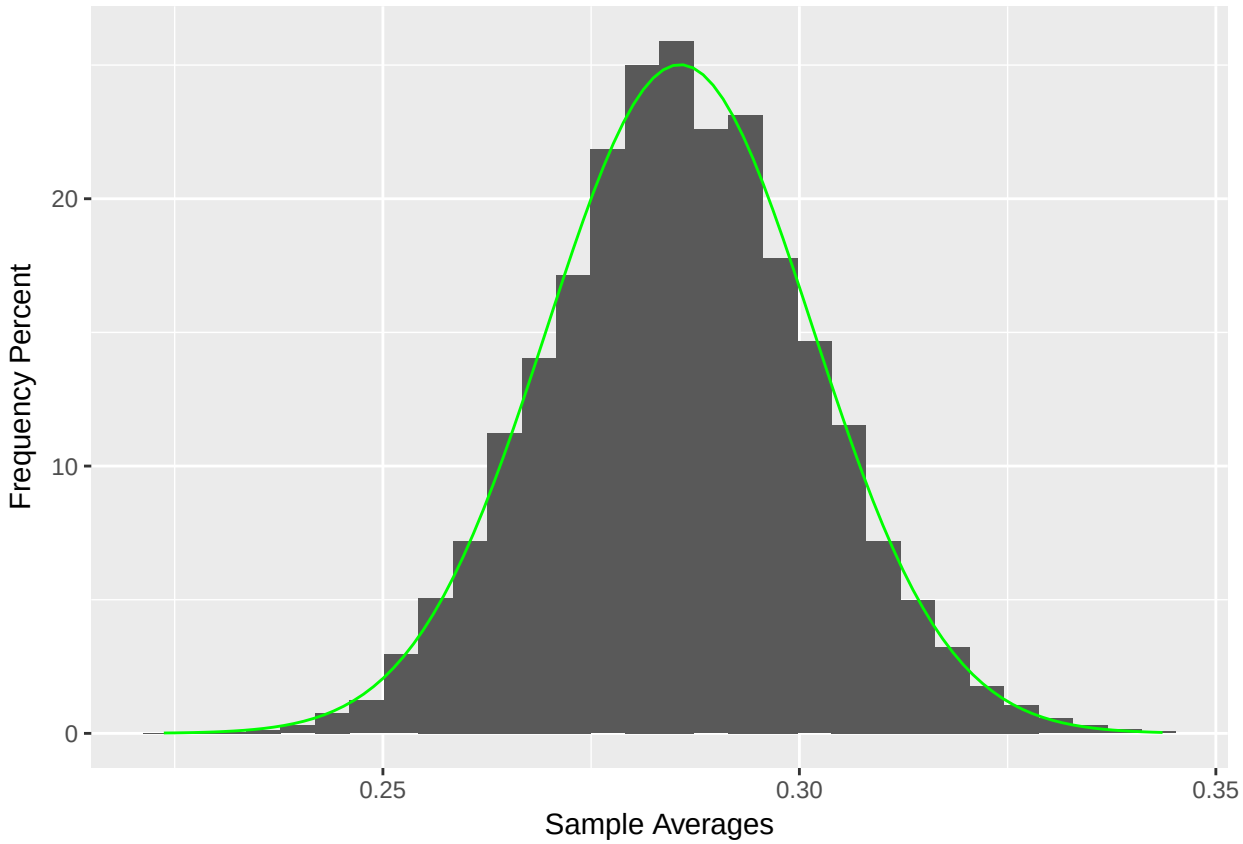
This statement is backbone of what we are going to see in hypothesis testing framework.Lets try to understand what it is saying with some example.

```
set.seed(1)
d=data.frame(X=rbeta(20000,2,5))
p=ggplot(d,aes(x=X))
p+geom_histogram(aes(y=..density..))+ylab("Frequency Percent")
```



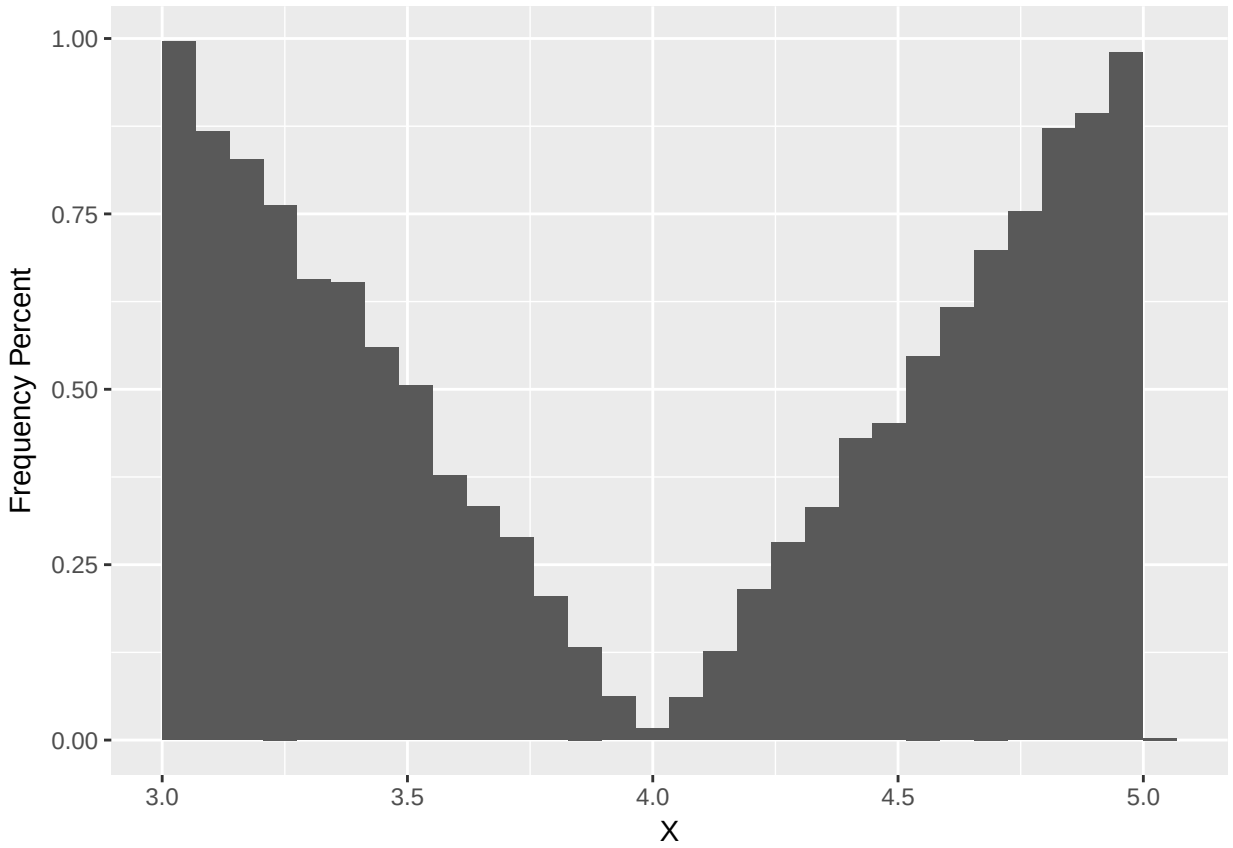
Clearly the distribution of the variable X here doesnt look normal or symmetric.What we are going to do next is to text 10000 samples of size 100 each from this population and then plot histogram for their **Averages** and see whether that looks like a normal distribution.

```
k=numeric(10000)
for (i in 1:10000){
  j=sample(1:20000,100)
  k[i]=mean(d$X[j])
}
d1=data.frame(k)
p=ggplot(d1,aes(x=k))
p+geom_histogram(aes(y=..density..))+ylab("Frequency Percent")+xlab("Sample Averages")+
  stat_function(fun=dnorm,
                args=list(mean=mean(d1$k),sd=sd(d1$k)),
                color="green")
```



