Business Statistics

Agenda - Estimation and Hypothesis Testing - Week 2

1. Sampling and Inference
   a. Simple random samples
   b. Sampling distribution
   c. Central Limit Theorem
2. Estimation
   a. Point estimation
   b. Interval estimation
3. Hypothesis Testing
   a. Introduction
   b. Hypothesis Formulation
4. Basic concepts of Hypothesis Testing
   a. Importance of null
   b. Importance of test statistic
   c. Type I and Type 2 errors
   d. Hypothesis testing template
5. Performing a Hypothesis Test
   a. Some key ideas
   b. Assumptions
   c. Critical point
   d. Rejection region approach
   e. p-value approach
6. One-Tailed and Two-Tailed Tests
7. Confidence Interval and Hypothesis Test

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

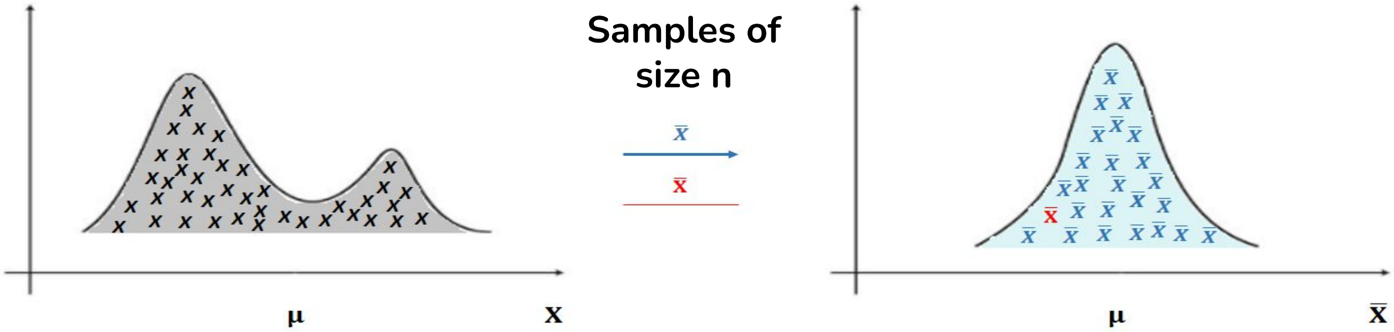
Why are simple random samples important?

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

Sampling Distribution

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

The mean of $\bar{X}$ is equal to μ

The standard deviation of $\bar{X}$ is equal to $\sigma/\sqrt{n}$ (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

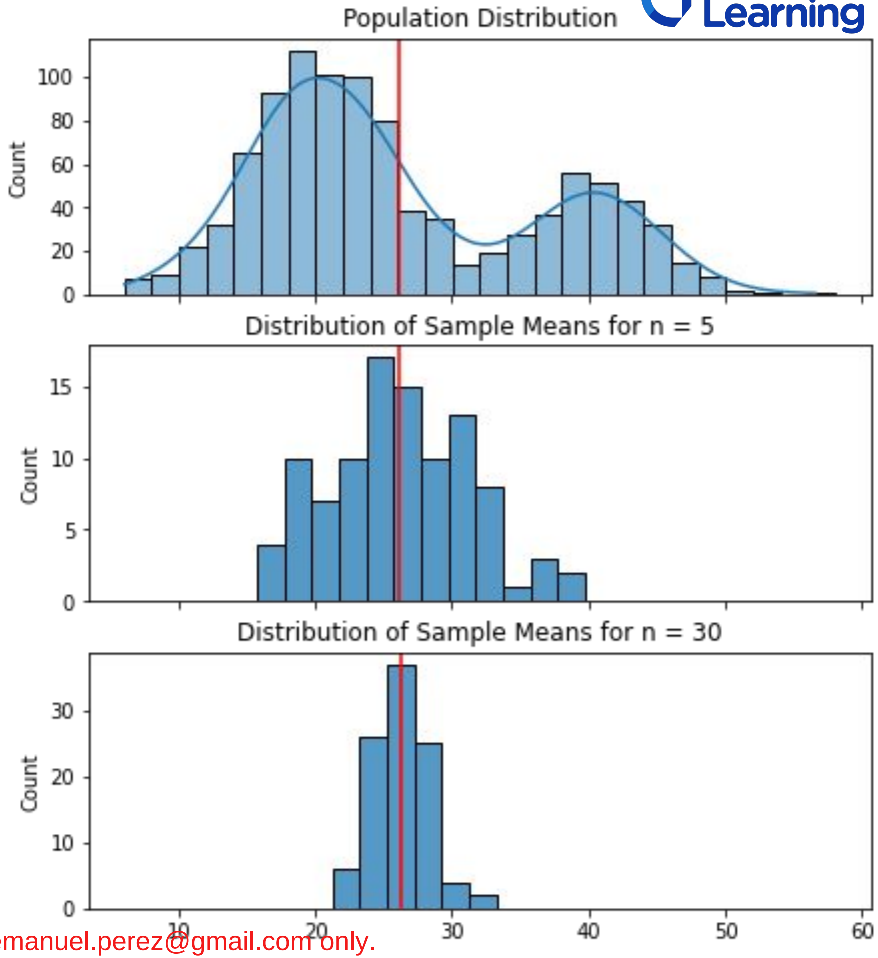
Data must be **randomly sampled**	Sample values must be **independent** of each other
Samples should come from the **same distribution**	Sample size must be **sufficiently large (≥30)**

