Solutions to Assignment #6

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

Solution: Going through the 10 ranked ages and asking the question: "Is the age 55 or higher?" we find that the 10 can be grouped into two groups depending on whether the answer to the question was "yes" (Y), or "no" (N).

$$\text{N, N, N, N, N, Y, Y, Y, Y, Y.}$$

Out of the group labeled Y, 3 were laid off, while out of the group labeled N, none were laid off. We then obtain the values shown in Table 2.

□

(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 3 in which the expected values are displayed.

Age⩾ threshold? \ Fired?	No	Yes	Total
Yes	2	3	5
No	5	0	5
Total	7	3	10

Table 2: Observed Values

Age⩾ threshold? \ Fired?	No	Yes	Total
Older	3.5	1.5	5
Younger	3.5	1.5	5
Total	7	3	10

Table 3: Expected Values

***Solution**:* Let X denote the number of workers in group Y that get selected for layoff in a random sample of size 3. The possible values for X are 0, 1, 2 and 3. To find the probability distribution for X, we compute

$$P(X = 0) = \frac{\binom{5}{3}}{\binom{10}{3}} = \frac{1}{12};$$

$$P(X = 1) = \frac{\binom{5}{1}\binom{5}{2}}{\binom{10}{3}} = \frac{5}{12};$$

$$P(X = 2) = \frac{\binom{5}{2}\binom{5}{1}}{\binom{10}{3}} = \frac{5}{12};$$

$$P(X = 3) = \frac{\binom{5}{3}\binom{5}{0}}{\binom{10}{3}} = \frac{1}{12}.$$

We then obtain the probability distribution for X to be

$$P(X=k)=\begin{cases}1/12 & \text{if } k=0;\\ 5/12 & \text{if } k=1;\\ 5/12 & \text{if } k=2;\\ 1/12 & \text{if } k=3.\end{cases} \tag{1}$$

The expected value for this random variable is

$$\begin{aligned} E(X) &= 0p_{_X}(0)+1p_{_X}(1)+2p_{_X}(2)+3p_{_X}(3)\\ &= \frac{5}{12}+2\frac{5}{12}+3\frac{1}{12}\\ &= \frac{18}{12}=\frac{3}{2}, \end{aligned}$$

or 1.5. Thus, the entry in the “Yes” column and “Yes” row in Table 3 is 1.5. Similarly, the entry in the “Yes” column and “No” row in the table should be 1.5.

To find the entries in the “No” column of the table, we may proceed as in the previous part of the solution, or we may reason as follows: Seven out of the 10 workers get selected at random to keep their jobs. Since there are an equal number of workers of age 55 or above as there are workers below that age, there is a 1/2 chance for a worker selected to keep her or his job to be under the age of 55. Thus, on average, we expect $\dfrac{1}{2}\cdot 7$ workers selected to keep their jobs to be under 55. A similar reasoning leads to 3.5 workers selected to keep their jobs to be 55 or above. We then get the values shown in Table 3.

□

Alternate Solution: We could also have obtained Table 3 as follows:

The entry in each cell of the table is obtained by multiplying the column total and row totals for the cell and dividing by the grand total for the table. For example, the entry for the cell in the first column and first row is

$$\frac{7\cdot 5}{10}=3.5$$

and the entry in the second column and second row is

$$\frac{3 \cdot 5}{10} = 1.5$$

□

(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.

Solution: Compute

$$X^2 = \frac{(2-3.5)^2}{3.5} + \frac{(3-1.5)^2}{1.5} + \frac{(5-3.5)^2}{3.5} + \frac{(0-1.5)^2}{1.5}$$

which is about 4.2857. □

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 4 below.

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

Table 4: Simulated values

Solution: We can select random samples of size 3, without replacement, from the vector

```
hourly2 <- c(25,33,35,38,48,55,55,55,56,64)
```

using the R command

```
LaidOff <- sample(hourly2,3,replace=F)
```

This simulates selecting three working at random to be laid off. Counting the ages in the sample LaidOff that are 55 or higher yields the entry in the "Yes" column and "Yes" row entry in Table 4 of simulated values. The other entries in the table may be obtained as follows:
Define an array, Sim, of simulated values as follows by typing

```
Sim <- array(dim=4)
```

Sim will be a vector where the first two entries are the entries in the first row, and the last two entries are the entries in the second row of the table.
Next, count the number of ages in LaidOff that are 55 or higher by means of the code:

```
Sim <- c(0,0,0,0)
L <- length(LaidOff)
for (i in 1:L)
{
    Sim[2] <- Sim[2] + (LaidOff[i]>=55)
}
```

The other entries in Sim are then obtained as follows:

```
Sim[1] <- 5-Sim[2]
Sim[3] <- 7-Sim[1]
Sim[4] <- 3-Sim[2]
```

This procedure can then be repeated as many times as needed. □

(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.

