Statistics for Data Science

Inferential Statistics

Agenda – Inferential Statistics

1. Inferential Statistics
2. Some fundamental terms first
    a. Random Variables
    b. Distribution and its types
3. Binomial Distribution
4. Uniform Distribution
5. Normal Distribution
6. Sampling and Inference
    a. Simple Random Samples
    b. Sampling Distribution
    c. Central Limit Theorem
7. Estimation
    a. Point Estimation
    b. Interval Estimation

Descriptive vs. Inferential Statistics

Summaries from data

Summaries give a sense of central tendency

Tell a lot about ‘what’s happening’

Mean, Std Dev, IQRs etc.

Is that enough?

Typically, we work with only ‘samples’ of data

It is not rough to make generalized statement about the population

Challenge - how to learn about population from the sample?

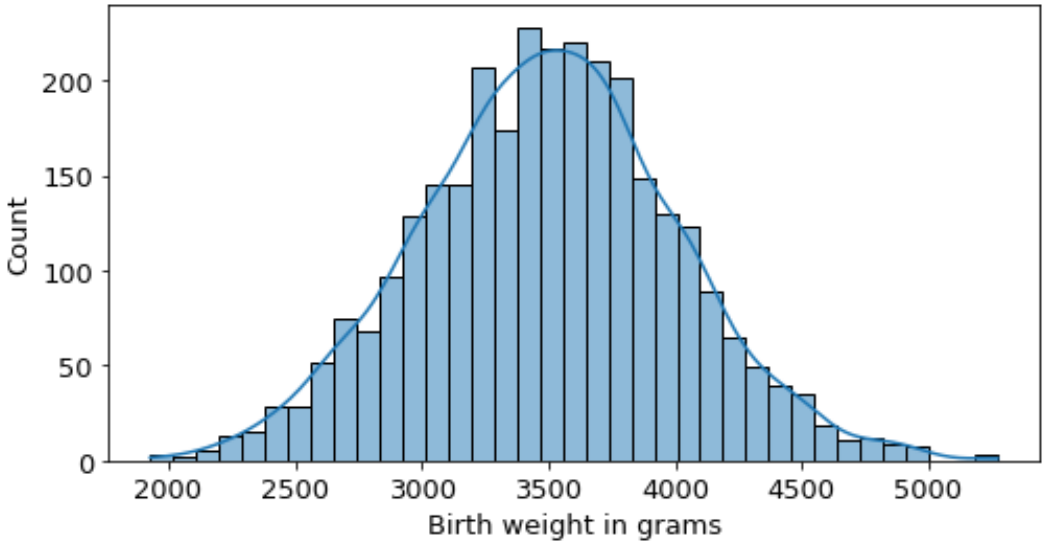
Inferential Statistics

Inferential statistics helps tackle that challenge

There are powerful methods to draw reasonable conclusions about the population from an observed sample

This becomes extremely critical in business decision making

Role of distributions in inferential statistics



Real World Problems

Quality Testing	Is the new manufacturing process better/more reliable than the old process?
Meteorology	What is the likelihood that temp will be more than 20 degree celsius on a specific day?
Human Resources	Does training the workforce improve sales?
Digital Marketing	What is the likelihood that the conversion rate of the website will be above x% next month?
...	...

We will learn about...

1. Some famous distributions & their characteristics
2. Central Limit Theorem - Powerful idea that enables a lot of inferential statistics
3. Estimations and Confidence Intervals
4. Hypothesis Testing

Some fundamental terms first

Random Variable

Suppose there are 1000 students in the university. What is the probability that 500 students will pass the upcoming exam?

There is a 50-50 chance that each student will pass or fail

The total number of students who pass can range from 0 to 1000

A random variable assigns a numerical value to each outcome of an experiment. It assumes different values with different probability.