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.



Estimation

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.

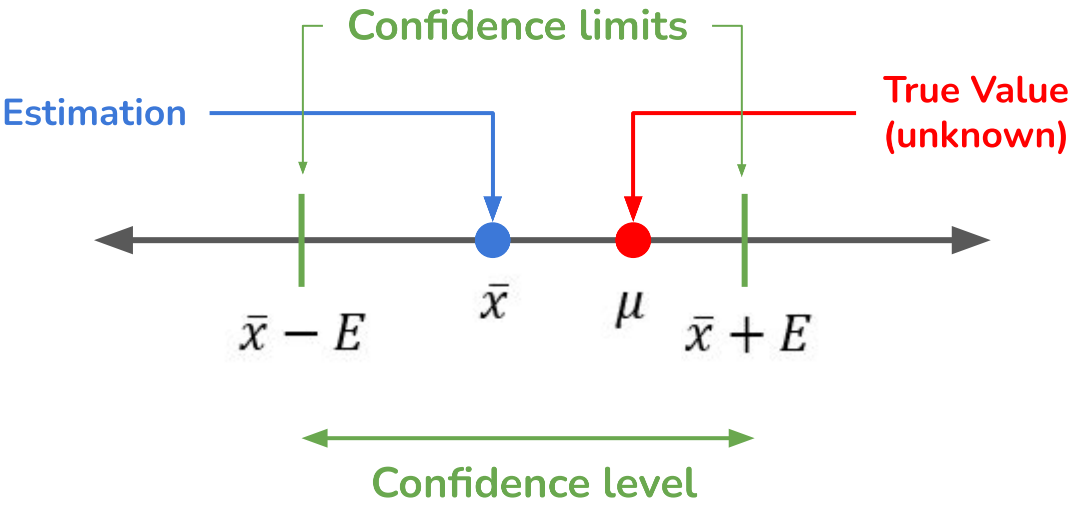
ESTIMATOR how to estimate	PARAMETER what to estimate
$\bar{x}$	μ
s^2	σ^2

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.*

Hypothesis Testing

Real World Problem

Suppose you are a quality analyst at a bulb manufacturing company and analyze the reliability of bulbs. Historically, 70% of the bulbs pass the reliability test.

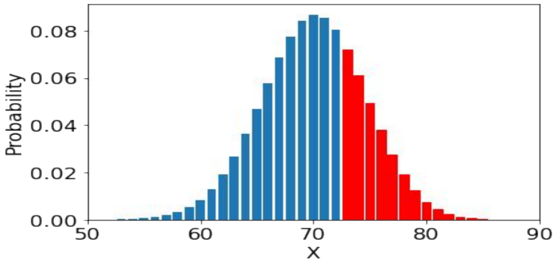
Now, a slightly altered manufacturing process(B) has been introduced to produce the bulbs.

Can you conclude whether the new process improves the reliability of the bulbs or not by checking the number of reliable bulbs in a sample?

Gathering evidence for statistical Inference

We selected a random sample of 100 bulbs out of which 73 are reliable. Does this provide strong evidence that the new manufacturing process is more reliable?

If the new manufacturing process was only as good as the current process - What is the probability of getting 73 or more reliable bulbs in a sample of 100 bulbs?