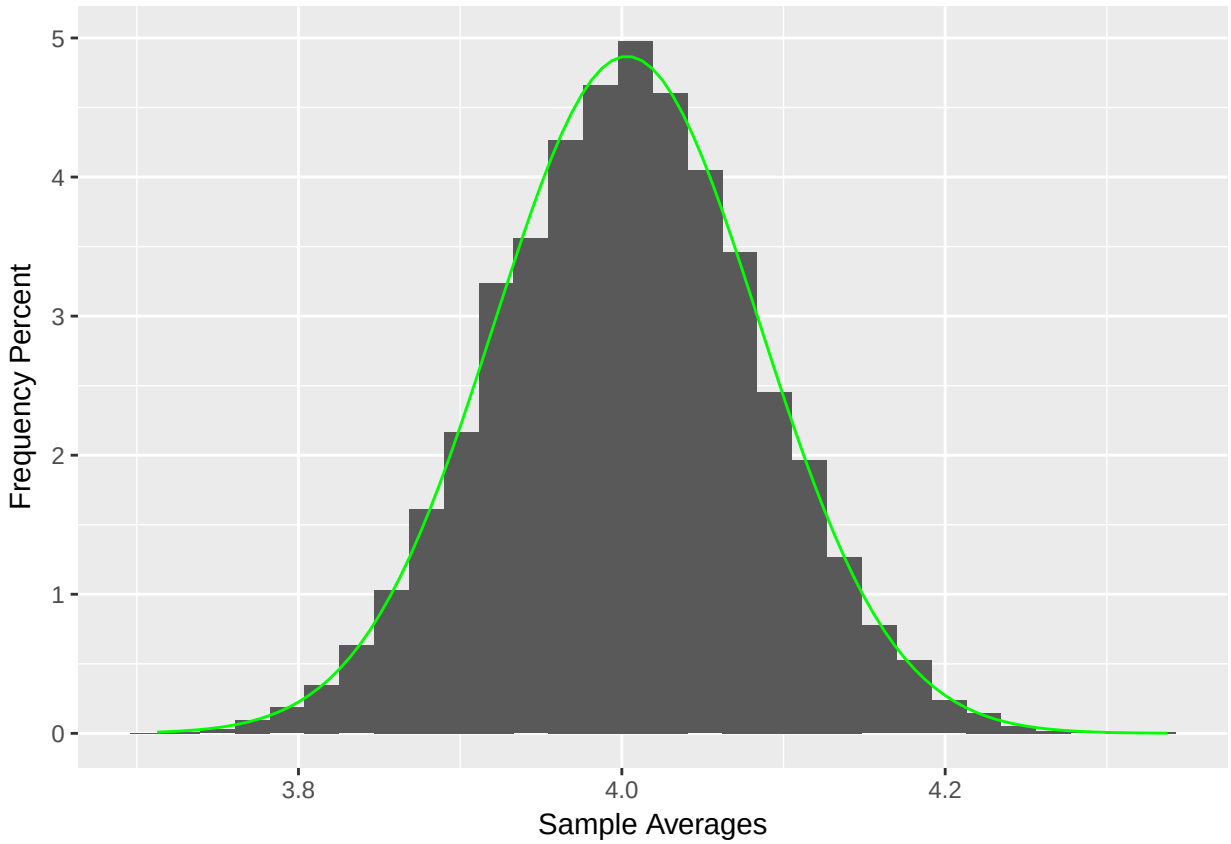
And it surprisingly does. Lets try this with some other more radical distribution. How about a parabolic distribution, lets see if sample averages still follow normal distribution.

```
set.seed(1)
t=runif(20000)
set.seed(2)
k=runif(20000)
X=ifelse(k>0.5,4+sqrt(1-1.332*t),4-sqrt(1-1.332*t))
d=data.frame(X=X)
p=ggplot(d,aes(x=X))
p+geom_histogram(aes(y=..density..))+ylab("Frequency Percent")
```



This again is distribution which is no where close to a normal distribution, lets see how sample averages behave in this case.

```
k=numeric(10000)
for (i in 1:10000){
  j=sample(1:20000,100)
  k[i]=mean(d$X[j],na.rm=T)
}
d1=data.frame(k)
p=ggplot(d1,aes(x=k))
p+geom_histogram(aes(y=..density..))+ylab("Frequency Percent")+xlab("Sample Averages")+
  stat_function(fun=dnorm,
                args=list(mean=mean(d1$k),sd=sd(d1$k)),
                color="green")
```



And again, here it is . Sample averages seem to follow normal distribution in this case as well. Without going into mathematical details of Central Limit Theorem, our take aways from here is that irrespective of underlying population distribution, samples averages follow normal distribution.

Hypothesis Testing

Now we have gone through all the tools that are necessary to take the next step. Consider that I believe Average Marks obtained by Highschool Students of Maharashtra Board is 85. This is my null hypothesis which i have great faith in.

$$H_0 : \mu = 85$$

Now somebody comes and proposes that this null hypothesis which I have great faith in is not true and Average Marks are not equal to 85. That is Alternate Hypothesis.

$$H_a : \mu \neq 85$$

To check that whether this claim of null hypothesis not being correct is true, I will take a sample of student scores. We know that this sample estimate will never be actually equal to 85 or other true mean if thats the case; owing to inherent errors associated with estimates.

Now question arises when would I refute the claim by H_a. Since I have great faith in H_0 , unless this sample presents an extreme evidence against H_0, I will stand by H_0 and reject claims by H_a.

What is an extreme evidence then? How do i draw the line. Due to CLT I know that sample averages follow normal distribution with $(\mu, \frac{\sigma^2}{n})$. Using this I can say build a 95% confidence interval for my H_0 and If

sample which i draw happens to have average marks fall outside of this interval, I'd consider that to be an extreme evidence against H_0 and reject H_0.

Lets say I drew a sample of size 100, average marks in the sample came out to be 83. Standard deviation of the population estimated from this population came out to be 10. *$[\sigma = 10]$.

Which means sample averages $\sim N(85, 1)$. I can convert this standard normal distribution. Standardized value of the sample Average would be:

$$= \frac{83 - 85}{1} = -2$$

For a standard normal distribution, 95% confidence interval is [-1.96,1.96]. Since the standardised value falls outside of even a 95% confidence interval, we conclude that our H_0 does not hold and we conlude in favour of H_a.

Acceptance & Rejection Region

There is no rule for taking a 95% CI as limit for accepting or rejecting H_0. It only depicts how comfortable or keen we are to accept/reject H_0. Lets call this 95% as *size* of my acceptance region. The *size* of my rejection region will be 5% [1-0.95]. This is denoted by α.

Choice of α is a subjective decion of business analyst and very much depends on the business process goal. Consider these extreme cases where we'll discuss what kind of α do we decide to take.

- I am manufacturing plant owner , I want to check whether the iron rings which i produce are deviating from the diameter specifications. If my tests conclude that there is deviation I'll to dismantle and re-assemble the entire which is a huge cost in comparison to go on with manufacturing of iron rings with small deviations. Which implies, unless i get a very extreme evidence against my null hypothesis *[being average diameter of iron rings= specifications]* I will not reject null hypothesis. In this case I would keep my α small.
- I am in the business of launching Rockets. Fuel efficiency is of critical importance and should not deviate from the specifications. I get fuel barrels from an external supplier. I want to check the quality of this fuel by measuring fuel efficiency. In this case I'd like to capture even small deviations. I'll keep my α relaxed.

Standard practice in industry is to keep $\alpha = 0.05$ *or* 5%.

One Side and Two Sided Test

Consider this alternate hypothesis again:

$$H_a \neq 85$$

In order to conclude in favor of this ,my extreme evidence can be on either side of the proposed mean *[85]* by my null hypothesis. Too small, or too large, both will be extreme evidences in favor of alternate hypothesis.

