STATISTICS FOR DATA SCIENCE

HYPOTHESIS and INFERENCE

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

UNIT-4 HYPOTHESIS and INFERENCE

Session-1

INTRODUCTION

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Introduction

- There will be many situations in which we need to Take decisions based on random data.
- For example how a companies decide to give advertisement for a TV channel. For a specific time slot.
- You all know it is based on the viewer ship and the channels will vary their advertisement rates based on that ,similarly how many companies can take decisions on the effectiveness of theirs new product that they are developing and decide on product prizing and positioning.
- Even if we take a recent example how various Govt. agencies decide on activities based on the data of covid-19 patients.
- There are statistical tools available to help making effective decisions and hypothesis test is one of such tools and we will learn how different types of data are analyzed by different methods.

Hypothesis Test:

Let us begin with an example:

- **A certain type of automobile engine emits a mean of 100 mg of oxides of nitrogen (NOx) per second at 100 horsepower. A modification to the engine design has been proposed that may reduce NOx emissions. The new design will be put into production if it can be demonstrated that its mean emission rate is less than 100 mg/s.**
- **A sample of 50 modified engines are built and tested. The sample mean NOx emission is 92 mg/s, and the sample standard deviation is 21 mg/s.**

Source: Text Book Chapter6

Hypothesis Test:

Let us begin with an example

- The population in this case consists of the emission rates from the engines that would be built if this modified design is put into production. If there were no uncertainty in the sample mean, then we could conclude that the modification would reduce emissions— from 100 to 92 mg/s. Of course, there is uncertainty in the sample mean. The population mean will actually be somewhat higher or lower than 92.
- The manufacturers are concerned that the modified engines might not reduce emissions at all, that is, that the population mean might be 100 or more.

Source: Text Book Chapter6

Hypothesis Test:

Let us begin with an example

They want to know whether this concern is justified. The question, therefore, is this: Is it plausible that this sample, with its mean of 92, could have come from a population whose mean is 100 or more?

- This is the sort of question that hypothesis tests are designed to address, and we will now construct a hypothesis test to address this question. We have observed a sample with mean 92. There are two possible interpretations of this observation:

Source: Text Book Chapter6

Hypothesis Test:

Let us begin with an example

1. The population mean is actually greater than or equal to 100, and the sample mean is lower than this only because of random variation from the population mean. Thus emissions will not go down if the new design is put into production, and the sample is misleading.
2. The population mean is actually less than 100, and the sample mean reflects this fact. Thus the sample represents a real difference that can be expected if the new design is put into production.

Source: Text Book Chapter6

Hypothesis Test:

Let us begin with an example

- These two explanations have standard names. The first is called the null hypothesis. In most situations, the null hypothesis says that the effect indicated by the sample is due only to random variation between the sample and the population. The second explanation is called the alternate hypothesis. The alternate hypothesis says that the effect indicated by the sample is real, in that it accurately represents the whole population.

Source: Text Book Chapter6

Hypothesis Test:

Let us begin with an example

- In our example, the engine manufacturers are concerned that the null hypothesis might be true. A hypothesis test assigns a quantitative measure to the plausibility of the null hypothesis. After performing a hypothesis test, we will be able to tell the manufacturers, in numerical terms, precisely how valid their concern is.
- To make things more precise, we express everything in symbols. The null hypothesis is denoted H_0. The alternate hypothesis is denoted H_1. As usual, the population mean is denoted μ. We have, therefore, $H_0: \mu \geq 100$ versus $H_1 : \mu < 100$

