Hypothesis Tests

Hypothesis testing is a widely used data analytics technique to assess effectiveness and impact of business decisions and to provide confidence for future business decisions. It is, in fact, the back bone of many predictive analytics algorithms.

In the Bank telemarketing case, using the data that is available, the Bank can test multiple hypothesis and report results with confidence that can help guide its strategy for improving its efficiency on response.

Let’s start with applications of single sample tests. Remember in a single sample test, we are looking at a sample mean and checking whether that sample mean is statistically different from an expected population mean.

What would be practical applications of a single sample test in the telemarketing example?

Now supposing the Bank wants to do some qualitative research of customers. It wants to understand more intensively behavior thought process of customers. Now it doesn’t want to call every customer. So what it will do is it will choose a smaller sample and run its questionnaire on the smaller sample.

So supposing the Bank takes a random sample of 100 customers from the underlying data. Now before it does any other analysis, the first thing the Bank should do is to make sure that the sample is representative. Remember you can pick representative samples by random sampling. But it is possible that sometimes you may end up with non-representative samples because of random chance variation. So the first thing to do is to check whether or not the sample that has been chosen is representative.

Now let's say that age is very important for this analysis. So we want to make sure that the sample age is representative of the population age. So is the sample average age the same as the population? If not, how much is the difference? And is the difference statistically significant?

So is an observed sample mean statistically different from an expected population mean?

To do this lets simulate choosing of 100 random customers. Now this is a sample of 100 rows using random sampling. I did this in excel by using the random variable. I have generated random numbers between 1 and 45211 and I have selected 100 random rows where the random value is greater than 15000. So this is just like random sampling.

When I do that lets say that these are the 100 rows that I selected. Now if I look at the average age of these 100 customers, I can see that the average is 42.22. In the population, the average is 40.94. This is obviously different from the population average. So the sample average 42 is different from the population average of 40.9. The question that we now have is- should I worry about this difference? Is it very small and therefore it doesn't matter? Or is it statistically different? Which means that I may have to re-sample. Remember to check this difference we will run a hypothesis test.

All hypothesis tests start with setting up of a null hypothesis. So the null hypothesis here will be that the sample age is the same as the population. There is no difference between sample and population. The alternate hypothesis is that there is a difference between the sample age and the population. This sample is unlikely to have come from this population.

Once you set up the null and alternate, we have to decide on a level of significance. How much random chance variation are we willing to tolerate? Let's use 5% which is the most commonly used cutoff. Now in order to actually decide which hypothesis we are accepting, we have to calculate the likelihood of seeing this observed sample age of 42.22 from a population where the average is 40.94 and that likelihood of probability is calculated using a distribution.

What is the test distribution that we will use here? We will use normal distribution because our sample size is greater than 30 and therefore according to the Central Limit theorem, the sample averages will approximate a normal distribution. So the test distribution we use will be normal. Once we know that we are using a test distribution which is a normal distribution, we calculate the test statistic which is (X-μ)/(σ/sq root of n).

Remember we have to divide the standard deviation by the square root of sample size because we are calculating the probability from a distribution of sample averages, not the distribution of the population. Now what we are actually calculating is the likelihood of seeing a sample average of 42.2 or greater from an underlying population with an average of 40.94 simply because of random chance variation. So now if you use this formula, we will say 1-normal distribution, 42.22 which is the outcome X, the next is μ which is the population average 40.94, standard deviation and we have gotten the standard deviation of age in the descriptive statistics if you remember this 10.62 divided by the square root of sample size, sample size is 100, so the square root of sample size is 10 and true.

Remember we are calculating what is the likelihood, but we will see a sample average of 42.2 or greater from a population with an average of 40.94 simply by random chance. We cannot calculate point probabilities here because this is a continuous distribution. We can only calculate cumulative probabilities. We are calculating the likelihood of greater than

42.2 which is 1- probability of less than equal to 42.2 which is 11.3%. So the likelihood that you will end up seeing a sample with an average of 42.2 from an underlying population which has an average of 40.9 simply because of random chance variation is 11.3%.