Discrete Random Variable

You work for an Auto insurance company. Suppose the number of insurance claims filed by a driver in a month is a random variable (X) described by

$$X = \begin{cases} 0, \textit{ with prob } 0.95 \\ 1, \textit{ with prob } 0.04 \\ 2, \textit{ with prob } 0.008 \\ 3, \textit{ with prob } 0.002 \end{cases}$$

All probabilities must be non-negative and sum to 1.

When all possible values the random variable can take can be listed, we call it a **discrete random variable**

Continuous Random Variable

Suppose the volume of soda in a bottle is described by a random variable.

Can we list all possible values?

498 mL, 499 mL, 500 mL, … … … What about 499.2129415 mL?

Sometimes it is just not possible to list all values a random variable can take

If the random variable can take any value in a given range, we call it a **continuous random variable**

Probability Distribution

Probability Distribution

Describes the values that a random variable can take, along with the probabilities of those values

Discrete Probability Distribution

Arises from discrete random variables

Has an associated **probability mass function**, which gives the probability with which the random variable takes a particular value

Continuous Probability Distribution

Arises from continuous random variables

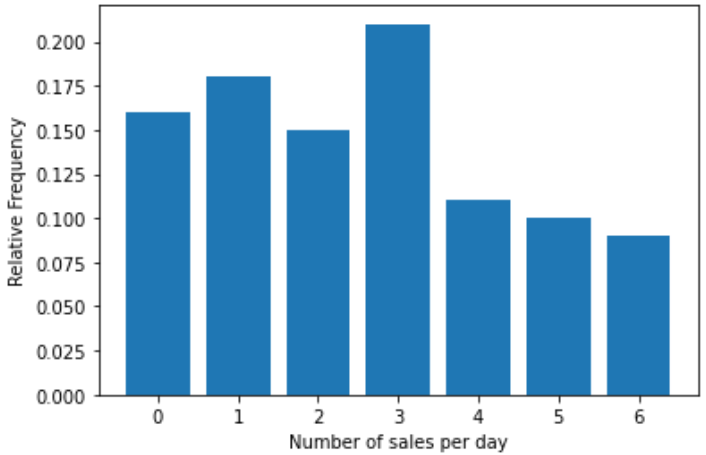
Has an associated **probability density function**, which helps determine the probability with which the random variable lies between two given numbers

Probability Distribution: Example

A company tracks the number of sales new employees make each day during a 100-day probationary period. The results for one new employee are shown. Construct and plot a probability distribution.

Sales	#Days
0	16
1	18
2	15
3	21
4	11
5	10
6	9

Sales	#Days	Relative Frequency
0	16	0.16
1	18	0.18
2	15	0.15
3	21	0.21
4	11	0.11
5	10	0.10
6	9	0.09



Distributions around us (commonly occuring)

Bernoulli	Company has introduced a new drug to cure a disease, it either cures the disease (it's successful) or it doesn't (it's a failure)
Binomial	The number of defective products in a production run.
Uniform	The number of of microwave oven sold daily at a electronic store
Normal	Income distribution of a country : middle-class population is a bit higher than the rich and poor population in the country

Basic distributions - Binomial

Bernoulli Distribution

Success and failure are non-judgemental. Any one outcome may be termed as success

It has only **two possible outcomes**, namely 1 (success) and 0 (failure), of **one single trial**.

$$X = \begin{cases} \textbf{1, with prob } \boldsymbol{p} \\ \textbf{0, with prob } \boldsymbol{1\text{-}p} \end{cases}$$

Very useful in many scenarios:

Manufacturing defective parts	Outcome of medical test

Binomial Distribution

Suppose we ask any adult who uses the app TikTok if he/she has ever posted a video on the app

The answer can be Yes or No (success or failure)

We can use the Bernoulli distribution to model this scenario

Now let us extend this into a survey of 25 adults chosen at random

We can define **a random variable *X* which counts the number of successes**
(say, the number of adults who responded Yes)

Binomial Distribution

In many situations an experiment may have only two outcomes - success and failure

These experiments can be modelled using the Binomial probability distribution.

Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

Probability Mass Function

$$P(X = x) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

Binomial Distribution : Assumptions

Number of trials (n) is fixed.

Each trial is independent of the other trials.

There are only two possible outcomes (success or failure) for each trial.

The probability of a success (p) is the same for each trial.

What happens if these assumptions are violated?

In a month of 30 days, what is the probability that it will rain on more than 10 days, if on average the chance of rain on a given day is 20%?

If we assume that:

1. The event of rain on a particular day is independent of it raining on the previous day.
2. The chance of rain does not increase or decrease over the duration of the month.

Then we can use the binomial distribution with n=30 and p=0.2 to calculate the probability.

Assumptions 1 and 2 in the example are not strictly valid, but they allow for a direct calculation that may be good enough for practical purposes.

Basic distributions - Uniform

Uniform Distribution

Suppose we roll a die. The outcomes of this event can be 1,2,3,4,5,6

https://pixabay.com/vectors/dice-games-play-1294902/

All of the outcomes have an **equal probability of occurrence** and are **mutually exclusive**

We can say that the probabilities of occurrence is **uniformly distributed**.

This is referred to as **Uniform Distribution**

Useful when we are interested in unbiased selection

Uniform Distribution

There are two types of Uniform Distribution

Discrete Uniform Distribution: Can take a finite number (*m*) of values and each value has equal probability of selection.

For example: Number of books sold by a bookseller per day can be uniformly distributed between 100 to 300.

Continuous Uniform Distribution: Can take any value between a specified range.

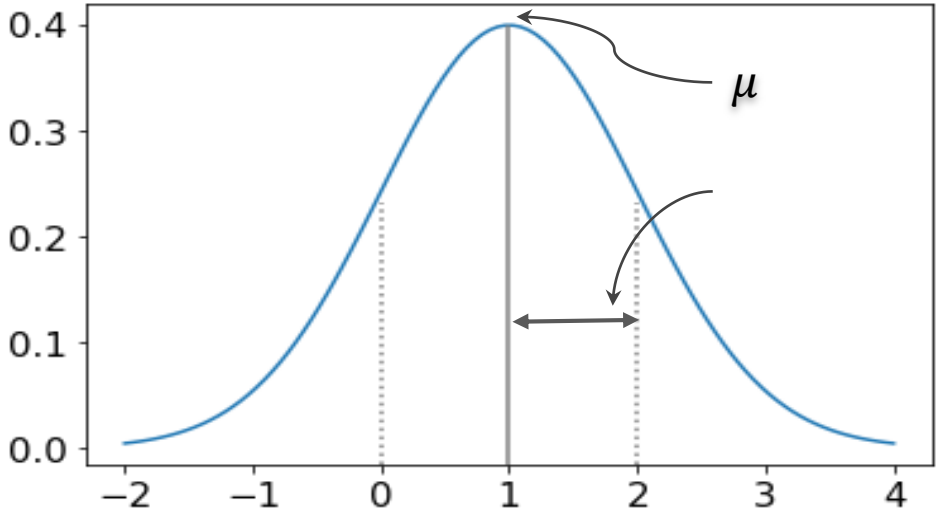
For example: Tomorrow's temperature in United states can be uniformly Distributed between 12 degree Celsius to 17 degree celsius

Basic distributions - Normal

Normal Distribution : Introduction

Normal distribution is the most common and useful continuous distribution

It is characterized by a **symmetric bell-shaped** curve having two parameters - mean (μ)and standard deviation (σ).



Normal Distribution : Why Normal

Why is it called the normal distribution?

