L01:Introduction to Statistics

Prof. Yiming QIN
Assistant Professor
School of Energy and Environment
City University of Hong Kong

Why Study Statistics?

"Without data, you're just another person with an opinion."

— W. Edwards Deming (an American Engineer, Statistician, and Professor)

However, when you get the data, how to analyze the data?

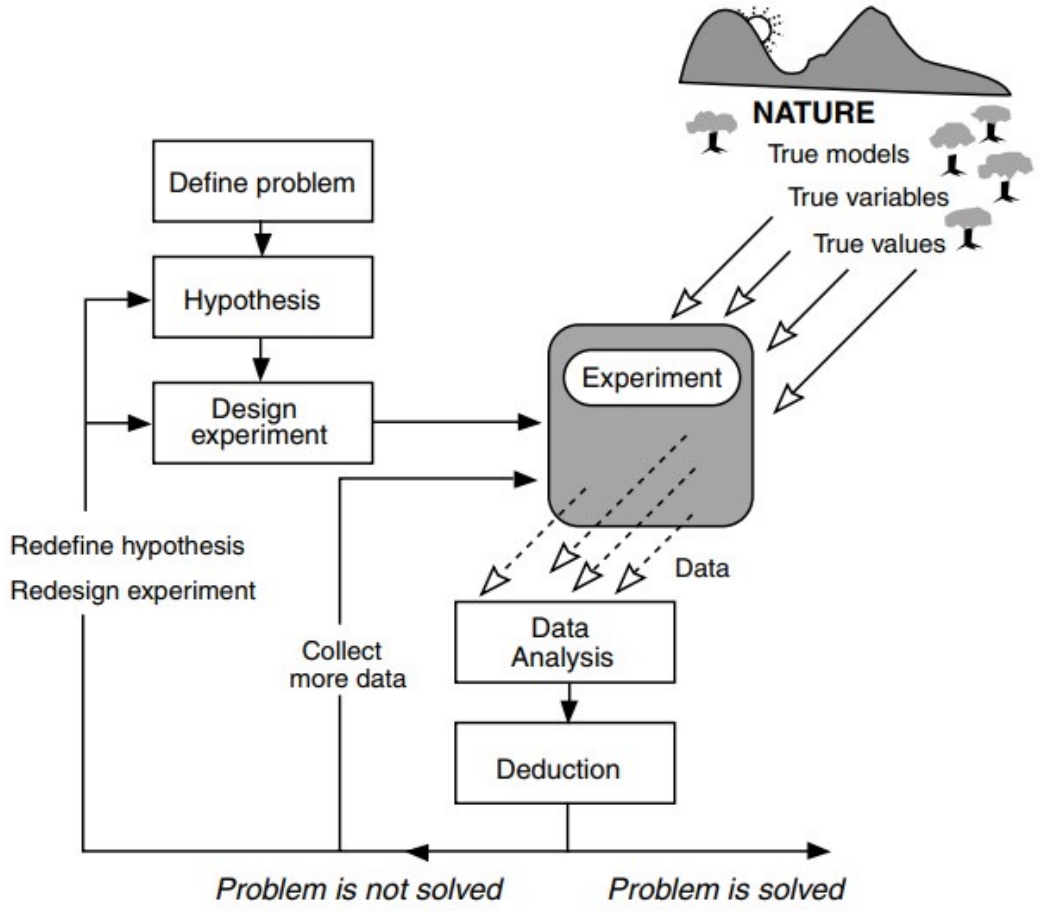
Why Study Statistics?

- To be informed…
  a) Extract information from tables, charts and graphs
  b) Follow numerical arguments
  c) Understand the basics of how data should be gathered, summarized, and analyzed to draw statistical conclusions

- Provide you the tools to evaluate the work you need to make informed judgments.
- Use the data to answer questions of interest and to evaluate decisions that affect your life

Why Study Statistics?

Nature is viewed through the experimental window.



- Knowledge increases by iterating between experimental design, data collection, and data analysis.
- In each cycle, the engineer may formulate a new hypothesis, add or drop variables, change experimental settings, and try new methods of data analysis.
- Statistics is important to data analysis.

Environmental Problems and Statistics

There are many aspects of environmental and energy problems: economic, political, psychological, medical, scientific, and technological.

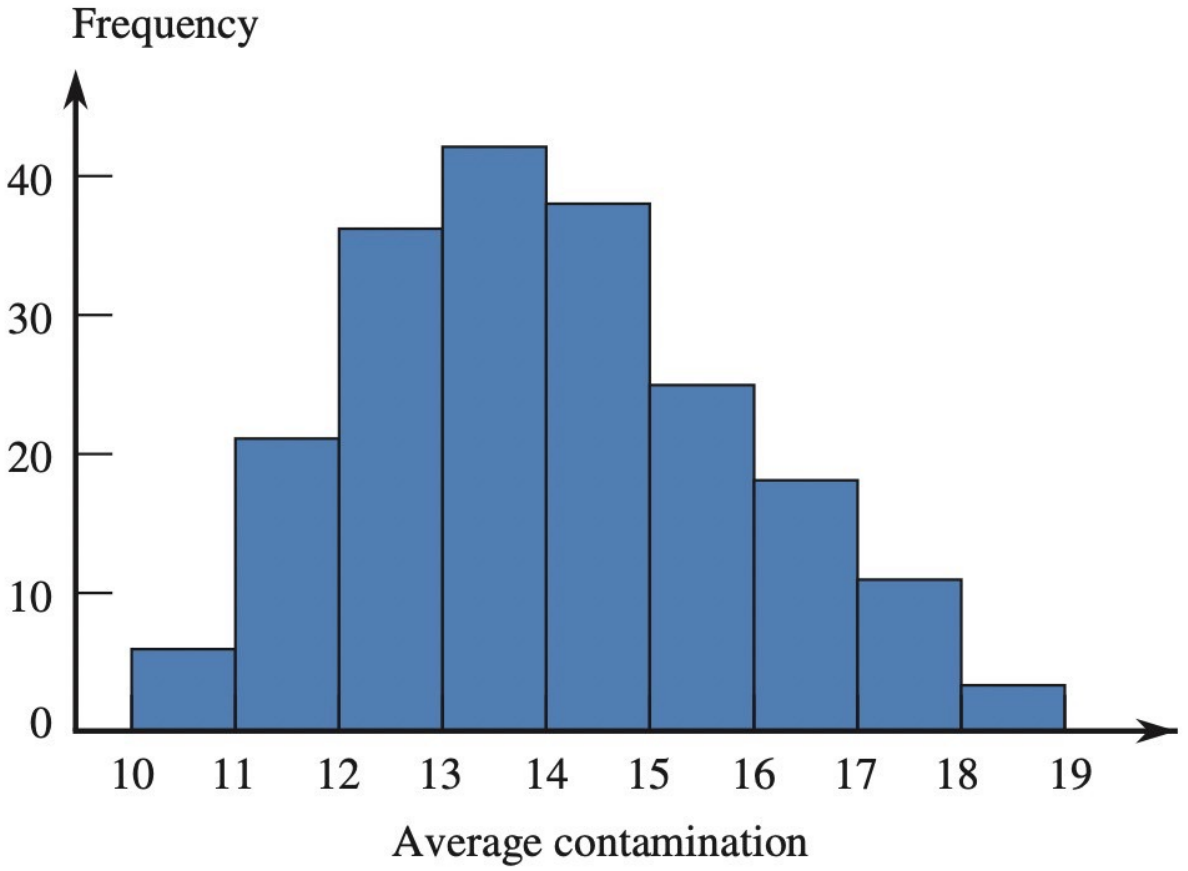
Understanding and solving such problems often involves certain quantitative aspects, in particular the acquisition and analysis of data.