Source: Text Book Chapter6

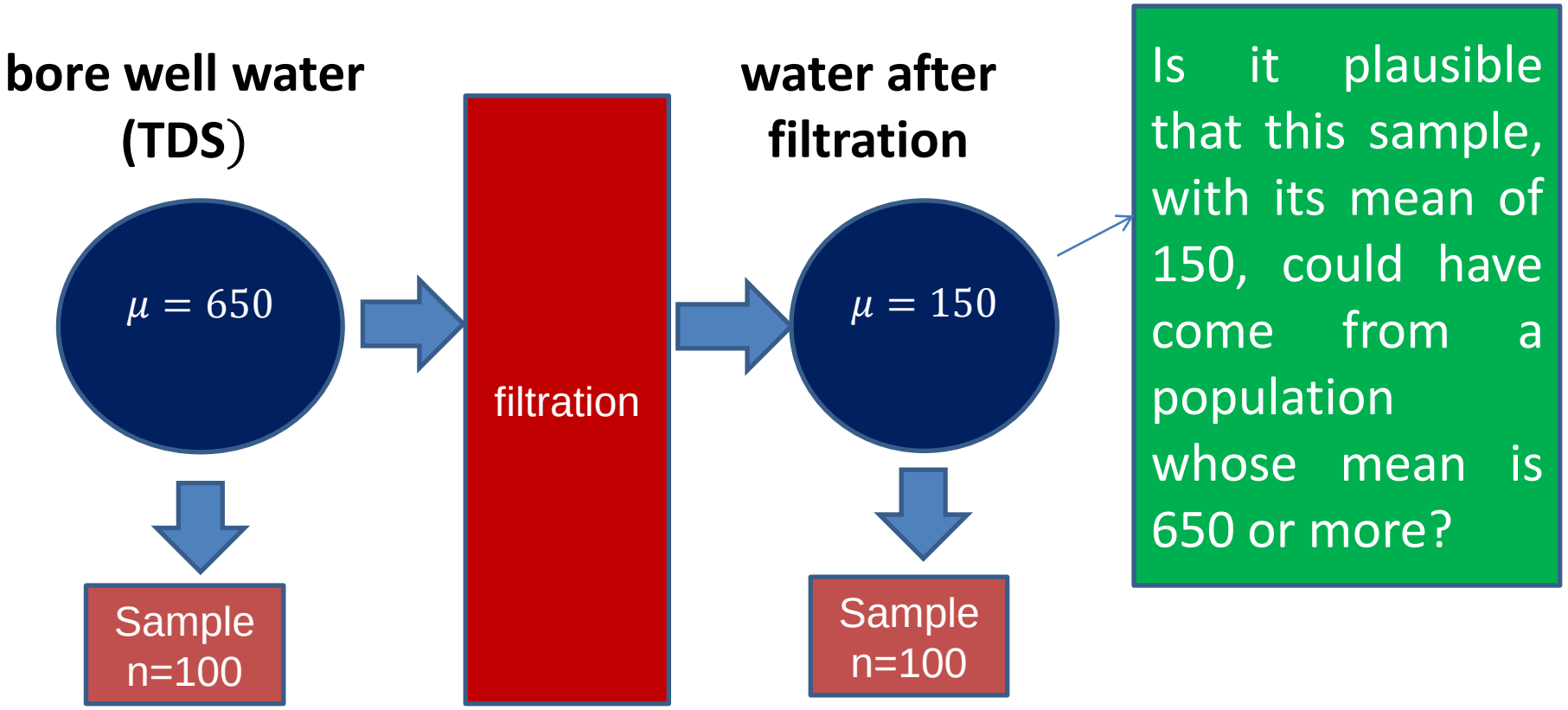
Hypothesis Test:

Let us begin with an example

- Consider another example that a company wanted to lunch a new water filter product that can help provide drinking water. In general any TDS < 600 is considered for drinking and we know there are expensive RO water filters available which can also give a TDS< 100. A sample of 100 bore well water is tested after filtration and the same is described in the diagram

STATISTICS FOR DATA SCIENCE

Introduction



STATISTICS FOR DATA SCIENCE

Introduction

- A hypothesis test produces a number between 0 and 1.
- That measures the degree of certainty we may have in the truth of a hypothesis about a quantity such as a population mean or proportion.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Introduction

- It turns out that hypothesis tests are closely related to confidence intervals.
- In general, whenever a confidence interval can be computed, a hypothesis test can also be performed, and vice versa.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Introduction

- **A Hypothesis is an assumption about population Parameter**

Types of Statistical Hypotheses:

There are two types of statistical hypotheses:

Null hypothesis:

The null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance

Types of Statistical Hypotheses:

Alternative hypothesis:

The alternative hypothesis, denoted by H_1 *or* H_a, is the hypothesis that sample observations are influenced by some non-random cause (investigator's belief).

STATISTICS FOR DATA SCIENCE

Introduction

Example:

Do men and women having different average salaries after graduating University?

Solution:

$$H_0: \mu_M = \mu_F$$

$$H_1: \mu_M \neq \mu_F:$$

Example:

A coin was flipped 50 times, resulting in 40 Heads and 10 Tails.

Solution:

$$H_0: p = 0.5$$

$$H_1: p \neq 0.5$$

Example:

A new type of battery will be installed in heart pacemakers if it can be shown to have a mean lifetime is greater than eight years.

Solution:

$$H_0: \mu \leq 8$$

$$H_1: \mu > 8$$

STATISTICS FOR DATA SCIENCE

Introduction

- **A parameter is the characteristic of the population Like its mean or variance**

- **The parameter must be identified before the analysis.**

I Assume the average weight Of this class is 58 KG

STATISTICS FOR DATA SCIENCE

Introduction

- The best way to determine whether a statistical hypothesis is true would be to examine the entire population.
- Since that is often impractical, researchers typically examine a random sample from the population.
- If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

STATISTICS FOR DATA SCIENCE

Introduction

In performing Hypothesis Test, we essentially put the

null hypothesis on trial.