Test, like this are called two sided test. We have already done this once above. Now, if alternate hypothesis was not equality, such as :

$$H_a < 85 \text{ or } H_a > 85$$

Then you need extreme evidence on left/right hand side only. These are called one sided tests.

P-value and Alpha

Once we have decided our α , we can conclude by checking whether sample mean falls beyond the CI limits to conclude in favor/against H_0.

There is another way to look at this. Consider this diagram for a one sided test.

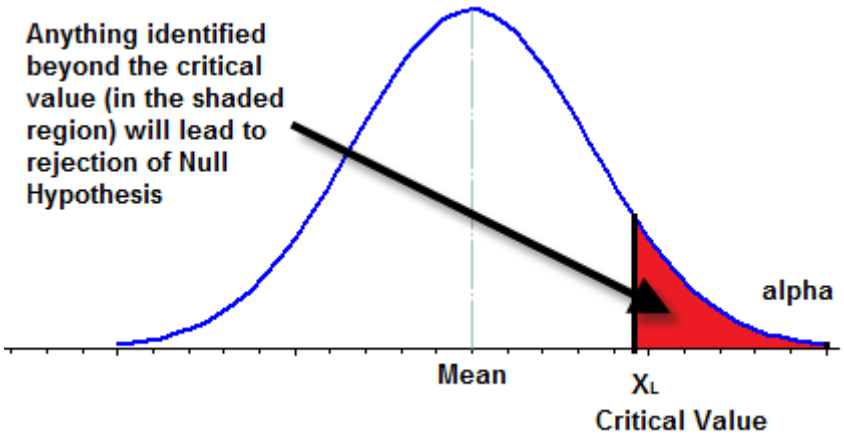


Figure 3: alpha

What this diagram tells us that essential α is nothing but area under the distribution curve beyond the interval limit [in this case X_L]. As mentioned in the figure if the sample estimate falls beyond X_L or in the shaded region, we'll conclude against H_0.

Consider diagram for anothe one sided test:

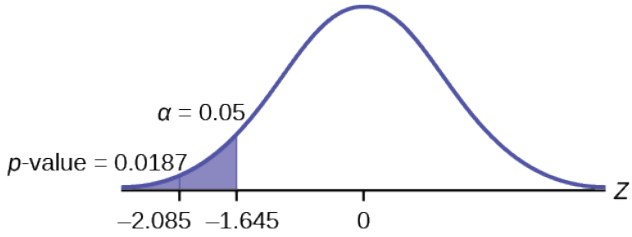


Figure 4: pval

You can see that sample estimate falls in the shaded region beyond the interval limits, hence we'll conclude against H_0 or reject H_0. You can also see that the area under the curve beyond *[-2.085]* will definitley be less than α *[area under the curve beyond -1.645]*. This area under the curve beyond *[-2.085]* is called p-value associated with the sample estimate. You can see that if :

$$p - value < \alpha : \ Reject \ H_0$$

$$p - value > \alpha : \ Failed \ to \ Reject \ H_0$$

Your software output for hypothesis tests will be in terms of p-values and using above mentioned guidelines, you can interpret the results.

Errors Associated with Hypothesis Testing

Hypothesis testing frame work that we just is of course not fool proof. In fact there are definite probabilities of error associated with every hypothesis test.

For example consider this pair of null and alternate hypothesis:

$$H_0 : \mu = \mu_0$$

&

$$H_a : \mu = \mu_a \; where \; \mu_a > \mu_0$$

Now even if the null hypothesis is true is there probability α that sample estimate will fall in the rejection region and you will end up rejecting H_0 despite its being true. This is called **Type I Error* . This is equal to α. The more you increase α , probability of Type I error goes up.

Similarly there is a proability that alternate hypothesis is true, but sample estimate falls in the acceptance region and you end up accepting null hypothesis despite its being not true. This is called **Type II Error* . This is depicted by β in literature. As clear from the figure given below, you can not decrease both Type I and Type II errors simultaneously. If you decrease one , its counterpart goes up. Also there is no credible way to exactly asses β because μ_a is unknown.

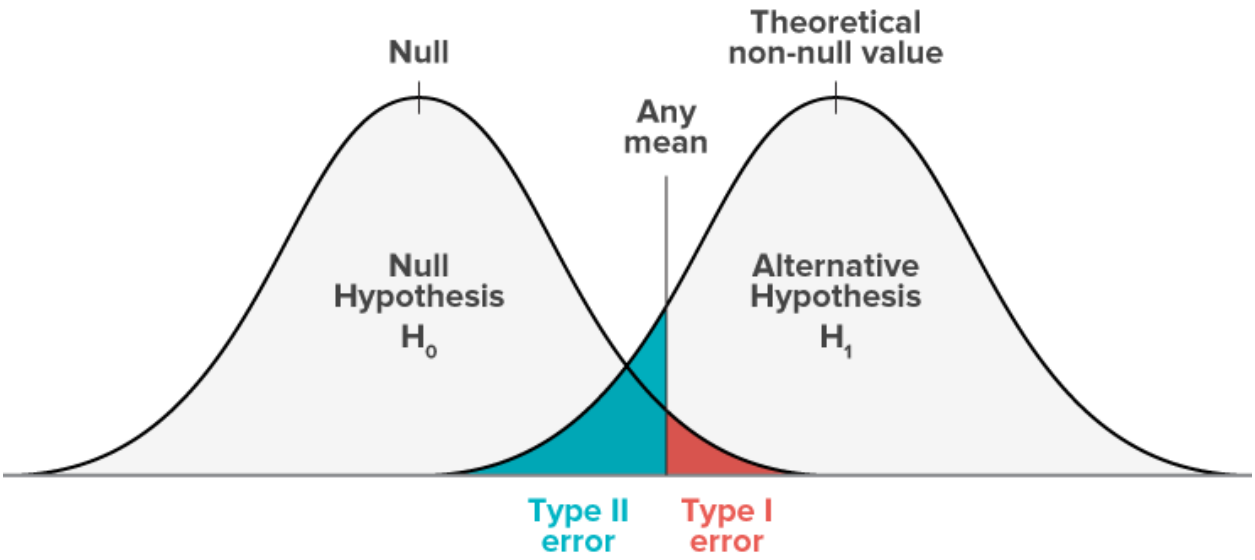


Figure 5: Type I & Type II Errors

Summarising Hypothesis Testing Framwork : Steps to interpret a test results

1. H_0 : You need to know null hypothesis associated with the test , so that you know what you are concluding in favor or against.
2. α : You decide this depending upon your business process. Industry standard is 0.05
3. Once you carry out the test using your sample, if resulting p-value $> \alpha$ you conclude in favor of H_0. if p-value $< \alpha$, you conclude against H_0

few things to note:

- α does not depend on anything accept your decision

- p-value does not depend on α. It depends on H_0/H_a and your sample data. p-value will change only when either of these change.

We'll now formally discuss various hypothesis tests without getting into mathematical details

One Sample T-test

We have infact already done this. Let me however clarify the name of the test. If you recall sample averages follow normal distribution for large enough sample size which is typically considere to be 30. If sample size goes below 30, sample averages instead follow T-distribution. T-distribution is very similar to normal, just a little thicker on tails. In fact for sample size more than 30, T-distribution and Normal distribution take almost identical values. You can use T-distribution through out the range of sample sizes and it wont make much difference.This is where the name T-test comes from.

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0 \; OR \; \mu > \mu_0 \; OR \; \mu < \mu_0$$

Paired sample T-test

Finding out whether the average statistic is equal to some value is not the only kind of business problem that we are intrested in solving. Paired sample t-test is used to find out difference in average value of population parameter "before and after". Although the term "before and after" might be misleading at times.

Let me give you few examples ,

- You want to check whether a medicine for lowering sugar levels in blood should be approved or not. You'll measure blood levels of say 100 patients before taking the medicine and after completing the medicine course. You'd check whether average sugar levels in these patients have gone down significantly after completing the medicine course or not.
- You want to check whether performance in Mathematics and English is very different for a certain school. You'd check if average scores in Mathematics and English are significantly different.