Therefore, because this 11.3% is higher than our cutoff of 5%, we cannot reject the null that the sample age is the same as the population. In other words we will conclude that this difference is within the accepted random chance variation that we are willing to work with. If the random chance variation was less than 5%, then we would be confident that this sample is very different from the population.

Let's look at another example of hypothesis testing but now with a small sample.

Now supposing we want to run a focus group process in another city with customers. Again, we will end up choosing a smaller sample because there is a cost associated with getting the customers to travel to another city and spending a whole day for example for the focus group. So let's say that the budget allows for 25 customers. Now in this random sample of 25 customers, the average age was 39.5 years with a standard deviation of 8.2. Now is the difference in age statistically significant?

Here it's the same hypothesis test except we are going to use a small sample test because the sample size is less than 30. What that means is the test distribution will now be a T distribution instead of a normal distribution. So the null hypothesis remains the same, the alternate hypothesis remains the same , that the sample average is less than the population average, different from the population average, the level of significance 5%, the test distribution is a T distribution , the test statistic is calculated as (X-µ)/ (σ/ sq root of n)- 5.41. So now the associated p value, which we can calculate in excel easily.

Remember this is the test statistic and this is the P value. The degrees of freedom are 25-1 so 24, the p value is 7E-06. Essentially this is 0. Six zeros followed by 7(0.0000007). Therefore, it's a very, very low p value. We will reject the null that there is no difference between the sample and the population. In fact, we are concluding that it is highly unlikely that you will get an average age of 36.5 years from a population which actually the average is 40 years in a random sample of 25. So this sample is definitely not representative.

Now let's look at a third example of hypothesis testing which is more frequently used which is Two Sample Test.

Supposing in this particular example, the company wants to check effectiveness of agents by checking the average time spent on the phone with a customer that results in a yes. In other words, of the customers that say yes, many agents call them. Now can some agents convert customers faster? So let's say they choose two agents and record their calls into 17 calls each that resulted in a conversion. For agent A, the average call duration is 1125 seconds. For agent B the average call duration is 1030 seconds. Can it be concluded that agent B is more efficient or is this variation simply because of random chance?

So again we can replicate this by taking 17 random calls. So I have taken 17 calls for agent A and 17 calls from agent B. You can see the averages are 1126 and 1031. We want to figure out is this difference statistically significantly different or not. Now this is a two sample test.

We can run two sample tests in excel using the data analysis tab. We will choose T test, two sample assuming unequal variance. If we run that, we can see the outcome here. What we are interested in is the p value. And if you look at the p value for a two tail test it is 0.56. In other words, we cannot reject the null hypothesis that there is no difference between the agents. So even though we see this difference, this difference is not large enough for it to be statistically significant. We cannot be confident that really agent B is more efficient than agent A. Because there is a large chance that this difference is driven by randomness.

So that's how hypothesis testing can be applied to test and validate multiple hypothesis.

We can run two sample test in excel using the data analysis tab. Go to data, data analysis and choose two sample assuming equal variances or assuming unequal variances. At this point you can choose either. Let's say we choose assuming equal variance. Just click on ok. Excel will ask you what is the range of variable 1. Now the range of variable 1 is the duration for variable 1. So I am going to say this is variable 1. Remember we don't want to include the average. The range of variable 2 which is the duration for variable 2.

The hypot size mean difference. Remember we are checking for whether there is a difference between A and B. So a null hypothesis is that there is no difference. So the hypothesis that we are testing is that A=B, therefore A-B=0. I don't have labels, I am going to leave it blank and just say ok.

So this is the output that we get. It shows us for agent A and agent B duration what is the average – we saw that 1126, 1030, the variance, the number of observations, etc. and then some other statistics. But ultimately what we are interested in is p value. Remember we are checking for whether or not there is a difference between A and B. In fact, is A greater than B. In that case, the one tail test is essentially the p value of the one tail test is

0.28. And because 0.28 is much higher than say a significance level of 5%, we fail to reject the null hypothesis.

In other words what we are concluding here is that even though it looks like agent A duration average is much higher than agent B's duration average, statistically they are not significantly different. That is we cannot be 95% confident that there is really a difference between A and B. This is an example of a two sample test.