Treating these quantitative problems effectively involves the use of statistics. Statistics can be viewed as the prescription for making the quantitative learning process effective.

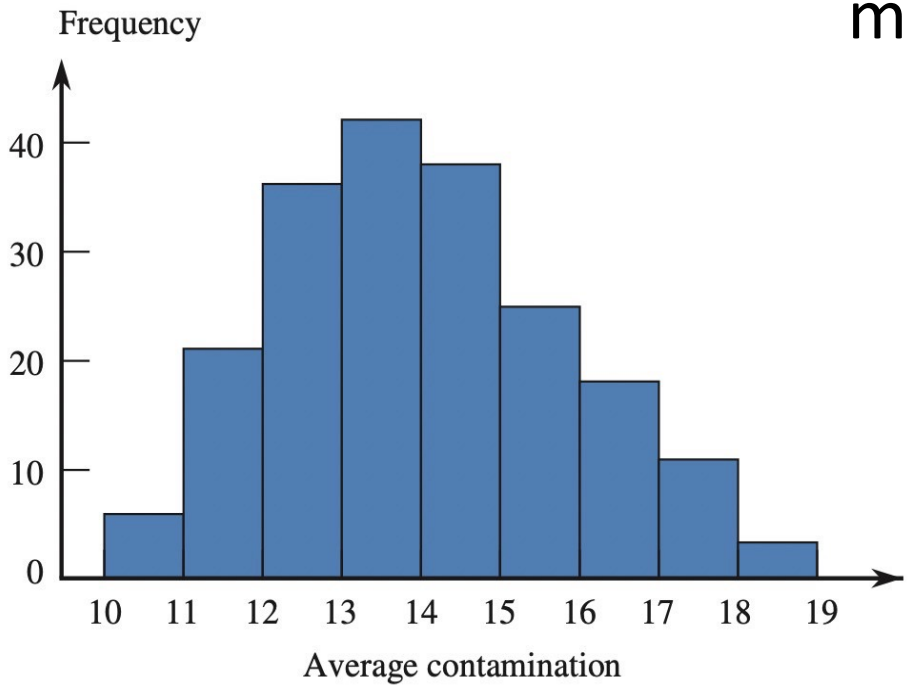
Example: Monitoring Water Quality

An environmental control board selects five water specimens from a particular well each day and calculated the *average contamination concentration in ppm*

The histogram summarizes the average contamination values for 200 days.



Example: Monitoring Water Quality



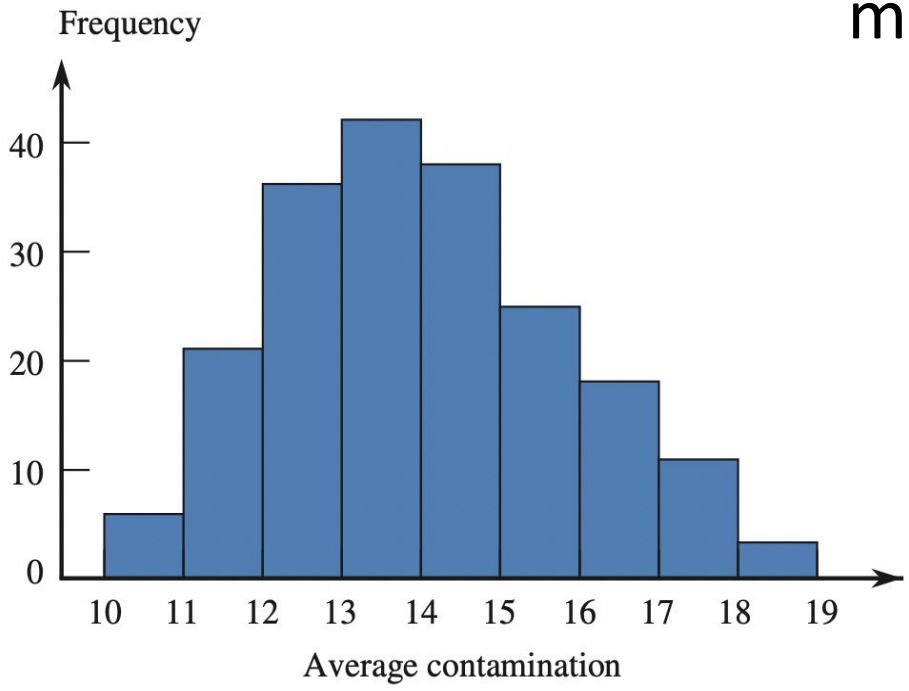
Average contamination values for 200 days before the spill

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well.

Five water specimens are collected from the well, and the average contamination is 15.5 ppm.

Do you think this is a convincing evidence that the well water was affected by the spill?

Example: Monitoring Water Quality



Average contamination values for 200 days before the spill

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well.

What if the calculated average was 17.4 ppm? 22.0 ppm?

- Conclusions required an understanding of statistics, and specifically in this case, the variability.

What is Statistics?

- A mathematical science that pertains to the data collection, data analysis, data interpretation, data explanation, and data presentation.

- A method for processing and analyzing the collected data so as to help reduce the uncertainty inherent in decision making.

Jobs that use statistics

Jobs that use statistics

Whether you're an entry- or senior-level professional, you're sure to find a statistics career that matches your interests and qualifications. Knowing what to expect regarding these careers can help you determine the particular career you want to pursue. Here are 17 jobs that use statistics, their average salaries and their primary duties. For the most up-to-date salary information from Indeed Salaries, click on the salary link by each job title below:

1. Meteorologist

National average salary: $65,357 per year

Primary duties: Meteorologists study meteorological phenomena such as the climate and weather to forecast weather conditions in a particular area. They use their knowledge of statistics to record and analyze data from radars, weather stations, satellites and remote sensors.

9. Research scientist

National average salary: $112,850 per year

Primary duties: Research scientists create and conduct experiments, record and analyze data, perform fieldwork and present their findings to a senior-level employee or a company's research staff. They may also create statistical forecasting models and write research papers and proposals.

11. Data scientist

National average salary: $123,010 per year

Primary duties: Data scientists create machine learning-based tools or processes and use software designed to help them perform statistical analysis. Their specific approach to data analysis depends on their industry and their company's needs.

