# Welcome

Name: Georgy Balayan

# Goals

Introduce the audience to the Central Limit Theorem (CLT)

Provide examples of Applications of the the Central Limit Theorem

# Prerequisites

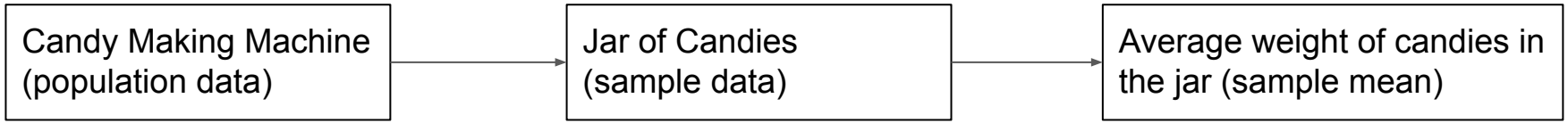
Basic Python coding skills

# Why central limit theorem

Represents the most confused and misinterpreted fundamental topics in Statistics

# Example: Candies

Estimating the distribution of weights of candies the machine produces




Notes:

- Number of candies in the jar = sample size
- Number of jars = number of samples

# 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 a sufficiently large number 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

# Environment setup

1. Recommended docker setup:

   Run the following commands in your terminal:

   ```
   git clone git@github.com:MLWorkshops/data-science-interview-workshop.git
   make pull
   make docker-run
   ```

   And then open http://127.0.0.1:8888/ in the browser

2. Colab access:

   Open this link in your browser:

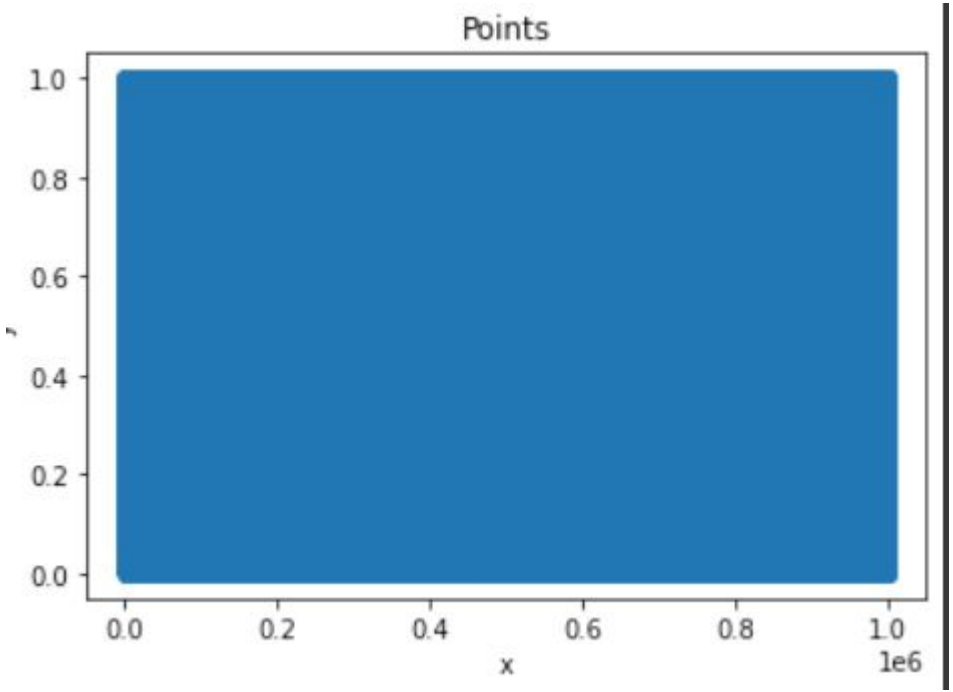
   https://colab.research.google.com/drive/1EBvCVN1JX_X-_9qkH3YpQpEYXxc9VbdY?usp=sharing

