Lord of the Rings Final Paper

Sophia McFadden-Keesling Jennifer Cho Dahwi Kim

June 02, 2017

Contents

1 Background

Rather than being born from love, the initial idea for the project was created from hate. I (Sophia) always disliked The Lord of the Rings (LOTR) series growing up. I explained this annoyance at LOTR and Tolkien's writing style stemming from his overuse of and focus on walking. While in literature it is argued that the journey a character (or characters) make is far more interesting than the end goal, I found the walking boring, tedious, and unnecessary at times. The rest of the original "we" in this group disagreed. Whether the walking in the series is necessary or not this project serves to show the frequency of words per page. The parameter we used is the rate a variation or reference to "walking" appears per page in each book. This parameter is represented as λ. (The list of words used for this project are found at the bottom of the page[1].)

2 Supporting Statistical Theory

We are assuming our data will have a Poisson distribution. We are interested in finding the mean probability of the number of times Tolkien references walking per page. The number of words referencing walking per page are independent of one another because of the assumed Poisson distribution. We are using Maximum Likelihood Estimator (MLE) for analyzing our data. The MLEs are those values of the parameter that maximize the likelihood function with respect to the parameter, λ, for each book.

We wanted to test that each estimator is unbiased. An estimator is unbiased if $E[\hat{\theta}] = \theta$. Since we assume our population is a Poisson distribution, we have $\hat{\theta} = \hat{\lambda} = \bar{X}$. Therefore, $E[\hat{\lambda}] = E[\bar{X}]$. Since $E[\bar{X}] = \mu$, we now conclude $E[\hat{\lambda}] = \lambda$.

$$Bias_1 = E[\hat{\lambda_1}] - \lambda_1 = \lambda_1 - \lambda_1 = 1.241 - 1.241 = 0.$$
$$Bias_2 = E[\hat{\lambda_2}] - \lambda_2 = \lambda_2 - \lambda_2 = 1.619 - 1.619 = 0.$$
$$Bias_3 = E[\hat{\lambda_3}] - \lambda_3 = \lambda_3 - \lambda_3 = 1.896 - 1.896 = 0.$$

As shown above, the estimator for each book is unbiased. Because our estimators are unbiased, MSE for our estimators are equal to the variance of our estimators.

$$MSE_1 = Var[\hat{\lambda_1}] = 1.326122$$
$$MSE_2 = Var[\hat{\lambda_2}] = 2.969796$$
$$MSE_3 = Var[\hat{\lambda_3}] = 3.614694$$
$$Var[\hat{\lambda_1}] < [\hat{\lambda_2}] < Var[\hat{\lambda_3}]$$

Notice that the MSE for the first estimator $\hat{\lambda_1}$ is the smallest, meaning Book 1's estimator is the most efficient.

The Poisson estimator is preferable than other possible estimators because we are interested in the rate at which words referencing walking occur on each page.

[1] walk, walks, walking, walked, run, runs, ran, ride, rode, rides, journey, journeys, march, marches, marched, pace, paces, paced, speed, sped, speeds, stroll, strolls

3 Data Collection Method

The first thing we did was download text files for each book from online. We programmed Java program that gets rid of all characters besides English alphabet letters and space between the words and writes out a new txt file that contains only letters (Appendix C). We used the Python programming language to read in the text file data (Appendix A). It read each line of the text file and found how many times the "walk" words appear per book. Since the text files were not broken into neat pages, we broke the text file into 500 word chunks because the average number of words per page in a book we assumed to be 500 words. Then it stores the number of occurrences on each page and exports the result into a CSV file, allowing us to collect our sample.

The population is the total number of times "walk" words appears throughout the whole book. We were dealing with three populations in total. We generated 50 random page numbers for each population using **R** programming language to form our sample of size $n = 50$ for each book. Then we got the number of occurrences per page of the 50 randomly selected pages for each population to calculate the sample mean (Appendix B), proved in Section 2 to be equivalent to estimated lambda when we assume a Poisson distribution.