Key terms in statistics

Population: A collection of all possible objects or individuals of interest.

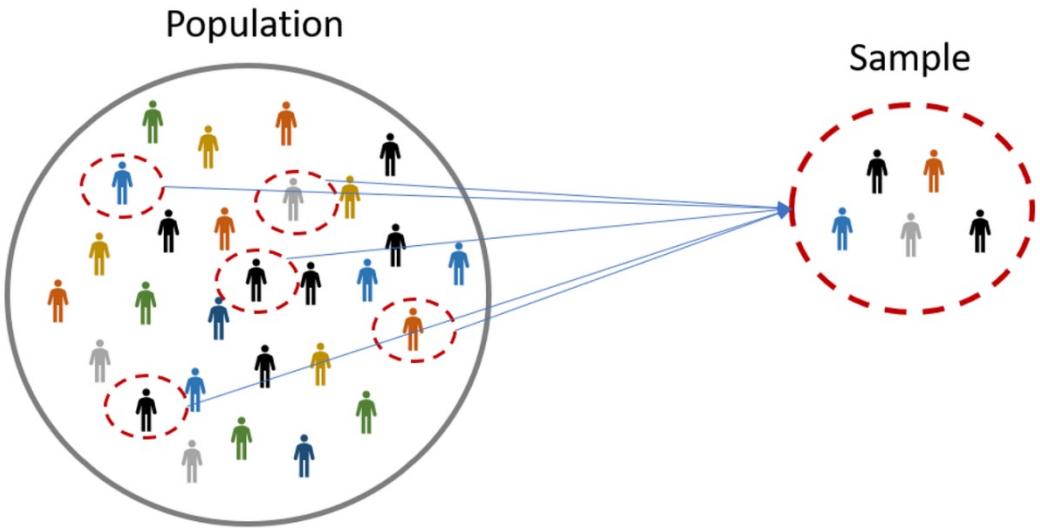
Sample: A part of a population selected for analysis.

Parameter: A numerical measures computed from a population and used to describe a characteristic of the population.

Statistic: A numerical measures computed from a sample and used to describe a characteristic of the sample.

Population & Sample

A population data set contains all members we want to know about



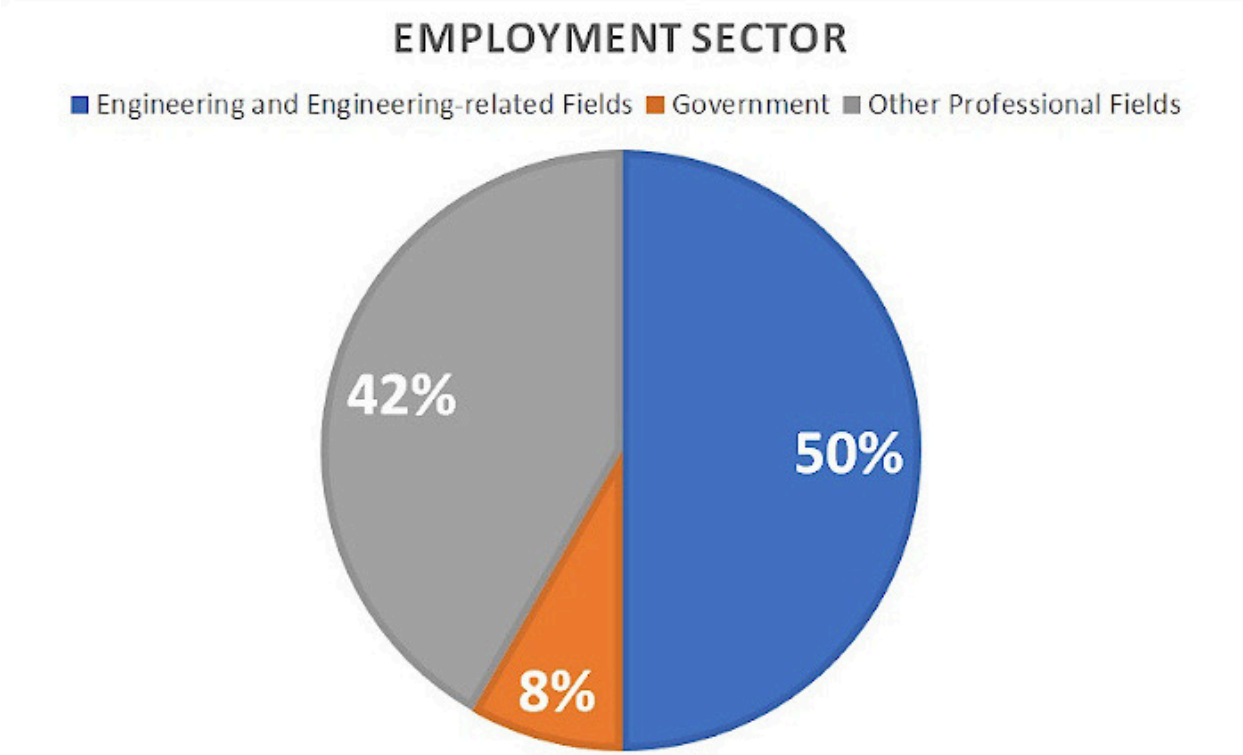
POPULATION	SAMPLE
Whole group	**Part** of the group
Group **I want to know** about	Group **I do know** about
Characteristics are called **parameters**	Characteristics are called **statistics**
Parameters are generally **unknown**	Statistics are always **known**
Parameter is **fixed**	Statistics **change** with the sample

A sample data set contains a part, or a subset, of a population.

Example

Graduate Employment Statistics of BEng Energy Science and Engineering in SEE, CityU

According to the Graduate Employment Survey 2021



Is it a sample or population? Why?

Example

Population	Sample
All likely voters in the next election	800 voters selected at random for interview
All parts produced today	A few parts selected for quality control
All sales receipts for September	Every 100th receipt selected for audit

Example

Population	Sample
All likely voters in the next election	800 voters selected at random for interview
All parts produced today	A few parts selected for quality control
All sales receipts for September	Every 100th receipt selected for audit

Why drawing a sample?

Example

Population	Sample
All likely voters in the next election	800 voters selected at random for interview
All parts produced today	A few parts selected for quality control
All sales receipts for September	Every 100th receipt selected for audit

Why drawing a sample?

It is generally impossible or impractical to examine the entire population, but we may examine a part of it (a sample from it) so that we can find out something (say, population mean) about the target population from the sample (say, use the sample mean) --- making a statistical inference regarding the entire (unknown) population.

Sampling Distributions

Many studies are conducted in order to generalize from a sample to the corresponding population. As a result, it is important that the sample be representative of the population.

To be reasonably sure of this, we must carefully consider the way in which the sample is selected.

Collecting data sensibly is very important !

Sample Distributions

Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

Sample Distributions

Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

How close is the sample mean to μ?