   And choose File -> Save a copy in Drive in your to be able to edit the notebook

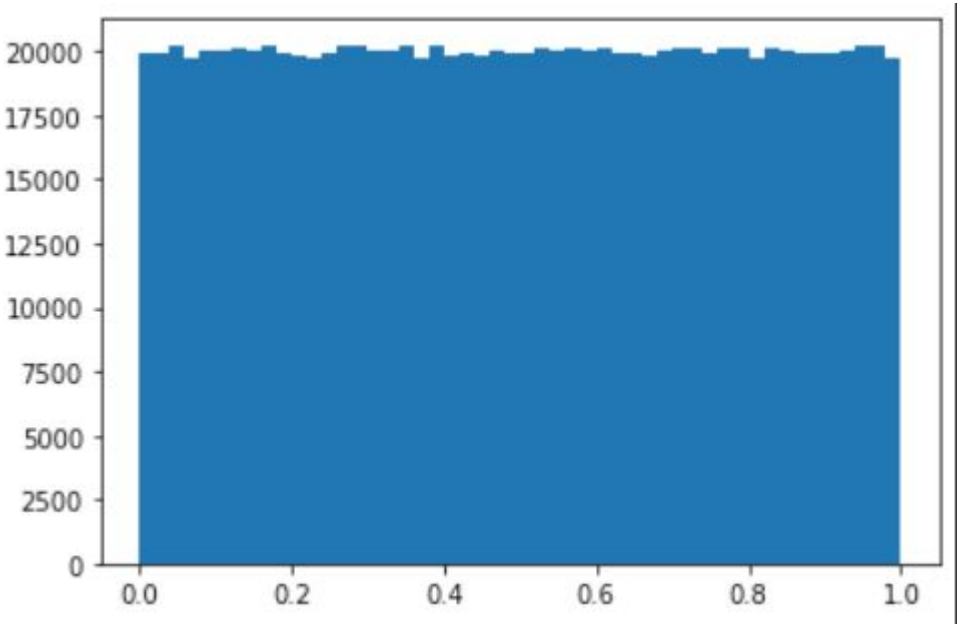
# Example: quantified

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

# Scatter plot of population (size of 1M)



# Histogram of the population

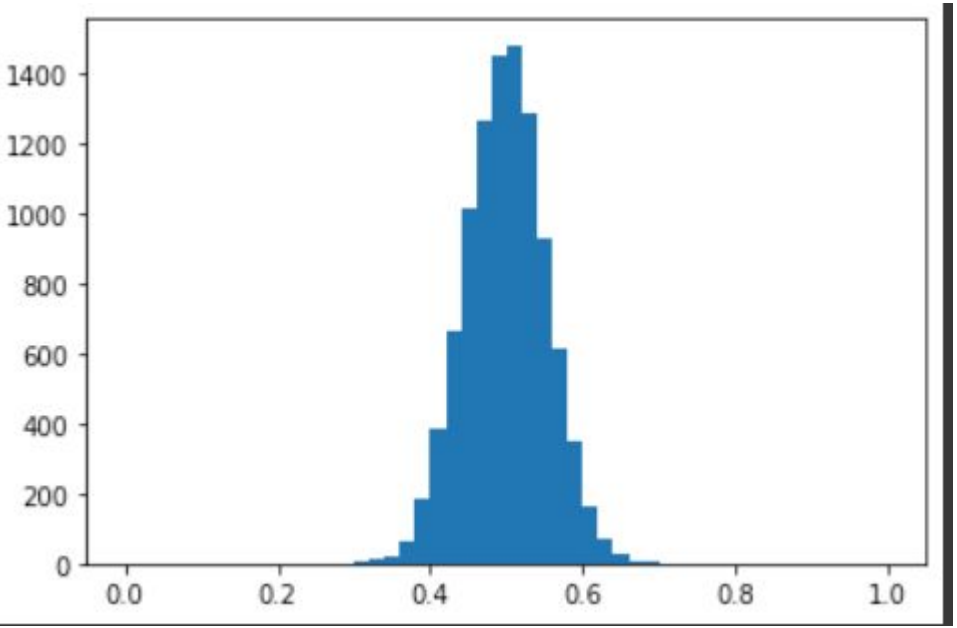



1m items
50 bins
20k = 1M/50

# 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) |
| | | | |