4 Results

Table 1: Table of numerical summary of collected data

Book	Mean	Standard deviation	chi-square value	p-value
1	1.241	1.394	13.905	0.003036
2	1.619	1.753	20.544	0.0003898
3	1.896	1.857	24.044	0.0005126

Table 1 gives a summary of the population means of the three books individually, their standard deviation, their chi-squared value, and their p-values from chi-square goodness of fit test. For our p-value calculation, the null hypothesis was the population of walking words per page in a given book is a Poisson distribution, while the alternative hypothesis said it is not. We reject the null hypothesis for each book based on the p-values being less than 0.05 for each book.

Table 2: Table of statistical analysis of collected data

Book	Sample Mean	95% Confidence Interval
1	1.02	(0.7758,1.3409)
2	1.64	(1.3213,2.0354)
3	2.24	(1.8617,2.6950)

Table 2 shows the sample means of 50 randomly selected pages from each book. It also provides the 95% confidence intervals for the population parameter, λ, for each book. The actual population mean (1st Column, Table 1), λ, falls within the confidence interval

calculated. We know $\mu = \lambda$ because of the proof on page 155 that $\bar{x} = \hat{\lambda}$ for a Poisson distribution, which we assumed for this project.

Table 3: Table of result of bootstrapping the difference of means

Book	95% confidence interval	Actual mean difference
1 and 2	(-0.24, 1)	0.378
1 and 3	(0.02, 1.30)	0.655
2 and 3	(-0.44, 1)	0.277

Table 3 shows the 95% bootstrapped confidence interval of the difference in population means simulated in **R** with a sample size of 50. The actual population mean difference is shown above as well and falls within the confidence interval. Since the confidence intervals of Book 1/Book 2 and Book 2/Book 3 contains the value zero, we cannot leave out the possibility that $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$. Between Book 1 and Book 2, the confidence interval is skewed to the right and so we are 95% confident that Book 2 contains more walking in the story. For the same reason, we are 95% confident that Book 3 contains more walking in the story than Book 1. And lastly, we are 95% confident that Book 3 contains more walking than Book 2.

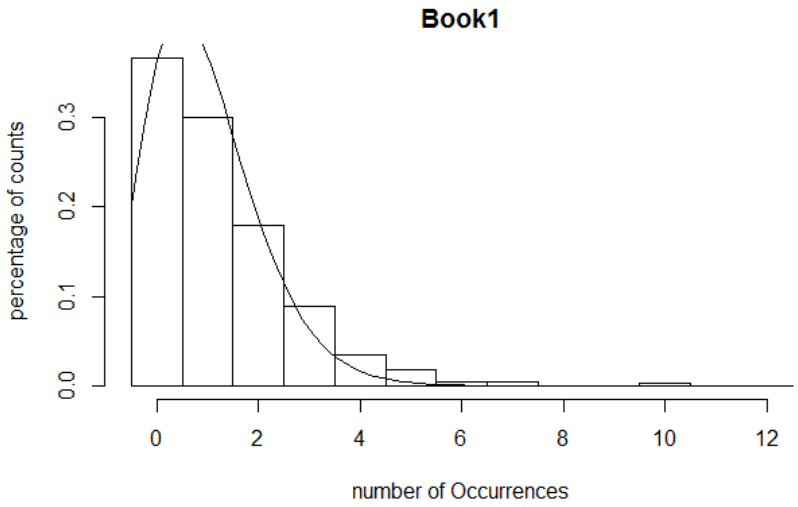

Figure 1: Histogram of the number of occurrences per page of the population of Book 1

Figure 1 represents the distribution of the population of walking words per page for Book 1. We are assuming a Poisson distribution, so the curved line is the Poisson distribution with lambda estimated for Book 1 from our sample of size $n = 50$. The curved line appears to be a really good fit on this Poisson graph, which is surprising because the chi-square goodness of fit test failed, showing that it is in fact not a Poisson distribution. The chi-squared is shown as failing for Book 1 because the p-value for the chi-squared goodness of fit is 0.003036 which is less than 0.05 (shown in Table 1). Therefore, we *reject* the null hypothesis that the population for Book 1 is a Poisson distribution.

