In the first part of the statistics module, we have looked at different kinds of statistics, descriptive and inferential. We have introduced the concept of random variables and probability distributions of random variables and we looked at how to calculate the probability of an outcome of a random variable using a probability density function or a probability distribution function.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

What we do now is to introduce the concept of hypothesis testing. Hypothesis testing is a basic building block of many statistical analysis and analytical algorithms.

- So we will talk about what is a hypothesis test and when to use it.
- What is the basic framework of a hypothesis test.
- How to actually implement the hypothesis test.
- We will talk about a very important concept with applications to hypothesis testing which is the Central Limit Theorem.
- We will review multiple kinds of hypothesis test.

Let’s understand first what exactly hypothesis test. In order to do that let’s go back to our airline no-show example.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Remember the no of no-shows for an airline on any given day for any particular flight is a random variable. Because the airline cannot be sure till the gate is closed how many people are not going to make it. Now it is a random variable but we also have historical data on the outcomes of the random variable. That allow us to have some information about this "expected behaviour".

So supposing in our airline example, we looked at historical data and it turns out that in the last 6 months the average number of no-shows was 5%. Now if you go and report this to your General Manager of Sales, the general manager says" you know, I want to check this for myself". So he wants to check on the basis of 10 random flights.

Every day you choose one random flight and you look at how many no-shows you have. Remember for simplicity sake, we are assuming that there are a 100 seats on your airline.

- Day 1, 3 people don't show up.
- Day 2, 3 people don't show up.
- Day 3, 4 people don't show up.
- Day 4, 7 people don't show up and so on.

If you look at an average for these 10 days you get an average of 3.7 and the standard deviation of 1.7. This 3.7 obviously is not equal to the 5% which is what was generated from looking at 6 months of data.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

If you were ask to talk about the average no. of no-shows for your airline. Which one will you pick? Will you pick 5% or will you pick 3.7%? Remember both of these numbers are coming from data. Which number is now right? Is it 5% or 3.7%. Some of us may say I am going to still go with 5% because it is based on 6months of data. Some of us may say I am going to go with 3.7% because it is from a more recent set of data points.

Remember no-shows is a random variable with an expected value of 5 per flight. However because this is a random variable we know that we will see variation simply because of randomness. Some days you will see fewer people than expected not showing up for a flight and some days you will see more people than expected did not showing up for a flight. Therefore when you see an average of 3.7% in a sample, there are two possible explanations for this 3.7%.

1. The reason your sample is showing you a number that is different from what you are expecting, which is the population average is because your sample is capturing behaviour that is different from the population. In this context it is possible that the number of no-shows have reduced over time and therefore your latest sample is capturing that behaviour.
2. But it turns out that there is another possible explanation for why you might see 3.7% and that could be not because that there is any change in behaviour. It is simply random chance variation.

If you pick multiple samples from an underlying population, do we expect that every sample to have the same average which is population average. No, you will see variation in the sample averages simply because of random chance variation. So it is possible that the number of no-shows are still 5%, they have not reduced. What you are seeing is because of random chance variation.

As a business decision, we have to choose between explanation 1 and explanation2. Because one explanation says there is a change in behaviour and the other explanation says there is no change in behaviour.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

How do we do this? Many of you may already have thought about this that you could potentially look at more samples or you could look at larger samples. Both of these are valid options. Instead of looking at one sample, why don't we look at 3 or 4 different samples? If all that samples show you a number that is less than 5% like 3.5%, 3.8% or so on then we can be pretty sure that there is a change in behaviour or even if you don't want to do multiple samples supposing that instead of looking at a sample size of 10, we look at 50 flights or a 100 flights and then look at an average.

Again with the larger sample, if the average is lower than 5% then we can be fairly sure that there is a change in behaviour. So certainly when you look at a sample outcome and you find that it is not equal to an expected population outcome you could look at increasing sample size or using multiple samples. But it turns out that in some business situations this may not be valid options.

