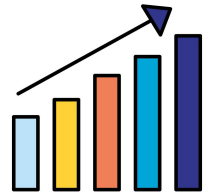
Intro to Applied Statistics and Hypothesis Testing

CSE 407 - Week 8

LEC RAIYAN RAHMAN
Dept of CSE, MIST
raiyan@cse.mist.ac.bd

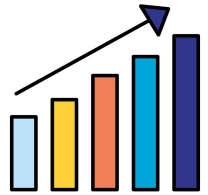
What's Statistics?



- The branch of mathematics that deals with the collection, organization, analysis, interpretation and presentation of data.

- "Statistics" as defined by the American Statistical Association (ASA) "is the science of learning from data, and of measuring, controlling and communicating uncertainty."

What's Statistics?



- A collection of well-documented data
- Methods to properly collect and organize data
- Methods to analyze gathered data
- Methods to interpret that analysis
  - or Study of the techniques and methods used to gather insights from that data.
- Then the presentation of the statistical data and results.
- It's where we study statistical inference, statistical modelling, probability, applied statistics, queuing theory etc.

What's Applied Statistics?

- Use of statistical tools, methods and techniques to analyze and get insights into a particular dataset.
- The root of modern data science and data analysis.
- Computers and Programming are widely used to analyze huge masses of data.

Examples

Data: Say we have the final marks of the 84 students of CSE 407. The average mark is 68.5 for CSE 17.

Question: How likely is it that the average mark of CSE 407 for CSE 18 will be within 65-70?

Data: You're working as maintenance engineer at Samsung. The new SSDs that they are making have an advertised read speed of 550mb/s. But after purchasing some users are reporting much lower read speeds at an avg of 535mb/s.

Question: Is the advertised speed by Samsung wrong? Do you need to revise it?

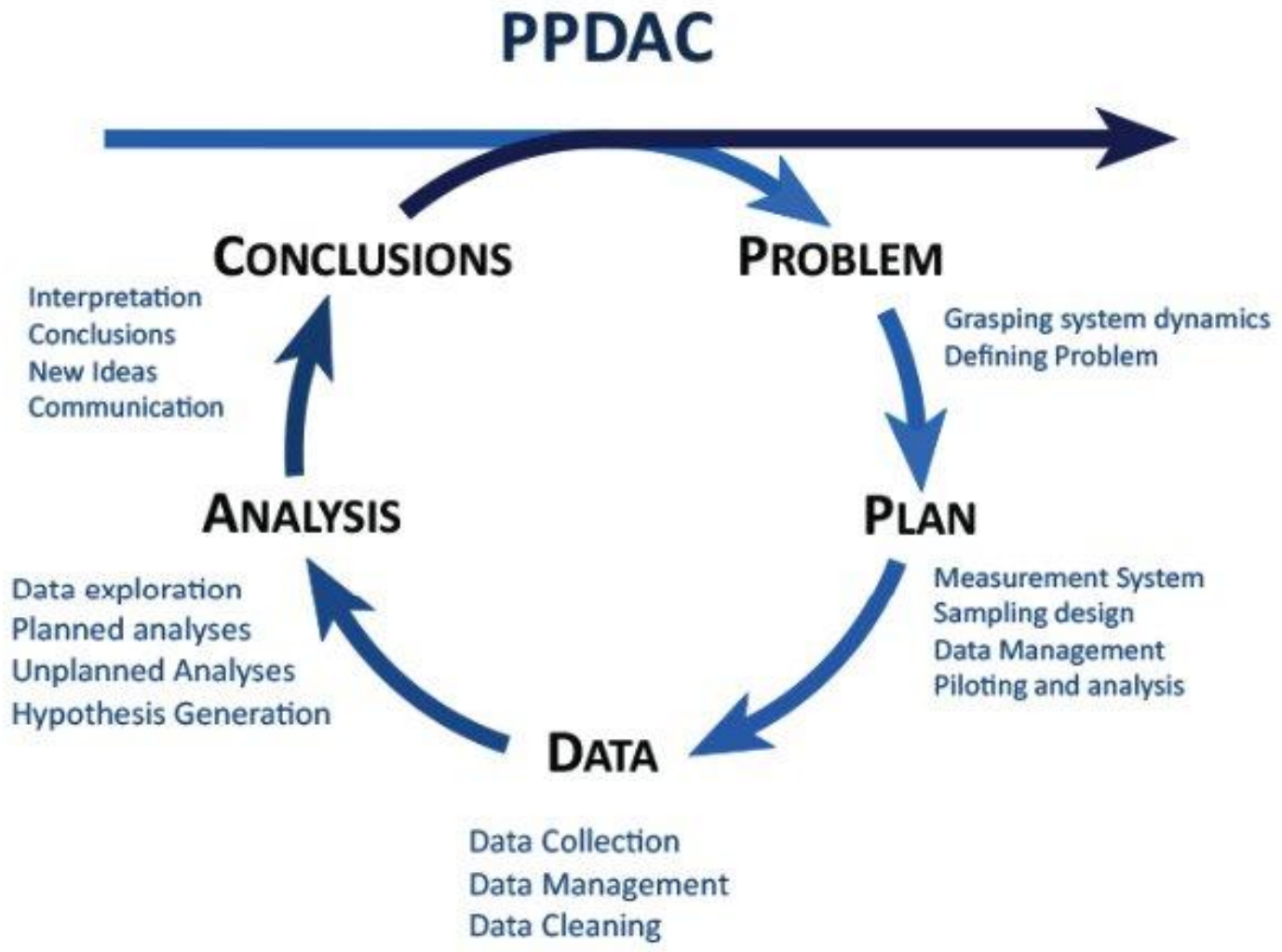
(this is what we'll study today)

