greatlearning

Exercise on Hypothesis Testing and Linear Algebra

Learning Objectives

- Exercise on Hypothesis Testing
- Hypothesis Testing Using Python

Linear Algebra

- Line concept
- Vectors
- Matrices

Exercise on Hypothesis Testing

greatlearning

Problem Statement

Sample size = 400, X-bar = 30000, sigma = $8000 , Population Income = $29,000

The product line will be adequately profitable only in markets where the mean household income is greater than $29,000. Should product line be introduced into the new market.

One – Tailed Hypothesis Test

It signifies that all – or z-values that would cause to reject H_0, are in just one tail of the sampling distribution.

- μ = Population Mean
- $H_0: \mu \leq \$29{,}000$
- $H_a: \mu > \$29{,}000$

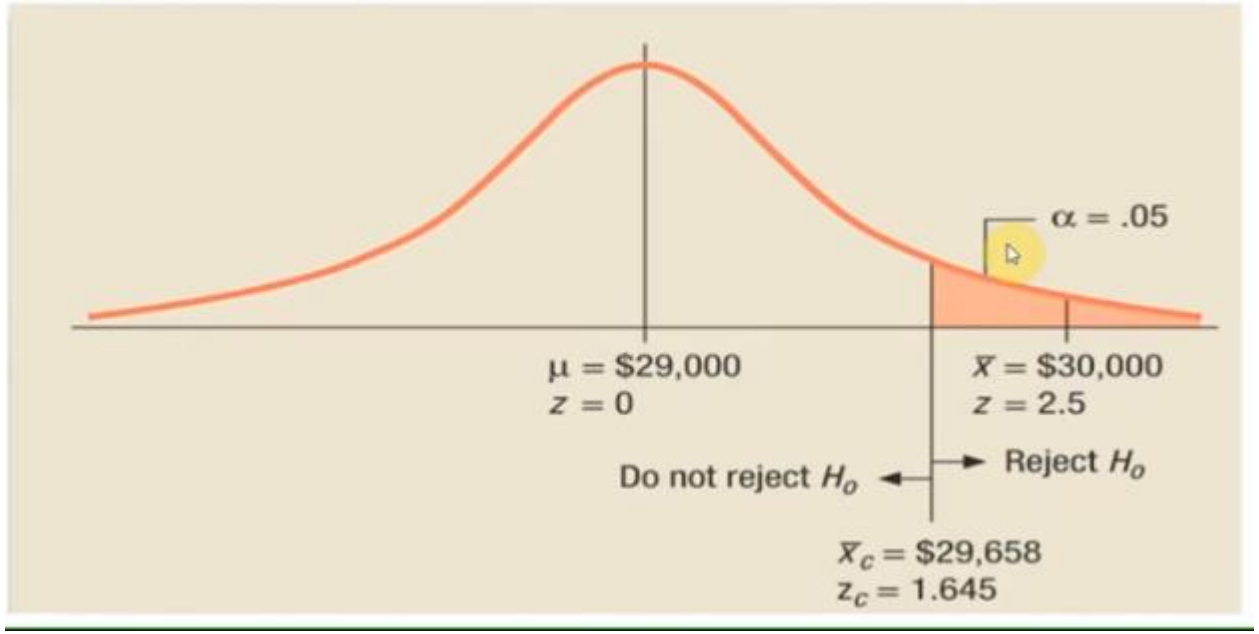
Note: The null hypothesis is tested only at the equality sign

Test Statistic

Substituting the values in the formula given below for the unknown terms, we get

z = 2.5

$$Z = \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$



Critical value for rejecting the Null Hypothesis

Notes:

- P value gives you the actual risk or level of significance by which the null hypothesis is rejected.

- Solution method 1: Whenever the p value < alpha then the null hypothesis is rejected.

- Solution method 2: If the computed z value falls in the rejection region then the null hypothesis is rejected.

Hypothesis Testing Using Python

greatlearning

One Sample Testing

Some important functions:

1. *t_statistic, p_value = ttest_1samp(daily_intake, n)*

Here n= sample number , daily_intake= array

2. *z_statistic, p_value = wilcoxon(daily_intake – n)*

Plotting

1. *plt.hist(daily_intake)*
2. *plt.boxplot(daily_intake)*

Two Sample Testing

Some important functions:

1. *t_statistic, p_value = ttest_ind(group1, group2)*
2. *u, p_value = mannwhitneyu(group1, group2)*
3. *t_statistic, p_value = ttest_1samp(post-pre, 0)*
4. *z_statistic, p_value = wilcoxon(post-pre)*
5. *levene(pre,post)*
6. *shapiro(post)*

greatlearning

Linear Algebra

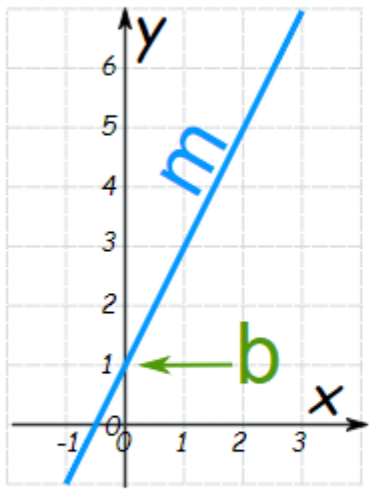
Line Concept

- A line is a mathematical object which is used for multiple purposes. One of the purposes that it is used for is to find the shortest distance between two points in mathematical space.

- It also gives a relation between two variables.

- The equation of a straight line on a graph is made up of a Y term, an X term, and a number written as Y=mx+c . The slope of the line is known as the gradient and is represented by m in the equation.

What does it stand for?



$$y = mx + b$$

Slope (or Gradient) — Y Intercept

y = how far up

x = how far along

m = Slope or Gradient (how steep the line is)

b = the Y Intercept (where the line crosses the Y axis)

Vectors

- A vector is n dimensional, where n is a real number.
- Direction and magnitude are the two important properties of vector.
- Vector operations include addition, subtraction, and three types of multiplication. The sum of two vectors is a third vector.
- When a vector is multiplied by a positive scalar (number), its magnitude is multiplied by the scalar and its direction remains unchanged.

Addition and Scalar Multiplication

1. $\vec{a} + \vec{b} = \vec{b} + \vec{a}$
2. $\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}$
3. $\vec{a} + \vec{0} = \vec{a}$
4. $\vec{a} + (-\vec{a}) = \vec{0}$
5. $c(\vec{a} + \vec{b}) = c\vec{a} + c\vec{b}$
6. $(c + d)\vec{a} = c\vec{a} + d\vec{a}$
7. $(cd)\vec{a} = c(d\vec{a})$
8. $1\vec{a} = \vec{a}$

Dot Product

The dot product is defined by

$$\vec{a} = (a_1, a_2, a_3), \ \vec{b} = (b_1, b_2, b_3)$$

$$\implies \vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

Matrices

- A matrix is defined as an ordered rectangular array of numbers.

- They can be used to represent systems of linear equations.

And a fully expanded m×n matrix A, would look like this:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

... or in a more compact form: $A = (a_{ij})$

Matrix Addition and Subtraction

DEFINITION: Two matrices A and B can be added or subtracted if and only if their dimensions are the same (i.e. both matrices have the same number of rows and columns. Take:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix} \quad and \quad B = \begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix}$$

Addition

If A and B above are matrices of the same type then the sum is found by adding the corresponding elements $a_{ij} + b_{ij}$.

Here is an example of adding A and B together.

$$A + B = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 3 & 5 \\ 2 & 0 & 5 \end{pmatrix}$$

Subtraction

If A and B are matrices of the same type then the subtraction is found by subtracting the corresponding elements $a_{ij} - b_{ij}$.

Here is an example of subtracting matrices.

$$A - B = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 2 & 1 & 2 \\ 1 & 0 & 3 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

Case Study 1: Calculate t Test Application One sample

Context: Experience Marketing Services reported that the typical American spends a mean of 144 minutes (2.4 hours) per day accessing the Internet via a mobile device. In order to test the validity of this statement, you select a sample of 30 friends and family. The result for the time spent per day accessing the Internet via mobile device (in minutes) are stored in Internet_Mobile_Time.csv file.

Q: Is there evidence that the population mean time spent per day accessing the Internet via mobile device is different from 144 minutes? Use the p-value approach and a level of significance of 0.05

Steps to follow

1. Import libraries
2. Get the data
3. Calculate Mean , Standard Deviation
4. Calculate t-statistics
5. Compare with the critical t-value
6. Find p-value after comparison with t

Case Study 2: Calculate Independent t-Test Two Sample

Context: A hotel manager looks to enhance the initial impressions that hotel guests have when they check in. Contributing to initial impressions is the time it takes to deliver a guest's luggage to the room after check in. A random sample of 20 deliveries on a particular day were selected each from Wing A and Wing B of the hotel. The data collated is given in Luggage.csv file.

Q: Analyze the data and determine whether there is difference in the mean delivery times in the two wings of the hotel.

Steps to follow

1. Import libraries
2. Get the data
3. Calculate ttest at 0.05 and 0.01 alpha threshold
4. Collaborate the hypothesis with Box Plot

greatlearning

HappyLearning

Questions?