Welcome

Name: Georgy Balayan
Link to colab notebook:
https://colab.research.google.com/drive/1JF3PVHAtANERpZkVkxYRgcLmhTuoFIzu?usp=sharing

Goals

Introduce students to the Central Limit Theorem (CLT)

Provide examples of Applications of the the Central Limit Theorem

Hypothesis Testing (TBD)

Prerequisites

Basic Python coding skills

Why central limit theorem

Example: performance profiling

Represents the most confused and misinterpreted fundamental topics in Statistics

Q: What is central limit theorem

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be normal.

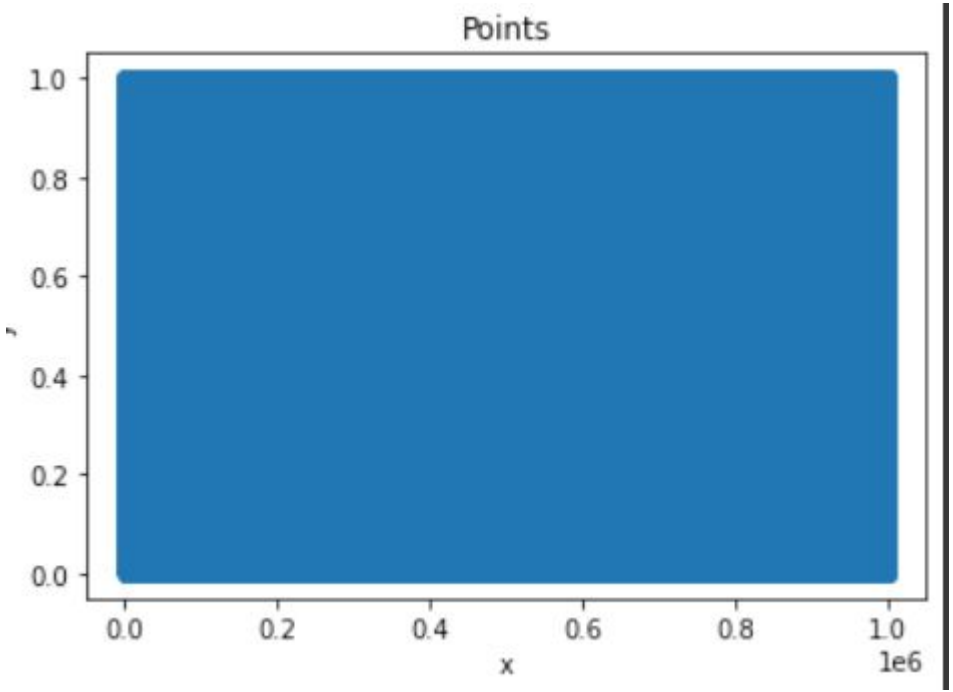
Properties of the distribution of sample means

Mean of the sample means = mean of the population
Standard deviation of the sample means = standard deviation of the population / sqrt(n), where n - sample size