Data: Say that Robi is targeting new 4GB,7 days internet packs at a value of BDT 99. They want to pick the right target group to advertise these pack to. From historical data, they know people who bought >2GB, 7 days pack are 70% likely to buy the new pack and those who bought >1GB packs are 50% likely to buy it?

Question: What should be ratio of this ad's targeted user group?

All of these can be answered with Applied Statistics.

The Investigative Statistical Cycle (PPDAC)



But first, Let’s get oriented with some terms and concepts of statistics.

Population

- The entire possible dataset (that i ***should*** base my decisions on).
- But more often than not, the entire dataset is not available.
- For example, Let's say Samsung already manufactured and shipped 2.5M SSDs. Now, to test the customer's claim (and to clear samsung's name!) I can't call back all 2.5M SSDs to test their actual speed again.
- However, in case of the internet packs, the telecom operator should have the entire dataset. So we should be able make decisions more accurately (depending on how good my model is).
- So although we ***should***, we ***can't*** always analyze the entire dataset.

Sub-Population

- Classification of the entire possible dataset based on some criteria.
- For example, in the first scenario, if i was calculating the average height of CSE 17 students instead of marks, it might be a good idea to divide the population of 84 into two sub-population, namely Male and Female, since height are more closely matched within that sub group.
- however, for marks, no such distinction (hopefully) will be there and no need to divide into sub-populations.

Sample

- A subset of the population.
- For example, since we can't call back all 2.5M SSDs. We'll do the speed testing on 1000 SSD cards ready to be shipped. We'll reject the customer's claim (or revise our claim/tweak production) based on the result from this "sample" of 1000 SSDs.
- In most production cases, this is more often the statistical scenario.

Population vs Sample Mean

- Mean is simply the average value of a particular parameter for a sample or population.
- We divide the sum of all values of a variable by the total number of available data.

Population vs Sample Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

Note: notice the different symbols used to denote the different means.

Population vs Sample Variance and Std Deviation

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n}$$

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X_{avg})^2}{n-1}}$$

Note: Variance is a measure of how varied the values are from the calculated mean.

Hypothesis Testing

Hypothesis Testing

- Let’s go back to the Samsung SSD example.
- We have contradictory claims by each group.
- How to resolve? The concept of hypothesis testing.

Testing A Hypothesis - Scenario

Data: You're working as maintenance engineer at Samsung. The new SSDs that they are making have an advertised read speed of 550mb/s. But after purchasing some users are reporting much lower read speeds with different values.

To verify their claims, you decided to gather 100 SSDs ready to be shipped to stores from your production line. After careful testing, the read speed of these test SSDs were found to be 548.9 mb/s. You know from previous tests that the population variance of your SSDs is 6.5 mb/s.