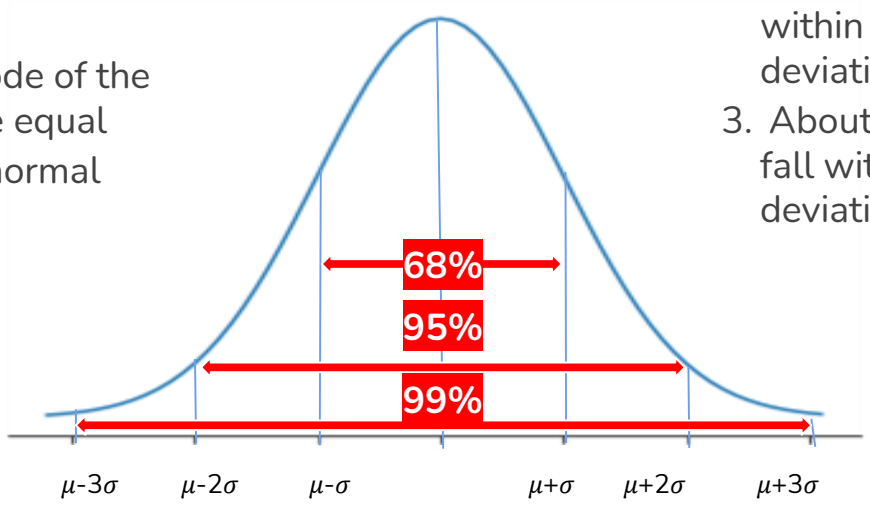
They are commonly found everywhere starting from nature to industry

Many useful datasets are approximately normally distributed

For example the height and weight of the adults, IQ scores, measurement errors, quality control test results etc.

Normal Distribution : Properties

1. The graph of the normal distribution is called the normal curve
2. Normal curve is symmetric around the mean
3. Mean, Median and Mode of the normal distribution are equal
4. Total area under the normal curve is 1



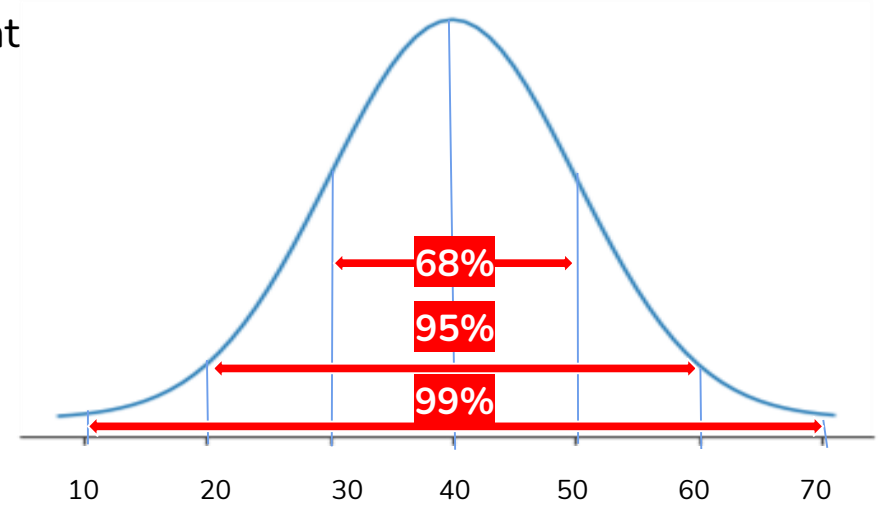
1. About 68% of the data fall within one standard deviation from the mean
2. About 95% of the data fall within two standard deviation from the mean
3. About 99.7% of the data fall within three standard deviation from the mean

Normal Distribution : Example

Assume that a food delivery service provider A has a mean delivery time of 40 minutes and a standard deviation of 10 minutes.

Using the Empirical Rule, we can determine that

- *About 68% of the delivery times are between 30-50 minutes (40 ± 10)*
- *About 95% of the delivery times are between 20-60 minutes (40 ± 2(10))*
- *About 99.7% of the delivery times are between 10-70 minutes (40 ± 3(10))*



This property is known as Empirical rule.

