Assignment Name: Hypothesis Testing, Correlation & Regression Analysis

Course Code: ICT-2107

Course Name: Statistics and Probability for Engineers

Submitted to:

Md. Fazlul Karim Patwary

Professor

Institute of Information Technology

Jahangirnagar University

Submitted By:

Md. Shaon Khan

Roll:1984

Session: 2022-2023

Year: 2nd Year 1st Semester

Institute of Information Technology

Jahangirnagar University

Submission Date: 28th July 2025

Question-01: What is Hypothesis? Why Do We need to draw Hypothesis?

Answer: A Hypothesis is a precise, testable statement or prediction about the relationship between two or more variable. A Hypothesis is formulated to be tested through scientific methods, particularly using sample data. It serves as the foundation for statistical inference, enabling researchers to draw conclusions about a population based on observations from a sample.

There are generally two types of Hypothesis in statistics:

1. Null Hypothesis (H_0): It proposes that there is no significant effect or relationship between variables.

2. Alternative Hypothesis (H_1 or H_a): It suggests that there is a statistically significant effect or relationship.

Importance of Drawing a Hypothesis:

1. **Provides Direction for Research:** A Hypothesis guides the researches in designing the study, collecting data, and selecting appropriate statistical tests. It clearly defines what needs to be investigated.

2. **Establishes the Basis for Testing:** Hypothesis form the basis for Hypothesis testing, allowing researchesrs to make objective decisions about the validity of assumptions regarding a population.

3. **Supports Scientific Decision-Making:** By testing a Hypothesis, statisticians can determine whether observed patterns in the data are due to chance or represent true population characteristics

4. **Enables statistical Inference:** Drawing a hypothesis allows the use of inferential statistical techniques to make generalizations about a population from a sample.

Question-02: What Do you Mean by null Hypothesis? When Do you reject our null hypothesis?

Answer: The null hypothesis, denoted as H_0 is a formal statement in statistics that assumes there is no effect, no difference or no relationship between the variables being studied. It serves as the default or starting assumption in hypothesis testing.

For example, in testing whether a new drug is effective, the null hypothesis might state that the drug has no effect on patients compared to a placebo.

Reject the NULL Hypothesis:

We reject the null hypothesis when the evidance from the sample data is strong enough to conclude that the observed effect or difference is unlikely to have occured by random chance alone.

The decision is based on the p-value and a significance level (α):

Significance level: A predefined threshold, commonly set at 0.05 (5%) which indicates the maximum probability of rejecting a true null hypothesis (Type I error)

P-value: The probability of obtaining a test statistic as extreme as, or more extreme than, the one observed, assuming the null hypothesis is true.

Rejection Rule:

If the p-value $\leq \alpha$, we reject

If the p-value $> \alpha$, we fail to reject the null hypothesis.

Question-03: What is your understanding on "T-Test" for single mean?

Answer: The t-test for a single mean is a statistical method used to determine whether the mean of a single sample significantly differs from a known or hypothesized population mean.

This test is applicable when:

1. The population standard deviation is unknown.
2. The sample size is small
3. The population is normally distributed or the sample is approximately normal.

Purpose:

To test whether the sample provides enough evidance to conclude that the population mean is different from a specified value.

Formula:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Where,

$\bar{x}$ = sample mean

μ_0 = hypothesis population mean

s = sample standard deviation

n = sample size

Question-09: When do we use paired sample T-Test?

Answer: A Paired sample t-test is used when we want to compare the means of two related groups to determine whether there is statistically significant difference between them.

This test is approximate when.

1. The data consists of paired observations, such as before and after measurements on the sample subjects.
2. The measurements are taken under two different conditions on the same individuals.
3. The difference between paired observations are appoximately normally distributed.

Common Seenarios:

- Measuring a patients blood pressure before and after taking a medication.
- testing student scores before and after a special training program.

Formula:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Where,

$\bar{d}$ = mean of the differences

s_d = standard deviation of the differences

n = number of pairs.

Topic: Correlation & Regressio

Question 1: What are the differences between correlation and regression?

