Hypothesis Testing

Hey I'm,

Pararawendy Indarjo

I am a,

- CURRENTLY | **Senior DM at AlloFresh**
- 20 - Jul 22 | **Senior DS at Bukalapak**
- 19 - 20 | **Data Analyst at Eureka.ai**

BSc Mathematics

Universiteit Leiden

MSc Mathematics

Linkedin :
https://www.linkedin.com/in/pararawendy-indarjo/
Blog : medium.com/@pararawendy19

Class Outline

1. Probability Distribution
2. Introduction to Hypothesis Testing
3. T-test
4. One-way ANOVA

Probability Distribution

What is probability distribution

Probability distribution is a statistical function that describes the **probability** of obtaining all possible values that a **random variable** can take.

Random variable: variable that can take specific values, whose probabilities follow a certain probability distribution.

Probability distribution:
- Discrete variable: Probability **mass** function
  E.g. coin flip, dice, etc
- Continuous variable: Probability **density** function
  E.g. Height, weight, temperature, etc

Probability Mass Function

Probability mass function (PMF): probability distribution for discrete variables

$p(x)$ = the probability that a random variable takes a specific value equal to x.

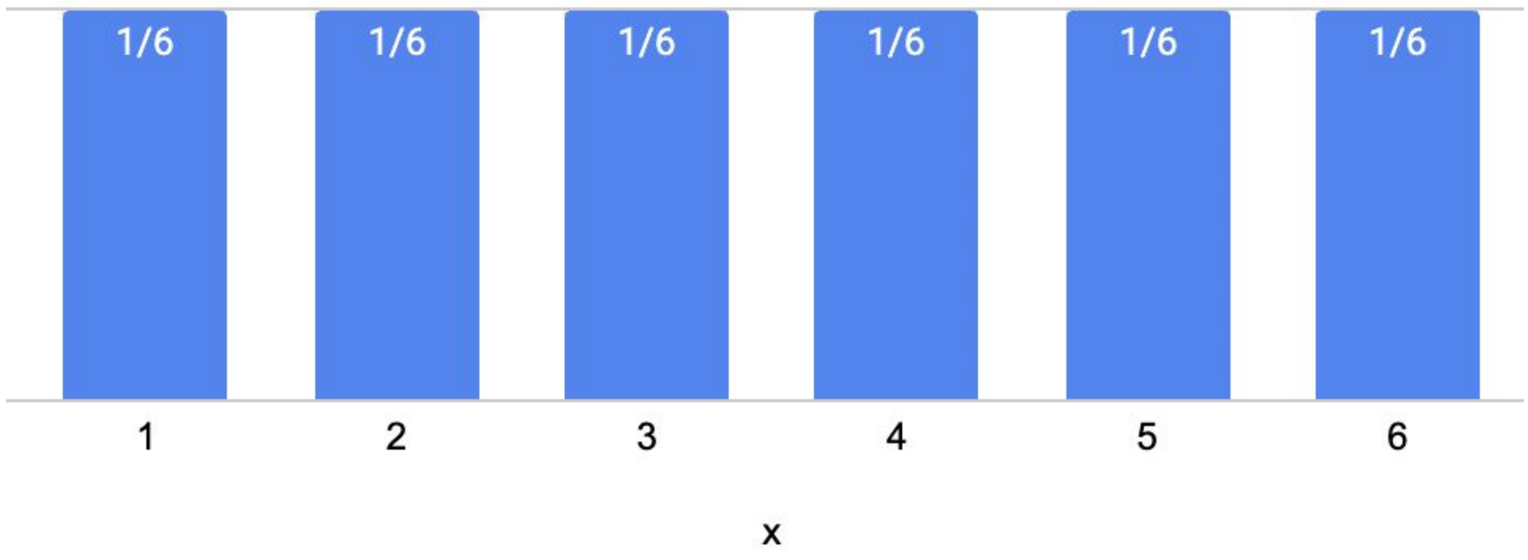
PMF Properties:

- $p(x) >= 0$ for all x
- sum of $p(x)$ for all x is 1

Sample PMF: dice distribution

The probability of rolling a specific number on a dice is 1/6

So, $p(x) = ⅙$, for x in (1,2,3,4,5,6)



Probability Density Function

If the variable can take an **infinite number of values** between any two values, then we have Continuous variables

E.g. height, weight, temperature, sales amount

Continuous variables also have probability distribution function p(x)

Here p(x) is only applicable when x is range/interval of values

E.g. $p(5<x<10)$ is defined as **area under the distribution plot** where $5<x<10$

In other words, $p(x=8)$ [or any number] is equal to zero (because it's only a point–not range)