# 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)

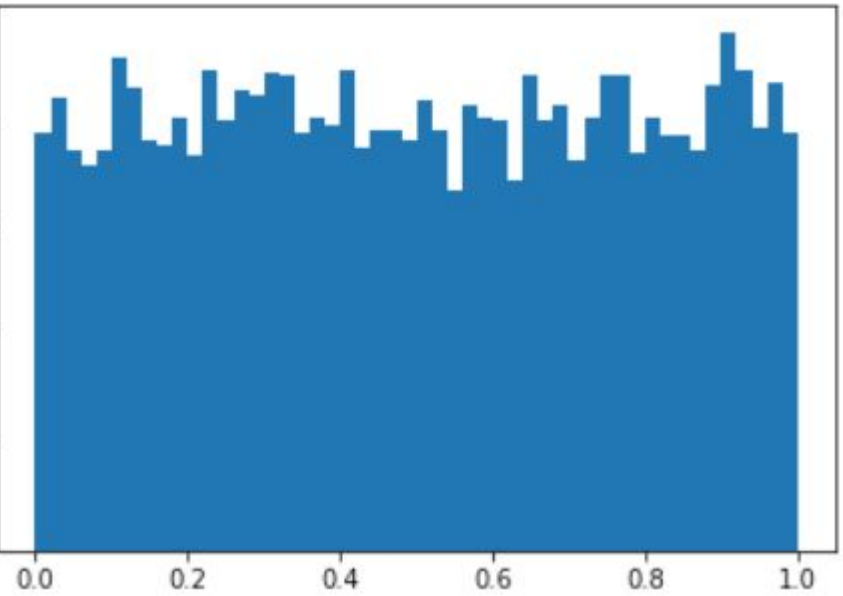
# Coding Challenge

# write a function that computes a variance of a given list of values

# 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)

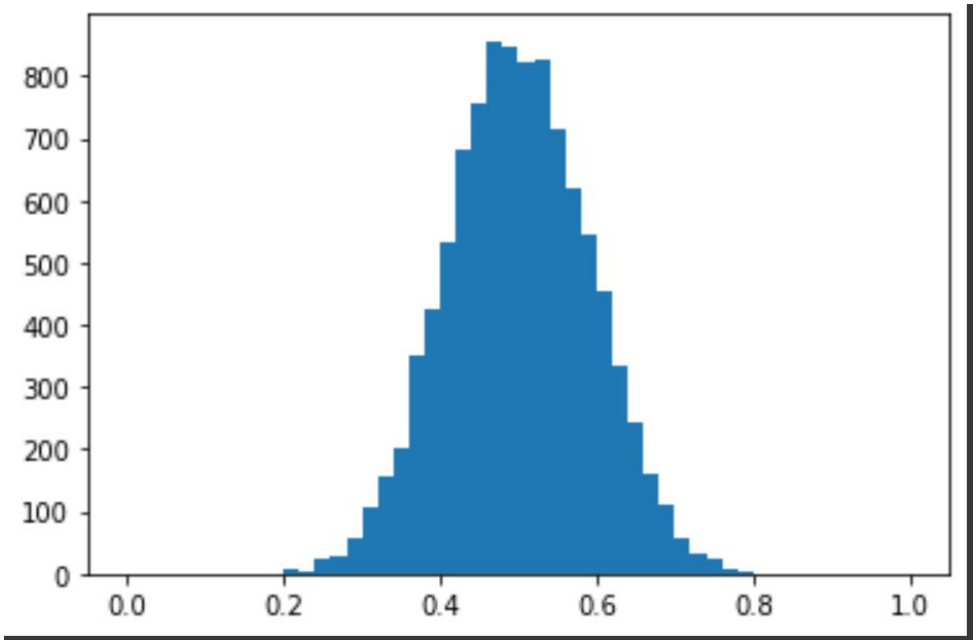



Number of samples = 10k
Sample size = 1
Mean = 0.5 (population mean)
Standard deviation = Standard deviation Variance / sqrt(1)