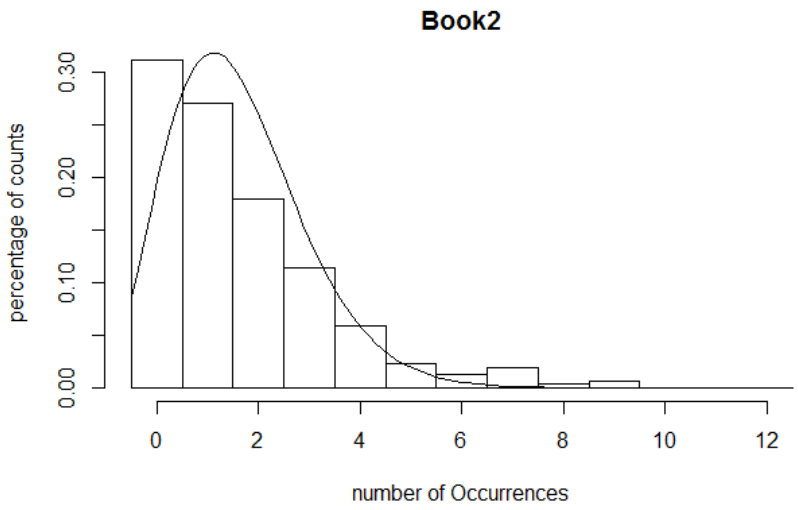

Figure 2: Histogram of the number of occurrences per page of the population of Book 2

Figure 2 represents the distribution of the population of walking words per page for Book 2. Assumed Poisson distribution so the curved line is the Poisson distribution with lambda estimated from our sample of size $n = 50$ for Book 2. The curved line appears to be a decent fit on this graph, which is surprising because the chi-square goodness of fit test failed since p-value for Book 2, 0.0003898, is less than 0.05 (shown in Table 2). Therefore, we *reject* the null hypothesis that the population of book 2 is a Poisson distribution.

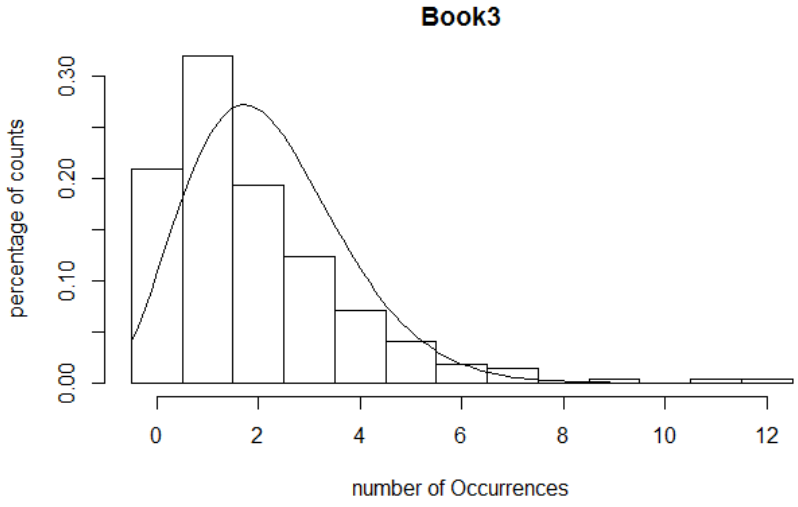

Figure 3: Histogram of the number of occurrences per page of the population of Book 3

Figure 3 represents the distribution of the population of walking words per page for Book 3. Again, we assumed Poisson distribution at the beginning so the curved line is the Poisson distribution with lambda estimated from our sample of size $n = 50$ for Book 3. The curved line appears to not be that great of a fit on this graph. The figure's rough fit is not all that

surprising because the chi-square goodness of fit test failed, showing that it is in fact not a Poisson distribution. Again, we know it failed because of the p-value being less than 0.05, it is 0.0005126 (shown in Table 3). Therefore, we *reject* the null hypothesis and conclude that the population of Book 3 is not a Poisson distribution.