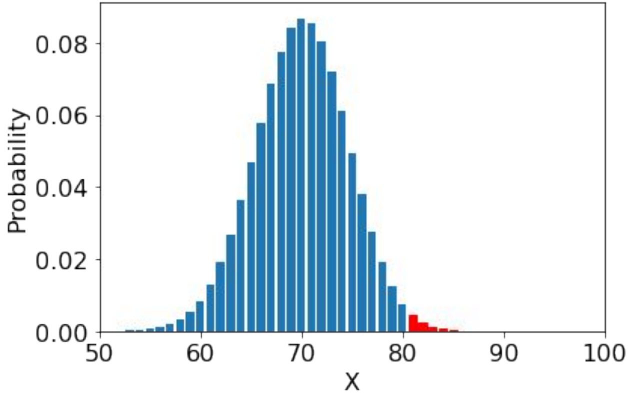


The probability of getting 73 or more reliable bulbs in a sample of 100 bulbs is ~0.30.

Thus, there is no strong evidence that the new process improves reliability

Gathering evidence for statistical Inference

A similar experiment was run with yet another manufacturing process (C). A sample of 100 bulbs produced using this process had 81 reliable bulbs.



The probability of getting 81 or more reliable bulbs in a sample of 100 bulbs is ~0.01.

Thus, there is strong evidence that the new process improves reliability

Why Hypothesis?

Estimation

The problem of estimation is considered, when there is no previous knowledge of the population parameter. The problem is simpler in that case. A random sample is taken, a sample statistic is computed and an appropriate point and interval estimate is suggested.

Hypothesis Testing

Often the interest is not in the numerical value of the point estimate of the parameter, but in knowing the plausibility of a hypothesis about the population parameter by using sample data. Estimation is not enough to arrive at a conclusion in such cases.

What is Hypothesis?

Often we are interested in population parameter(s)

A hypothesis is a conjecture about the population parameter(s)

For example, a bulb manufacturing company is interested in knowing whether the new manufacturing process improves reliability of the bulbs.

The objective of the Hypothesis Testing is to SET a value for the parameter(s) and perform a statistical TEST to see whether that value is tenable in the light of the evidence gathered from the sample.

Overview of Applications

Applications of Hypothesis Testing

- **Testing Research Hypotheses**
  - e.g. a new automobile system increases the mean mpg performance
- **Testing the validity of a claim**
  - e.g. a manufacturer claims that 1L soft drink bottles are filled with an average of at least 0.99L
- **Testing the business decisions**
  - e.g. new online ad has resulted in higher online conversion rates for an E-commerce website

Stating the Hypothesis

Null and Alternative Hypotheses - Two mutually exclusive statements about the population parameter(s)

Null Hypothesis (H_0)

The presumed current state of the matter or status quo.

E.g. The new process for manufacturing bulbs does not improve reliability.

Alternative Hypothesis (H_a)

The rival opinion or research hypothesis or an improvement target.

E.g. The new process for manufacturing bulbs improves reliability.

Null & Alternative Formulation : Example

Mean length of lumber is specified to be 8.5m for a certain building project. A construction engineer wants to make sure that the shipments she received adhere to that specification.

The population parameter about which the hypothesis will be formed is **population mean μ**.

The hypotheses are

$$H_0 : \mu = 8.5$$

$$H_a : \mu \neq 8.5$$

Null & Alternative Formulation : Example

There is a belief that 20% of men on business travel abroad brings a significant other with them. A chain hotel claims that number is too low.

The population parameter about which the hypothesis will be formed is **population proportion π.**

The hypotheses are

$$H_0 : \pi = 0.2$$

$$H_a : \pi > 0.2$$

Tips to formulate Null & Alternative

Am I testing a status quo that already exists?

Null Hypothesis

Negation of the research question

Always contains equality (=, >= , <=)

Am I testing an assumption or claim that is beyond what I know?

Alternate Hypothesis

Research question to be proven

Doesn't contain equality (≠, >, <)

Basic Concepts of Hypothesis Testing

Importance of Null

Null hypothesis is assumed to be true unless reasonably strong evidence to the contrary is found.

Based on a random sample a decision is made whether there exists reasonably strong evidence against the null hypothesis.

Evidence is strong (satisfies the predetermined decision rule) → **Reject the null hypothesis** in favour of alternative hypothesis

Evidence is not strong (does not satisfy the predetermined decision rule) → **Fail to reject the null hypothesis** in favour of alternative hypothesis

Importance of Test Statistic

The test statistic is calculated from the sample data and tested against the predetermined Decision Rule.