# 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=10 (t student)

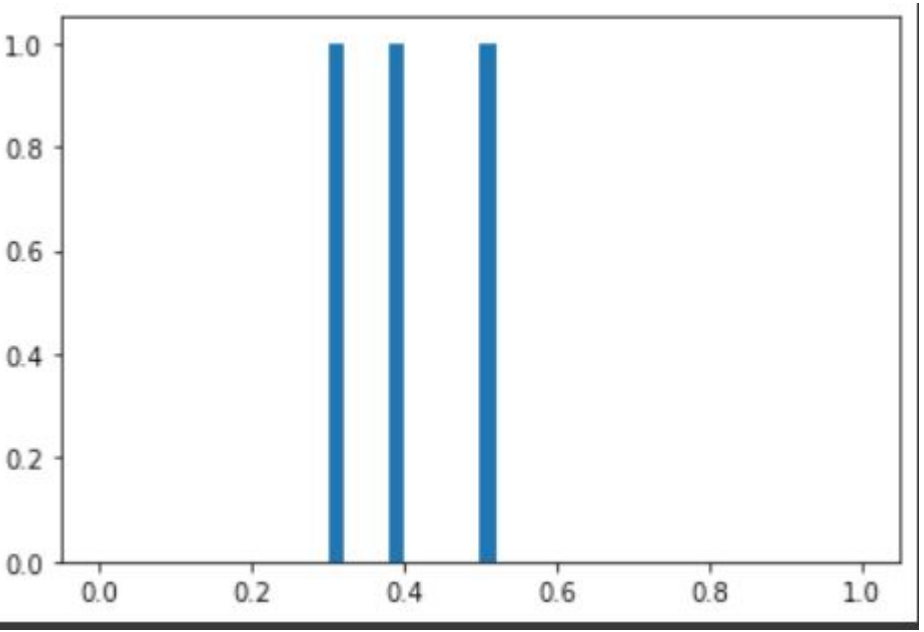



Number of samples = 10k
Sample size = 10
Mean = 0.5 (population mean)
Standard deviation = Standard deviation Variance / sqrt(10)

# 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=3 and sample size=10 (unknown)

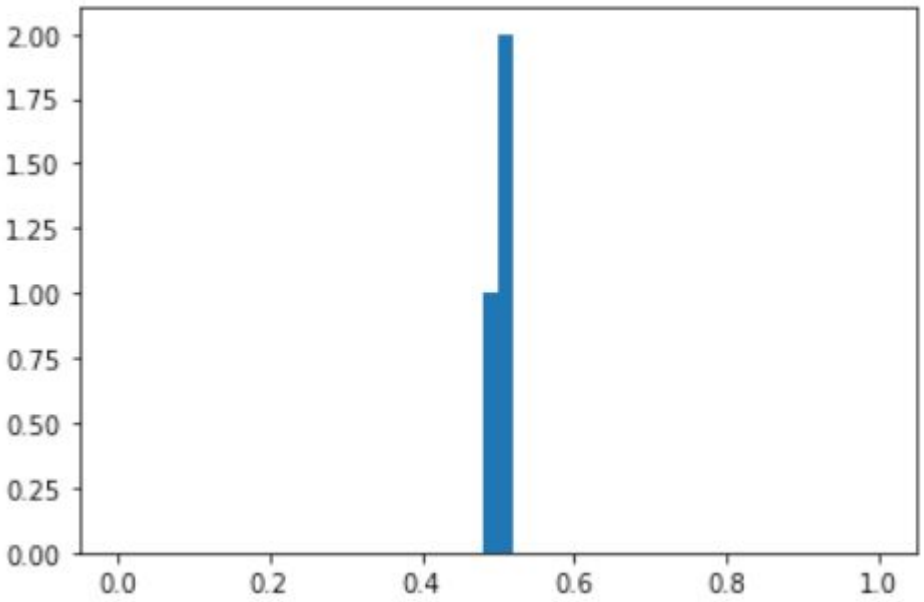



Number of samples = 3
Sample size = 10
Mean = 0.5 (population mean)
Standard deviation = Population
Standard deviation / sqrt(3)

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




Number of samples = 3
Sample size = 3000
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) |
| | | | |

# Application: 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 *experiment* 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)*

# Coding Challenge

# write a function that computes a z-score for a given number assuming

# that the values follow a normal distribution

# z-score is is the number of standard deviations by which the value of a

# raw score (i.e., an observed value or data point)

# is above or below the mean value of what is being observed or measured.
[https://en.wikipedia.org/wiki/Standard_score]

Thank you for your time and attention