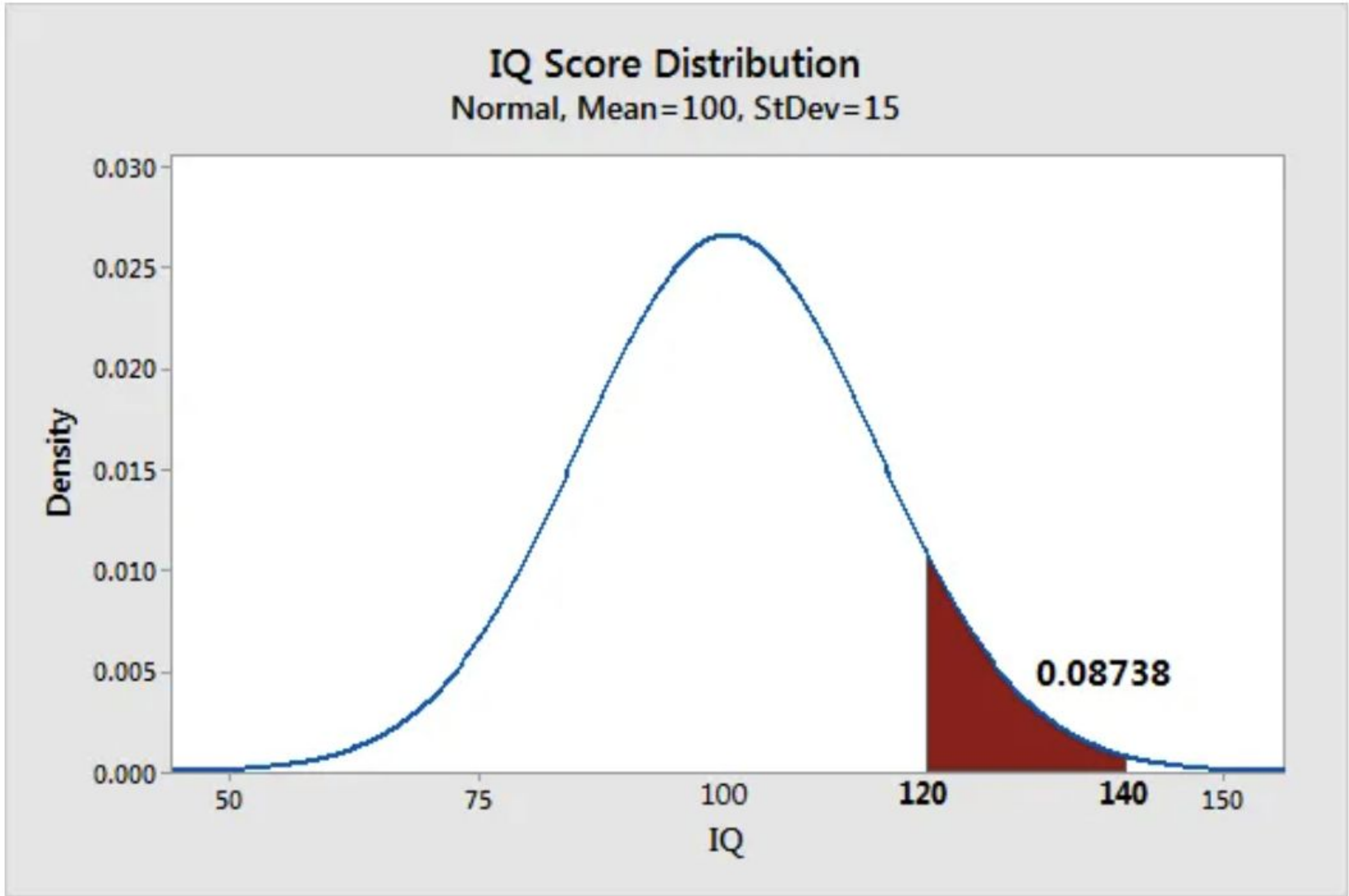
On the distribution plot, the entire area under the curve equals 1.

Also, the probability for a particular value or range of values must be between 0 and 1.

Probability Density Function

Given IQ follows Normal distribution with mean = 100 and std = 15.

- P(-inf < IQ < inf) = 1
  - This is whole area under the PDF
- P(120 <= IQ <= 140) = 0.0874
- P(IQ = 100) = 0
  - Single point probability is zero