Normal Distribution : Area under Density Curve

As with any continuous probability distribution, the area under the density curve between two points indicates the probability that the variable will fall within that interval.

μ and σ are the parameters that decide the center and spread of the normal curve

To find the area, we need Calculus

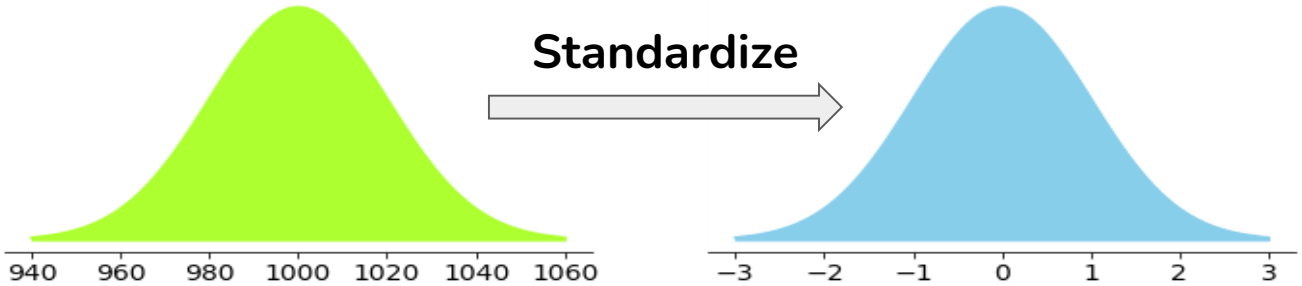
But, there is an easier way to do it in Python (or other softwares).
It provides us the necessary functions to calculate the area.

Normal Distribution : Standard Normal Distribution

A standard normal distribution is used to compare two normal distributions with different parameters (μ, σ)

The standard normal variable is denoted by Z and the distribution is also known as Z distribution

It always has a mean of 0 and standard deviation of 1



Normal Distribution

Standard Normal Distribution

Normal Distribution : Z-Score

A normal variable can be converted to standard normal variable by subtracting the mean (μ) and dividing the standard deviation (σ):

$$Z = \frac{X - \mu}{\sigma}$$

where,

- X is the observed data point
- Z (**Z-Score/Standard score**) is the measure of the number of standard deviations above or below the mean that data point falls

Sampling and Inference

Revisiting the need for sampling..

In many of the situations, what we have available to us is a sample of data.

The data we have is finite.

Till now, the goal was to find ways of describing, summarizing and visualising the sample data only

Moving ahead, we want to make inferences about the “entire” population using the sample data.

Sampling : Simple Random Sampling

A sampling technique where every item in the population has an equal chance of being selected

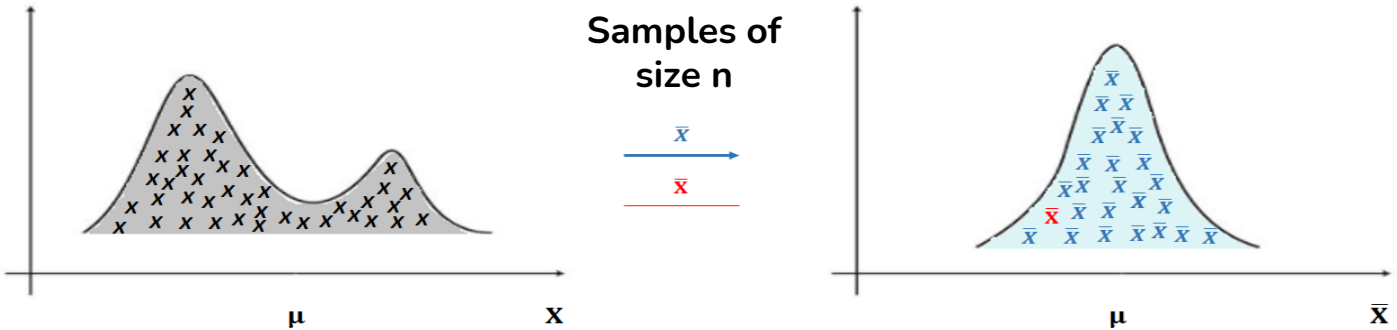
Why are simple random samples important?