Would the sample mean be a good estimate of μ?

Will other samples of n = 50 have the same sample mean?

To answer these questions, we will examine the sampling distribution, which describes the long-run behavior of sample statics.

Sample Distributions

Statistic

- A number that that can be computed from sample data
- Some statistics we will use include
  - sample mean, $\bar{x}$
  - standard deviation, s

The variability is called sampling variability

- The observed value of the statistic depends on the particular sample selected from the population and it will vary from sample to sample.

Population Mean — μ or mu

The Greek letter μ (mu) is the symbol for a population mean.

$$\mu = \frac{\sum X}{N}$$

In this equation, the numerator sums all values in the population. That's the ∑X in the numerator. In the denominator, N is the total number of values in the population. This formula is comprehensive, encompassing every single data point in the population.

Crucially, calculating the population mean μ is generally impossible.

Sample Mean — $\bar{x}$ (x bar)

We can use the sample mean $\bar{x}$ to estimate it when using random samples with the following formula:

$$\bar{x} = \frac{\sum X}{n}$$

This formula is structurally similar to that of the population mean. We're still summing all the values but only those in our subset. Additionally, notice the lowercase n in the denominator.
This n represents the number of values in the sample rather than the population.

So, we sum all values in our subset and divide by the sample size.

Population standard deviation

When you have collected data from every member of the population that you're interested in, you can get an exact value for population standard deviation. The population standard deviation formula looks like this：

Formula	Explanation
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$	• σ = population standard deviation • $\sum$ = sum of... • X = each value • μ = population mean • N = number of values in the population

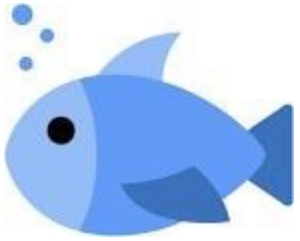
Sample standard deviation

When you collect data from a sample, the sample standard deviation is used to make estimates or inferences about the population standard deviation.

Formula	Explanation
$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$	• s = sample standard deviation • $\sum$ = sum of... • X = each value • $\bar{x}$ = sample mean • n = number of values in the sample

We use (n – 1) because we just like to make the "spread" (or deviation) a little larger to reflect the fact that, since we are using a sample, not the entire population, we have more uncertainty.

Sample Distributions



Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches. What is the sample mean?

Sample Distributions

Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches. What is the sample mean?

$\overline{x}$ = 7.27 inches

Let's catch two more samples and look at the sample means

Sample Distributions

Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches. What is the sample mean?

$\overline{x}$ = 7.27 inches

Sample Distributions

Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.

$\overline{x}$ = 7.27 inches

2nd sample - 8.5, 4.6, and 5.6 inches.

3rd sample – 10.3, 8.9, and 13.4 inches.

Let's catch two more samples and look at the sample means

Sample Distributions

Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.

$\overline{x}$ = 7.27 inches

2nd sample - 8.5, 4.6, and 5.6 inches.

$\overline{x}$ = 6.23 inches

3rd sample – 10.3, 8.9, and 13.4 inches.

$\overline{x}$ = 10.87 inches

Let's catch two more samples and look at the sample means

Sample Distributions

4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

There are 1140 ($_{20}C_3$) different possible samples of size 3 from this population. If we catch all those different samples and calculate the mean length of each sample, we would have a distribution of all possible $\overline{x}$

This would be the sampling distribution of $\overline{x}$

Sample Distributions

The distribution that would be formed by considering the value of a sample statistic for every possible different sample of a given size from a population.

Sample Distributions

The distribution that would be formed by considering the value of a sample statistic for every possible different sample of a given size from a population.

In this case, the sample statistic is the sample mean $\overline{x}$

Sample Distributions

Suppose there are only 5 fish in the pond. The lengths of the fish are given below

6.6	11.7	8.9	2.2	9.8

What is the mean and standard deviation of this population?

$$\mu_x = 7.84$$

$$\sigma_x = 3.262$$

Sample Distributions

6.6	11.7	8.9	2.2	9.8

Let's find all the samples of size 2.

Pairs	6.6 & 11.7	6.6 & 8.9	6.6 & 2.2	6.6 & 9.8	11.7 & 8.9	11.7 & 2.2	11.7 & 9.8	8.9 & 2.2	8.9 & 9.8	2.2 & 9.8
$\bar{x}$	9.15	7.75	4.4	8.2	10.3	6.95	10.75	5.55	9.35	6

What is the mean and standard deviation of this population?

$$\mu_{\bar{x}} = 7.84$$

$$\sigma_{\bar{x}} = 1.998$$

Sample Distributions

6.6	11.7	8.9	2.2	9.8

Let's find all the samples of size 3.

Pairs	6.6 & 11.7 & 8.9	6.6 & 11.7 & 2.2	6.6 & 11.7 & 9.8	6.6 & 8.9 & 2.2	6.6 & 8.9 & 9.8	6.6 & 2.2 & 9.8	11.7 & 8.9 & 2.2	11.7 & 8.9 & 9.8	11.7 & 2.2 & 9.8	8.9 & 2.2 & 9.8
$\bar{x}$	9.067	6.833	9.367	5.9	8.433	6.2	7.6	10.133	7.9	6.967

$$\mu_{\bar{x}} = 7.84$$

$$\sigma_{\bar{x}} = 1.332$$

Sample Distributions

What do you notice?

- The mean of the sampling distribution EQUALS the mean of the population

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

- As the sample size increases, the standard deviation of the sampling distribution decease.