5 Calculations

Assuming that the distribution of the number of occurrences of words referencing "walk" per page is a Poisson distribution, the estimator for each book is $\hat{\lambda}$. We proved the Maximum Likelihood Estimator (MLE) of the Poisson distribution to be $\hat{\lambda} = \bar{X}$. The proof is below:

$$\begin{aligned} f(x) &= P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \\ L(\lambda) &= f(x_1; \lambda)....f(x_{50}; \lambda) \\ &= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!}....\frac{\lambda^{x_{50}} e^{-\lambda}}{x_{50}!} \\ &= \frac{e^{-50\lambda}\lambda^{(x_1+.....+x_{50})}}{(x_1!)(x_2!)...(x_{50}!)} \end{aligned}$$

Taking the natural log of $L(\lambda)$ yields $ln(L(\lambda)) = -50\lambda + \ln(\lambda^{\sum x_i}) - \ln(x_1!....x_{50}!)$. Then we take a derivative with respect to λ and set it equal to zero to find the maximum:

$$\frac{d\ln(L(\lambda))}{d\lambda} = -50 + \frac{\sum x_i}{\lambda} = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum x_i}{50} = \bar{X} \text{ (since } n = 50)$$

Because we have shown $\hat{\lambda} = \bar{X}$ for the MLE of the Poisson Distribution, our estimated lambda is equal to our sample mean. Actual calculation for MLE was implemented using **R** programming language (Appendix B, Code 2).

Using the Central Limit Theorem (CLT) approximation to find a 95% confidence interval for the population parameter (λ) we were able to find the 95% confidence interval for lambda for each book. The work is shown below:

$$q = \frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}}$$

$$\Rightarrow \lambda - \bar{X} + q\sqrt{\frac{\lambda}{n}} = 0$$

Let $\lambda = y^2$, then we get

$$y^2 + y\frac{q}{\sqrt{n}} - \bar{X} = 0$$
$$\Rightarrow y = \frac{\frac{-q}{\sqrt{n}} \pm \sqrt{\frac{q^2}{n} + 4\bar{X}}}{2} = \sqrt{\lambda}$$
$$\Rightarrow \lambda = \frac{1}{4}(\frac{q^2}{n} \pm 2\sqrt{\frac{q^4}{n^2} + \frac{4q^2\bar{X}}{n}} + \frac{q^2}{n} + 4\bar{X}) = \bar{X} + \frac{q^2}{2n} \pm \sqrt{\frac{q^4}{4n^2} + \frac{q^2\bar{X}}{n}}$$

As shown above, we constructed a confidence interval for λ where $\bar{X}$ is the sample mean, n is the sample size, 50 in our case, q is a 0.975th quantile from standard normal distribution, 1.96. Actual calculation for confidence intervals of λ was implemented using **R** programming language (Appendix B, Code 3).

95% bootstrapped confidence interval of the differences in population means is calculated using **R** programming language (Appendix B, Code 4).

While calculating the chi-squared goodness of fit we needed expected values and observed values. We wanted our expected values to be greater than 5. Because some of our expected values were less than 5 in the lower tails we added up all the values in the lower tails so that our rightmost value (the smallest value) became greater than 5. We manipulated the observed data corresponding to the expected data so that the length of observed and expected were the same. Then we calculated the p-value using **R** programming language (Appendix B, Code 5).

6 Conclusions

In the end, our final project was slightly different from the original proposal. Instead of only using Book 1 and 2 and comparing the MLEs for those books as compared to the series as a whole, we took the MLEs for all three books and compared them to each other.

Looking at our results, the MLE for each book was unexpected as the Book 3 had the greatest MLE, $\hat{\lambda_3} = 2.2$. This was surprising because we expected the second book to have the largest MLE as it is in the middle of the three stories about journeys and therefore should be the center of walking, but it was not. What does fit our thought process is that the MLE for Book 1 was the smallest, $\hat{\lambda_1} = 1.02$.

The population means of the books, which are the population parameters (λ), fall within our calculated 95% confidence interval using Central Limit Theorem, sample mean and sample size, $n = 50$. The 95% Confidence Intervals for Book 1,2, and 3 can be seen in Table 2, and are respectively $(0.7758, 1.3409), (1.3213, 2.0354)$, and $(1.8617, 2.6950)$. The population means for each book are found in Table 1: $\mu_1 = 1.241, \mu_2 = 1.619, \text{and} \mu_3 = 1.896$. As can plainly be seen the population means fall into the 95% Confidence Intervals for each book.