Answer:

The difference between correlation and regression are give below:

Aspect	Correlation	Regression
Purpose	Measures the strength and direction of the linear relationship between two variables.	Models the relationship between a dependent variable and one or more independent variables to predict or explain the dependent variable.
Nature	Descriptive statistic that quantifiers association.	predictive/statistical modeling technique.
Variables	Both variables are treated symmetrically; no distinction between dependent variable.	Distinguishes between dependent and independent variables.
Output	Correlation coefficient (r), a value between −1 and 1.	Regression equation coefficients and predictions.
Interpretation	Indicates how strongly and in what direction variables are related.	Shows how much the dependent variable changes with changes in independent variable.
Causation	Doesn't imply causation, only association.	Can be used to infer causation if assumptions are met.

Question 2 : What is correlation coefficient?

Answer: The correlation coefficient is a numerical measure that quantifiers the strength and direction of the linear relationship between two variables. It is usually denoted by r and ranges from -1 to $+1$.

- When $r = +1$, it indicates a perfect positive linear relationship.
- When $r = -1$, it indicates a perfect negative linear relationship.
- When $r = 0$, it means there is no linear relationship between the variables.

Question 3 : age: 12, 23, 14, 20, 23, 25, 18, 28, 30, 34

weight: 67, 89, 45, 78, 56, 34, 34, 55, 66, 77

(a) Calculate correlation and explain

(b) Construct a regression line $wt = a + b * age$

(c) Using line: if a persons age is 38 what is the predicted weight(wt) of that person.

Answer:

We know the formula of correlation is,

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Social n	Age(x)	Wt(y)	xy	x^2	y^2
1.	12	67	804	144	4489
2.	23	89	2047	529	7921
3.	14	45	630	196	2025
4.	20	78	1560	400	6084
5.	23	56	1288	529	3136
6.	25	34	850	625	1156
7.	18	34	612	324	1156
8.	28	55	1540	784	3025
9.	30	66	1980	900	4356
10.	34	77	2618	1156	5929
	$\Sigma x = 227$	$\Sigma y = 601$	$\Sigma xy = 13929$	$\Sigma x^2 = 5587$	$\Sigma y^2 = 39218$

$$r = \frac{10(13929) - (227 * 601)}{\sqrt{[10 * 5587 - (227)^2][10 * 39218 - (601)^2]}}$$

$$= \frac{12863}{\sqrt{4341 * 30979}}$$

$$= 0.2469$$

The correlation coefficient $r \approx 0.2469$ indicates a weak positive linear relationship between age and weight in this data set.

This means as age increases, weight tends to increase slightly, but the relationship is not very strong.

(b) From a. we get

$\Sigma x = 227, \quad \Sigma Y = 601, \quad \Sigma xy = 13929$

$\Sigma x^2 = 5587, \quad n = 10$

using the regression formulas,

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \frac{\Sigma y - b\Sigma x}{n}$$

$$\text{so, } b = \frac{10 * 13929 - 227 * 601}{10 * 5587 - (227)^2}$$

$$= \frac{2863}{4371} \approx 0.655$$

$$a = \frac{601 - (0.655) * 227}{10}$$

$$\approx 45.23$$

So final regression line $\hat{y} = 45.23 + 0.655x$

(c)

From (b) we get,

Regression line, $\hat{Y} = 45.23 + 0.655X$

Here,

$X = 38$ [the person's age]

So,

$Y = 45.23 + 0.655 * 38$

$= 70.12 kg$

So, if a persons age is 38 then the predicted weight of that person is 70.12kg.

Question4: When do we use Rank Correlation?

Answer: Rank correlation is used when:

1. The data is ordinal:
   When variables are ranked (e.g. 1st, 2nd, 3rd) rather than measured on a numerical scale.

2. The data is not normally distributed:
   When the assamptions of Pearson's correlation are not satisfied.

3. There are outliers:
   Rank correlation is less sensitive to outliers than pearson's correlation.

4. The relationship is monotonic but not linear:
   If two variables tend to increase or decrease together, but not at a constant rate, rank correlation is more appropriate.

5. Data is in form of preferences or rankings:
   For example, comparing judges rankings in a contest or customer preferences.