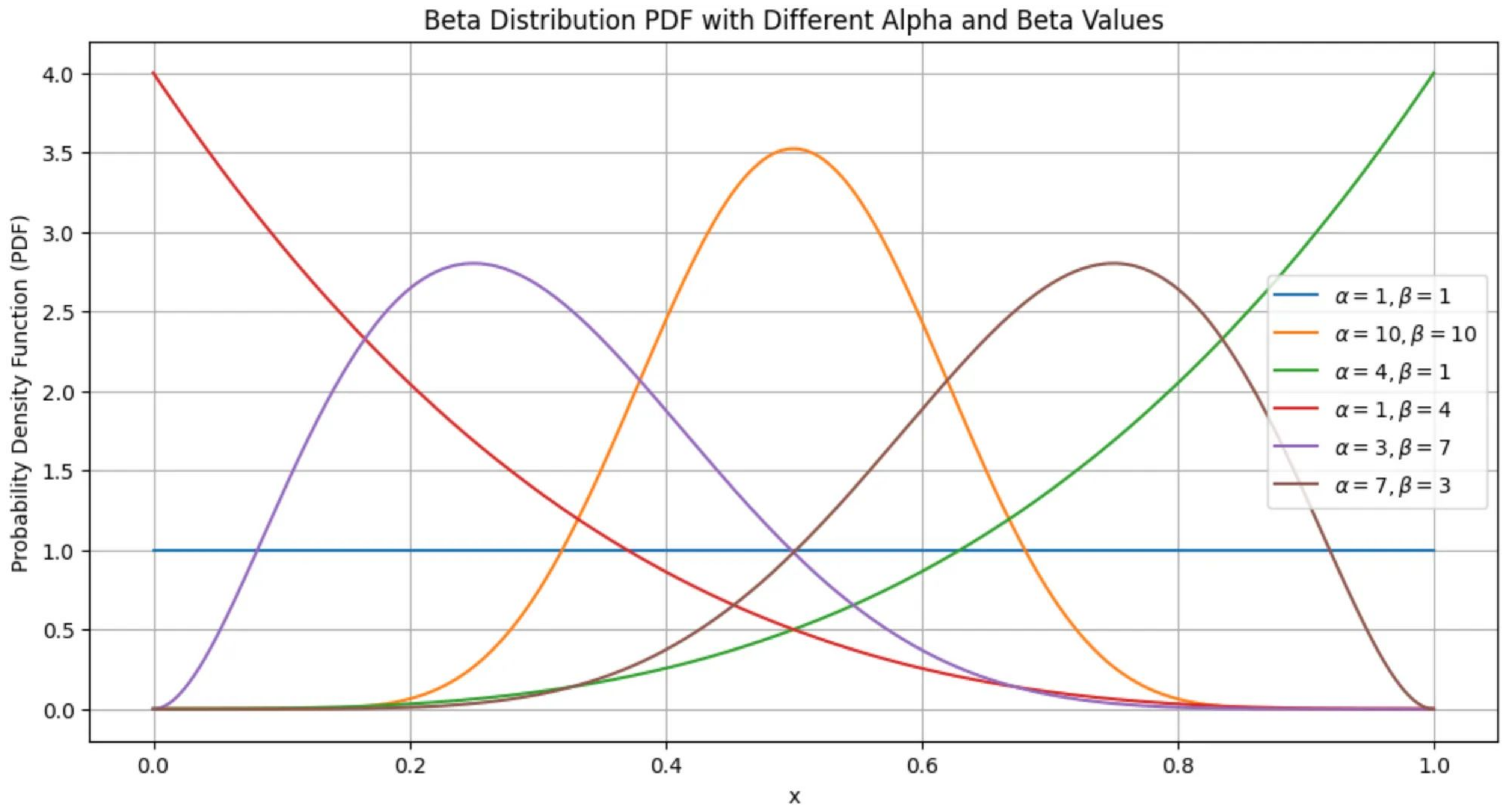


Beta Distribution (just FYI)

Another continuous distribution is Beta Distribution

It can be used to represent probability distribution of a random variable with range between 0 and 1. E.g. Conversion rates, Proportion metrics, etc

Beta distribution is defined by two parameters: alpha and beta.



Any

Question

Hypothesis Testing

Between two search algorithms, which one has a higher conversion?

Hypothesis Testing

An exercise to test hypotheses.

Given two competing hypotheses, which one is more likely to be true?
(re: more supported by the data)

Concepts and Steps in Hypothesis Testing

#1. Determine your competing **hypotheses.**

- Hypotheses: The statements that explain metrics condition on different experiment treatments.
    - We want to know which statement is more supported by the data

- There are always two hypotheses:
    a. H0 (null hypothesis): The default state/condition
    b. H1 (alternative hypothesis): The opposite of H0
        i. Usually the hypothesis that we want to prove to be true

Concepts and Steps in Hypothesis Testing

#2. Specify your **alpha**

- Alpha = significance level
- This is similar to ask “How strong the evidence (re: experiment data) should be, so that we can prefer H1 over H0?”
- Common values; 0.05, 0.01
    i. Smaller alpha = we need stronger evidence to prefer H1 (re: not believing H0)

Concepts and Steps in Hypothesis Testing

#3. Work out the appropriate **statistical testing method** (see later slides) to get the corresponding **p-value**

- P-value: assuming H0 is the truth, what is the probability that we obtain the observed experiment data or more extreme than it

- Example:
  - H0 = “IQ follows Normal(100,15)”.
  - Experiment data: IQ = 115 (note: Z-score = 1)
  - The P-value of this data is 1-0.84 = 0.16

Concepts and Steps in Hypothesis Testing

#4. **Conclude** the experiment!

- Reject H0 (or, prefer H1) IF p-value < alpha
- It means our experiment data is so rare/extreme to occur IF we assume that the true distribution is H0.
- Hence, we can reject H0 and prefer H1

Statistical testing

Method to analyze hypothesis testing

Consider the following competing hypotheses

- H0: Conversion rates of two search algorithms are the **same**
- H1: Conversion rates are **different**

