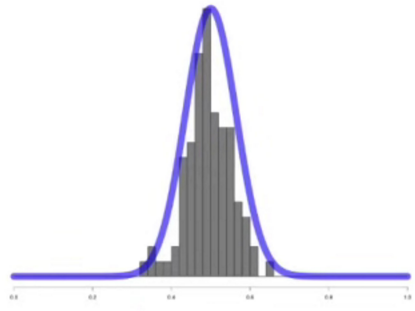
Day 29

Central limit theorem (CLT)

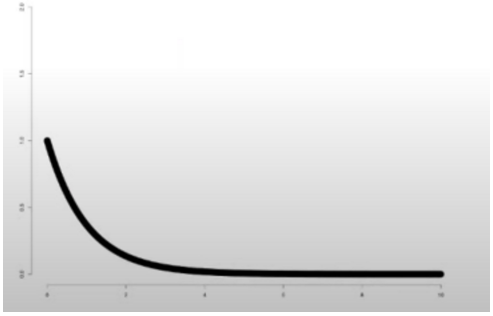
Definition: The theorem states that, regardless of the original distribution of the variables, the distribution of their sum (or average) approaches a normal distribution as the number of variables grows.

In other words:
For any distribution (uniform distribution, exponenetial distribution, etc.) the means are normally distributed
E.g.
Uniform distribution:

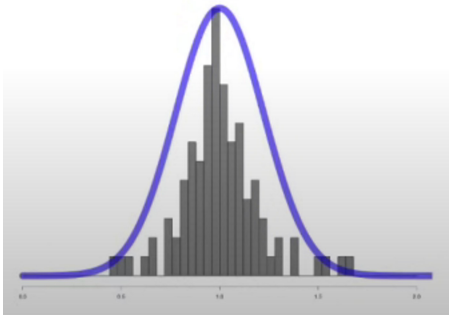
The mean of the above graph is normally distributed (and not uniformly distributed):



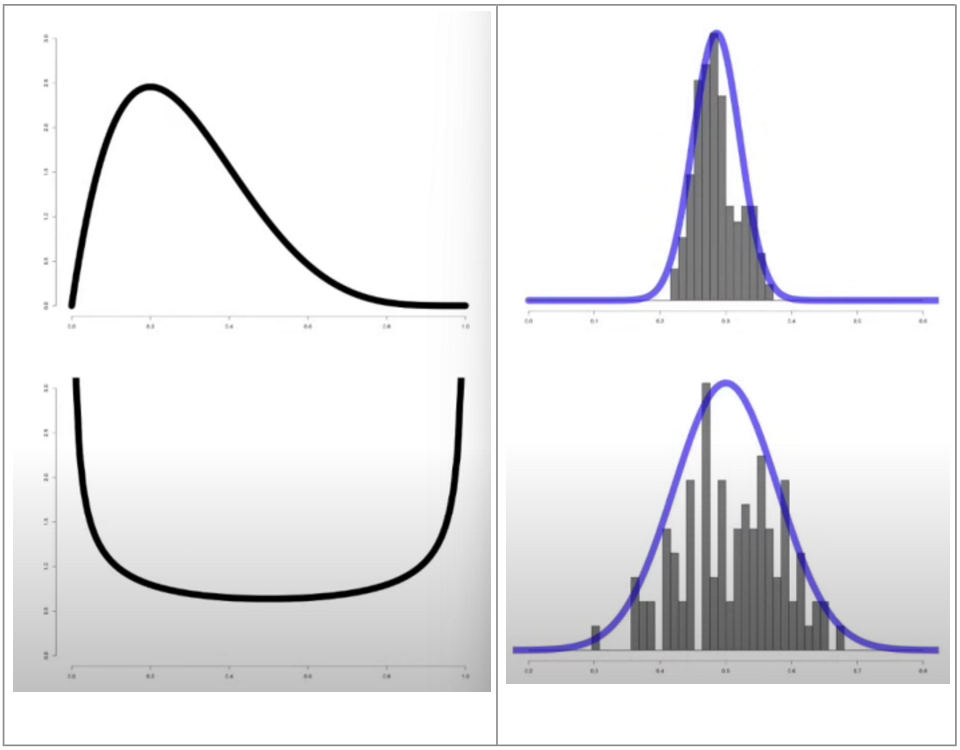
Exponential graph:



The mean of the above graph is normally distributed (and not exponentially distributed):



Similarly,



Statistical methods that rely on central limit theorem principle:

1. Confidence Intervals for the Mean
Confidence intervals provide a range of values within which the true population mean is likely to fall. The formula for constructing confidence intervals for the mean assumes that the distribution of the sample mean follows a normal distribution due to the CLT.

2. Hypothesis Testing for the Mean
Common hypothesis tests for population means, such as t-tests and z-tests, rely on the assumption that the sampling distribution of the test

statistic is approximately normal. This assumption is justified by the CLT when sample sizes are large.

- **One-sample t-test**: Tests if the mean of a single sample is significantly different from a known or hypothesized population mean.
- **Two-sample t-test**: Compares the means of two independent samples to determine if they come from populations with the same mean.
- **Paired t-test**: Compares means from the same group at different times (repeated measures) or under two different conditions.

3. ANOVA (Analysis of Variance)
ANOVA is used to compare the means of three or more groups to see if at least one group mean is different from the others. The F-statistic calculated in ANOVA relies on the assumption that the sampling distribution of the group means is approximately normal, which is supported by the CLT.

4. Regression Analysis
In linear regression, hypothesis tests on the regression coefficients (e.g., t-tests for individual coefficients, F-tests for overall model significance) assume that the sampling distribution of the estimators is normal. This assumption is justified by the CLT when the sample size is large.

5. Proportion Testing
When dealing with proportions (e.g., the proportion of successes in a sample), the sampling distribution of the sample proportion will be approximately normal for large sample sizes due to the CLT. This principle is used in constructing confidence intervals and performing hypothesis tests for proportions.

- **Z-test for proportions**: Tests whether the proportion in a sample differs from a known or hypothesized population proportion.
- **Two-proportion z-test**: Compares the proportions from two independent samples to see if they differ significantly.

6. Chi-Square Tests
While the chi-square distribution is not normal, the chi-square test for independence and the chi-square goodness-of-fit test assume that the test statistic approximates a chi-square distribution, which in turn relies on the CLT for large sample sizes. These tests compare observed and

expected frequencies in categorical data.

7. Bootstrapping

Bootstrapping is a resampling technique used to estimate the distribution of a statistic by repeatedly sampling with replacement from the data. The CLT justifies that the distribution of the bootstrap sample means will approximate a normal distribution as the number of resamples increases.

8. Central Limit Theorem-Based Approximations

Various statistical methods rely on normal approximations for discrete distributions. For example, the normal approximation to the binomial distribution is used when the sample size is large, as per the CLT.

- **Normal Approximation to the Binomial Distribution**: Used for calculating probabilities and constructing confidence intervals for binomially distributed data when sample sizes are large.