$$\text{As } n \uparrow, \sigma_{\bar{x}} \downarrow$$

Sample Distributions: General Property

Rule 1: $\mu_{\bar{x}} = \mu$

Rule 2: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

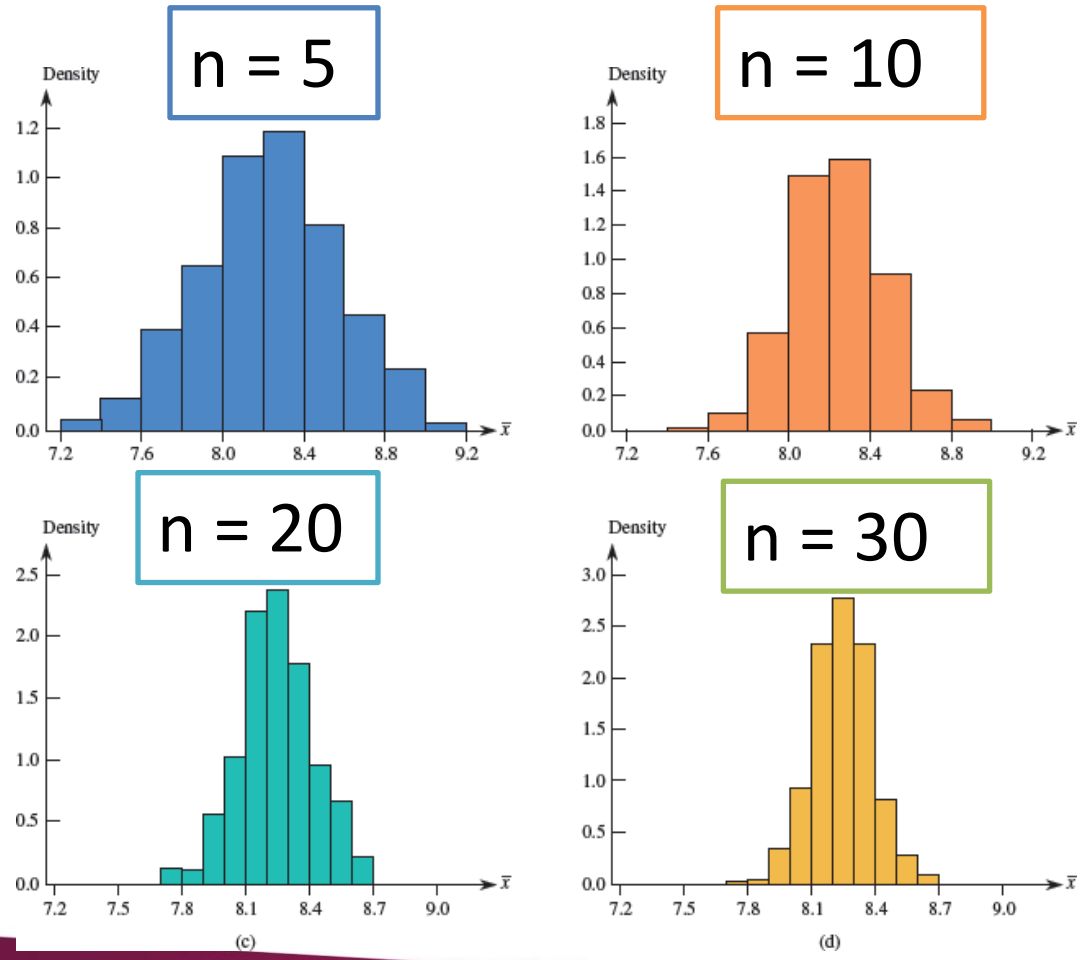
The mean of the sampling distribution EQUALS the mean of the population.

The standard deviation of the sample means is equal to the population standard deviation divided by the square root of n.

The standard deviation of the sampling distribution of the sample means is called the standard error of mean.

Example

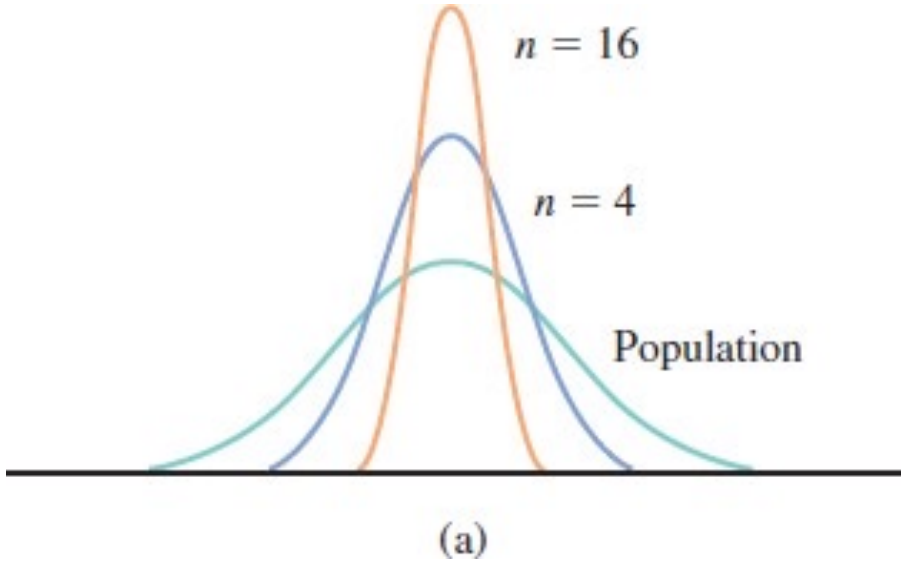
Let's generate 500 random samples of $n = 5$, 10, $n = 20$, and $n = 30$. The density histograms below display the results for each of the given sample sizes.



What do you notice about the shape, mean, and standard deviation of these histograms?

Sample Distributions: General Property

Rule 3: When the population distribution is normal, the sampling distribution of *x* is also normal for any sample size *n*.



$$\mu_x = \mu$$

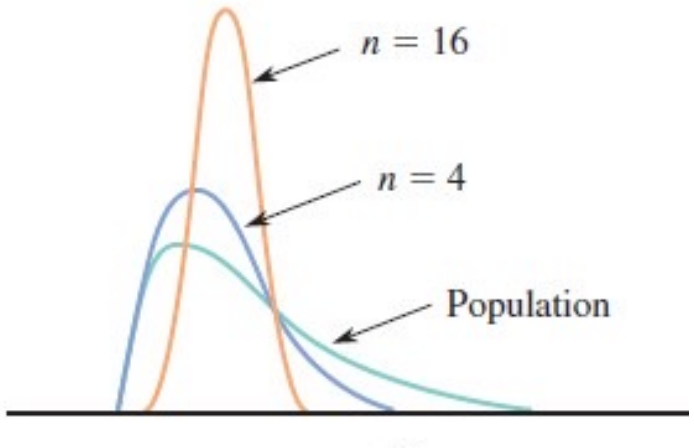
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Population distribution and sampling distributions of x of symmetric population

Sample Distributions: General Property

Rule 4: Central Limit Theorem

When *n* is sufficiently large, the sampling distribution of *x* is well approximated by a normal curve, even when the population distribution is not itself normal.



