Statistics for Data Science – Part III: Statistical Inference

This article is a third of a series and I will cover the parts of probability that are related to data science. Statistical inference is defined as the second major branch of statistics and very important for the data scientist. The target will be making more meaningful estimates by specifying an interval of values on a number line, together with a statement of how confident you are that your interval contains the population parameter.

I will try to give information about the following subjects:

- Central Limit Theorem
- Confidence Intervals,

You may find the first article of this series here and the second part is here.

Central Limit Theorem

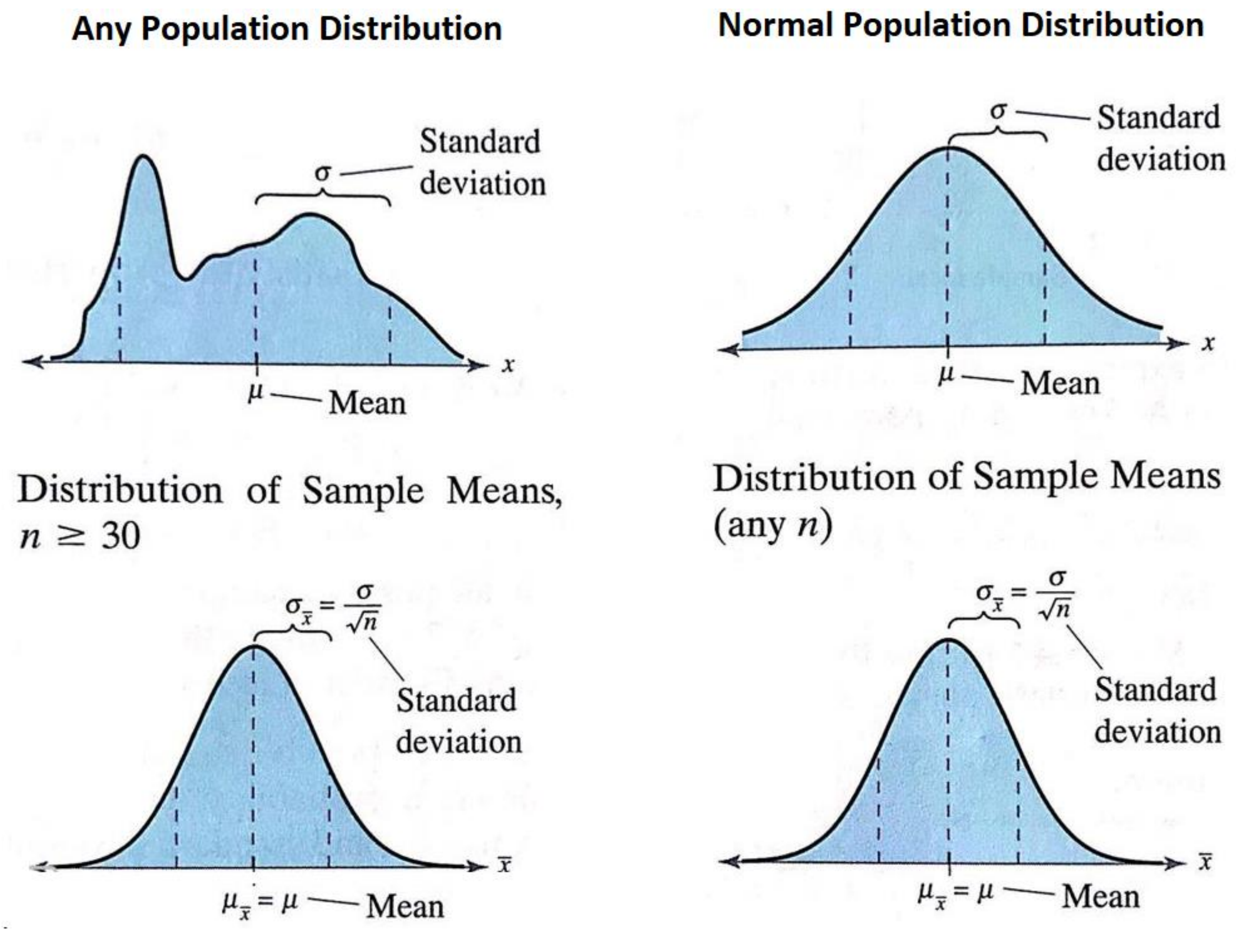
The **Central Limit Theorem (CLT)** describes the relationship between the sampling distribution of sample means and the population that those samples were taken from. Basically:

1. If samples of size n, where $n \geq 30$, are drawn from any population with a mean μ and a standard deviation σ, then the sampling distribution of sample means approximates a normal distribution. Increasing the sample size will give a better approximation.
2. If the population is normally distributed, the sampling distribution of sample means is normally distributed for *any* sample size *n*.

In either case, the sampling distribution of sample means has a mean equal to the population mean:

$$\mu_{\bar{x}} = \mu$$

The standard deviation of the sampling distribution of the sample means, $\sigma_{\bar{x}}$ is also called the **standard deviation of the mean**.