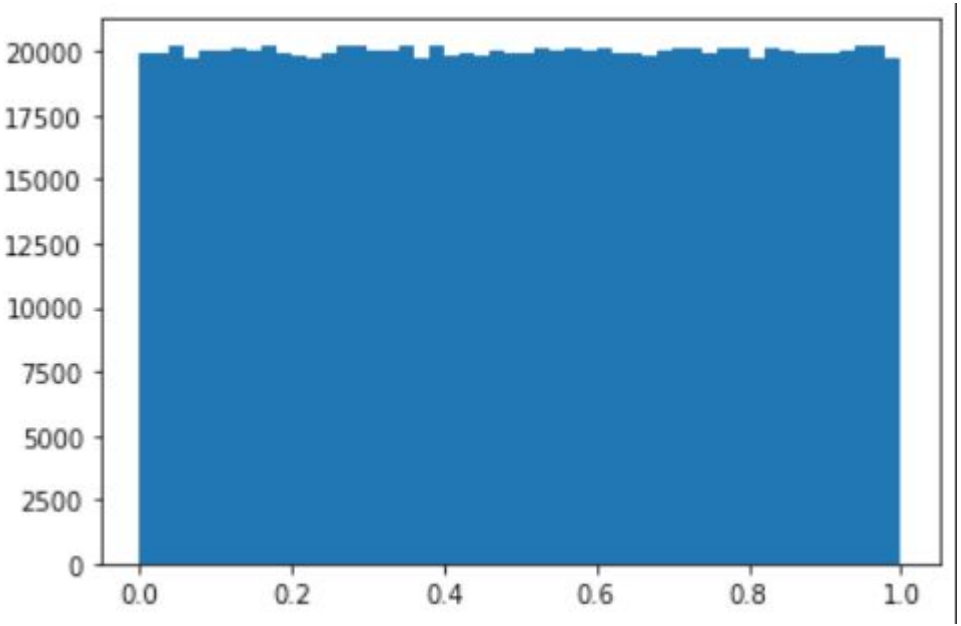
Example

Generate population of 10K floating point numbers uniformly distributed in the range of 0 to 1

Scatter plot of population (size of 1M)