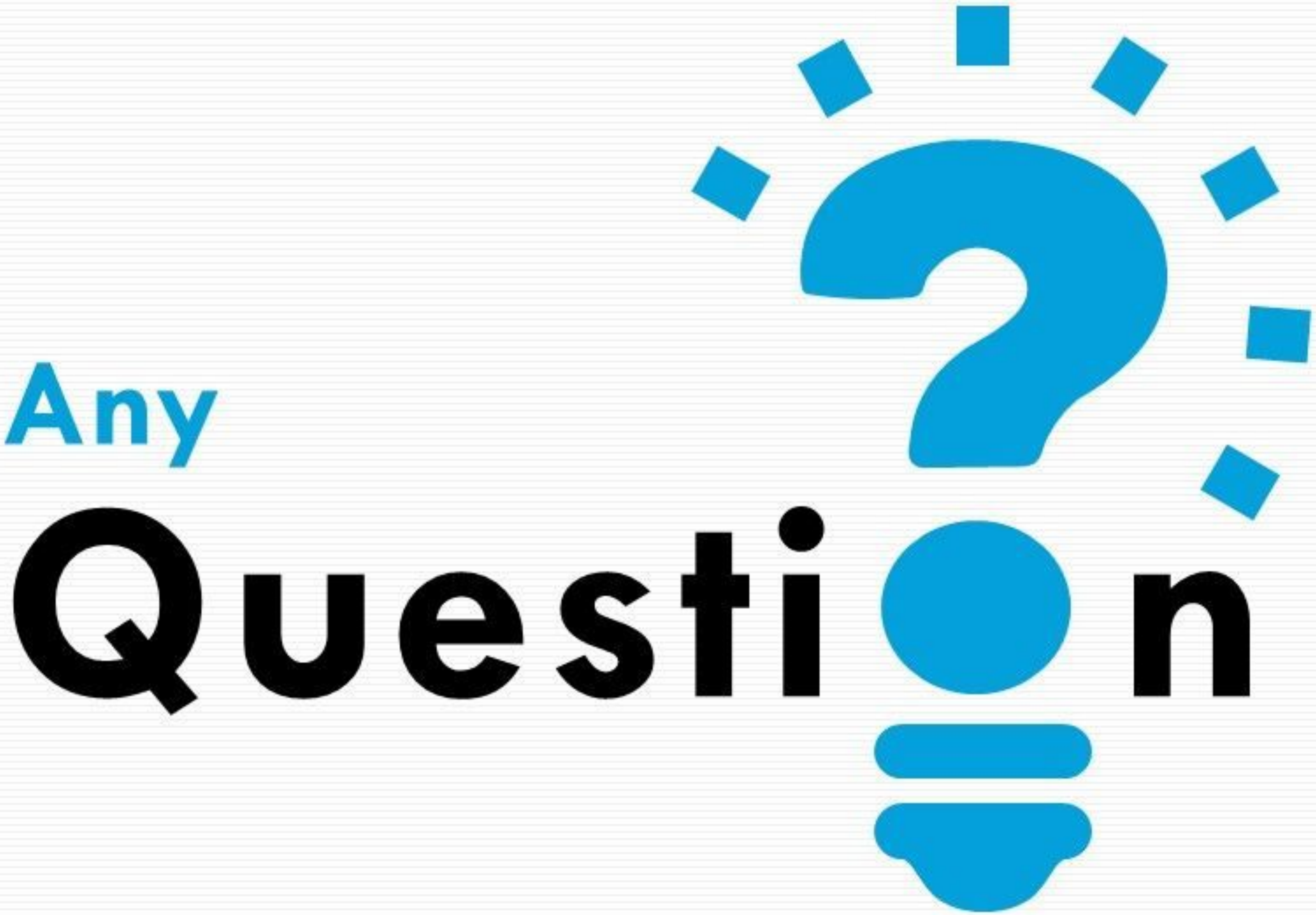
Given the observed data/evidence, we use an appropriate statistical test to determine which hypothesis is more likely to be true

There are hundreds of statistical testing methods, each use depends on different aspects

- Type of data, number of experiment groups, etc

Today we will discuss two popular methods

- **T-test**
- One-Way **ANOVA**


Any
Question

T-Test

T-test

- T-tests are hypothesis tests that **assess the means of one or two groups.**
- It can determine whether two group means are statistically different.
- Hypotheses considered
  - H0: Means between the two groups are the same
  - H1: Means between the two groups are NOT the same
- The term “t-test” refers to the fact that these hypothesis tests use t-values as **test statistics** to evaluate the experiment data

Test statistics

- A test statistic is a value that hypothesis tests calculate from the experiment data.
- Hypothesis tests use the test statistic to compare experiment data to the null hypothesis.
- Test statistic measures the overall **difference between experiment data and the null hypothesis**
    - Higher == more against H0
- So, if the test statistic is high enough, this indicates that the experiment data are so NOT compatible with the null hypothesis
    - Therefore we can reject the null hypotheses

Sampling Distribution

- Sampling distribution represents the test statistic distribution if the null hypothesis is true
- We use sampling distributions **to calculate probabilities for how unusual our sample statistic is if the null hypothesis is true**.
- Recall in t-test
  - H0: two means are the same (mean2 - mean1 = 0)
  - H1: tow means are different (mean2 - mean2 != 0)
- In t-test, the sampling distribution is called the t-distribution