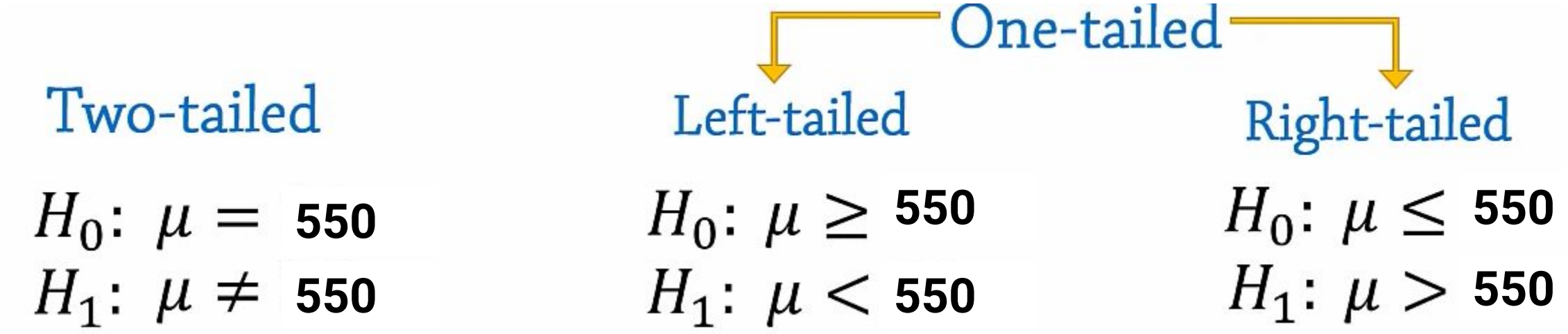
Question: Test the claim by the customers with 95% confidence and establish if the advertised read speed of 550 mb/s is acceptable or not.

Null and Alt Hypothesis

1. Null hypothesis (H_0):
   a. It's the currently claimed or accepted value of a parameter.
   b. In our example, 550mb/s is the Null hypothesis.

2. Alternate Hypothesis (H_1 or H_a):
   a. It's the claim different than the current claim.
   b. May be the exact opposite of a range. but never equal to the H_0 value.

Null and Alt Hypothesis



This is meaningless for our case.

Possible Outcomes for H_0

1. Accepted or “fail to reject”: We couldn’t gather enough evidence from our sample dataset to reject the Null hypothesis. Hence, we accepted it.
2. Rejected: We concluded from our sample dataset that the original claim (H_0) was wrong.

(correction from class)

Level of Confidence

- However, we are testing on a sample dataset.
- So We can't be that rigid.
- We need to give some leeway (instead of a value, we'll test for a range).
- Hence we have the concept of "level of confidence" (C).
- The smaller the sample dataset the higher we should set the val for C.
- α (next slide) should be smaller and C should be higher for a smaller sample size. Since a higher val for C gives us a wider acceptance region and that is what we want for smaller sample sizes (we shouldn't be too rigid if n is small).

Level of Significance

- The complement of C.
- Denoted by α
- $\alpha = 1 - C$
- This defines the critical values or the “area of rejection”.

