**Distributions of Sampling Statistics**

Table of Contents

[Approximate Distribution of the Sample Mean 3](#_Toc63699731)

[Sample Variance 5](#_Toc63699732)

[Joint Distribution of and 7](#_Toc63699733)

We have yet to define what a statistic is. Any function that has been created based on data is a statistic. For example, we saw that the sample mean can be found by the following formula

Thus, is a statistic.

Our goal has been to try to find how such statistics are distributed. We use random variables to do this, but we also need to keep in mind that not all random variables are observable. Some are observable, while others are hypothetically observable. For example, we cannot actually observe the exact heights of everyone in a country. By the time we finish measuring, many of the people we measured will be dead while many more will have been born. Even if we invest an infinite amount of resources, we cannot get an exact result.

Instead, we want to use observable random variables to predict the values of hypothetically observable random variables. We collect random samples of the data and use that to make estimates about the values of different parameters. We do this using the CLT. To use the CLT, we need a large amount of observed data. Otherwise, the approximation is not applicable.

When there is a single parameter, we sometimes use to represent it. For multiple parameters, we might use .

## Approximate Distribution of the Sample Mean

According to the CLT, for a very large number of random variables,

This means that no matter what the individual distributions of the random variables of each of the pieces of data we collect are, the sum will be approximately normally distributed.

Since , the sample mean should also be normally distributed. Thus,

However, we cannot use a table to solve this. So, we have to normalize it.

Example

The delay of a population of data packets has a mean value of and standard deviation of . For a sample of size , find the approximate probability that the sample mean is between and .

Details of how values for standard normal distributions work can be found towards the end of Lecture 3.

The second part of this question tells us to increase the sample size to . Then, we will find that the resulting probability becomes approximate . The probability tells us that what the chances are that the sample we took has a mean value within the range of and . Thus, the larger the sample we take, the better the results.

Now the question becomes, how large does have to be for our approximation that to be correct? The value is not specified by the CLT. In reality, it is found that for most cases, if , the approximation is applicable. Even values as low as or are found to be sufficient in reality.

## Sample Variance

The distribution of the sample variance is rather difficult, so we will not be going into details of the derivation.

The sample variance is given by

where is the standard deviation.

It can be mathematically proven that

The actual proof is available in the book.

Thus,

Since we are working with IID random variables, this can be written as

x

We know that

Putting this in the previous equation,

Just like the expected value of the sample mean was equal to the population mean, the expected value of the sample variance is also equal to the population variance.

## Joint Distribution of and

The derivation for this equation is given in the book for anyone interested.

All the random variables we use here are IIDs and random samples taken from a Normal population, meaning the population has a normal distribution with parameters and . Thus, we can replace the left-hand side equation:

One of the equations we learn for was

Thus, we can replace the second term in the right-hand side equation:

We can further replace the two terms we have as chi-square random variables with and degrees of freedom respectively. Thus,

One last thing that we did not know before is that if we add two chi-square random variables with and degrees of freedom respectively, the result is a chi-square random variable with degrees of freedom.

Using this in our equation, we get

To make the last change, we essentially said

We can re-write this a little like this:

The middle term here is just the sample variance, so we can say

One thing we need to remember in all of this is that the chi-square random variables must be independent.

In conclusion, if are a random sample from a Normal population with parameters and , and are independent.