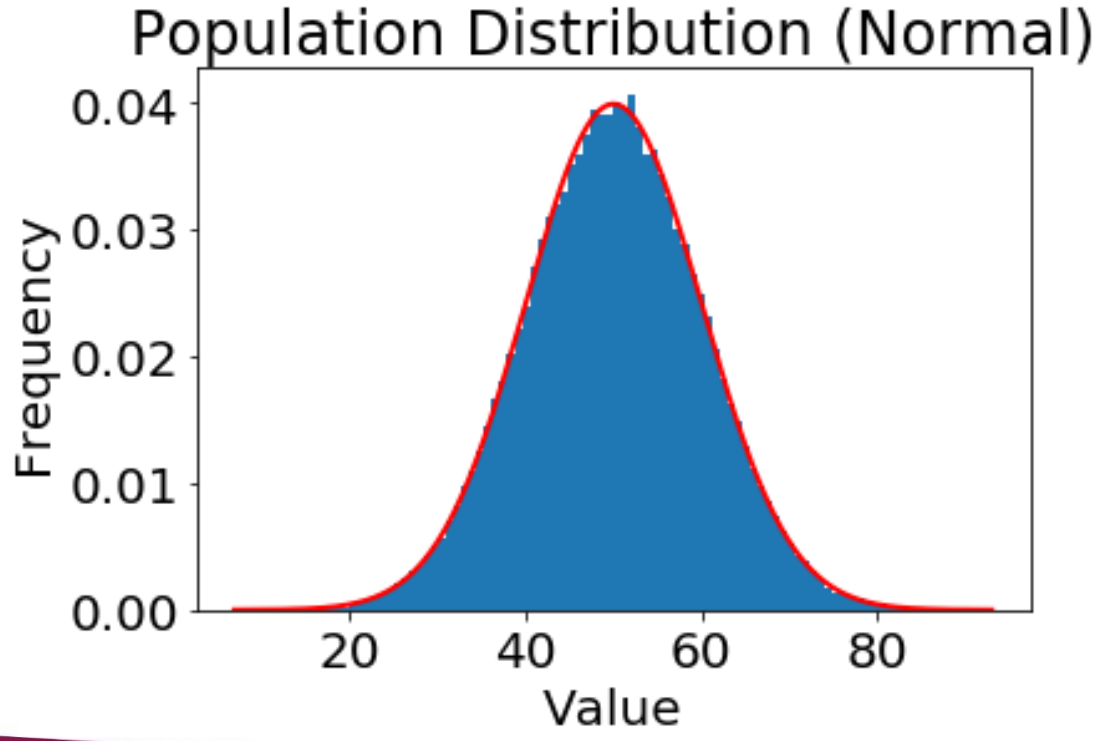
This result has enabled statisticians to develop procedures for making inferences about a population mean m using a large sample, even when the shape of the population distribution is unknown.

Population distribution and sampling distributions of x of skewed population

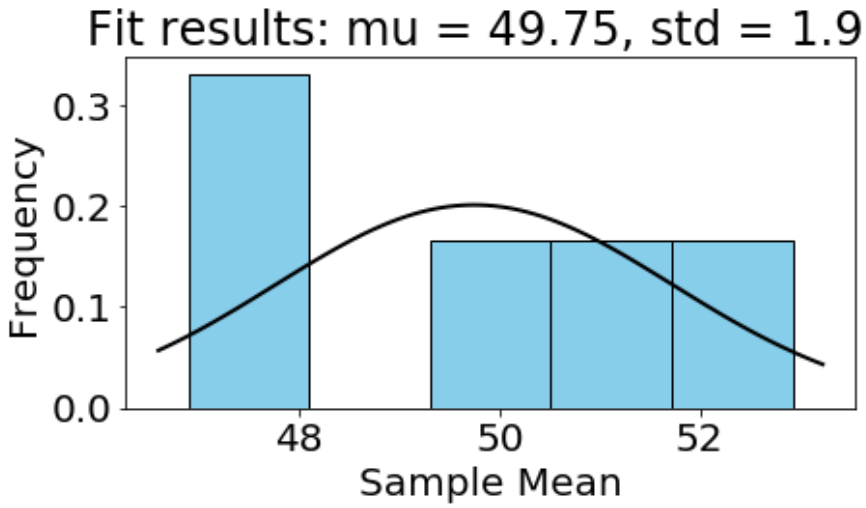
Python test for the Central Limit Theorem

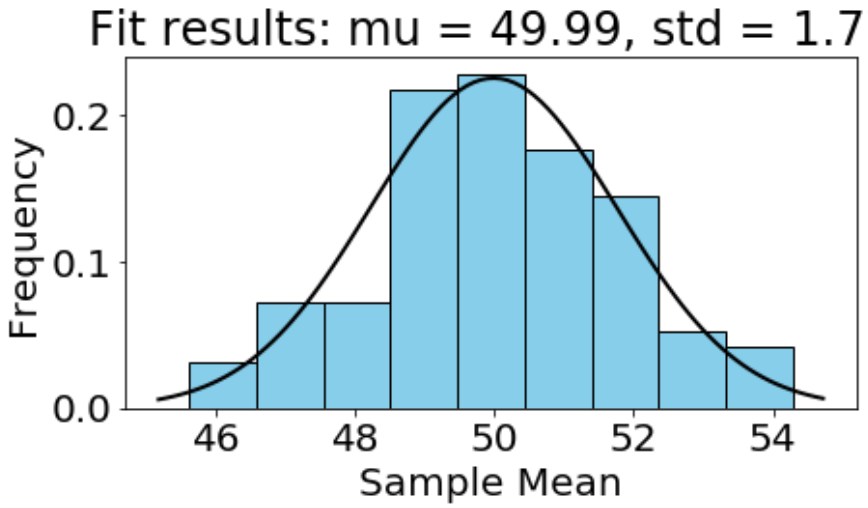
- Generate a normal distribution with the size of one hundred thousand, a mean of 50 and a standard deviation of 10

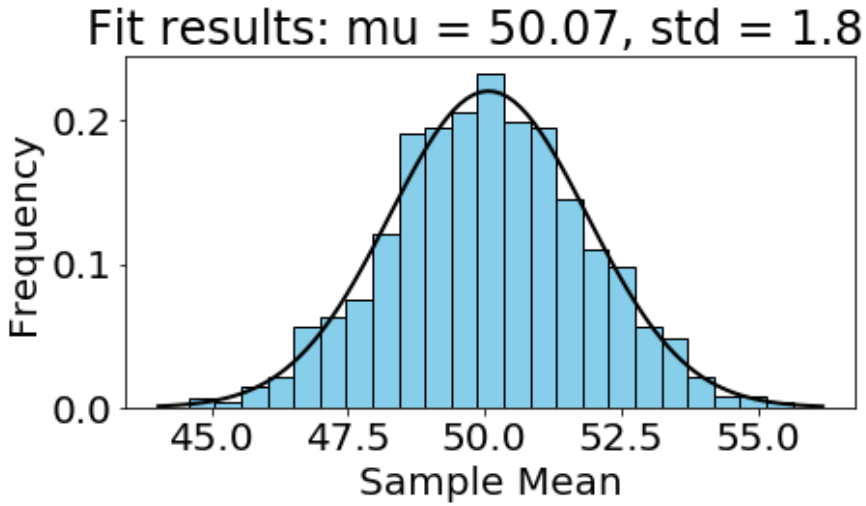
- Select sample size of 30 and number of samples from 10, 100, 1000, 10000

