Netshock
Burn Bright

CENTRAL LIMIT THEOREM

WHAT IS CENTRAL LIMIT THEOREM?

If a population distribution is a non-normal distribution; we can normalise it by taking the mean of many sample populations from the original dataset. When we plot the means of the samples (and the number of samples with that mean), it will result in an approximately normally distributed dataset.

Wikipedia defines it as: “when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.”

Onlinemathlearning.com defines it as: “If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means are approximately normally distributed for sufficiently large sample sizes (n ≥ 30) regardless of the shape of the population distribution. If the population is normally distributed, the sample means are normally distributed for any size sample.”

Netshock
Burn Bright

WHY IS IT USEFUL?

Before we get started, some points to remember.

With CLT, the following applies

1. The mean of the approximately normal distribution will be the same as the population mean.
2. The approximately normal distribution variance is (population variance / sample size). The larger the sample, the lower the variance & hence the closer the approximation will be. We require at least 30 observations for central limit theorem to be effective.

The central limit theorem allows us to perform tests against a normal distribution. This is important because the normal distribution has the most applicability to many statistical models (such as confidence intervals).

In short, central limit theorem allows us to make inferences from a normally distributed dataset even if the original dataset is not normal.

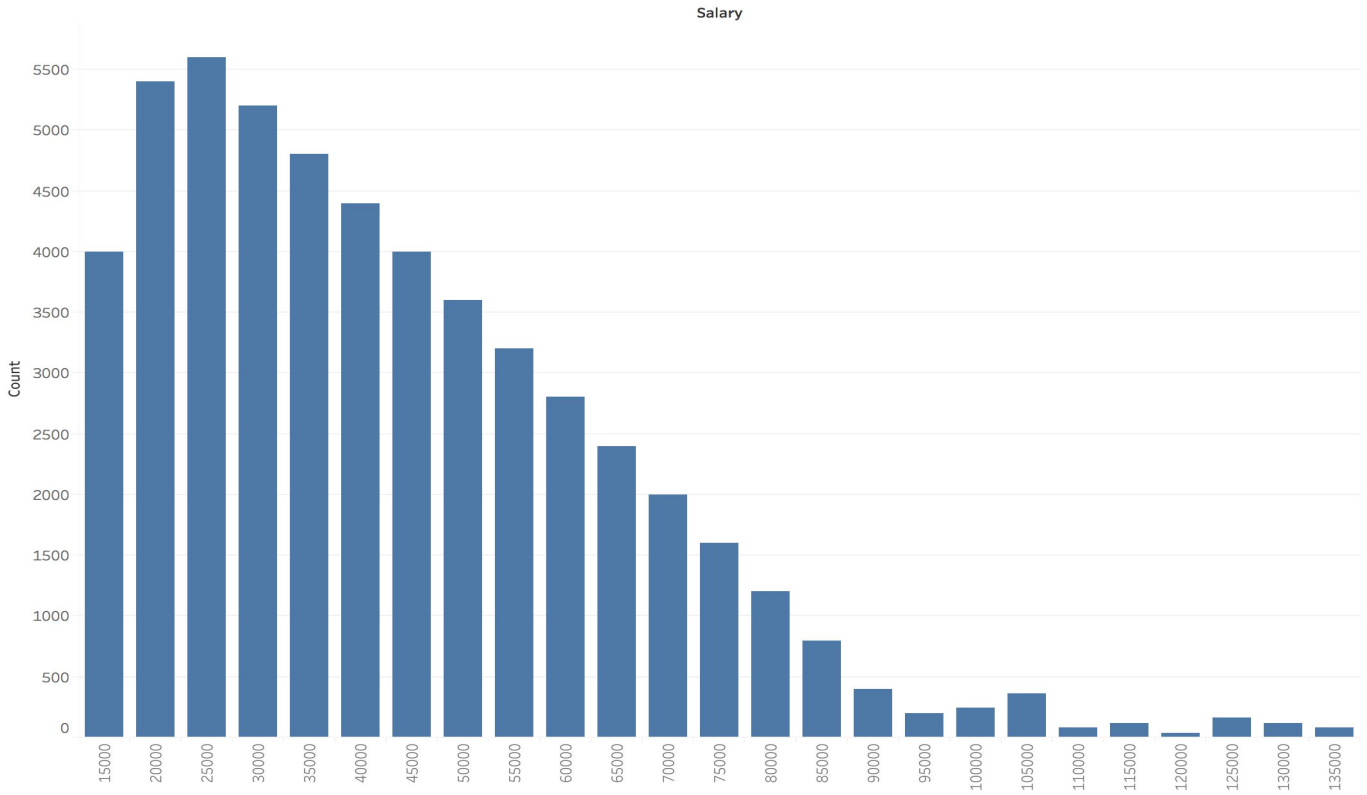
WHAT DOES THAT MEAN?

Let's say, we have income values. The original dataset looks like this:

This dataset has 52,800 observations.

If we split this into 1,760 samples and have 30 observations per sample and take the mean salary of each, we should see a much more normal distribution.

The mean salary of the sample will be plotted on the X axis. The number of samples where that was the mean, will be the frequency on the Y axis.



Netshock
Burn Bright