Lecture #10 **Chapter 10 Correlation and Regression**

The main focus of this chapter is to form inferences based on sample data that come in pairs. Given such paired sample data, we want to determine whether there is a relationship between the two variables and, if so, to identify what the relationship is. We call this relationship "correlation".

10-2 Correlation

The main objective of this section is to analyze a collection of paired sample data (sometimes called **bivariate data**) and determine whether there appears to be a relationship between the two variables.

A **correlation** exists between two variables when one of them is related to the other in some way.

We can often see a relationship between two variables by constructing a graph called a **scatterplot**, or **scatter diagram**.

A **scatterplot** is a graph in which the paired (x, y) sample data are plotted with a horizontal x-axis and a vertical y-axis. Each individual (x, y) pair is plotted as a single point.

Example 1: a sociologist conducted a study to determine whether there is a linear relationship between family income level (in thousands of dollars) and percent of income donated to charities. The data are listed in the table. Display the data in a scatterplot and determine the type of correlation.

| | | | | | | |
|---|---|---|---|---|---|---|
| Income Level (in $1000_s$), x | 42 | 48 | 50 | 59 | 65 | 72 |
| Donating Percent , y | 9 | 10 | 8 | 5 | 6 | 3 |

**Linear Correlation Coefficient**

Interpreting correlation using a scatterplot can be subjective. A more precise way to measure the type and strength of a linear correlation between two variables is to calculate the linear correlation coefficient.

The **linear correlation coefficient, r**, measures the strength of its straight line trend and the direction of the association between the paired x- and y-values in a sample.

$$r = \frac{n\sum xy - \sum x \cdot \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}$$

Where n is the number of pairs of data.

**Round r to three decimal places.**

Use $\rho$ (rho) for population linear correlation coefficient.

Properties of the linear correlation coefficient r:

1. $-1 \leq r \leq 1$
2. The closer r is to $\mp 1$, the closer the data points fall to a straight line, and the stronger is the linear association. In this case, we conclude that there is a significant linear correlation between x and y.
3. If r is close to 0, we conclude that there is no significant linear correlation between x and y.
4. A positive correlation indicates a positive association, and a negative correlation indicates a negative association.
5. The value of the correlation does not depend on the variables unit.

Example 2: Calculate the linear correlation coefficient for the income level and donating percent data given in example 1.

**Hypothesis testing for a population correlation coefficient:**

Once you have calculated the sample linear correlation coefficient, r, you will want to determine whether the population linear correlation, $\rho$, is significant.

You can do this by performing a hypothesis test. A hypothesis test for $\rho$ can be one tailed or two tailed. The null and alternative hypotheses for these tests are as follows.

$H_0$: $\rho = 0$ (No significant correlation) Two-tailed test

$H_1$: $\rho \neq 0$ (significant correlation)

$H_0$: $\rho = 0$ (No significant correlation) Left-tailed test

$H_1$: $\rho < 0$ (significant negative correlation)

$H_0$: $\rho = 0$ (No significant correlation) Right-tailed test

$H_1$: $\rho > 0$ (significant positive correlation)

**The t-test for the correlation coefficient**

A **t-test** can be used to test whether the correlation between two variables is significant. The test statistic is $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

The sampling distribution for r is a t-distribution with n-2 degrees of freedom.

**Guidelines: Using the t-test for the correlation coefficient**

1. State $H_0$ and $H_1$.
2. Specify$\alpha$.
3. Determine the degrees of freedom. d.f. = n -2
4. Find the critical value(s) from table A-3 with n-2 degrees of freedom and identify the rejection region(s).
5. Find the test statistic. $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

6. Make a decision to reject or fail to reject the null hypothesis. If the absolute value of the test statistic, t is in the rejection region, reject $H_0$ . Otherwise, fail to reject $H_0$.
7. Reject $H_0$ and conclude that there is a linear correlation. Fail to reject $H_0$ and conclude that there is not sufficient evidence to conclude that there is a linear correlation.

Example 3: In example 2, we use the data to find r. Test the significance of this correlation coefficient. Use $\alpha$ = 0.05.

**10-3 Regression**

The main objective of this section is to describe the relationship between two variables by finding the graph and equation of the straight line that represents the relationship. This straight line is called the **regression line**, and its equation is called the **regression equation**.

Given a collection of paired sample data, the regression equation $\hat{y}$**=b$_0$ +b$_1$x** algebraically describes the relationship between the two variables. The graph of the regression equation is called the **regression line** (or line of best fit, or least-squares line).

Equation of the regression line: $\hat{y}$**=b$_0$ +b$_1$x**

Formula: slope:

$$b_1 = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - (\sum x)^2}$$

y-intercept: $b_0 = \bar{y} - b_1 \bar{x}$

**Round b$_0$ and b$_1$ to three significant digits**.

Example 4: find the equation of the regression line for the income level and donating percent data used in example 1, 2, 3.

**Applications of regression equations**:

After finding the equation of a regression line, you can use the equation to predict y-values over the range of the data.

Example 5: Use the equation in example 4 to predict the expected donation percent for the following income levels in 1000s).

a) 52
b) 69

Prediction values are meaningful only for x-values in (or close to) the range of the data.

**Procedure for Predicting**

1. Calculate the value of r and test the hypothesis that $\rho=0$.
2. Is $\rho=0$ rejected (so that there is linear correlation)?
    i) If the answer to step (2) is yes, then use the regression equation to make predictions. Substitute the given value in the regression equation.
    ii) If the answer to step (2) in no, then the best predicted value for any given value is the sample mean of the other variable.

**(If there is not a linear correlation, the best predicted y-value is the mean of the y-values.)**