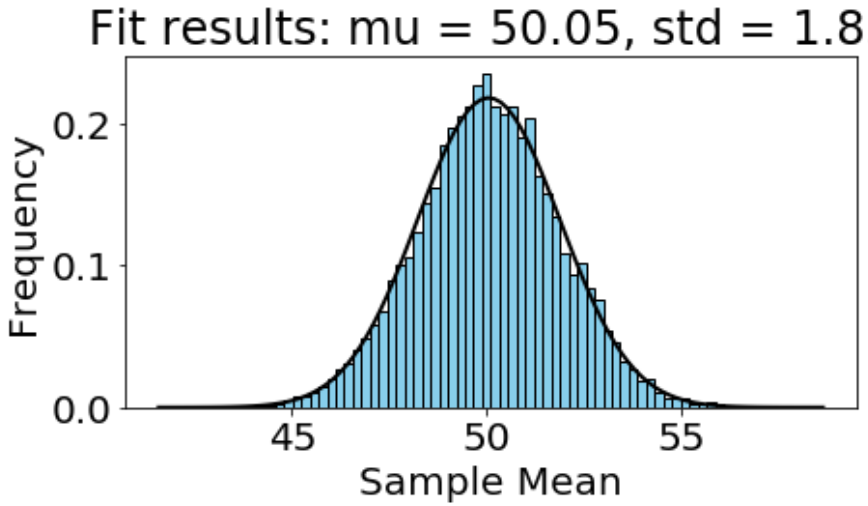


Python test for the Central Limit Theorem



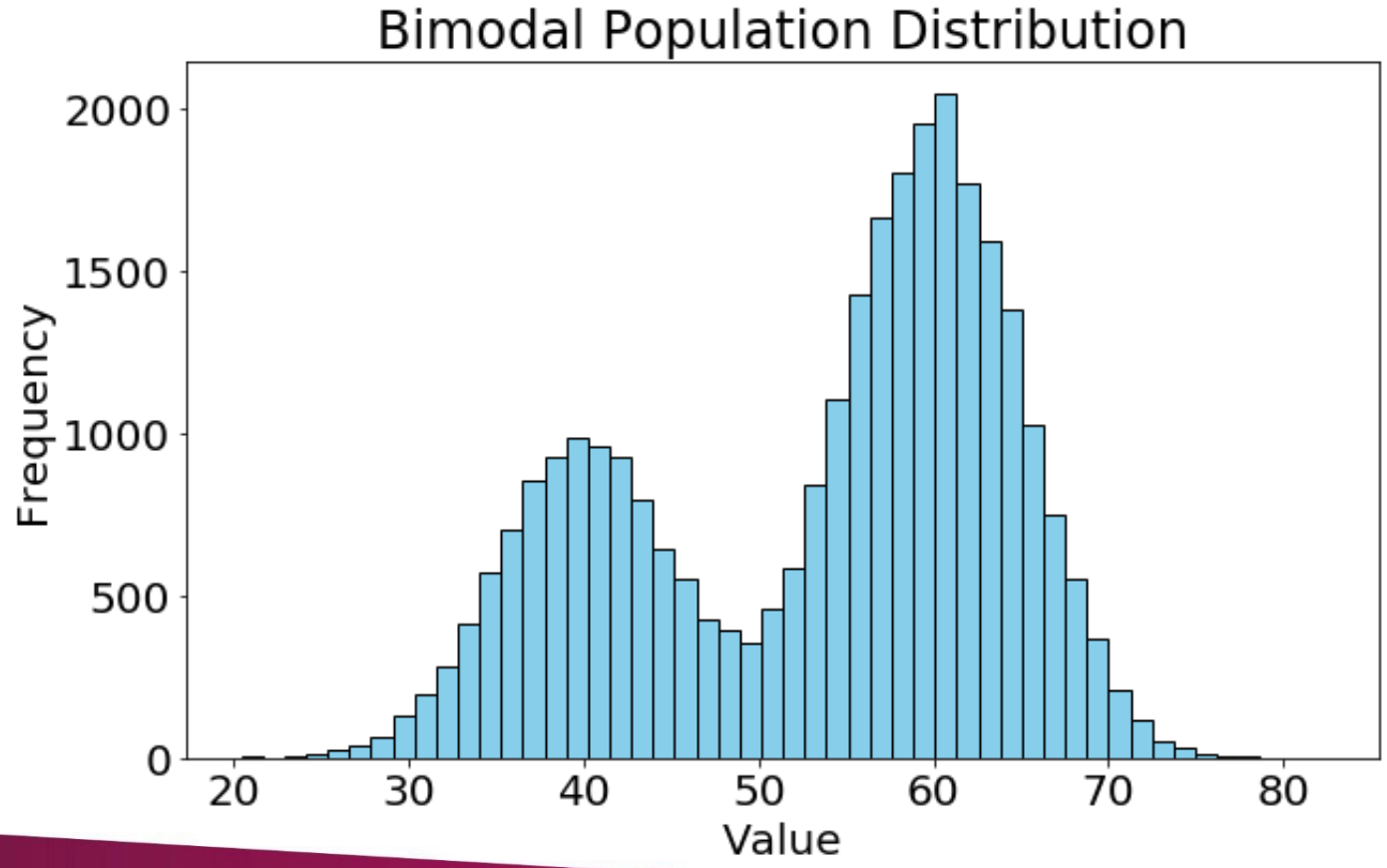




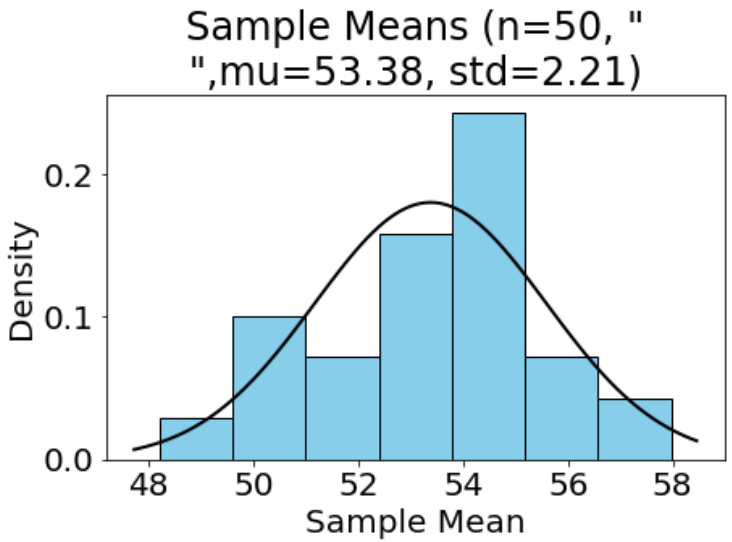


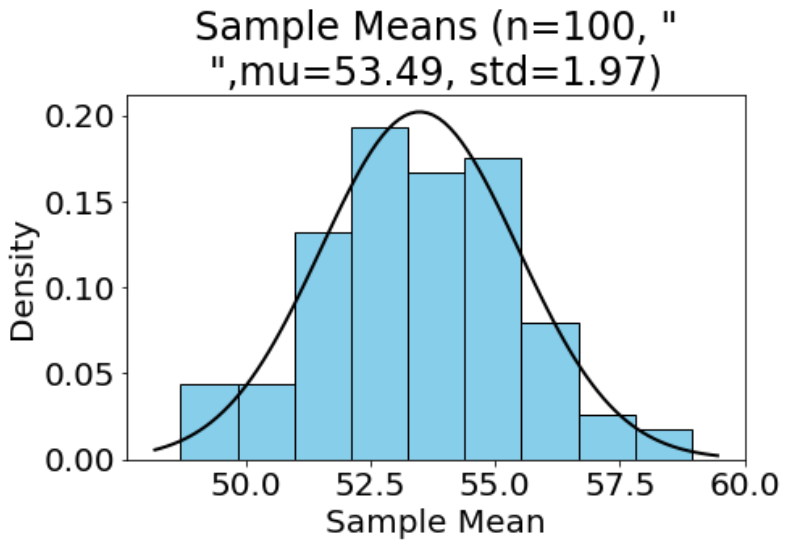
Python test for the Central Limit Theorem