What makes it "paired" is that both kind of observation have same source. In first example , before and after sugar levels belonged same patients. In second example , scores in Mathematics and English belonged to same students.

$$H_0 : \mu_1 = \mu_2 \; OR \; \mu_1 - \mu_2 = \delta = 0$$
$$H_a : \delta \neq 0 \; OR \; \delta > 0 \; OR \; \delta < 0$$

Unpaired Two Sample T-test

This is used when you want to check whether two group means are significantly different or not. For example :

- Whether corporate salaries for same designations are different for Males and Females. You'll take samples from Male workers and Female workers separately and check whether their average salaries are different or not.

Note that here observations are not paired , they do not have same source. Hypothesis remains same just that the underlying statistics methods change slightly. We don't need to worry about that.

$$H_0 : \mu_1 = \mu_2 \; OR \; \mu_1 - \mu_2 = \delta = 0$$
$$H_a : \delta \neq 0 \; OR \; \delta > 0 \; OR \; \delta < 0$$

ANOVA : Analysis of Variance

We'll first discuss what do we use ANOVA for. We'll get into a little bit of mathematics involved with ANOVA which you can skip if you want to.

The T-test is limited to 2 groups at a time in your data. Now lets say you want to check if average yield of a rice field is varying across states in India or not, so that you can accordingly plan farmer subsidies given to rice farmers across state by the central government. You can not use T-test for it. Hypothesis for ANOVA are as follows:

$$H_0 \ : \mu_1 = \mu_2 = \mu_3 ..... = \mu_p = \mu_0 \ where \ p \ is \ number \ of \ groups \ present \ in \ the \ data$$
$$H_a \ : \mu_i \neq \mu_0 \ for \ atleast \ one \ i \ in \ [1, 2, 3, ...p]$$

If the p-value for the test comes out to be $< \alpha$ then you might be interested in which group is differing in mean. You can check that by doing bon-feroni test in conjunction with ANOVA. We'll directly see an example for it towards the end.

Now the time for a little mathematics. You can consider ANOVA to be very similar to a linear regression model with all categorical variables. *[If you havent gone through linear regression yet , you can always revisit to understand better]*

Consider Total variance in the data without any groups for variable X

$$Total \ sum \ of \ squares = SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

If all groups means are equal then total within group variance is going to be very close to SST. and Between group variance is going to be close to zero. and

$$SST = SSW + SSB$$

If group means are equal then $\frac{SSB}{SSW}$ will be close to zero. This ratio statistic follows F-distribution .We can a build a confidence interval around 0 and see whether this statistic is significantly different from zero or not. If p-value for this F statitic comes out to be $< \alpha$ then we'll conclude that atleast one of the group means are different.

Chisq Test

So far we have discussed just numeric variables. How about categorical variables. First we need to understand what do we mean categorical variables affecting each other. Consider this cross table of counts student's preference for learning programing across genders

\	Male	Female
Yes	30%	29%
No	70%	71%

You can see that preference for programing among both the genders is roughly same close to 30:70. Only slightly different from females. Now when this slight difference is "significant", can be assesed through chisq test.

$$H_0 \ : \ Categorical \ variables \ have \ no \ effect \ on \ each \ other$$
$$H_a \ : \ Categorical \ variables \ do \ affect \ each \ other$$

Examples with R

We'll be looking at 4 kind of hypothesis tests. Key to understand results of any hypothesis test result is to know what is null hypothesis and how to conclude by looking at p-value.

In short you need to do this:

1. Know what is the null hypothesis H_0. Complimentary to this would be Alternate hypotheis H_a
2. If p-value of the test is smaller than alpha [standard value is 0.05, you can take 0.1,0.01 etc] then conclude against H_0; otherwise in favor of H_0.

Lets import the data

```
wq=read.csv("~/Dropbox/0.0 Data/winequality-white.csv",sep=";")
```

One Sample T-test.

you are checking whther mean of the variable in question is equal to the specified value [H_0=6.10] or not. Changing α does not change p-value ofthe test as that depends on the data only however it does change the confidence intervals being displayed.

In the example given below. Null hypothesis is that mean of fixed_acidity=6.10

```
t.test(wq$fixed.acidity,mu = 6.10)
```

```
##
##  One Sample t-test
##
## data:  wq$fixed.acidity
## t = 62.598, df = 4897, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 6.1
## 95 percent confidence interval:
##  6.831149 6.878426
## sample estimates:
## mean of x
##  6.854788
```

You can see that the p-value is $<2.2 * 10^{-16}$. Which is very low , we'll conclude in favor Alternate hypothesis which is stated as

```
alternative hypothesis: true mean is not equal to 6.1
```

You can also see confidence interval of the mean according to alternate hypothesis. [6.831149 6.878426]. Mean which you propose in the null hypothesis does not fall in this, this is another indicator that null hypotehsis is not true.

Lets provide value of confidence interval other than default 0.05 or 5% as well

```
t.test(wq$fixed.acidity,mu = 6.10,conf.level = 0.01)
```

```
##
##  One Sample t-test
##
## data:  wq$fixed.acidity
## t = 62.598, df = 4897, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 6.1
## 1 percent confidence interval:
##  6.854637 6.854939
## sample estimates:
```

```
## mean of x
##  6.854788
```

Every thing else remains same just the displayed confidence interval changes to [6.854637 6.854939] which is a little shorter than the earlier as you have increase value of α. Remember this confidence interval is coming from alternate hypothesis thats why it shrinks.

You can change alternative hypothesis from the default of ineqaulity also

```
t.test(wq$fixed.acidity,mu = 6.10,alternative ="less" )
```

```
##
##  One Sample t-test
##
## data:  wq$fixed.acidity
## t = 62.598, df = 4897, p-value = 1
## alternative hypothesis: true mean is less than 6.1
## 95 percent confidence interval:
##      -Inf 6.874625
## sample estimates:
## mean of x
##  6.854788
```

You can see here that p-value is large, in fact it is very close to 1, we'll take against the alternate hypotehsis that mean is less than 6.1.

```
t.test(wq$fixed.acidity,mu = 6.10,alternative ="greater" )
```

```
##
##  One Sample t-test
##
## data:  wq$fixed.acidity
## t = 62.598, df = 4897, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 6.1
## 95 percent confidence interval:
##  6.834951      Inf
## sample estimates:
## mean of x
##  6.854788
```

Here your decision will be in favor of alternate hypothesis that is mean is greater than 6.1

Paired Two Sample T-test :

Data for this is in SAS format. We'll use the function `read.sas7bdat` from the package `sas7bdat`.

In the example given below, H_0: Mean of Diffence between means of scores in write and read=0

```
library(sas7bdat)
d=read.sas7bdat("~/Dropbox/0.0 Data/hsb2.sas7bdat")
glimpse(d)
```

```
## Observations: 200
## Variables: 11
## $ id      <dbl> 70, 121, 86, 141, 172, 113, 50, 11, 84, 48, 75, 60, 95...
## $ female  <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ race    <dbl> 4, 4, 4, 4, 4, 4, 3, 1, 4, 3, 4, 4, 4, 4, 3, 4, 4, 4, ...
## $ ses     <dbl> 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 1, 1, 3, 2, ...
```

```
## $ schtyp  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, ...
## $ prog    <dbl> 1, 3, 1, 3, 2, 2, 1, 2, 1, 2, 3, 2, 2, 2, 2, 1, 2, 1, ...
## $ read    <dbl> 57, 68, 44, 63, 47, 44, 50, 34, 63, 57, 60, 57, 73, 54...
## $ write   <dbl> 52, 59, 33, 44, 52, 52, 59, 46, 57, 55, 46, 65, 60, 63...
## $ math    <dbl> 41, 53, 54, 47, 57, 51, 42, 45, 54, 52, 51, 51, 71, 57...
## $ science <dbl> 47, 63, 58, 53, 53, 63, 53, 39, 58, 50, 53, 63, 61, 55...
## $ socst   <dbl> 57, 61, 31, 56, 61, 61, 61, 36, 51, 51, 61, 61, 71, 46...
```

```
t.test(d$read,d$write,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  d$read and d$write
## t = -0.86731, df = 199, p-value = 0.3868
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7841424  0.6941424
## sample estimates:
## mean of the differences
##                  -0.545
```

p-value of test is high [0.3868], we'll be concluding in FAVOUR of H_0 and say that means of the variable write and read are NOT different.