T distribution

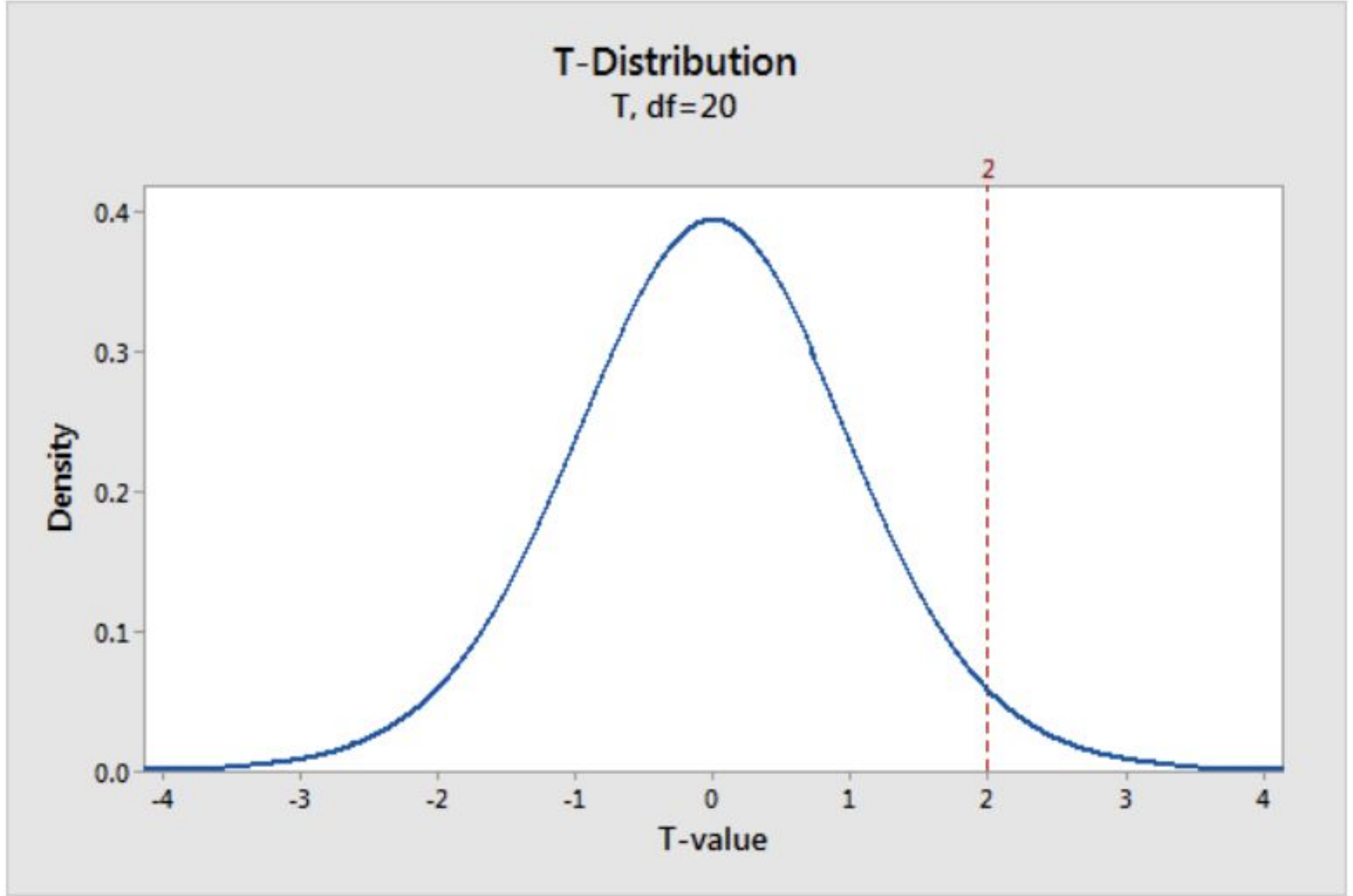
- The T distribution, also known as the Student's t-distribution, is a distribution that is **similar to the standard normal distribution** with its bell shape but has heavier tails.
  - which means it tends to produce values that fall far from its mean.

- When the null is true, the experiment is most likely to obtain a small t-value (near zero)
  - and vice versa if the null is wrong (hence H1 is true)

- T-distribution has one parameter: degree of freedom (df)
  - Higher df == closer to normal distribution
  - df = N1 + N2 - 2
    - Where N1 and N2 are number of data points in each group

T distribution

- Below is displayed the t-value = 2 from a hypothetical experiment data
- Under the assumption that the null is true, the t-distribution indicates that the t-value is not the most likely value (it has low density likelihood)
    - I.e. it's rare enough to occur