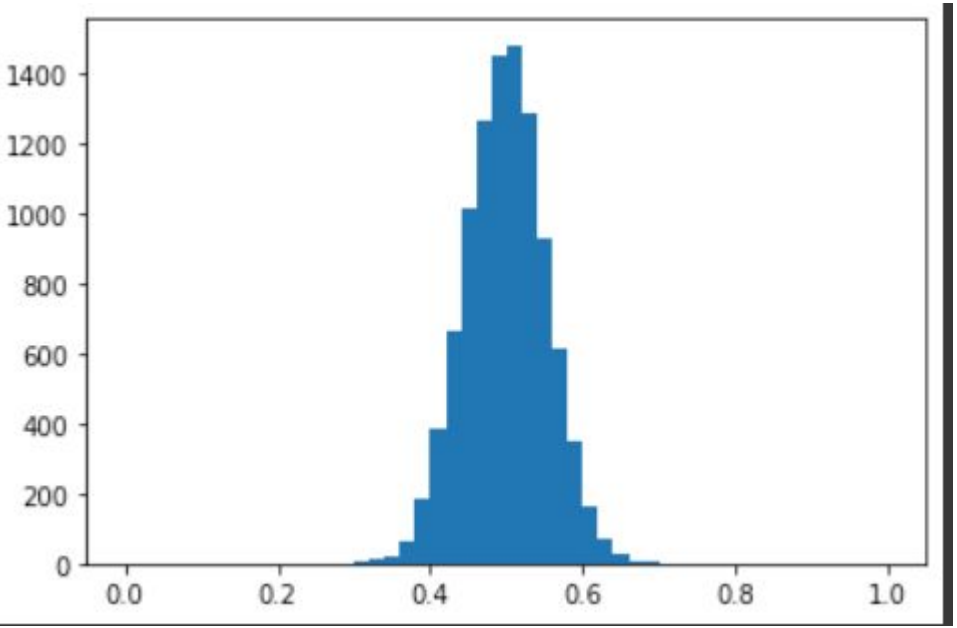


Histogram of the population



1m items
50 bins
20k = 1M/50

Let’s build a distribution (histogram) of the sample means

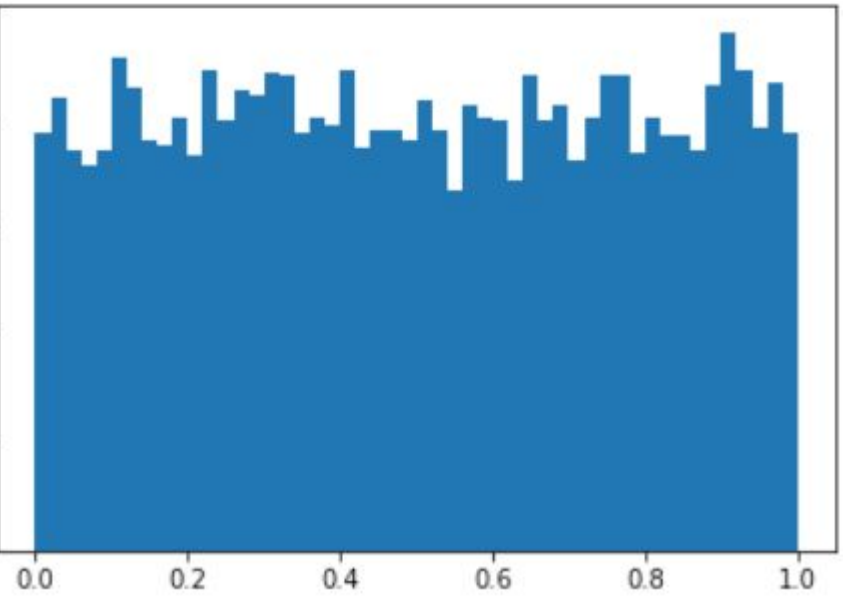


Number of samples = 10k
Sample size = 30
Mean = 0.5 (population mean)
Standard deviation = Population Standard deviation / sqrt(30)

Q: Change the number of samples and sample size in a systematic way

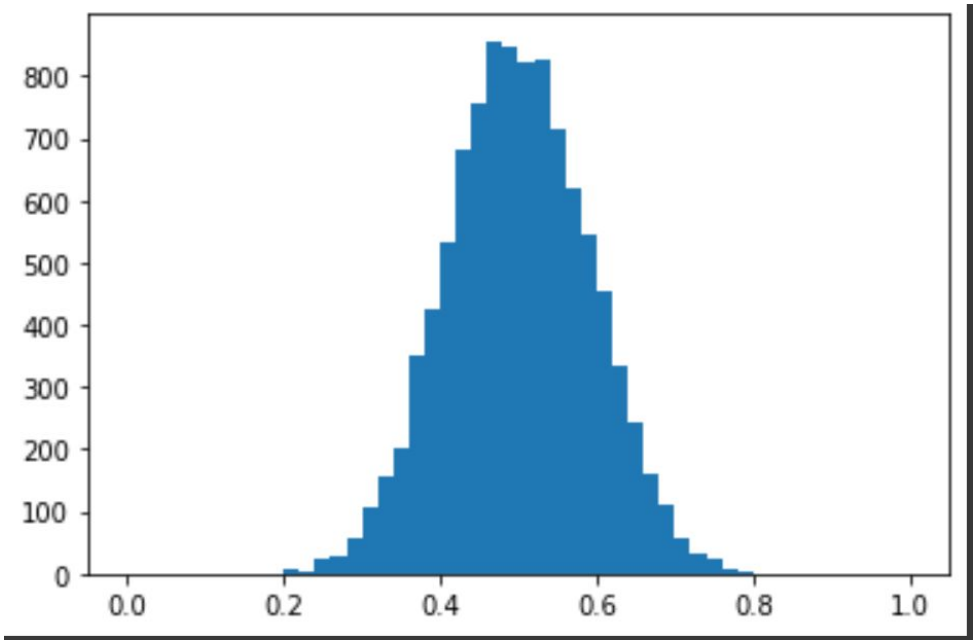
Number of samples\Sample size	1	10	30+
High	Population	Student	Normal
Low	Unknown	Unknown	Unknown (centered around mean)

Number of samples=10k and sample size=1 (the same distribution)