Allows all the entities in the population to have an equal chance of being selected and so the sample is likely to be representative of the population

Sampling Distribution

The sampling distribution of a statistic is the probability distribution of that statistic when we draw many samples

For example sampling distribution of the mean, sampling distribution of variance etc.
To a great extent, statistical inference techniques are based on sampling distribution of a statistic



Population Distribution

Sampling Distribution of means

Some Important Points

Suppose we are sampling from a population with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. Let $\bar{X}$ be the random variable representing the sample mean of n independent observations.

The mean of $\bar{X}$ is equal to $\boldsymbol{\mu}$

The standard deviation of $\bar{X}$ is equal to $\boldsymbol{\sigma/\sqrt{n}}$ (Also called the 'standard error' of $\bar{X}$)

Even the population is not normally distributed, then for sufficiently large n $\bar{X}$ is also normally distributed.

Central Limit Theorem

The sampling distribution of the sample means will approach normal distribution as the sample size gets bigger, no matter what the shape of the population distribution is.

Assumptions

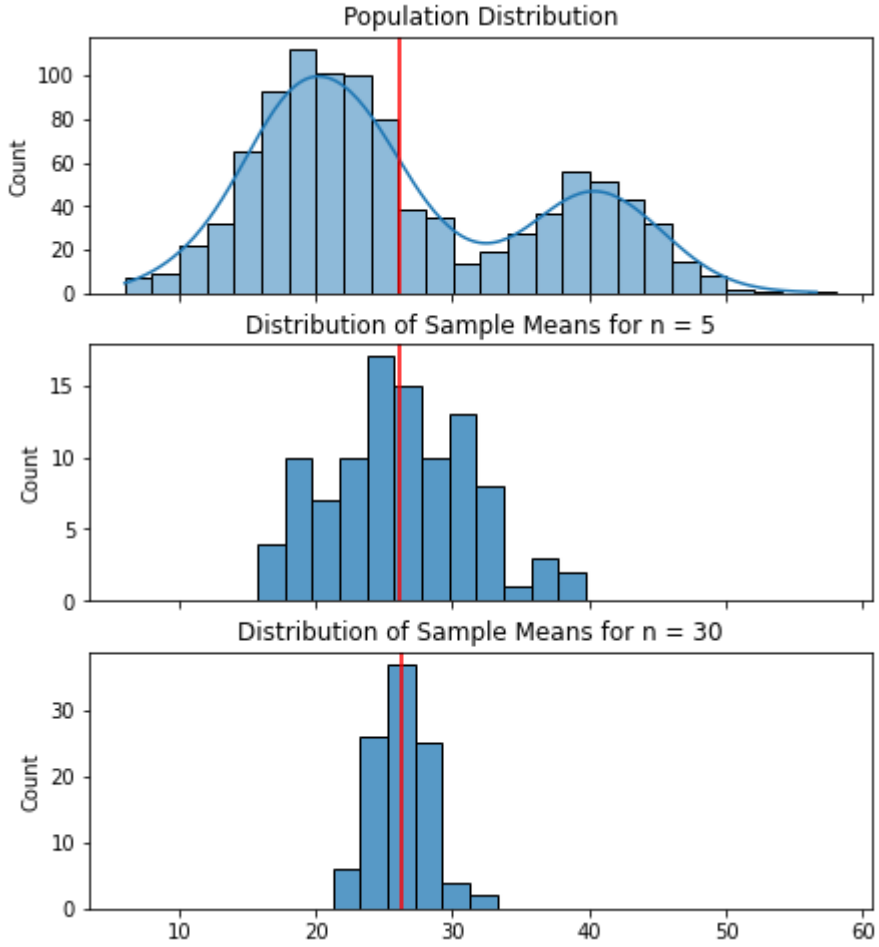
- Data must be **randomly sampled**
- Sample values must be **independent** of each other
- Samples should come from the **same distribution**
- Sample size must be **sufficiently large (≥30)**

Central Limit Theorem

Large sample size provides better estimate of the population mean.