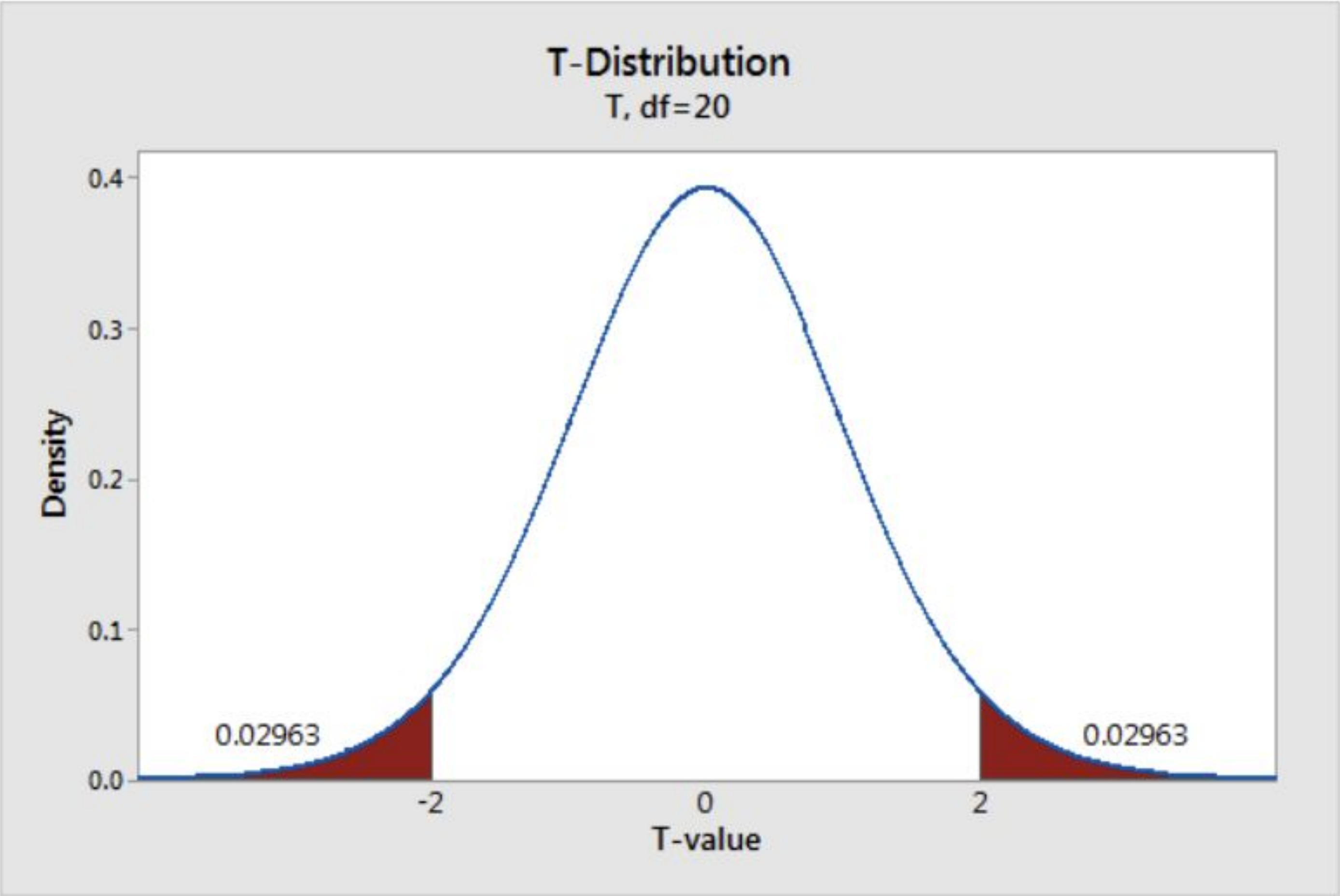


T distribution and P-value

- The graph below finds the probability associated with t-values less than -2 and greater than +2 using the area under the curve

- The probability distribution plot indicates that each of the **two shaded regions** has a probability of 0.02963—for a total of 0.05926 (~6%)

- This graph shows that t-values fall within these areas almost 6% of the time when the null hypothesis is true.
  - it's the **p-value**!



Two ways to check your results

Given alpha = 5% (0.05), and df = K

We can reject H0 IF:

- T-statistic > T-critical at alpha 0.05 and df = K, OR
- P-value < 0.05

Example: suppose alpha = 0.05, df = 20, and we got T-statistic = 5.17

- T-critical at alpha 0.05 and df = 20 → 2.08
  - From where? → two-tail T table!
  - Conclusion: Reject H0, because T-statistic > T-critical

- P-value = 0.000047
  - From where? → Online T-value to P-value calculator
  - Conclusion: Reject H0, because P-value < 0.05

Formula

$$t_{stat} = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where

- X with bar: average of data points in certain group
- s^2 : variance of data points in certain group
- n : number of data points in certain group

Any

Question

One-Way ANOVA

One-Way ANOVA

- Use one-way ANOVA (Analysis of Variance) to determine whether the means of **at least three groups** are different.

- One-way ANOVA requires **one categorical factor** for the independent variable and a continuous variable for the dependent variable.

- For example, if **fertilizer type** is the categorical variable, you can assess whether the differences between **plant growth means** for at least three fertilizers are statistically significant.