- $\alpha = 1 - 0.95 = 0.05$

Null and Alt Hypothesis

One-tailed

Two-tailed

$H_0: \mu = 550$

$H_1: \mu \neq 550$

Left-tailed

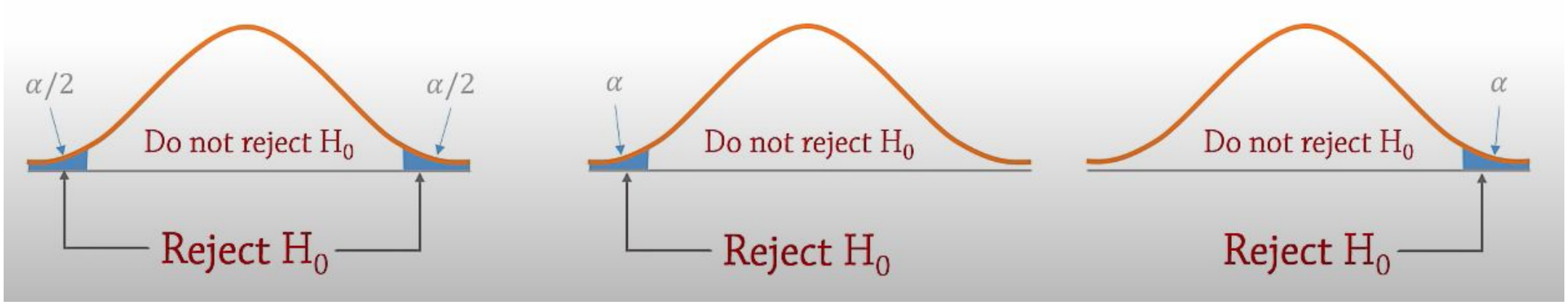
$H_0: \mu \geq 550$

$H_1: \mu < 550$

Right-tailed

$H_0: \mu \leq 550$

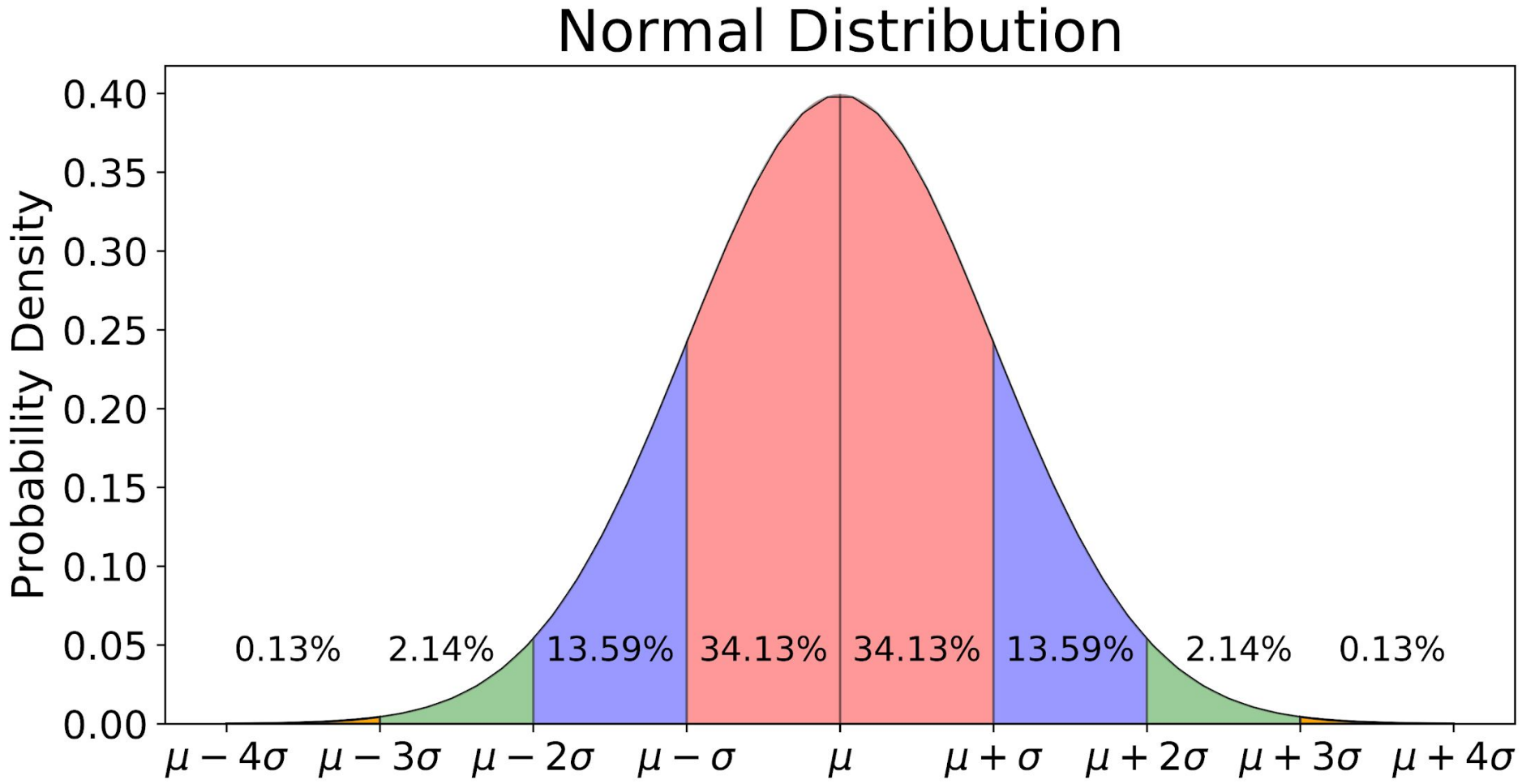
$H_1: \mu > 550$



Prelude: Normal Distribution

- Most real life means are distributed this way.
- We have a very specific graph for standard normal distribution (avg=0 and std deviance=1).
- We'll use this to test the hypothesis assuming the data is normally distributed.
- Let's see what it looks like.

Prelude: Normal Distribution



Thank You!