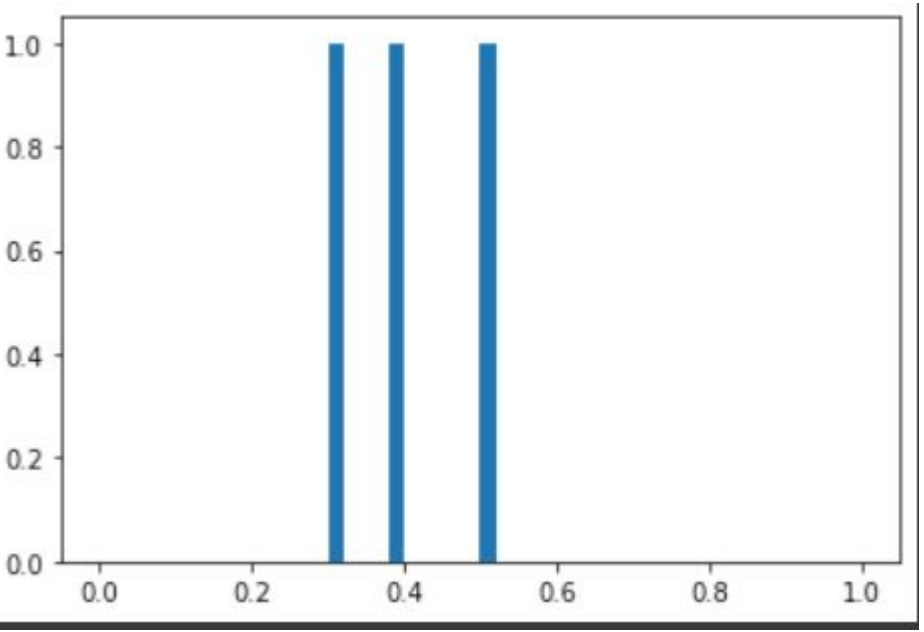
Number of samples = 10k
Sample size = 1
Mean = 0.5 (population mean)
Standard deviation = Standard deviation Variance / sqrt(1)

Number of samples=10k and sample size=10 (t student)



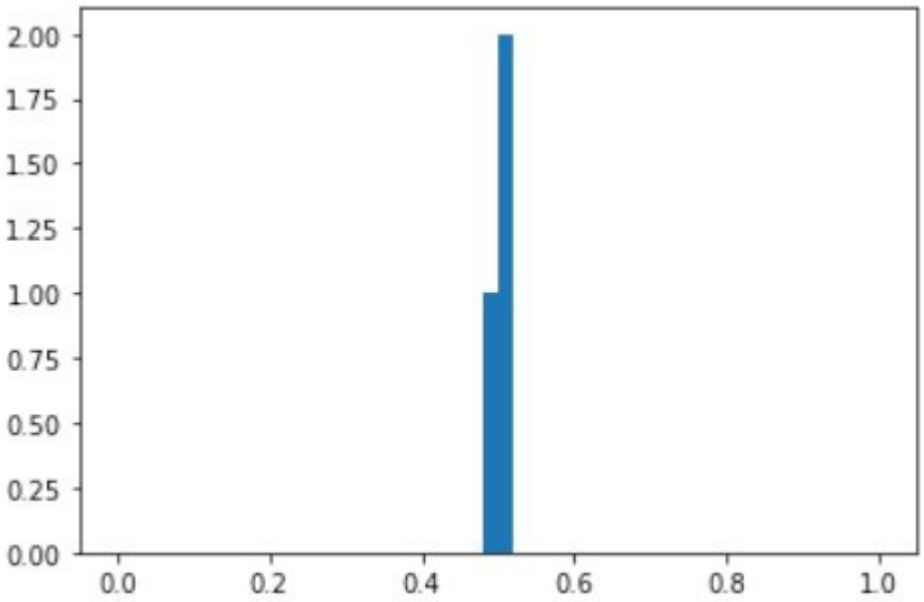
Number of samples = 10k
Sample size = 10
Mean = 0.5 (population mean)
Standard deviation = Standard deviation Variance / sqrt(10)

Number of samples=3 and sample size=10 (unknown)



Number of samples = 3
Sample size = 10
Mean = 0.5 (population mean)
Standard deviation = Population
Standard deviation / sqrt(3)

Number of samples=3 and sample size=3000 (unknown, but centered around the population mean)