By default this assumes that H_0 is $\delta = 0$, you can change that specifying `mu=` some other value.

```
t.test(d$read,d$write,paired = TRUE,mu=-0.50)
```

```
##
##  Paired t-test
##
## data:  d$read and d$write
## t = -0.071612, df = 199, p-value = 0.943
## alternative hypothesis: true difference in means is not equal to -0.5
## 95 percent confidence interval:
##  -1.7841424  0.6941424
## sample estimates:
## mean of the differences
##                  -0.545
```

you can see that p-value for the test is now very close to one, because the null hypothesis which you are suggesting is very close to result of the sample [mean of differences being -0.543]

you cal also use `conf.level` and `alternative` option here to ahiceve the same thing. I'm leaving that for you to try on your own.

Unpaired Two Sample T-test

```
unique(wq$quality)
```

```
## [1] 6 5 7 8 4 3 9
```

We need to find out whether alcohol percentages in wine vary across quality ratings . Remember that you can do this test only if number of classes are two.for more classes you'll have to use ANOVA.

Here variables values are from the same variable but belonging to different classes. They are not "paired" . In this example

H_0: Difference between means of alchol in class quality=3 and quality=9 is equal to zero

However before we go ahead and do the unpaired T-test for these two groups , we need to know another thing. The underlying distribution changes slightly depending upon whether variance of these two groups are same or not. We can find that out by doing a variance equivalence test first. [This is also known as F-tes]

```
var.test(wq$alcohol[wq$quality==3],wq$alcohol[wq$quality==9])
```

```
##
##  F test to compare two variances
##
## data:  wq$alcohol[wq$quality == 3] and wq$alcohol[wq$quality == 9]
## F = 1.459, num df = 19, denom df = 4, p-value = 0.7784
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1701394 5.1921582
## sample estimates:
## ratio of variances
##           1.459002
```

p-value for the test is very high, we'll conclude against the stated alternate hypothesis which says true raio of variances is not equal to 1 or in other words , variances are not equal. Once we know that we can do our unpaired t test.

```
t.test(wq$alcohol[wq$quality==3],wq$alcohol[wq$quality==9],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  wq$alcohol[wq$quality == 3] and wq$alcohol[wq$quality == 9]
## t = -3.0837, df = 23, p-value = 0.005246
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.0659873 -0.6040127
## sample estimates:
## mean of x mean of y
##    10.345    12.180
```

Here the p-value for the test is pretty low, we'll conclude in favor of stated alternate hypothesis which is true difference in means is not equal to zero or in other words, average alcohol content for quality rating wine 3 and 9 are singnficantly different.

In the above test we checked whether avaerage alcohol content is statistically different for classes defined by quality=3 or quality=9. Next you might be interested in whether alcohol averages are different across all the classes defined by quality variable [not just two]

ANOVA

In Case of ANOVA, test is based on F distribution, Test statistic is called F-statistics. Way to conclude in favour/against the Ho remains same

H_0 : means of variable in question is same in all the classes H_a : mean in Atleast one class is different from the rest

```
fit=aov(alcohol ~ quality ,data=wq)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
```

```
## quality        1   1407  1407.0    1146 <2e-16 ***
## Residuals   4896   6009     1.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table here look a bit different because of test being based on F-distribution. We'll be looking at p-value below `Pr(>F)`. We see that p-value [$<2 * 10^{-16}$] is very small. We conclude that Atleast one class mean is different.

Now to figure which class mean is differnt we need to do a pairwise bonferroni test.

```
pairwise.t.test(wq$alcohol, wq$quality, p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  wq$alcohol and wq$quality
##
##   3       4       5       6       7       8
## 4 1.00000 -       -       -       -       -
## 5 0.60054 0.00278 -       -       -       -
## 6 1.00000 3.6e-05 < 2e-16 -       -       -
## 7 0.00068 < 2e-16 < 2e-16 < 2e-16 -       -
## 8 1.1e-05 < 2e-16 < 2e-16 < 2e-16 0.06126 -
## 9 0.01566 0.00086 2.5e-05 0.02080 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

by looking at this table we can see that p-value for the test tell us that

- alcohol content for quality rating 9 is singnificantly different from rating (3,4,5,6) [all p-values are low] where as it is not so different from rating (7,8)

we have made comparison of alcohol content of quality rating 9 with rest of the quality ratings already, as we move forward quality rating 9 comaprison need not be done. same will happen with subsequent quality ratings as well.

- alcohol content for quality rating 8 is singnificantly different from rating (3,4,5,6,7) [all p-values are low]
- alcohol content for quality rating 7 is singnificantly different from rating (3,4,5,6) [all p-values are low]
- alcohol content for quality rating 6 is singnificantly different from rating (4,5) [all p-values are low] and is similar to quality rating 3
- alcohol content for quality rating 5 is singnificantly different from rating 4 [all p-values are low] and is similar to quality rating 3
- alcohol content for quality rating 3 and 4 are similar too.

Now you might be wondering how is it possible that 3 and 4 are similar , 3 and 5 are similar but 4 and 5 are significantly different.

Consider this scenario, you decide that a difference of 50 points in score is significant, but anything less than that is not a significant difference. Say A scored 40 points , B scored 70 points and C scored 100 points. According to the criterion A and B dont have significantly different scores , same goes for B and C taken together , but when you consider A and C, difference is 60 points which know is significantly different indicator. Same happened in the scenario above.

Chisq Test

tests for categorical variable relative frequencies.

```r
chisq.test(table(d$race),p=c(0.1,0.1,0.1,0.7))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(d$race)
## X-squared = 5.0286, df = 3, p-value = 0.1697
```

This tells you whether different categories in the variable race are similar to your assumption of relative frequencies which here is (10% 10% 10% 70%).

p-value for the test comes out to be 0.1697 which is greater than standard α value of 0.05 , hence we conclude that relative frequency distribution of different categories of the variable race is different than (10% 10% 10% 70%). You can play around with passing different test realtive frequencies and see how the p-value changes.

```r
chisq.test(table(d$schtyp,d$female))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(d$schtyp, d$female)
## X-squared = 0.00054009, df = 1, p-value = 0.9815
```

P-value is 0.9815 which is larger than α. We conclude that relative frequency distribution of schtype is not affected by gender and vice versa.

Next we want to do the same test for race Vs Socio Economic Status.

```r
chisq.test(table(d$race,d$ses))
```

```
## Warning in chisq.test(table(d$race, d$ses)): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(d$race, d$ses)
## X-squared = 18.516, df = 6, p-value = 0.005064
```

At the bottom of the result , you see a warning that `Chi-squared approximation may be incorrect`. This happens due to some cross tables having counts less than 5. Lets check the cross table.

```r
table(d$race,d$ses)
```

```
##
##      1  2  3
##   1  9 11  4
##   2  3  5  3
##   3 11  6  3
##   4 24 73 48
```

In such scenario we can fisher's test instead. Underlying hypothesis remains same. Lets do fisher exact test.

```r
fisher.test(table(d$race,d$ses))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(d$race, d$ses)
## p-value = 0.007329
```

```
## alternative hypothesis: two.sided
```

This produces result which tells you p-value [0.007329] is very low, you conlude against H_0, which means Socio Economic Status is affected by race.