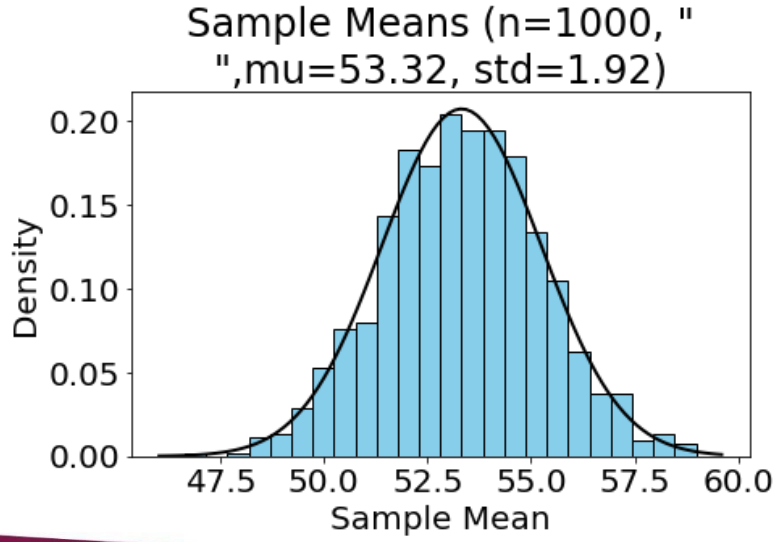
- Generate a population with bimodal distribution
- Select sample size of 30 and number of samples from 50, 100, 1000, 10000

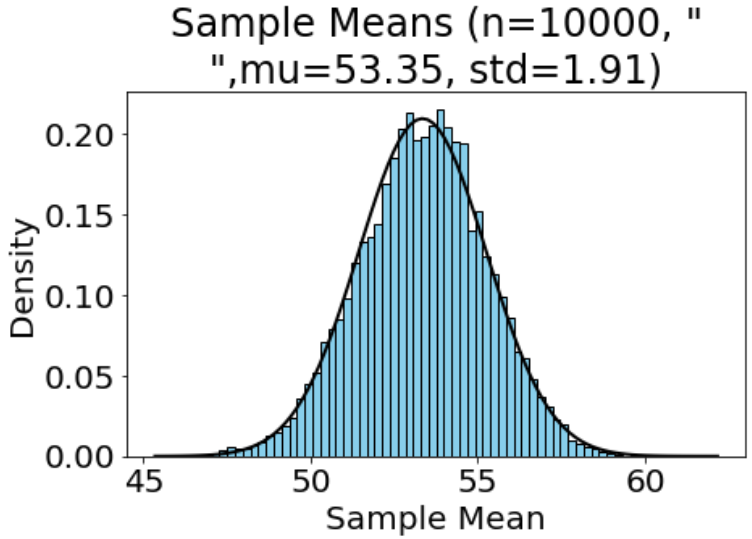


Python test for the Central Limit Theorem



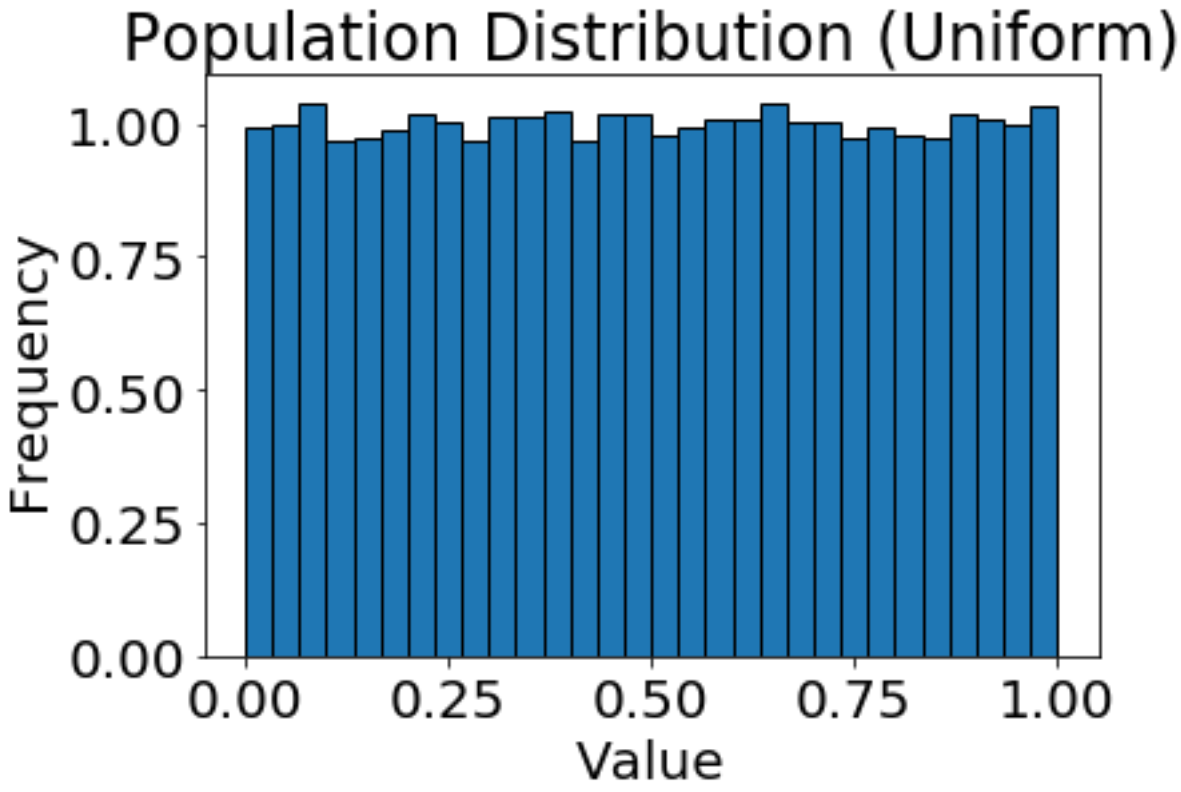






Python test for the Central Limit Theorem

- Generate a uniform distribution
- Select sample size of 30 and number of samples from 50, 100, 1000, 10000



Python test for the Central Limit Theorem