For example, supposing we are running a drug test may be to cure leukaemia in small children under the age of 5. I may want ideally to have 100 small children tested. But within a given time frame and in a certain location which is attached to a hospital, it may be very hard for me to be able to actually find 100 children that I want to test. You simply may not have that many children under the age of 5 who suffer from leukaemia that you can enrol in the test.

You may be able to source 15 children or 25 children or 10 children. You still have to test the effectiveness of the drug. So you may not have the option of using the other sample or increasing sample size because of budget constraints, infrastructure constraints, and time constraints and so on.

How do you decide? It turns out that you could look at a third outcome which is calculate the random chance probability of the observed outcome. We are taking multiple samples from an underlying population. If you see a sample mean that is different from the population mean it could be because this sample mean is here simply because of random chance variation. All your sample means will not be equal to the population mean. There will be a distribution of sample means and therefore this sample is still coming from this population.

The fact that this average is less than population average is simply because of random chance variation. The shaded area is the probability of seeing a sample outcome less than or equal to this outcome simply by random chance variation.

How do we actually calculate this probability? That is what we were doing when we were looking at probability distributions.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

When we calculate this probability and let's just say that we calculated this for this particular example of seeing 3.7% or less and I find it is 40%.

What does that imply? In this particular context that essentially says that there is a 40% chance that when I pick a random sample from a population with a mean of 5%. I will see a sample mean of 3.7% or lower. So this shaded area is 40%. Is this likely 40% chance that what you are seeing is simply random chance variation?

What would you conclude? If there is a 40% chance that when I pick random sample from a population with the mean of 5% I will see a sample average of 3.7% or lower. I am unlikely to conclude that there is really a difference between the sample and the population. In other words there is a very high probability that the difference I am seeing is driven by random chance variation. Therefore the average number of no-shows are still 5%.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

What if we had calculated this probability and had found that this probability was only 15% then would our conclusion change? Probably, we would say there is only a 15% chance that what we are seeing is driven by random chance variation. Therefore it is pretty likely that this is not driven by random chance variation and it is because the average of the sample is capturing the fact that the behaviour has changed. The number of no-shows have actually reduced. They are not 5% anymore.

Remember what we are doing is taking a sample outcome and we are comparing the sample outcome to an expected population outcome. If the sample outcome is different from an expected population outcome we are essentially saying, look, it can be because of two reasons. Either this is random chance variation or there is a difference between the sample and the population. The way we decide which of these explanations that we are going to go with is by calculating the likelihood of seeing that particular outcome, the sample outcome simply because of random chance variation.

If the probability of random chance variation is very high, we are going to say the sample is not really different from the population. That difference that we are observing in the sample is simply because of randomness. If on the other hand, the probability of random chance variation is very low, we will say it is much more likely that this variation is being driven by something and it is not being driven simply by randomness.

In other words, it is very likely that the sample is very different from the population. That is the a test.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Remember a high likelihood of random chance variation essentially means we are not going to conclude that the sample is different from the population. In the airline example if we had seen a high likelihood of the probability, I would say the average is still 5%.

On the other hand, if we had calculated the probability and seeing that it is a very low number, we will conclude that the average is different from 5%. It is not 5% anymore.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Let's just recap using another example. Let's just say that we have a manufacturing process and the manufacturing process produces auto parts with an expected weight of 2.5 lbs and the standard deviation of 0.12 lbs. When you look at historical data on the weights of the product produced you find that it is normally distributed with the mean of 2.5 and standard deviation of 0.12 lbs.

Let's say that I take a random sample of 45 units and I find now that in that random sample of 45 units the average weight is 2.68 lbs. Is there a problem with my production process? There could be two reasons for why you see an average weight of 2.68 lbs in a random sample.

1. Your production process is not working very well. It is generating more weight than expected.
2. Your production process is fine. What you are seeing is simply random chance variation. Because you know when you take multiple samples. Every sample is not going to have an average of 2.5 lbs even if the production process is perfect.