Solution: Putting together the R commands in the previous part, we obtain a code that can be used to generate ChiSqr:

```
hourly2 <- c(25,33,35,38,48,55,55,55,56,64)
Nrep <- 10000
Exp <- c(3.5,1.5,3.5,1.5)
ChiSqr <- array(dim=Nrep)
for (k in 1:Nrep)
{
    Sim <- c(0,0,0,0)
    LaidOff <- sample(hourly2,3,replace=F)
```

```
        L <- length(LaidOff)
        for (i in 1:L)
        {
            Sim[2] <- Sim[2] + (LaidOff[i]>=55)
        }
        Sim[1] <- 5-Sim[2]
        Sim[3] <- 7-Sim[1]
        Sim[4] <- 3-Sim[2]
        ChiSqr[k] <- sum((Sim-Exp)^2/Exp)
    }
```

A histogram of `ChiSqr` is shown in Figure 1.

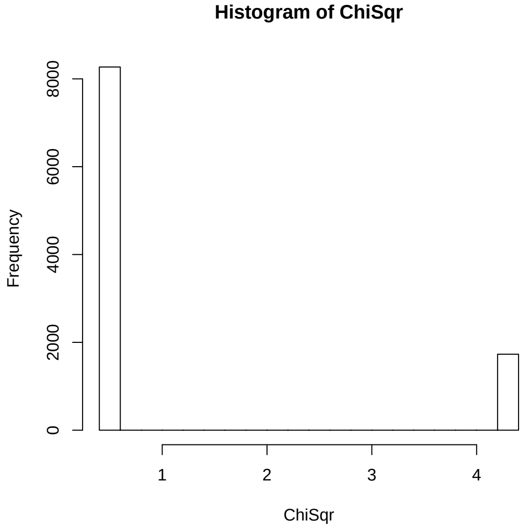

Figure 1: Histogram of `ChiSqr`

The histogram suggests that the Chi–Squared statistic has only two values. Typing `table(ChiSqr)` in R yields the table of counts:

```
> table(ChiSqr)
ChiSqr
0.476190476190476  4.28571428571429
             8270              1730
```

So, X^2 takes on, approximately, the values: 0.4762 and 4.2857. The approximate probabilities for these values are therefore 0.827 and 0.173, respectively. □

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

> ***Solution:*** p–value $\approx$ `pHat(ChiSqr,4.2857)` ≈ 0.173 or 17.3%. There is high chance that Chi–Squared statistic is 4.2857 or higher; consequently, the data do not provide enough evidence to reject the hypothesis that the company did the section at random. □

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 5.

Group	1	2	3	4	5
Count					

Table 5: Standard normal counts

Solution: Type `rnsn <- rnorm(400,0,1)` in R to obtain a sample of 400 random numbers from a standard normal distribution which are stored in a vector called `rnsn`.
To obtain the counts in each group, run the code:

```
Group <- rep(0,5)
L <- length(rnsn)
# Loop to get Group 1
for (i in 1:L)
{
    Group[1] <- Group[1] + (rnsn[i]<=-0.6)
}
#
# Loop to get Group 2
for (i in 1:L)
{
    Group[2] <- Group[2] + (rnsn[i]>-0.6 & rnsn[i]<=-0.1)
}
#
# Loop to get Group 3
for (i in 1:L)
{
    Group[3] <- Group[3] + (rnsn[i]>-0.1 & rnsn[i]<= 0.1)
}
#
# Loop to get Group 4
for (i in 1:L)
{
    Group[4] <- Group[4] + (rnsn[i]> 0.1 & rnsn[i]<= 0.6)
}
# We don’t need to loop to get Group 5
Group[5] <- L - sum(Group[1:4])
```

The entries in the vector `Group` are

```
 94  77  31  75 123
```

Putting this values in Table 5, we obtain the table

Group	1	2	3	4	5
Count	94	77	31	75	123

Table 6: Standard normal counts

□

(b) Compute the expected values for each category.