Normality Test

Although I personally am not a big fan of normality tests [Hypothesis Tests for normality], I am including them for completeness sake anyway. Here are two reasons for my reservations:

1. Normality tests are not very good at capturing what they should, I will demonstrate that , later in this section.
2. Slight deviations from normality dont hurt as much as some people come to believe over time.

You can use function `shapiro.test` to test normality of a variable [containing less than 5000 observations]. For larger number , you can do anderson-darling test using function `ad.test` found in package `nor.test`.

quick examples are given below.

```
x=runif(400)
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95982, p-value = 5.357e-09
```

Null hypothesis for all normality test is that the underlying distribution for the variable in question is Normal. Small p-value for the above example indicates that distribution is not Normal.

If the size is larger than 5000, you get an error.

```
y=rbeta(6000,2,8)
shapiro.test(y)
```

```
## Error in shapiro.test(y): sample size must be between 3 and 5000
```

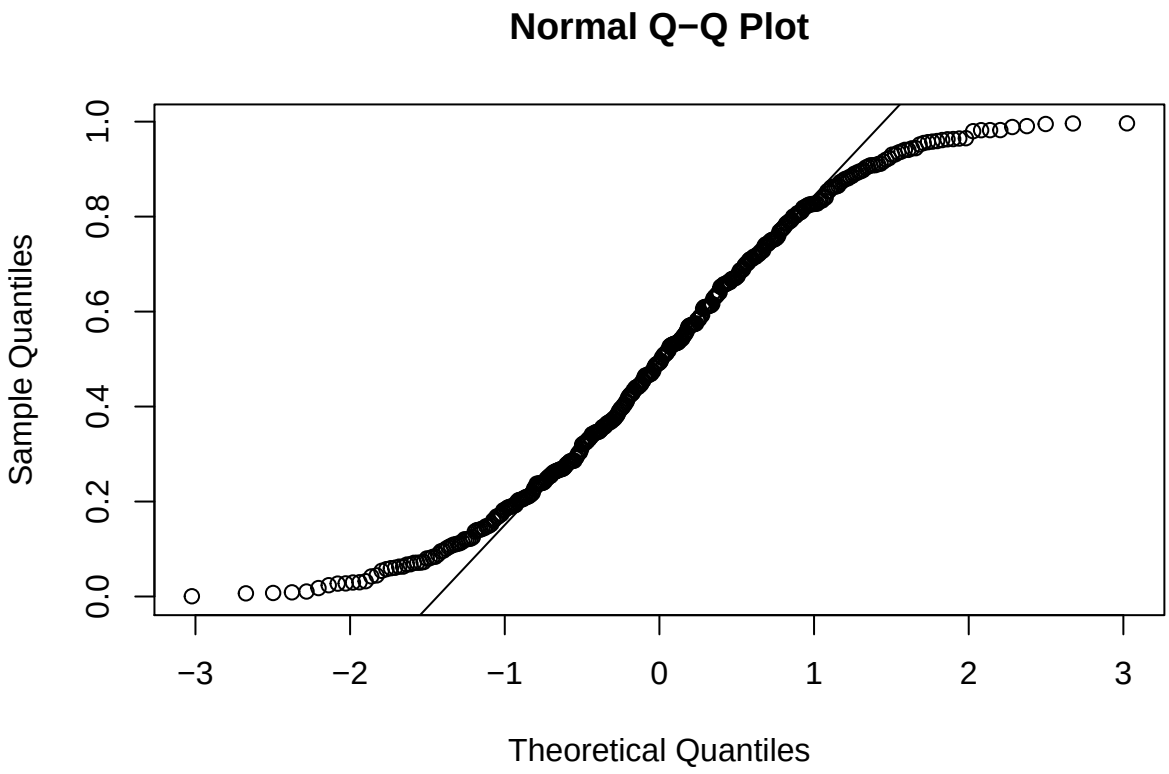
```
library(nortest)
ad.test(y)
```

```
##
##  Anderson-Darling normality test
##
## data:  y
## A = 72.407, p-value < 2.2e-16
```

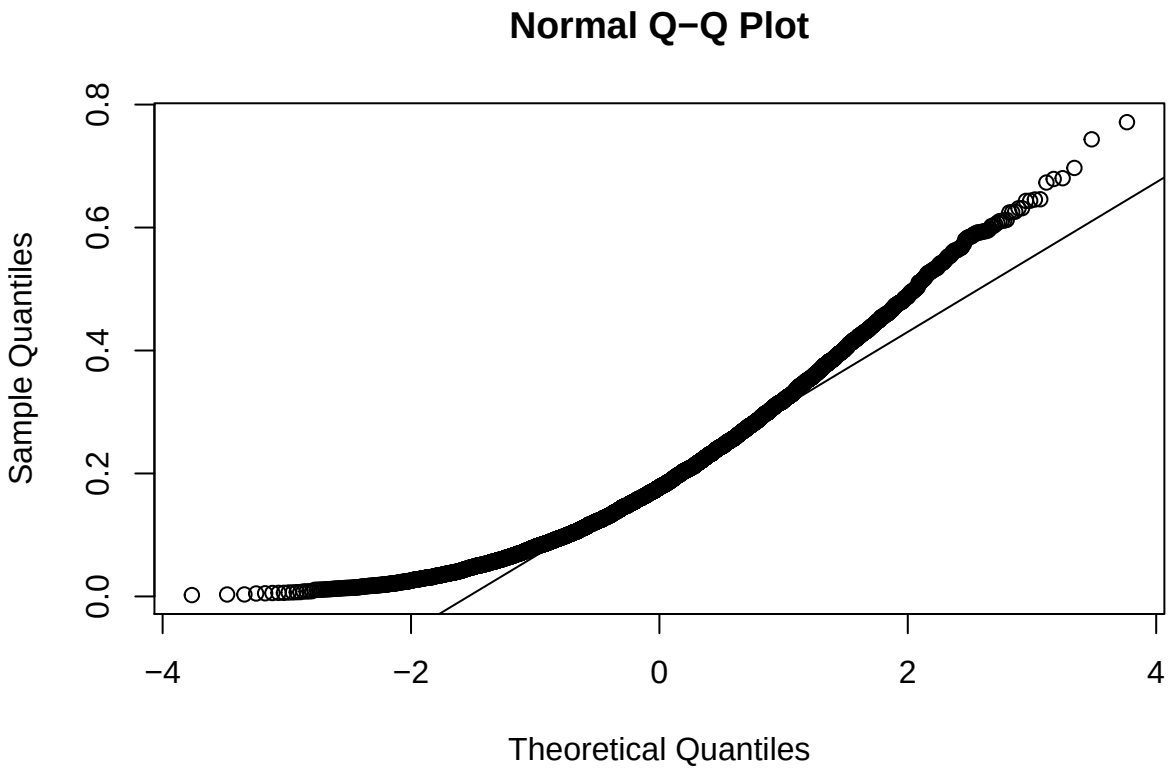
Anderson-Darling test reveals that , again underlying distribution is not normal. You see that these normality test are giving proper results for good deviations from normality. You must be wondering, why then i warned you against them. I will talk about that in a minute but first let me tell you another simpler method to check if data is normal or not.