Any Population Distribution

Normal Population Distribution

Distribution of Sample Means, $n \geq 30$

Distribution of Sample Means (any n)

Confidence Intervals for the Mean (Large Samples)

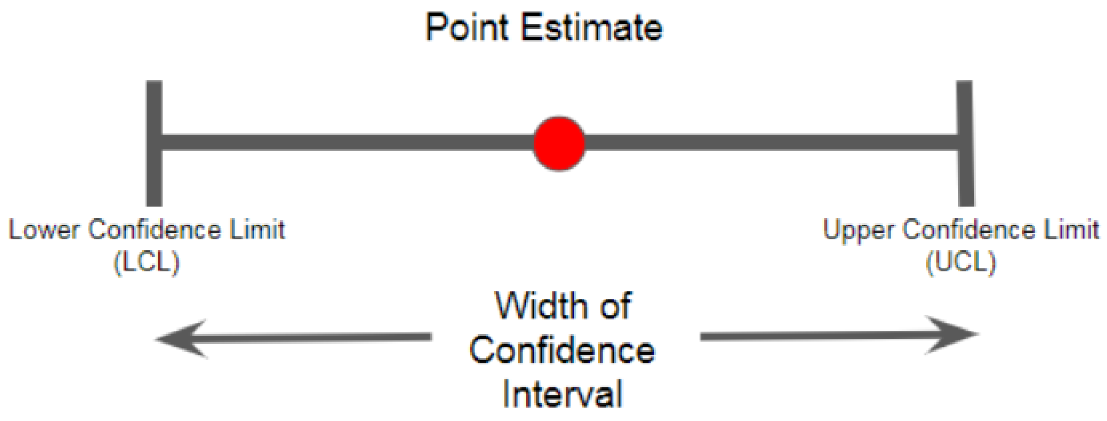
An important technique of statistical inference is using sample statistics to estimate the value of an unknown population parameter. When the sample size is at least 30 or when the population is normally distributed and the standard deviation is known, an estimate of the population parameter μ can be made. The first step should be finding a point estimate.

A **point estimate** is a single value estimate for a population parameter. The most unbiased point estimate of the population mean μ is the sample mean $\bar{x}$.

The validity of an estimation method increases if a sample statistic is unbiased and has low variability. To be unbiased, the statistic should not underestimate or overestimate the population parameter.

An **interval estimate** is an interval, or a range of values, used to estimate a population parameter.

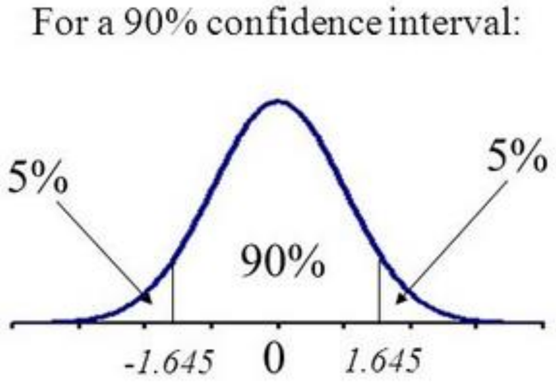
The point estimate may not be equal to the actual population mean, but it will probably be close to it. To form an interval estimate, most often, the point estimate is used as the center of the interval and a margin of error is added and subtracted.


Point Estimate
Lower Confidence Limit (LCL)
Upper Confidence Limit (UCL)
Width of Confidence Interval

Before finding a margin of error for an interval estimate, the level of confidence that the interval estimate contains the population mean (μ) should be determined

The **level of confidence *c*** is the probability that the interval estimate contains the population parameter.

As mentioned before, according to the CLT, when $n \geq 30$, the sampling distribution of sample means is (tends toward) a normal distribution. The level of confidence *c* is the area under the standard normal curve between the **critical values**, $-z_c$ and z_c. The graph shows that *c* (in this case 0.90) is the precent of the area under the normal curve between $-z_c$ and z_c. The remaining area is 1 – c, so the area in each side is ½(1 – c). In this example, for a *c* value of 90 %, $-z_c = -1.645$ and $z_c = 1.645$.



The difference between the point estimate and the actual parameter value is called the **sampling error**. When μ is estimated, the sampling error is the difference $\bar{x} - \mu$. In most cases, µ is unknown and $\bar{x}$ varies from sample to sample. However, if the level of confidence and the sampling distribution is known, a maximum value for the error can be calculated.

Given a level of confidence ***c***, the **margin of error** (maximum error of estimate) ***E*** is the greatest possible distance between the point estimate and the value of the parameter it is estimating.

Interpretation of Confidence Intervals

After constructing a confidence interval, it is important to interpret the results correctly. Considering a case, where the standard deviation is known and a confidence interval of 90 % is constructed, it will be ***correct*** to say:

> If a large number of samples is collected and a confidence interval is formed for each sample, approximately 90 % of these intervals will contain μ.

The ***incorrect*** way will be:

> There is a 90 % probability that the actual mean is in the interval.