The test statistic is a random variable that follows a standard distribution such as Normal, T, F, Chi-square etc. Sometimes the tests are named after the test statistic

Since hypothesis testing is done on the basis of sampling distribution, the decisions made are probabilistic.

Hence, it is very important to understand the errors associated with hypothesis testing.

Type I and Type II Error

Type I and Type II Errors

Level of significance

Power of the test

	H_0 is True	H_0 is False
Reject H_0	**Type I Error** **Prob = α**	**Correct decision** **Prob = 1 - β**
Fail to reject H_0	**Correct decision** **Prob = 1 - α**	**Type II Error** **Prob = β**

Type I and Type II Errors : Example

Null Hypothesis: The patient doesn't have cancer

Alternate Hypothesis: The patient has cancer

- **Type I error (false positive)**: "The patient doesn't have cancer but doctors says she does"
- **Type II error (false negative)**: "The patient does have cancer but report says she doesn't"

Template for Hypothesis Testing

Hypothesis Testing Template

1	Identify the key question	*What is the research question that you are trying to answer?*
2	Establish the hypotheses	*What is the metric of interest? Define the Null and Alternate Hypothesis.*
3	Understand and prepare data	*What data do you have? Do you understand what it means? Can it be used directly?*
4	Identify the right test	*Choose the method for testing based on the last three points*
5	Check the assumptions	*Ensure that data satisfies the assumption for the test.*
6	Perform the test	*Get to conclusion based on the results (p-value)*

Performing a hypothesis test

Some key ideas first

Level of Significance (α)

- Probability of rejecting the null hypothesis when it is true
- Fixed before the hypothesis test.

p-value

- Probability of observing test statistic or more extreme results than the computed test statistic, under the null hypothesis.
- Depends on the sample data. Alpha is pre-fixed but p-value depends on the value of the test statistic

Acceptance or Rejection Region

- The total area under the distribution curve of the test statistic is partitioned into acceptance and rejection region
- Reject the null hypothesis when the test statistic lies in the rejection region, Else we fail to reject it

Let’s start simple

Consider the following questions in hypothesis testing

What are the null and alternative hypotheses?	What is an appropriate test statistic?
What is preset level of significance?	How to check whether the data is giving significant evidence against the null hypothesis or not?

Let’s see an example and understand the significance of the above questions

For simplicity, we will assume that the population standard deviation is known and the sample size is more than 30.

Example

It is known from experience that for a certain E-commerce company the mean delivery time of the products is 5 days with a standard deviation of 1.3 days.

The new customer service manager of the company is afraid that the company is slipping and collects a random sample of 45 orders. The mean delivery time of these samples comes out to be 5.25 days.

Is there enough statistical evidence for the manager's apprehension that the mean delivery time of products is greater than 5 days.

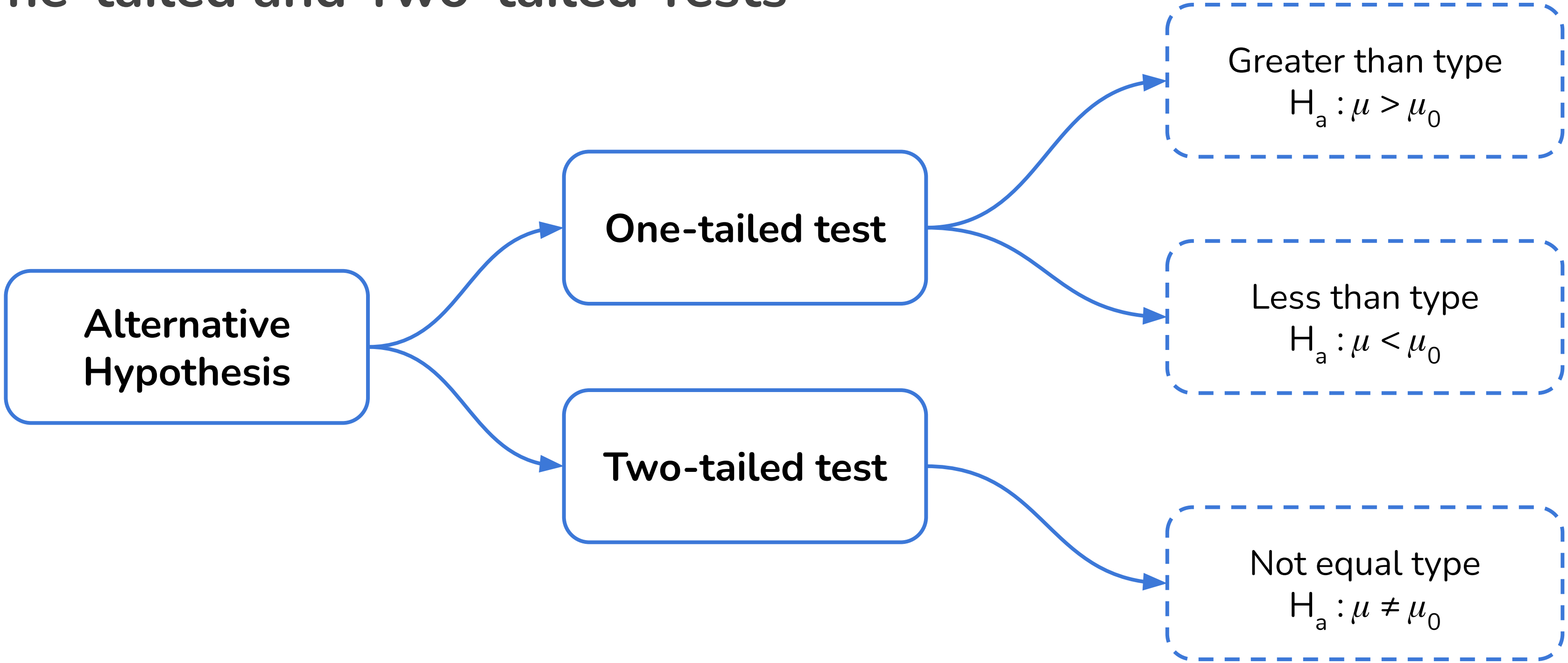
This is clearly a one-tailed test, concerning population mean μ, the mean delivery time of products.