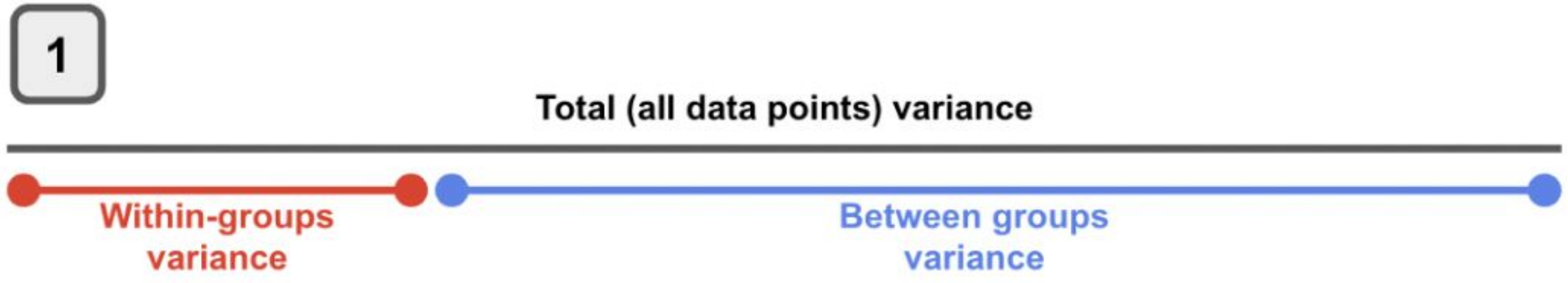
ANOVA Intuition

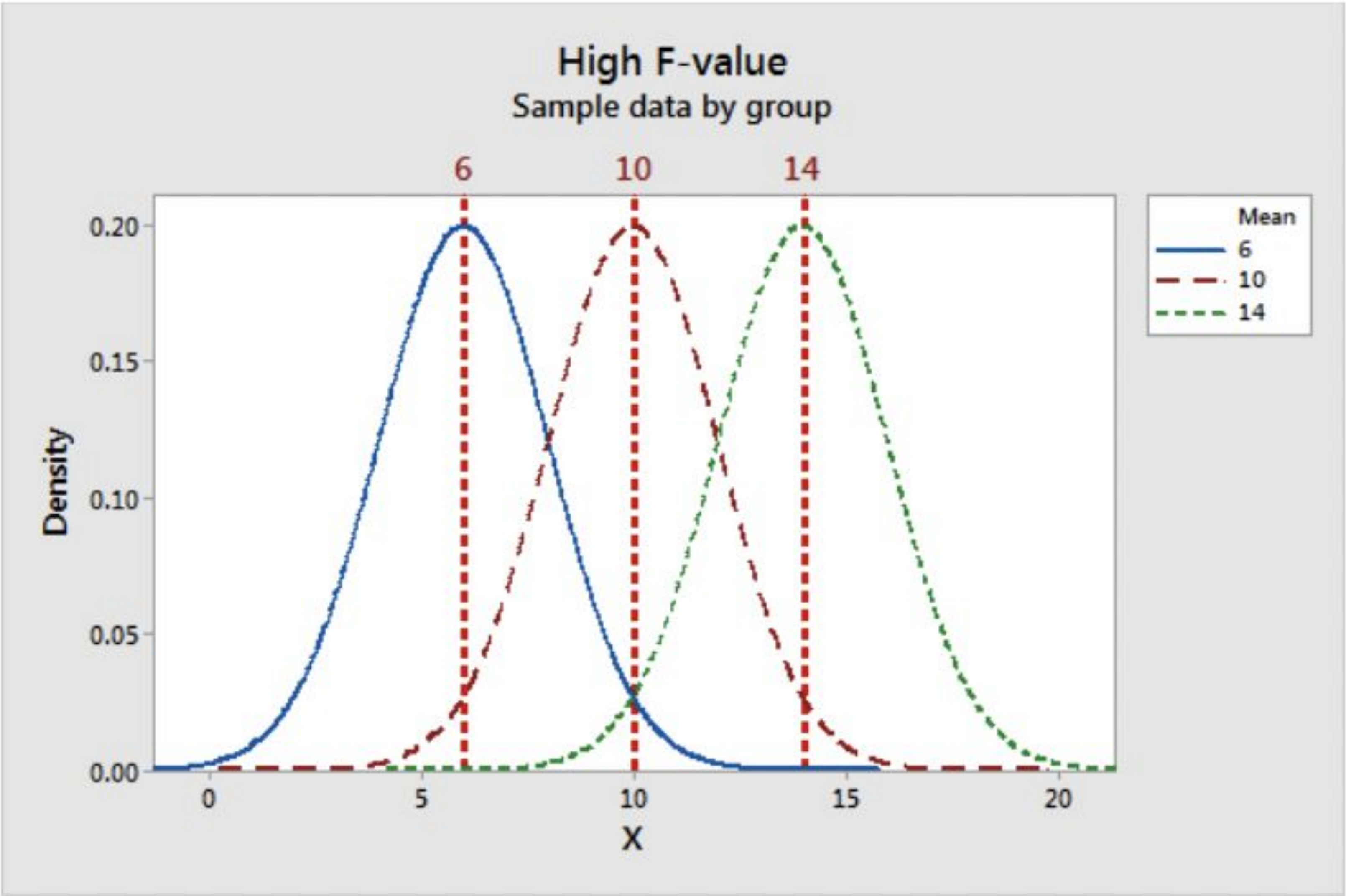
- In a nutshell, ANOVA compares the means (averages) of different groups by determining how likely it is that they are coming from the same distribution
  a. How? by analyzing their variances

- First, we group the variance from the overall data points into **two** components:
  a. variance between data points within the same group (within-group)
  b. variance between different groups' means (between-groups)

ANOVA Intuition

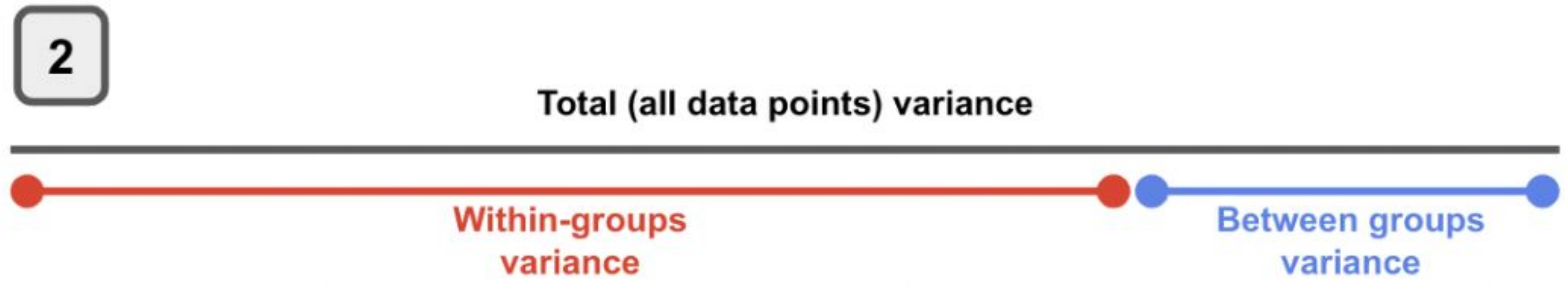
- If within-group variance is smaller, we can conclude that the means of the groups differ
- because the distribution of data points from different groups is well separated
    a. Hence we can conclude that they come from different population distributions — with different averages.