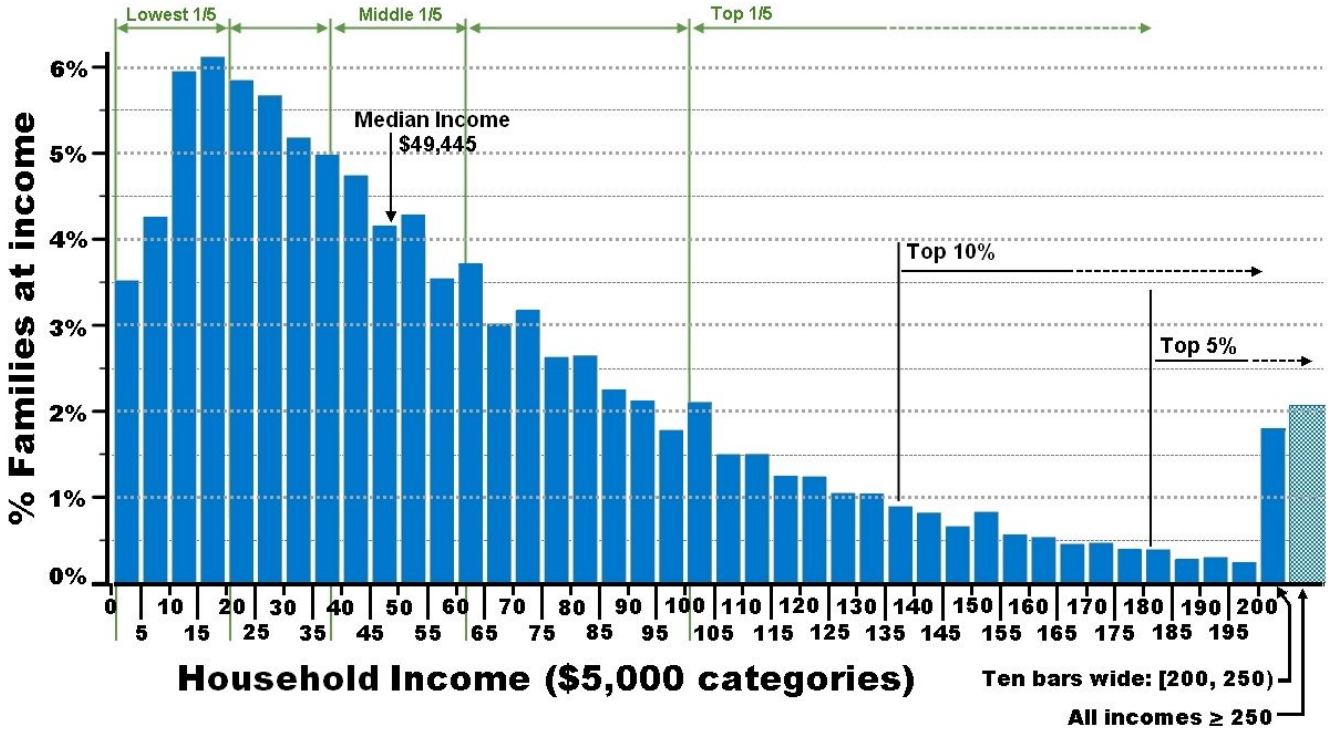
Number of samples = 3000
Sample size = 3
Mean = 0.5 (population mean)
Standard deviation = Population Standard deviation / sqrt(3)

Summary

Number of samples\Sample size	1	10	30+
High	Population	Student	Normal
Low	Unknown	Unknown	Unknown (centered around mean)

Applications 1: population distribution is not normal

Estimating the distribution of household income: mean and standard deviation



Applications 2: population distribution is normal

Measurement errors: performance profiling. Mean and Standard deviation can give us even more information about the underlying distribution.

Sample size = number of measurements

Number of samples = number of experiments

One experiments is comprised of multiple measurements

Mean of the population = Mean of sample means

Standard deviation of the population = standard deviation of the sample means * sqrt(sample size)

Hypothesis Testing (TBD)

Thank you for your time and attention