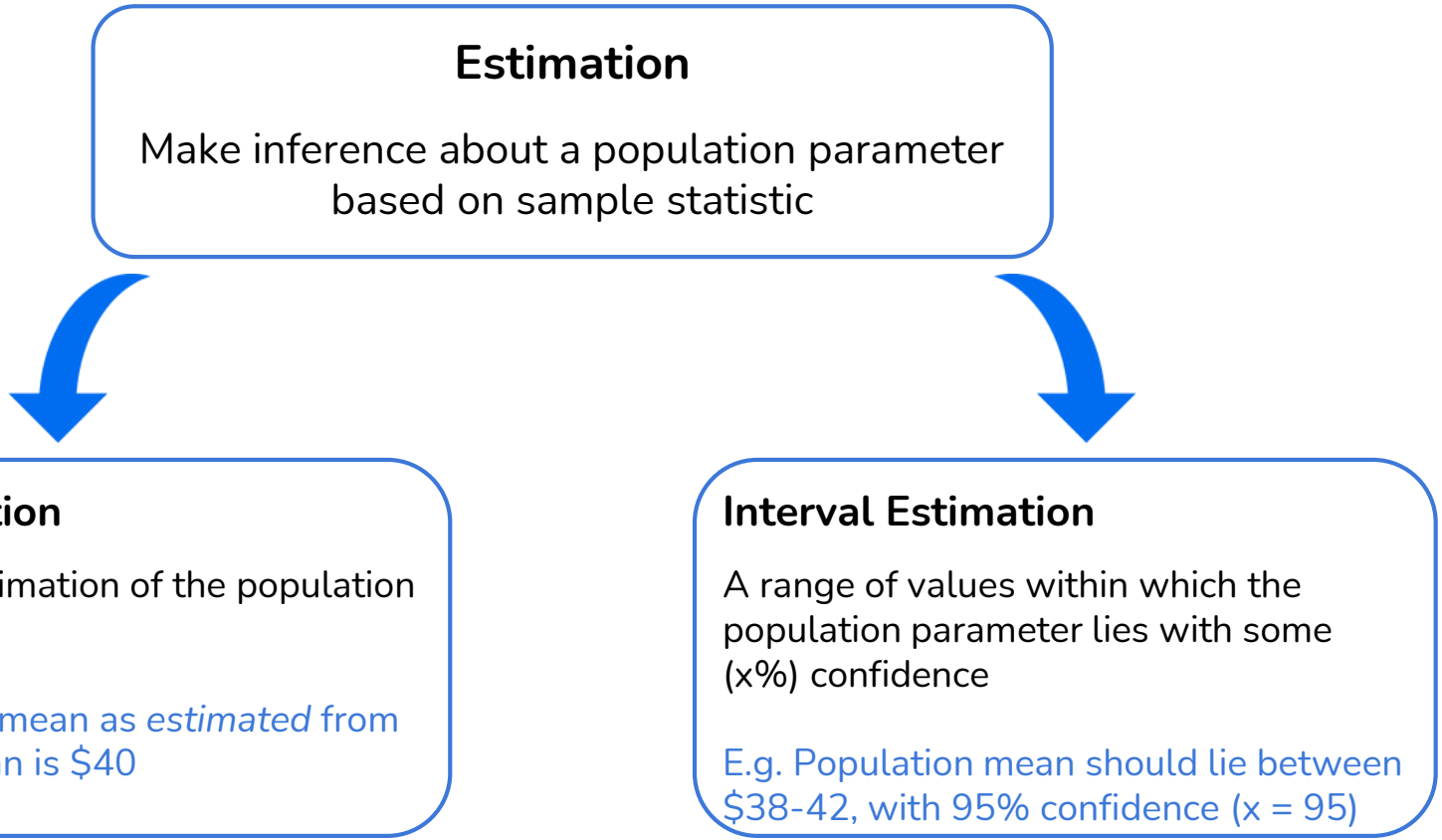
For sample size n = 5, the mean of sample means pile up around the population mean.

