Assignment #6

Due on Wednesday, November 19, 2008

Read Chapter 5 on *Goodness of Fit* in the class notes at the course webpage at `http://pages.pomona.edu/~ajr04747`

Read Section 9.3 on *Goodness of Fit* in Moore, McCabe and Craig.

Do the following problems

1. The ages of the hourly paid workers at Westcaco involved in the second round of layoffs that the Envelope Division of the company went through in 1991 are listed here below in increasing order.

$$25, 33, 35, 38, 48, 55, \underline{55}, \underline{55}, 56, \underline{64}$$

The underlined numbers are the ages of the workers that were laid off in the second round.

(a) Use the median of the ages of the workers that were laid off as a threshold to separate the workers into two classes: those whose age is above or equal to that value and those whose age is below the threshold. Based on this splitting, complete the Table 1

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes	2		
No	5		
Total			

Table 1: Observed Values

(b) If the company did the selection at random, how many ages would you expect to see in each category in the table? Make a table like that shown in Table 2 in which the expected values are displayed.

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes			
No			
Total			

Table 2: Expected Values

(c) Compute the Chi–Squared statistic

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

based on the observed and expected counts in the previous two parts of this problem.

2. Refer to the setup given in the previous problem.

   Testing the hypothesis that the assignments of ages to the fired or not fired column was done at random.

   (a) Use R to perform 10,000 simulations of random assignments of the ages of the 10 workers to the fired or not fired columns. Each one of these simulations will generate values that can be assigned to the cells in a 2×2 table. For example, one such simulation might yield the values displayed in Table 3 below.

Age$\geqslant$ threshold? \ Fired?	No	Yes	Total
Yes	4	1	5
No	3	2	5
Total	7	3	10

Table 3: Simulated values

   (b) Compute the Chi–Squared statistic for each one of the 10,000 simulations in the previous part and store them in a vector called `ChiSqr`. Plot a histogram of `ChiSqr`.

   (c) Use the `pHat()` function that we defined previously in the course to estimate the p–value for the test. What do you conclude?

3. The `rnorm(n,m,sd)` command in R generates n random numbers that should look like they are coming from a normal distribution with expected value `m` and standard deviation `sd`.

   (a) Generate 400 random numbers that look like they are coming from the standard normal distribution. Categorize the random numbers into five groups: Group 1: values $\leqslant -0.6$, Group 2: $-0.6 <$ values $\leqslant -0.1$, Group 3: $-0.1 <$ values $\leqslant 0.1$, Group 4: $0.1 <$ values $\leqslant 0.6$, and Group 5: values > 0.6. Determine the counts for each category and put them in a table like Table 4.

Group	1	2	3	4	5
Count					

Table 4: Standard normal counts

(b) Compute the expected values for each category.

4. Refer to the set up given in the previous problem.

   Use R to generate 10,000 random samples of size 400 from a standard normal distribution like the one you generated in the previous problem. For each one of the samples compute the Chi–Squared statistic and store the results in a vector called `ChiSqr`.

   (a) Plot a density histogram of `ChiSqr`.

   (b) Use the command `lines(density(ChiSqr),col="red")` to overlay a plot of the density function obtained in the previous part.

   (c) Estimate the p–value associated with a test of the hypothesis that the values obtained in the previous problem truly came from a standard normal distribution. What do you conclude?

5. (Is there a random distribution of trees?[1]) The Wade Tract in Thomas County, Georgia, is an old–growth forest if long–leaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. Foresters who study the trees are interested in how the trees are distributed in the forest. Is the distribution of trees random? We can examine this question by dividing the tract into four equal parts, or quadrants, in the east–west direction. Call the four parts Q_1, Q_2, Q_3 and Q_4. Suppose we take a random sample of 100 trees and count the number of trees in each quadrant. The data that we obtain is shown in the table below:

Quadrant	Q_1	Q_2	Q_3	Q_4
Count	18	22	39	21

   Perform a goodness of fit test for these data to determine if the trees in the sample are randomly scattered. Explain the reasoning, methodology and assumptions that you use to perform the test.

[1] Adapted from Exercise 9.40 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 557