Solution: Since the values in the random sample are supposed to come from the standard normal distribution, Z, the probabilities that the lie in a given group are estimated in R by

$$\begin{aligned} p_1 &= P(Z \leqslant -0.6) \\ &\approx \texttt{pnorm(-0.6,0,1)} \\ &\approx 0.27425, \end{aligned}$$

$$\begin{aligned} p_2 &= P(-0.6 < Z \leqslant -0.1) \\ &\approx \texttt{pnorm(-0.1,0,1)-pnorm(-0.6,0,1)} \\ &\approx 0.18592, \end{aligned}$$

$$\begin{aligned} p_3 &= P(-0.1 < Z \leqslant 0.1) \\ &\approx \texttt{pnorm(0.1,0,1)-pnorm(-0.1,0,1)} \\ &\approx 0.07966. \end{aligned}$$

By the symmetry of the standard normal distribution, we get that

$$\begin{aligned} p_4 &= 0.18592 \\ p_5 &= 0.27425. \end{aligned}$$

To find the expected numbers in each group, we multiply the respective probabilities by $n = 400$. In R, we may write

```
n <- 400
Prop <- c(0.27425, 0.18592,0.07966,0.18592,0.27425)
Exp <- round(n*Prop)
```

This yields expected values

```
110  74  32  74 110
```

which are displayed in Table 7. □

Group	1	2	3	4	5
Expected	110	74	32	74	110

Table 7: Expected Values

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`.

   ***Solution**:* We repeat the procedure outlined in Problem 3(a) `Nrep =10000` times. Here is the code:

```
# Author: Adolfo J. Rumbos
#
# Date: November 16, 2008
#
    # Simulations for problem 4 in Assignment 6
#
    Nrep <- 10000 n <- 400
    Prop <- c(0.27425, 0.18592,0.07966,0.18592,0.27425)
        # Proportions predicted by standard normal assumption
    Exp <- round(n*Prop)
        # Computes expected values based on Prop and n
    ChiSqr <- array(dim=Nrep)
        # sets up array ChiSqr
    for (k in 1:Nrep)
    # sets up loop to compute ChiSqr
    {
        rnsn <- rnorm(n,0,1) # samples from standard normal
        Group <- rep(0,5)      # initializes Group counts
        # Loop to get Group 1
        for (i in 1:n)
        {
            Group[1] <- Group[1] + (rnsn[i]<=-0.6)
        }
    #
        # Loop to get Group 2
        for (i in 1:n)
        {
            Group[2] <- Group[2] + (rnsn[i]>-0.6 & rnsn[i]<=-0.1)
        }
#
        # Loop to get Group 3
        for (i in 1:n)
        {
            Group[3] <- Group[3] + (rnsn[i]>-0.1 & rnsn[i]<= 0.1)
        }
#
        # Loop to get Group 4
        for (i in 1:n)
        {
            Group[4] <- Group[4] + (rnsn[i]> 0.1 & rnsn[i]<= 0.6)
        }
        # We don’t need to loop to get Group 5
        Group[5] <- n - sum(Group[1:4])
        ChiSqr[k] <- sum((Group-Exp)^2/Exp)
            # Computes Chi-Squared Statistic
}
```

□

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

Solution: Type `hist(ChiSqr,freq=F, main = "Density Histogram` of ChiSqr) to obtain the density histogram shown in Figure 2.

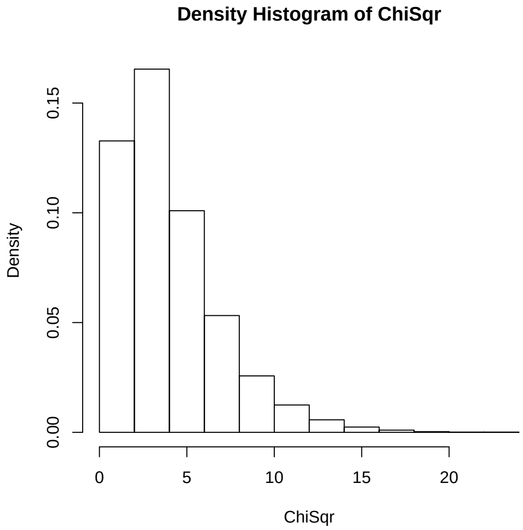

Figure 2: Density Histogram of ChiSqr

□

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

Solution: The density histogram along with density function is shown in Figure 3.