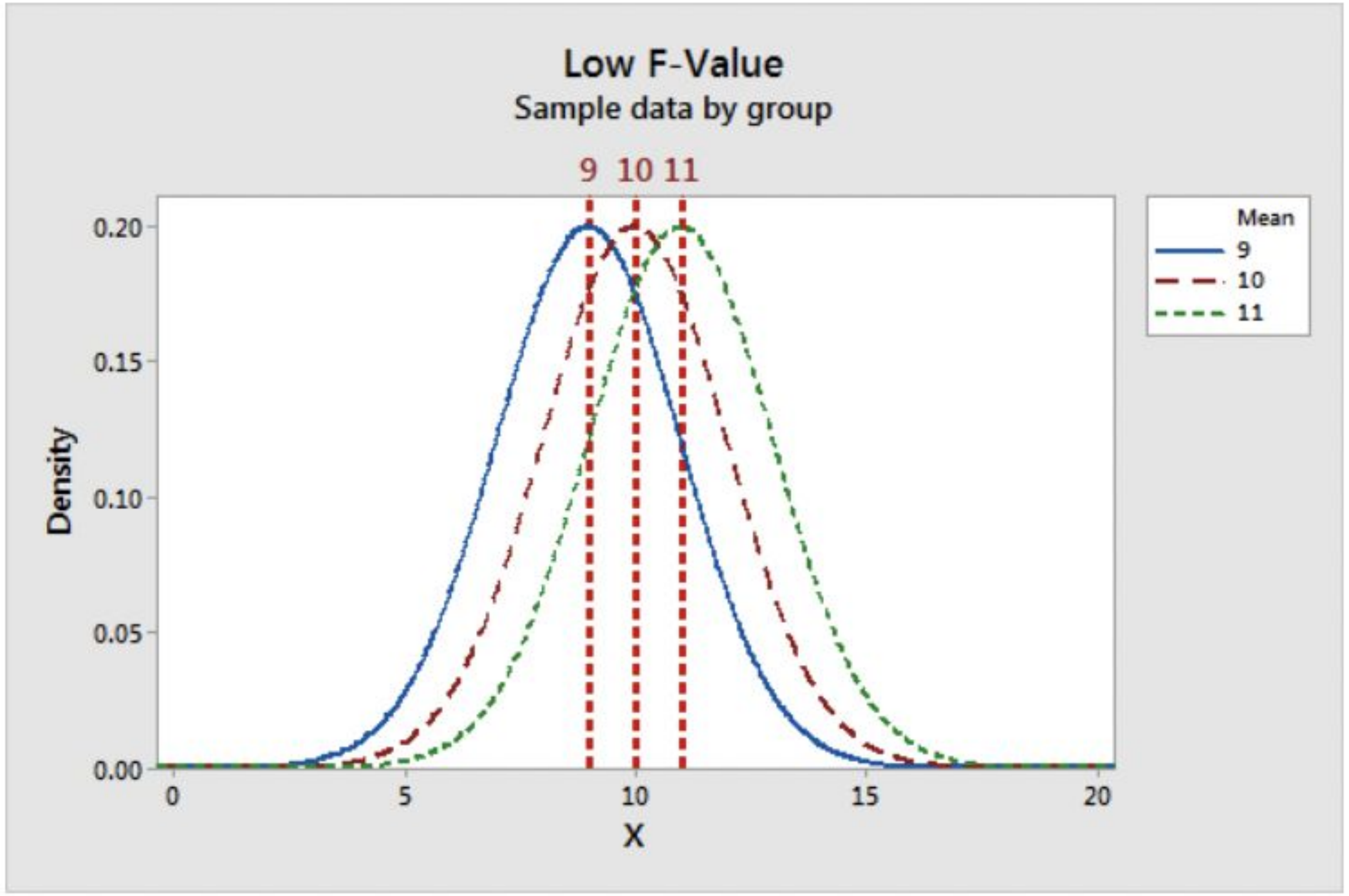




ANOVA Intuition

- Vice-versa, if most of the total variation is contributed from within-groups variation, we can say that the groups' means are perhaps not different in the first place





One-Way ANOVA

- The standard hypotheses for one-way ANOVA are the following:
  - H0: All group means are equal.
  - H1: Not all group means are equal.

- Test statistic is F statistic, which is the ratio between two sums of squared error (SSE), normalized by their degree of freedom
  - between-group SSE and within-group SSE.

- Sampling distribution is the F-distribution with two parameters
  - df treatment = number groups - 1
  - df error = number all data points - df treatment - 1

Formulas

$$\text{between-group SSE} = n_j \sum (X_j - \bar{X}_{..})^2$$

$$\text{within-group SSE} = \sum (X_{ij} - \bar{X}_j)^2$$

- Total SSE = between-group SSE + within-group SSE
- Mean square (MS)
  - between-group MS = between-group SSE / df treatment
  - within-group MS = within-group SSE /df error
- F statistic = between-group MS / within-group MS

Demo Using Google Sheet

- Consider an ABC test about different promotion strategies.
- There are three treatments (experiment groups):
  - Control: the existing promo strategy
  - Variant1: the first challenger promo strategy
  - Variant2: the second challenger promo strategy

- The metric of interest is the average transaction amount, and alpha is set as 5%
- Below is the obtained experiment results (transaction amount)

group	data_1	data_2	data_3	data_4	data_5	data_6	data_7	data_8	data_9	data_10	average
control	77,200	81,600	87,800	72,400	90,800	56,600	83,600	86,500	78,600	79,100	79,420
variant1	83,200	71,200	74,300	83,900	85,800	95,600	97,400	73,100	84,000	74,400	82,290
variant2	80,300	90,100	103,100	84,900	102,600	94,900	94,800	88,800	83,500	87,400	91,040

One-way ANOVA results

- Hypotheses statement
  - H0: All transaction amount means (averages) are the same
  - H1: There is at least one transaction amount mean that differs from the rest

- Using the formulas presented before, we got the following tabulation

Source	SSE	DF	MSE	F-statistic	F-table	Conclusion
between group	732,746,000	2	366,373,000	4.617	3.354	Reject H0
within group	2,142,649,000	27	79,357,370			
total	2,875,395,000	29				

Any Question

Hands-On

- Open today's Jupyter notebook on your Google Colab!
- Make sure you have uploaded the required CSV files to your google drive
    - Remember the file path!

Thank you

Assignment

- Dataset : https://www.kaggle.com/imakash3011/customer-personality-analysis

- What to submit? Google colab link (don’t forget to share access to me: pararawendy19@gmail.com)
  - Format notebook name: HW_HIPOTEST_<YOUR COMPLETE NAME>

- Instructions:
  - Extract & interpret relevant descriptive statistics
  - Carry appropriate hypothesis testing to validate if education affects income

Instructions

- Extract & interpret relevant descriptive statistics
- Carry appropriate hypothesis testing to validate if education affects income