First test - z-test for One Mean

Significance of the test	Assumptions	Test Statistic Distribution
Test for population mean $H_0 : \mu = \mu_0$	• Continuous data • Normally distributed population or sample size > 30 • Known population standard deviation σ • Random sampling from the population	Standard Normal distribution

One-tailed and Two-tailed Tests

One-tailed and Two-tailed Tests

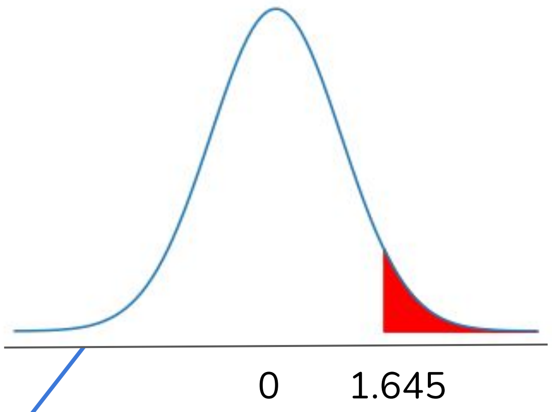


Choice of One tailed vs Two tailed depends on the nature of the problem, not on the sample data!

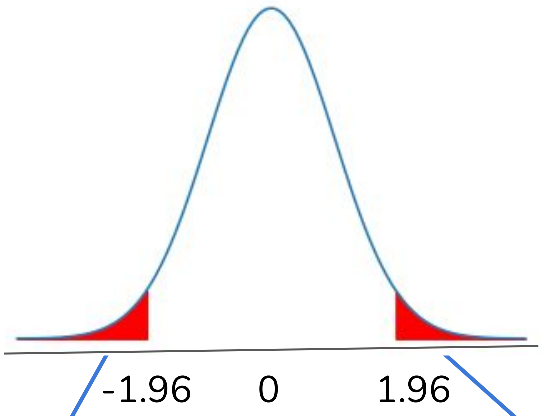
Difference between One-tailed and Two-tailed Tests

Test statistic value **does not change** for two-tailed or one-tailed test.

Only the critical value(s) / p-value associated with the test statistic changes



The difference is not tested on this side and the hypothesis test has greater power on the other side



The difference is tested on both the sides.

Connecting the dots with Confidence Intervals

Confidence Interval vs Hypothesis Testing

Suppose we calculate the (100 - 5)% confidence interval for the mean

We also conduct the Z-test for the mean with a 5% significance level.

The hypotheses of the Z-test are

$$H_0 : \mu = \mu_0 \text{ against } H_a : \mu \neq \mu_0$$

Is there any relationship between the estimated confidence interval and the hypothesis test?

The confidence interval contains all values of μ_0 for which the null hypothesis will not be rejected.