LECTURE 1 – **Hypothesis Testing**

Let’s start with *sampling.*

We want to estimate some parameter θ in a population and we need a method to estimate it. A sample is a subset of the population, and as the sample size increase, the sample mean converges to the population mean (Law of Large Numbers).

Diagram

Description automatically generatedFor example:

As we can see, the higher the sample size, the closer to the population parameter (which will always remain unknown).

Additionally, If two equally sized samples are taken from the same population, their sample mean might be different (sampling error)

When it comes to choosing the sample size it is very important to consider how rare the event is. In the example below, the probability of finding a person that is both blind and schizophrenic is 1 on 500,000. Therefore, a sample size of 100,000 would not be appropriate (more like 20M).

What is the *Central Limit Theorem*?

If sampling randomly and independently, the sample means distribute normally as the sample size increases, regardless of how the underlying population is distributed.

To illustrate how it works, we will draw independently and randomly (random sampling) from the same population, t times for each sample size n. We will then take the sample mean for each sample and plot the distribution.

After sampling t times and plotting each sample mean, we generated a normal distribution!

The mean of the normal distribution plot will be an approximation of the real values.

The standard deviation of the sample means is called SEM (standard error of the mean)

Diagram

Description automatically generatedAs t increases we learned that the mean of the sample means approaches its true value, but every t represents another sample of size n, which is expensive. Normally, Data Scientists only have one sample, the dataset.

Therefore, you want a reasonable chance that the sample mean of the one sample you have is good (enough) estimate of the population mean. For this to happen, we need to consider how scattered these sample means are. This is captured by the standard deviation of the sample means (SEM).

Shape, arrow

Description automatically generatedA picture containing diagram

Description automatically generatedText

Description automatically generated with low confidenceText

Description automatically generatedHow large does our sample size n need to be, in order for the sample mean to be a close estimate of the population mean?

We take the square root because we are interested in the standard deviation, not the variance.

What does *statistical significance* mean?

It means that an observed result is unlikely to be due to chance alone. Similarly, the probability of the data, assuming that the null hypothesis is true, is less than the chosen significance data.

Steps to perform a *Null Hypothesis Significance Testing*

Text

Description automatically generatedWe first assume that our hypothesis is false (that whatever we do to the treatment group will not differentiate them from the control group) but we do expect that there will be a difference between the treatment and control group which will be 100% sampling error.

Now, if the differences between both groups are too large for it to be due to sampling error, then we reject our own assumption (null hypothesis).