Confidence Intervals for the Mean (Small Samples)

In many real-life situations, the population standard deviation is unknown. Moreover, because of various constraints such as time and cost, it is often not practical to collect samples of size 30 or more. In those cases, if the random variable is normally (or approximately normally) distributed, t-distribution (or Student's t-distribution) may be used.

You can reach detailed information about the t-distribution here.

The Chi-Square Distribution

In manufacturing, it is very important to control the variation of the process. An automobile part manufacturer produces thousands of parts to be used in the manufacturing process. It is crucial that the parts vary little or not at all.

Measuring and consequently controlling the amount of variation in the produced parts starts with a point estimate.

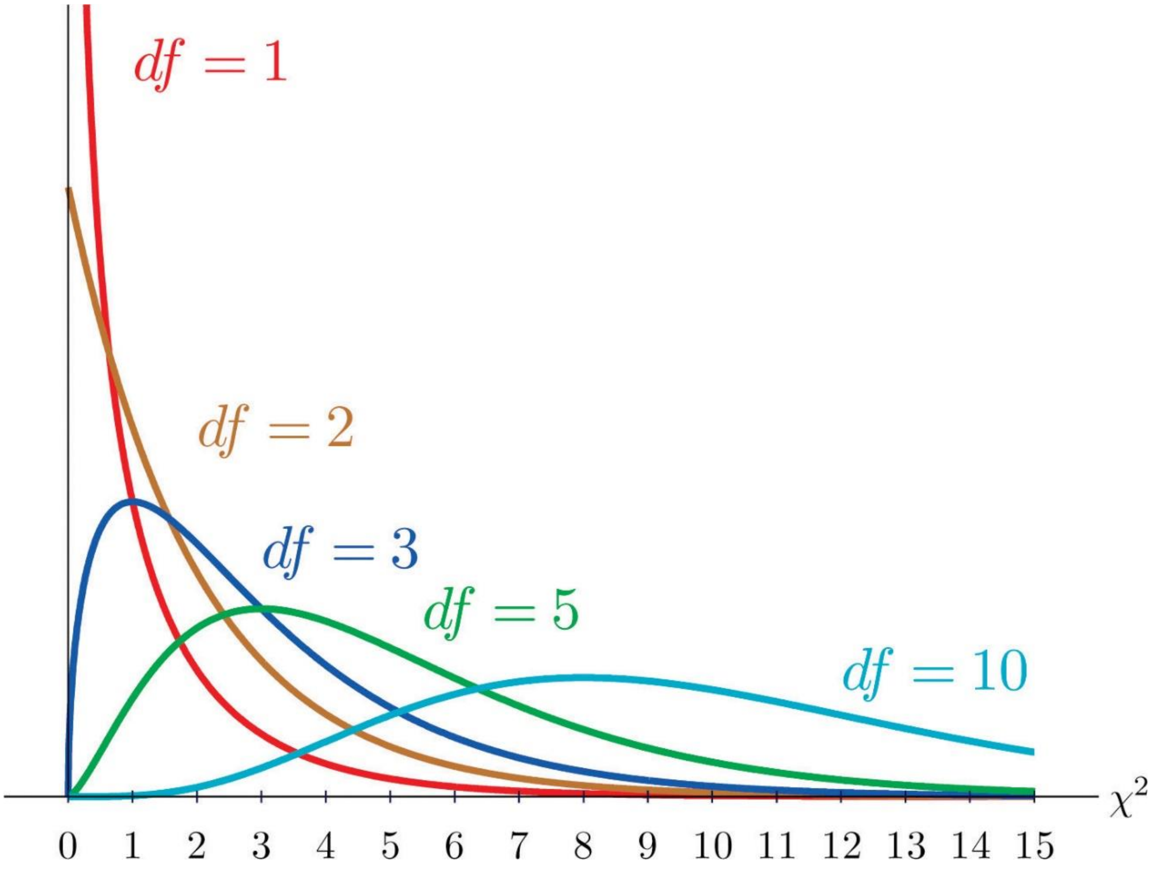
The **point estimate for σ^2** is s^2 and the **point estimate for σ** is s.

It is possible to use chi-square distribution to construct a confidence interval for the variance and standard deviation.

The chi-square distribution (also called the chi-squared distribution) is a special case of the gamma distribution; A chi square distribution with n degrees of freedom is equal to a gamma distribution with

$$a = n / 2 \quad \text{and} \quad b = 0.5 \text{ (or } \beta = 2\text{).}$$

When you have a random sample taken from a normal distribution, the chi square distribution will be the distribution of the sum of these random samples squared. The **degrees of freedom (df)** are equal to the number of samples being summed. For example, if you have taken 10 samples from the normal distribution, then **df** = 10.



If the random variable *x* has a normal distribution, then the distribution

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

forms a **chi-square distribution** for samples of any size n > 1.

List of Abbreviations

σ	Population Standard Deviation
μ	Population Mean
n	Sample Size
s	Sample Standard Deviation
$\bar{x}$	Sample Mean
c	Level of Confidence

You can access to this article and similar ones here.

Photo by Chris Liverani on Unsplash