Therefore how do I decide whether or not I have a problem? I can calculate how likely is it that I will end up with an average of 2.68 or greater from a production process that is expected to generate products with the mean of 2.5 lbs.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

In other words the probability of seeing a sample mean of 2.68 or greater when I am expecting a mean of 2.5. The data is normally distributed so we can actually calculate this probability. One minus the normal distribution of 2.68 which is the outcome, 2.5 which is the mean, 0.12 which is the standard deviation and true. So if you do that you will get a probability of seeing 2.68 or higher to be 0.07.
Now there is a 7% chance that what we are seeing is because of random chance variation.

What would you conclude? Remember there is a 7% chance at what we are seeing in because of random chance variation. Therefore it is pretty unlikely that what we are seeing is driven by randomness. It is much more likely that what we are seeing is because the production process is not in control. It is generating parts with the way higher than 2.5. Because when I calculated the random chance probability of seeing an observed sample outcome, I found that to be fairly low, only 7%.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Remember we are trying to choose between two possible explanations for why we see 2.68. One says it is driven by randomness and the other says it is not driven by randomness. There must be some difference. So if the random chance probability is very high, we will not conclude that there is a difference between the sample and the population.

If on the other hand, that random chance probability that we calculate is very low, we will conclude that the sample is different from the population. But what is high and what is low. What is that cut-off value?

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

It turns out that the boundary between high and low tends to be very subjective. It depends on our risk appetite. Some of us may decide to go ahead with the decision if there is a 70% chance of success. While some of us may say 70% chance is not high enough. For me to take this decision, I want to be assure that there is a 90% chance of success.

Similarly for low probability I may say I am not comfortable taking a decision when there is a 20% chance of this being driven by randomness while another person says 20% is fairly low. 80% chance that this is not random, 20% chance that this is random. So it turns out that this is a very subjective decision. It depends on individual risk appetite. It also to some extent depends on what is at stake, what is the outcome. If the outcome at stake is very, very critical, very, very expensive, you may want to have high chances of success. You want to make sure that what you are seeing is not random chance driven. Whereas on the other hand if what is at stake is not really that expensive, it is a fairly low grade decision. You may be ok having higher levels of uncertainty around whether or not this driven by random chance differences.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

To avoid subjectivity, both in academic studies and business applications, most often a cut-off of 5% for low probability is commonly used. This is simply an accepted cut-off that everyone agrees with. However there is no magic to the 5%.

What does the 5% imply? It says only if the random chance probability of seeing sample means as extreme or more extreme than the observed population mean is less than 5%, will you conclude that the sample is really different from the population.

In other words, if this low probability is less than 5%, we will conclude that the sample is different from the population. If this probability is more than 5% we will reach this conclusion cannot conclude that the difference between the sample and the population. We are saying, we want to make sure that the probability of the difference being because of random chance variation is 5% or less.

In other words, if you look at this figures, we want to make sure that the probability of an observed sample outcome is less than 5%. Only then we will conclude that there is a difference between the sample and the population. This is a very, very strict criteria. We are essentially saying we want to be very, very sure that there is a difference before we conclude that.

How sure do we want to be? We want to be 95% sure. If there is a 5% chance that this is because of randomness. There is a 95% probability that it is not because of randomness. Most often we want to be 95% sure before reporting a difference. It's a very, very strong criteria, very strict criteria.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

In the quality control example, remember we calculated the random chance probability to be 7%. What will you conclude? Remember the sample outcome probability is 7%. We will only conclude that there is a difference between the sample and the population if the probability is less than 5%.

Therefore in the quality control example, if we wanted to be 95% sure which is commonly accepted. We cannot conclude that there is a difference between the sample and the population. In other words we cannot conclude that there is a problem with the production process. Because there is a 7% chance that this could be driven by randomness and we will only conclude that there is the difference if the probability of the randomness is less than 5%.

That was a hypothesis test. We look at a framework of a hypothesis test in much more detail in the next module.

>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>