For sample size n = 30, the mean of sample means are much closer to the population mean.



Estimations

Estimation


Estimation
Make inference about a population parameter based on sample statistic
Point Estimation
Single point estimation of the population parameter
E.g. Population mean as *estimated* from the sample mean is $40
Interval Estimation
A range of values within which the population parameter lies with some (x%) confidence
E.g. Population mean should lie between $38-42, with 95% confidence (x = 95)

Point Estimation

A point estimate of a population parameter is a single value of a statistic

For example: The sample mean $\bar{X}$ is a point estimate of the population mean μ. Similarly, the sample standard deviation s is a point estimate of the population standard deviation σ.

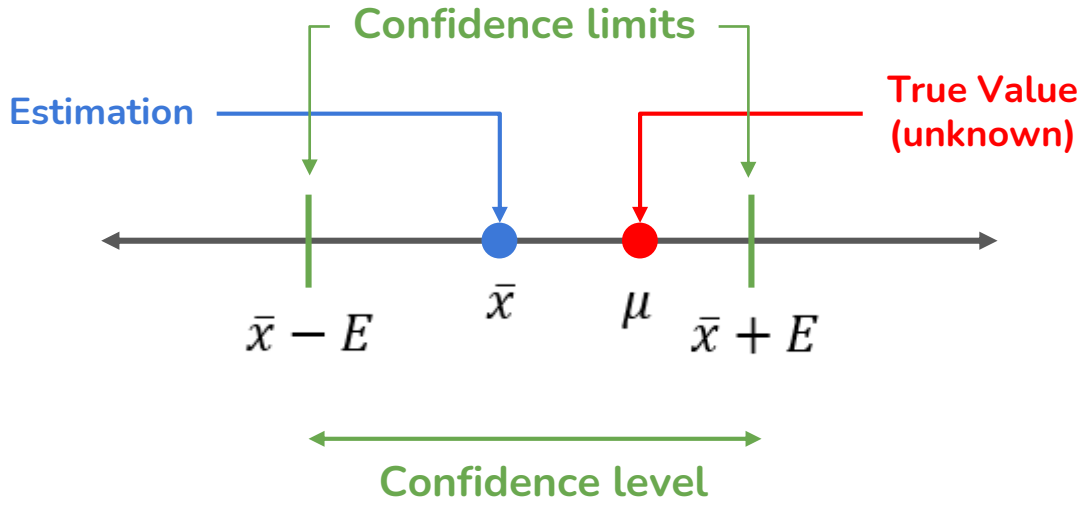
ESTIMATOR how to estimate	PARAMETER what to estimate
$\bar{x}$	μ
s^2	σ^2

Point estimates vary from sample to sample. Often an interval is used to provide a range of values the parameter can take, instead of a single point estimate

Interval estimation - Confidence interval

Confidence interval provides an interval, or a range of values, which is expected to cover the true unknown parameter.

The upper and lower limits of the interval are determined using the distribution of the sample mean and a multiplier which specifies the ‘confidence’



Confidence Interval for Mean μ

Interpretation of 95% Confidence Interval

- *The interpretation of a 95% confidence interval is that, if the process is repeated a large number of times, then the intervals so constructed, will contain the true population parameter 95% of times.*

Why not 100% Confidence Interval?

- *A 100% confidence interval will include All possible values.*
- *Hence there will be no insight into the problem.*