□

(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?

Solution: The Chi–Squared statistic value for the data in Table 6 is obtained in R as follows:
Define the vector of observed values by

```
Obs <- c(94,77,31,75,123)
```

The expected values were computed in the previous part; then,

$$X^2 = \text{ sum((Obs} - \text{Exp}) \wedge 2/\text{Exp}) \approx 4.03$$

□

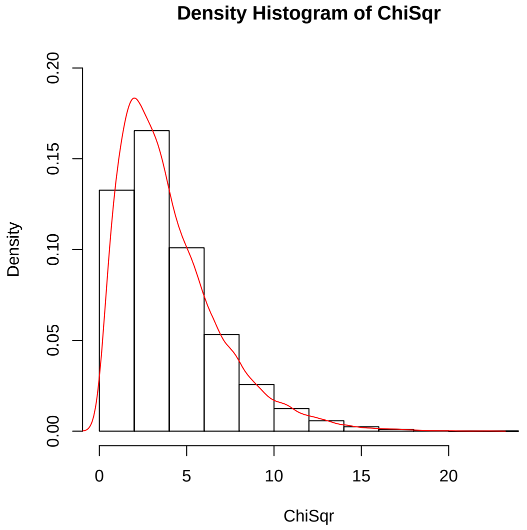

Figure 3: Density Function for `ChiSqr`

The p–value associated with the data in Table 6 is then the probability that the Chi–Squared statistic is 4.03 or above. This can be estimated applying the `pHat` function on `Chisqr`:

$$p\text{–value} \approx \text{pHat(ChiSqr,4.03)} \approx 0.40.$$

The p–value is therefore too high for us to reject the null hypothesis that the data obtained in problem 3(1) comes from a standard normal distribution.

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:

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

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.

***Solution**:* The null hypothesis in this case is that the trees are equally distributed in the four quadrants. Therefore the expected values are

```
Exp <- c(25,25,25,25)
```

The observed values are given in table:

```
Obs <- c(18,22,39,21)
```

The Chi–Squared test statistic is then

```
Xsqr <- sum((Obs-Exp)^2/Exp)
```

or 10.8.

To estimated the p–value, we may use R to collect samples of size 100, with replacement, form the vector

```
quadrants <- 1:4
```

made up of the digits 1 through 4, where each number corresponds to a quadrant. Each sample may be stored in a vector called `trees`:

```
trees <- sample(quadrants,100,replace = TRUE)
```

We then count how many trees are in each quadrant. This can be done in R using the `table` function. Typing `table(trees)` in R yields

```
trees
 1  2  3  4
25 20 28 27
```

Thus, three are 25 trees in the first quadrant, 20 in the second, 28 in the third and 27 in the fourth.

These values can be stored in an vector of simulated values, `Sim`, by typing

```
Sim <-  table(trees)
```

We can then obtain the value of the Chi–Squared statistic by typing

```
sum((Sim-Exp)^2/Exp)
```

This procedure can be repeated many times and the Chi–Squared values stored in a vector called `ChiSqr`. Here is the code:

```
# Author: Adolfo J. Rumbos
#
# Date: November 16, 2008
#
# Simulations for problem 5 in Assignment 6
# Nrep <- 10000
n <- 100
Prop <- c(0.25,0.25,0.25,0.25)
Exp <- round(n*Prop)
# Computes expected values based on Prop and n
quadrants <- 1:4          # Labels of groups
ChiSqr <- array(dim=Nrep) # sets up array ChiSqr
for (k in 1:Nrep)
    # sets up loop to compute ChiSqr
{
    trees <- sample(quadrants,100,replace = TRUE) # permutation
    Sim <-  table(trees) # Counts in each group
    ChiSqr[k] <- sum((Sim-Exp)^2/Exp) # Computes Chi-Squared Statistic
}
```

We can then use the `pHat` function to estimate the p–value; that is, the probability that we will see a Chi–Squared value of 10.8 or higher under the assumption that the tress distribute themselves equally among the quadrants. We obtain

$$p\text{–value} \approx \text{pHat(ChiSqr,10.8)} \approx 0.0131.$$

or 1.31%. We can therefore reject the null hypothesis at the 5% significance level and conclude that the trees are not randomly scattered.
□