STUDY GUIDE FOR MODULE NO. 7

HYPOTHESIS TESTING

MODULE OVERVIEW

A **hypothesis** is an educated guess about something in the world around you. It should be testable, either by experiment or observation. For example: A new medicine you think might work. A way of teaching you think might be better. A fairer way to administer standardized tests. It can really be anything at all as long as you can put it to the test. The main purpose of statistics is to test a hypothesis.

MODULE LEARNING OBJECTIVES

At the end of the lesson, you should be able to:

- Define the Ho, Ha, level of significance, test statistic, and statistical significance
- Define the Type I and Type II errors
- Distinguish the difference between the One-tailed and two-tailed tests.
- Solve problems and interpret results involving one-sample / two-sample hypothesis testing using statistical tools
- Discuss correlation and regression.

LEARNING CONTENTS

Hypothesis Testing

- Hypothesis testing and estimation are used to reach conclusions about a population by examining a sample of that population.
- Hypothesis testing in Statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Two types of Hypothesis:

1. Null Hypothesis (Ho)
   - It is the hypothesis to be tested which one hopes to reject.
   - It shows the equality or no significant difference or relationship between the variables.
2. Alternative Hypothesis (Ha)
   - It generally represents the idea which the researcher wants to prove.

Two Type of Hypothesis Testing:

1. **One-Tailed Test** - It is a directional test with the region of rejection lying on either left or right of the normal curve.
   a. *Right-Directional Test*
   (Ha uses comparatives such as greater than, more than, higher than, better than, upper than, superior to, exceeds, etc..)
   b. *Left- Directional Test*

(Ha uses comparatives such as smaller than, less than, lower than, inferior to, below, etc..) to : =

2. **Two Tailed Test** - It is a non- directional test with the region of rejection lying on both tails of the normal curve. (Ha uses words such as not equal to, significantly different, etc)

Level Of Significance

- The level of significance, α, is a probability and is, in reality, the probability of rejecting a true null hypothesis.
- For example, with 95% confidence intervals, $\alpha = 0.05$ meaning that there is a 5% chance that the parameter does not fall within the 95% confidence region. This creates an error and leads to a false conclusion.

Type I Error and Type II Error

When doing hypothesis testing, two types of mistakes may be made and we call them Type I error and Type II error.

Decision	*Reality*	
	H_0 is true	H_0 is false
Reject H_0	Type I error	Correct
Accept H_0		Type II error

- If we reject H_0 when H_0 is true, we commit a Type I error.
- If we accept H_0 when H_0 is false, we commit a Type II error.

Steps in Conducting the Hypothesis Testing

Every time we perform a hypothesis test, this is the basic procedure that we will follow:

Step 1. Check the conditions necessary to run the selected test and select the hypotheses for that test.
Step 2. Decide on the significance level, α.
Step 3. Compute the value of the test statistic:
Step 4. Find the appropriate critical values for the tests using statistical tables.
Step 5. Check to see if the value of the test statistic falls in the rejection region.
If it does, then reject H_0 (and conclude H_a). If it does not fall in the rejection region, do not reject H_0.
Step 6. State the conclusion in words.

LEARNING CONTENTS (HYPOTHESIS TESTING FOR SINGLE SAMPLE)

Test on Mean when Variance is Unknown

When the sample size is small, usually $(n < 30)$, and assuming that the sample is approximately normal, t-test in reference to the t-distribution is more appropriate to employ in testing the hypothesis for the mean.

In most cases of hypothesis testing, the variance of the population is unknown. In this case, the standard error should be estimated by the sample standard deviation. Associated with this is the divisor $n-1$ which we are going to call the degrees of freedom. Denoting the degrees of freedom we use $df = n - 1$.

Test Statistic:

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

T-test Example:

Ten years ago, the average height of young adult women living in a certain city was 63 inches. The standard deviation is unknown. A researcher wants to determine whether the height of young adult women differs significantly from 63 inches. She randomly sampled eight young adult women currently residing in her city and measures their heights. The following data are obtained: [64, 66, 68, 60, 62, 65, 66, 63].

Solution:
Following the steps in hypothesis testing we have the following results.

1. $H_0: \mu = 63$ inches
   $H_a: \mu \neq 63$ inches (two tailed)
2. For this particular problem, we can set the level of significance to be at 0.05.
3. Solving for the test statistic, we have

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

$$t = \frac{64.25 - 63}{\frac{2.55}{\sqrt{8}}}$$

$$t = 1.39$$

4. Critical value: Since $n = 8$, we have $df = 7$. Referring to the tables of critical values for t-distribution, the critical value is $t = 2.365$.