- We begin by assuming H_0 is true.
- The random sample provides the evidence.
- The Hypothesis test involves measuring the strength of disagreement between the sample and H_0 .

Two methods :

1) Traditional Method : Rejection region approach

2) P-value approach (Used in book – Commonly used)

STATISTICS FOR DATA SCIENCE

Introduction

Hypothesis Testing

Null Hypothesis: H_0

- State the Hypothesized Value of the parameter Before Sampling
- The assumption we wish To test.

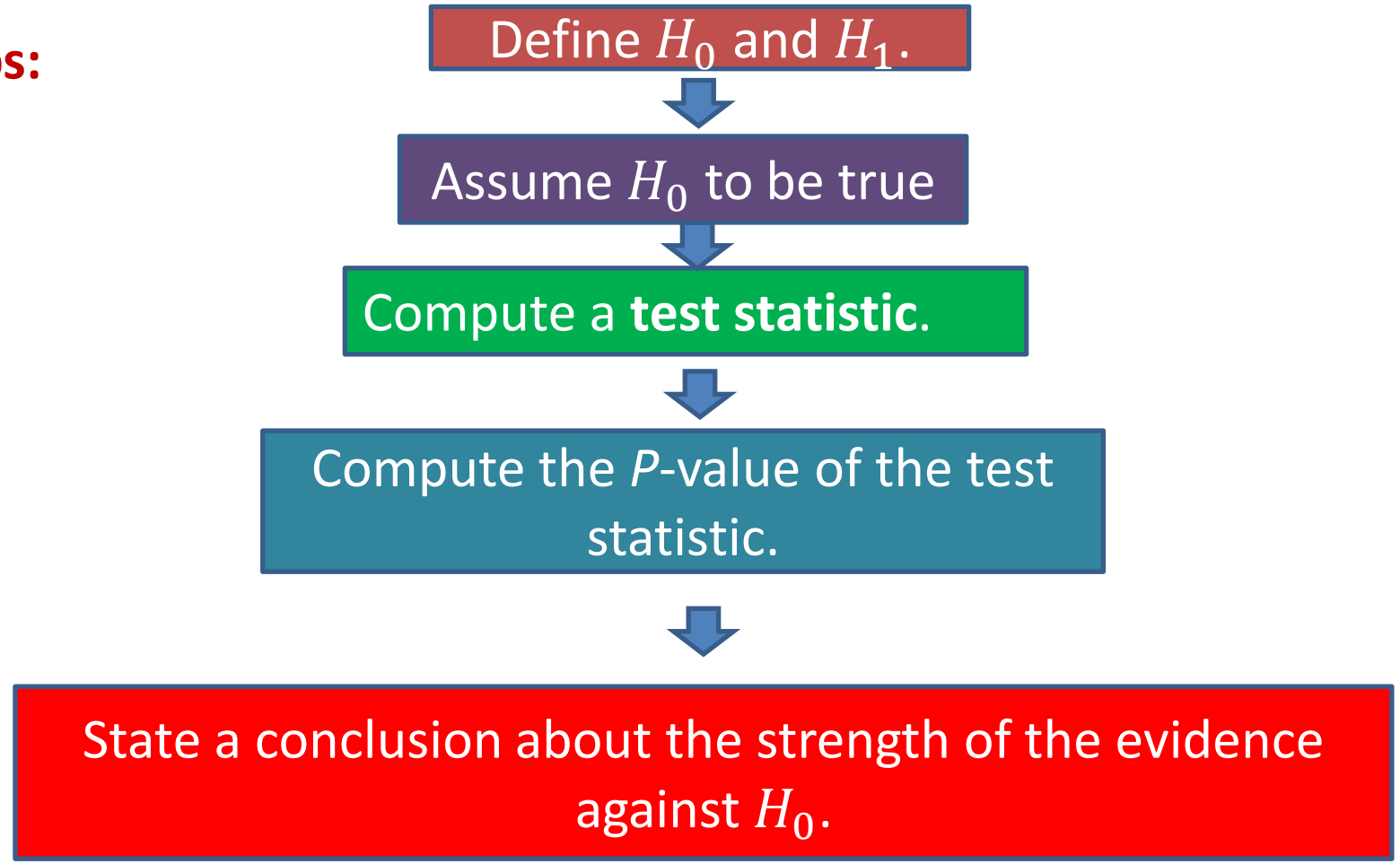
Alternate Hypothesis: H_1

- All [possible alternatives other than the null Hypothesis

STATISTICS FOR DATA SCIENCE

Introduction

Steps:



Dr. Deepa Nair

Department of Science and Humanities

deepanair@pes.edu