NORTHWESTERN UNIVERSITY

SCHOOL OF CONTINUING STUDIES

Handout: Problem Set #8 Solutions

PREDICT 401: Introduction to Statistical Analysis

These problem sets are meant to allow you to practice and check the accuracy of your work. Please do not review the solutions until you have finalized your work. Although these problem sets are not submitted and graded, treat them as if they were. It is to your great benefit to work on and even struggle with the problem sets. Looking at the solutions before finalizing your work will, quite simply, make for a less meaningful learning experience.

1. Imagine that you are analyzing information about visitors to a Web site designed to help individuals self-diagnose medical conditions. You hypothesize that older people visit the site more frequently, but you only have a small sample of data from which to draw. Specifically, you have the following information:

	Age			
# Web site Visits	**Old**	**Medium**	**Young**	**Total**
Low	1,000	1,030	1,050	3,080
Medium	525	520	515	1,560
High	475	450	435	1,360
Total	2,000	2,000	2,000	6,000

   a. State the null and alternative hypotheses.

Null Hypothesis (H_0): There is no relationship between number of Web site visits and age.

Alternate Hypothesis (H_1): There is a relationship between number of Web site visits and age

   b. To test the null hypothesis, imagine that you calculate the chi-square statistic for independence based on the data shown, obtaining a value of 3.13. Can you reject the null hypothesis at the 0.05 level?

Degrees of freedom = (3-1) + (3-1) = 2 + 2 = 4
Critical chi-square for 4 degrees of freedom = 9.49 at α = 0.05 (5%).

Since the calculated chi-square (3.13) is less than the critical value (9.49), the null hypothesis would not be rejected. That is, we can't rule out the possibility that age and Web site visits are unrelated.

2. As an avid blogger of politics and health care, you recently surveyed 100 of your readers to identify any potential relationships between political affiliation and physical health. You came up with the following data:

	Health Status			
Affiliation	**Good**	**Medium**	**Poor**	**Total**
Democrat	25	10	5	40
Republican	10	35	5	50
Independent	5	5	0	10
Total	40	50	10	100

   a. State the null and alternative hypotheses.

Null Hypothesis (H_0): There is no relationship between affiliation and health.

Alternate Hypothesis (H_1): There is a relationship between affiliation and health.

   b. Calculate the chi-squared statistic and test the above hypothesis at the 0.01 level.

First, we must find the expected frequencies for each cell, as if there were no relationship between affiliation and health:

This new table is shown here:

	Health Status			
Affiliation	**Good**	**Medium**	**Poor**	**Total**
Democrat	16	20	4	40
Republican	20	25	5	50
Independent	4	5	1	10
Total	40	50	10	100

As an example of how we arrived at these numbers, let's look at the most upper-left cell of 16. To arrive at 16, we follow a simple 2-step process:

1. **Determine that Democrats represent 40% of the total sample here (40 Democrats / 100 total = 40%).**
2. **Multiply the actual number of Democrats (40) by that 40%. So, 40 x 40% = 16.**

We can apply this method to each of the cells, arriving at the table above.

Second, we now calculate the chi-square by the following equation:

$$\sum_{1}^{9} \frac{(\mathbf{Oi} - \mathbf{Ei})^2}{\mathbf{Ei}}$$

$$X2 = \sum_{1}^{9} \frac{(Oi - Ei)^2}{E_i} = \frac{(25 - 16)^2}{16} + \frac{(10 - 20)^2}{20} + \frac{(5 - 4)^2}{4} + \frac{(10 - 20)^2}{20} + \frac{(35 - 25)^2}{25} + \frac{(5 - 5)^2}{5} + \frac{(5 - 4)^2}{4} + \frac{(5 - 5)^2}{5} + \frac{(0 - 1)^2}{1} = 20.56$$

Third, we now compare the value above, 20.56 to the appropriate value of chi-square in a table. The number of degrees of freedom is (3-1) + (3-1) = 4, and the confidence level is set at 0.01. This corresponds to 13.28 in the table. Since 20.56 > 13.28, then we can reject the null. That is, we believe that there is a relationship between political affiliation and health.