5. Decision: The computed t-value lies within the non-rejection region; thus we cannot reject the null hypothesis.

6. Conclusion: Therefore, the height of young adult women does not differ significantly from 63 inches.

Example:
A researcher read a paper introducing a new test and found that 51 is the average score on the new test. The researcher believes that patients with schizophrenia will score less than the normal average on this test of executive function. He tested 22 patients with schizophrenia. The average score for the patients is 39 with a standard deviation of 4.3. Is there significant evidence at the 5 % level to support the researcher's claim?

CORRELATION AND REGRESSION

CORRELATION

- Correlation is a statistical method used to determine whether a relationship between variables exist.
- Simple linear correlation can be positive or negative. A positive relationship exists when either variables increase at the same time or both decrease at the same time. On the contrary, in a negative relationship, as one variable increases, the other variables decreases or vice versa.

Pearson Product-Moment Correlation

- It is the most widely used in Statistics to measure the degree of the relationship between the linear related variables.
- It requires both variables to be normally distributed.
- Correlation refers to the departure of two random variables from independence.
- Correlation coefficient (Pearson's r) is a measure of the linear strength of association between two variables. It is founded by Karl Pearson.

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

- The value of the correlation coefficient varies between -1 and +1.
- Guilford's suggested interpretation for the value of r.

r Value	Interpretation
Less than 0.20	Slight; Almost negligible relationship
0.20-0.40	Low correlation, definite but small relationship
0.41-0.70	Moderate correlation; substantial relationship
0.71-0.90	High correlation; marked relationship
0.91-0.99	Very high correlation; very dependable relationship

Test of Significance

- A test of significance for the coefficient of correlation may be used to find out if the computed Pearson's r could have occurred in a population in which the two variable are related or not.
- The test statistic follows t-distribution with $n-2$ degrees of freedom, where n is the total number of pairs.
- The significance is computed using the formula below.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- Hypotheses: $H_o: \rho = 0$, there is no significant relationship between the two variables
  $H_a: \rho \neq 0$, there is a significant relationship between the two variables

Example:
The owner of a chain of fruit shake store would like to study the correlation between atmospheric temperature and sales during the summer season. A random sample of 12 days is selected with the results given as follows.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Days**
79	76	78	84	90	83	93	94	97	85	88	82	Temperature (F^o)
147	143	147	168	206	155	192	211	209	187	200	150	Sales

Compute the coefficient of correlation. Determine at the 0.05 significance level if there is a significant relationship between atmospheric temperature and sales.

Solution:

$r = 0.93$, There is a very high positive relationship between the sales and temperature. It means that as the temperature increases, the sales will also increase or vice versa.

1. $H_o: \rho = 0$, there is no significant relationship between the two variables
 $H_a: \rho \neq 0$, there is a significant relationship between the two variables
2. $\alpha = 0.05 \; or \; 5\%$
3. Test Statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.93\sqrt{\frac{12-2}{1-0.93^2}} = 8.001$$

4. $Critical\ Value$: 2.228
5. DECISION: Since 8.001 can be found in the rejection region, REJECT null hypothesis
6. CONCLUSION: there is a significant relationship between the temperature and sales.

SIMPLE LINEAR REGRESSION

- Regression analysis is a statistical method used to describe the nature of the relationship between variables.
- It is a simple statistical tool used to model the dependence of a variable on explanatory variables.
- This functional relationship may then be formally stated as an equation, with associated values that describe how well this equation fits the data.

The equation for a fitted line is:

$$\hat{y} = a + bx$$

where

$\hat{y}$ = *predicted value*
a = *y-intercept*
b = *slope of the regression line*
x = *the value of x to be predicted*

To find the slope b:

$$b = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2}$$

To find the value of a:

$$a = \bar{y} - b\bar{x}$$

where :

$\bar{y} = mean\ of\ value\ of\ Y$
$\bar{x} = mean\ of\ value\ of\ X$

Example:

1	2	3	4	5	6	7	8	9	10	11	12	Days
79	76	78	84	90	83	93	94	97	85	88	82	Temperature (F^o)
147	143	147	168	206	155	192	211	209	187	200	150	Sales

a. Find the regression line equation of the data from the previous example.
b. Use the model to predict the sales if the temperature is 75F^o.