You can instead plot qqplots, if your data points seem to follow for most of the case on a straight line [dotted line shown in the plot], then it is normal enough for you to relax. thats all.

```
qqnorm(x);qqline(x);
```

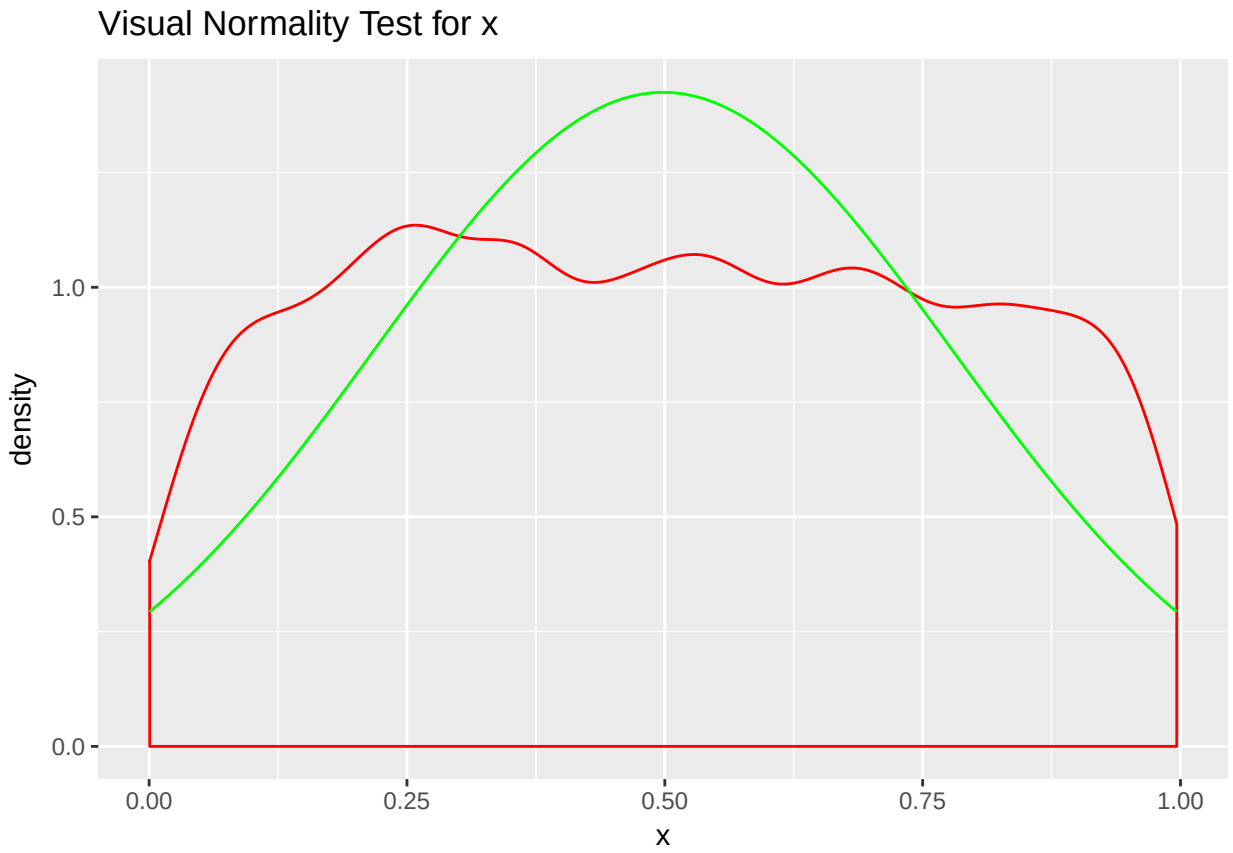



```
qqnorm(y);qqline(y);
```

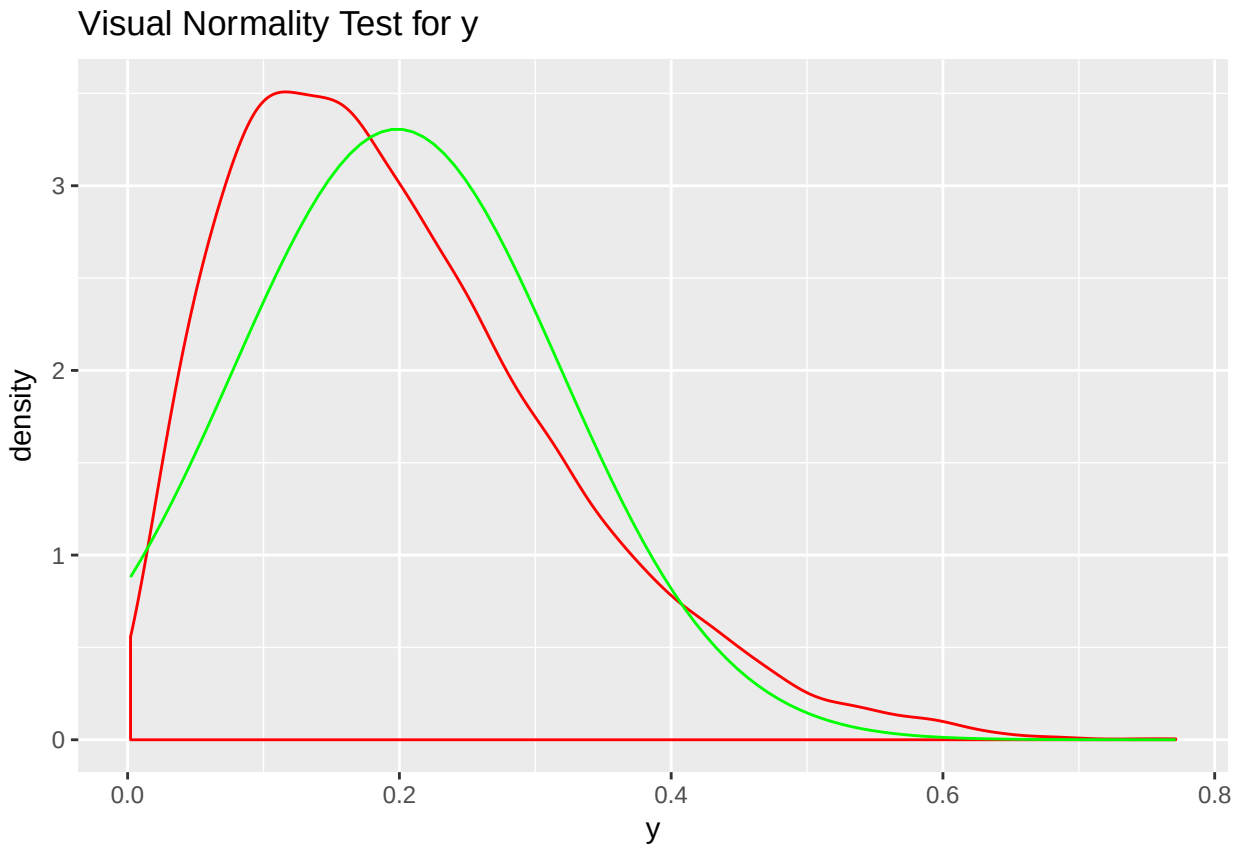


you can see that in both the cases a large portion of the data points do not fall on the straight line. You can also plot density curve and a normal curve as discussed in data viz module.

```
library(ggplot2)
df=data.frame(x,y)
ggplot(df,aes(x))+geom_density(color="red")+
  stat_function(fun=dnorm,args=list(mean=mean(df$x),sd=sd(df$x)),color="green")+
  ggtitle("Visual Normality Test for x ")
```




```
ggplot(df,aes(y))+geom_density(color="red")+
  stat_function(fun=dnorm,args=list(mean=mean(df$y),sd=sd(df$y)),color="green")+
  ggtitle("Visual Normality Test for y ")
```



you can see that these as well tell you that there is significant deviation from normality.

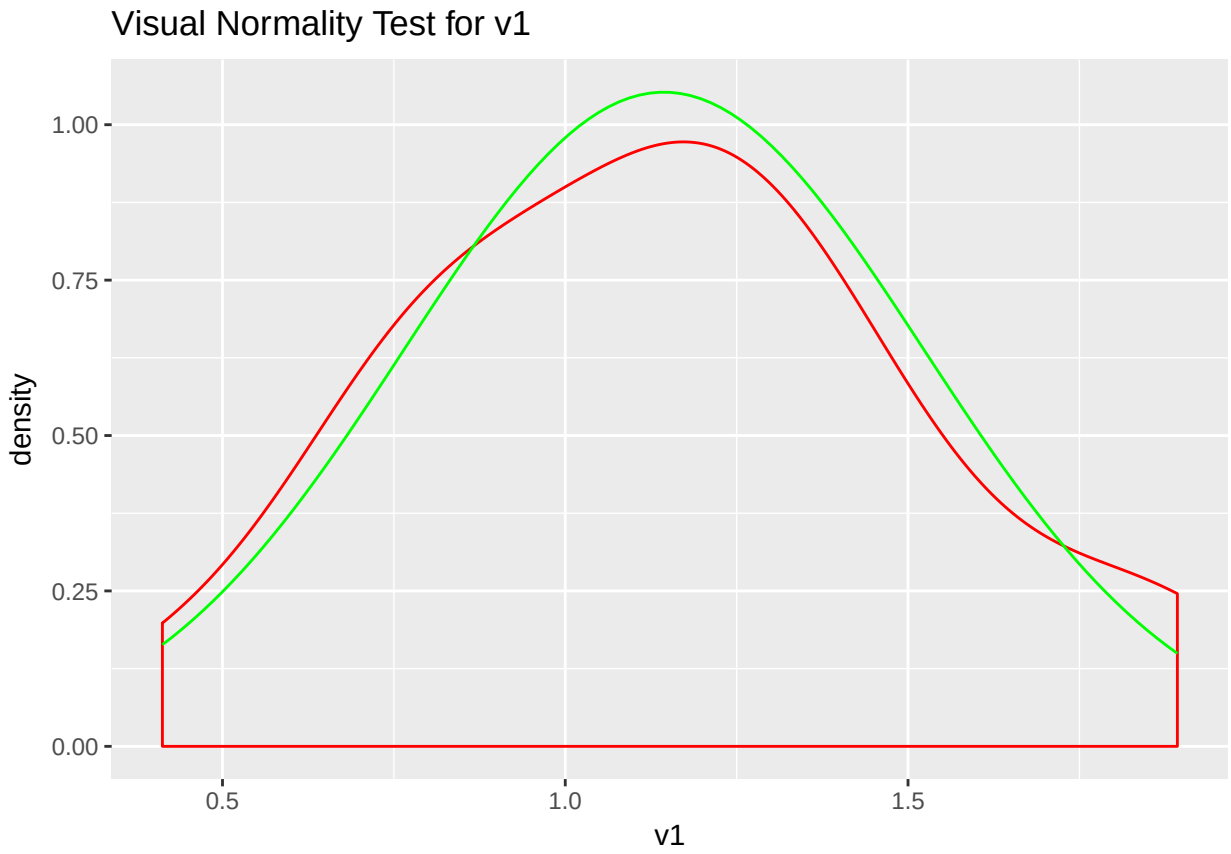
Now lets talk about my concerns about these tests. These tests are sensitive to sample size. If sample size is small, they might fail to catch large deviation, where as if sample size is large, they'll catch even a tiny deviation which might not be worth the hassle anyway.Lets look at these examples to better understand.

```
set.seed(1)
v1=rlnorm(20,0,0.4)
shapiro.test(v1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  v1
## W = 0.98049, p-value = 0.9403
```

Clearly the data is from log normal distribution , not Normal, but the test tells you that data is normal. If you looked at it visually instead, you would have concluded better.

```
df=data.frame(v1)
ggplot(df,aes(x=v1))+geom_density(color="red")+
  stat_function(fun=dnorm,args=list(mean=mean(df$v1),sd=sd(df$v1)),color="green")+
  ggtitle("Visual Normality Test for v1 ")
```



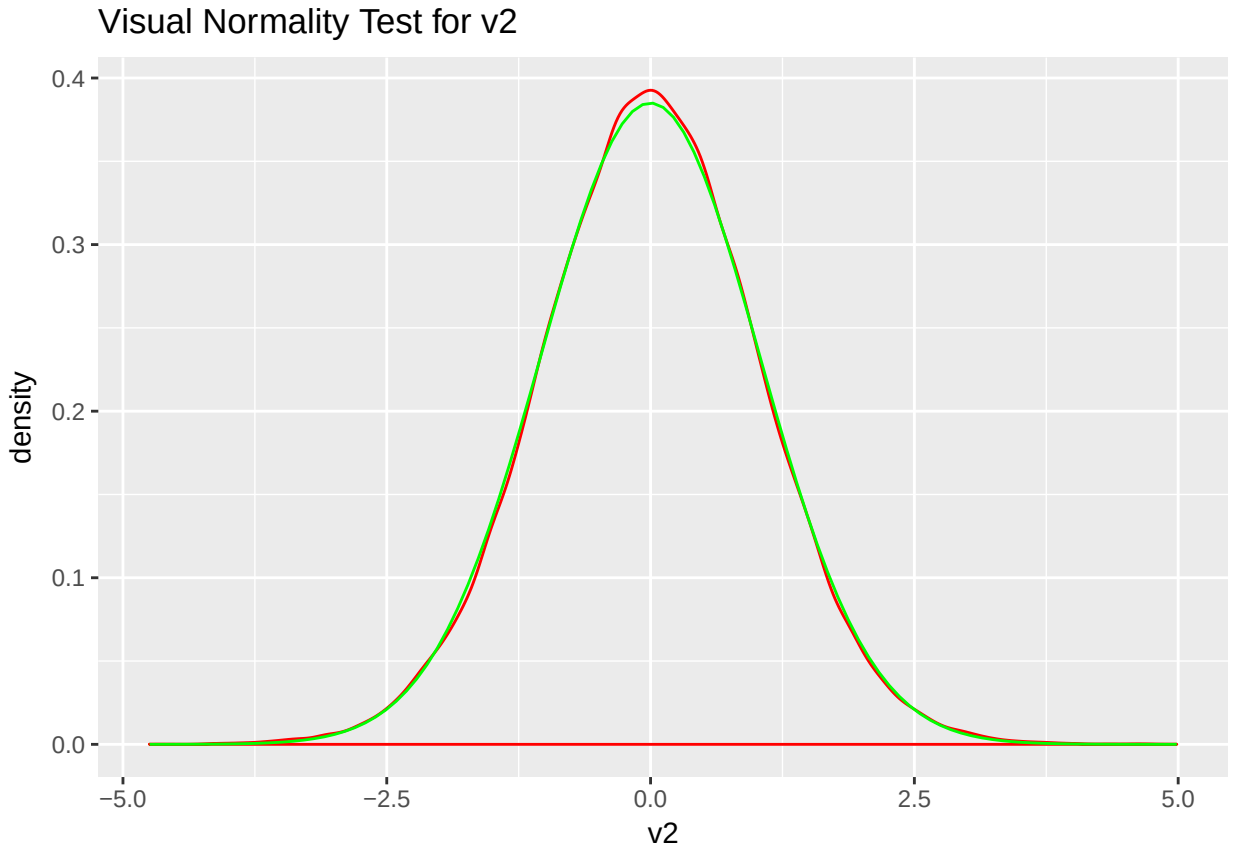
Now lets look at an example where you have a large sample which is as close to normal as it gets for all practical purposes.

```
set.seed(1)
v2 = rt(60000,29)
ad.test(v2)
```

```
##
##  Anderson-Darling normality test
##
## data:  v2
## A = 3.5228, p-value = 8.465e-09
```

This tells you that data is not Normal. T-distribution with degrees of freedom 29 is very close to Normal distribution.

```
df=data.frame(v2)
ggplot(df,aes(x=v2))+geom_density(color="red")+
  stat_function(fun=dnorm,args=list(mean=mean(df$v2),sd=sd(df$v2)),color="green")+
  ggtitle("Visual Normality Test for v2 ")
```



Look at those density curves, how much more close do you think they can be for all practical purposes, yet your normality test tells you that data is not normal.

We'll conclude here. You can play around checking various kind of hunches with your data. Lets me know if you face any issue.

Prepared by : Lalit Sachan

Contact: lalit.sachan@edvancer.in

In case of any doubts/question regarding contents of study material, please post them on Q&A Forum in LMS.