The 95% confidence interval for the difference in population means between Book 1 and Book 2, $(-0.24, 1)$, is skewed more to the positive side. Thus, we can conclude that we are 95% confident there is more walking within the pages of Book 2. Similarly, the 95% confidence interval for the difference in population means between Book 1 and Book 3, $(0.02, 1.30)$, does not include zero, so we can conclude that we are 95% confident there is

more walking in Book 3 than Book 1. Lastly, the 95% confidence interval for the difference in population means between Book 2 and 3, $(-0.44, 1)$, is skewed more to the positive side as well. Thus, we are 95% confident there is more walking involved in Book 3 than Book 2. Since both the confidence interval of Book 1/Book 2 and Book 2/Book 3 contains zero, we cannot rule out the possibility that the difference in population means are the same.

The results from the confidence interval was expected because we found that the mean value of the population data of Book 3, $\mu_3 = 1.896$, was greater than the mean population data of both Book 1 and Book 2 which are $\mu_1 = 1.241$ and $\mu_2 = 1.619$ respectively.

As for limitations, although we randomly sampled the population of the books our data is not truly random. It is not truly random because of the purpose of the sampling in the first place. We sampled the words associated with "walk" in order to prove a point. If we had sampled a more random word like "the" or "is" we would have been using a more randomly spread out word. Random in the sense of how it is used in the English language. Because the books are made up of stories the word "walk" is going to be clumped together in certain sections of the story that involve a journey. There will be clumps of where the words are found most often making the data we collected from be considered not as random as we originally hoped.

A Data Collection Source Code - Python

Code 1: DataReader.py source code.

```
1  import numpy as np
2  import re
3  import codecs
4
5  def readin(data, list):
6          newlist = [word.lower() for word in list]
7          #ans = dict()
8          freqTable = []
9          with open(data, "r", encoding = 'utf-8') as fo:
10                 count = 0
11                 freq = 0
12                 for line in fo:
13                         #freq = 0
14                         # if count == 500:
15                                 # freqTable.append(freq)
16                                 # count = 0
17                                 # freq = 0
18                         #split the sentence when there is space or
                               new line
19                         #count += 1
20                         a = re.compile("[ !\\.,\\?;\\-:\\(\\)\\\\]
                               ")
21                         newdata=a.split(line)
22                         #print(len(newdata))
23                         temp = []
24                         for val in newdata:
25                                 val = val.lower()
26                                 if count == 500:
27                                         #print(val)
28                                         freqTable.append(freq)
29                                         count = 0
30                                         freq = 0
31                                 count += 1
32                                 if val in newlist:
33                                         freq += 1
34                         #freqTable.append(freq)
35         return freqTable
36
37
38 data="fotr-fixed2.txt"
39 data2="twoTowers-fixed.txt"
```

```
40 data3="Returnfixed1.txt"
41 list = ["walk","walking","walked","run","runs","walks","ran","ride
       ","rode","rides","journey","journeys","march","marches","
       marched","pace","paces","paced","speed","speeds","sped","stroll
       ","strolls","strolled"]
42
43
44 output = open("result1.csv", "w")
45 output.write("pageNum, numOfOccur\n")
46 ans = readin(data, list)
47 counter = 1
48 for i in ans:
49         output.write(str(counter)+","+str(i)+"\n")
50         counter += 1
51
52 output2 = open("result2.csv", "w")
53 output2.write("pageNum, numOfOccur\n")
54 ans2 = readin(data2, list)
55 counter = 1
56 for i in ans2:
57         output2.write(str(counter)+","+str(i)+"\n")
58         counter += 1
59
60 output3 = open("result3.csv", "w")
61 output3.write("pageNum, numOfOccur\n")
62 ans3 = readin(data3, list)
63 counter = 1
64 for i in ans3:
65         output3.write(str(counter)+","+str(i)+"\n")
66         counter += 1
```

B Data Analysis Source Code - R

Code 2: randomPage.R source code.

```
1 data1 <- read.csv("result1.csv")
2 data2 <- read.csv("result2.csv")
3 data3 <- read.csv("result3.csv")
4
5 #get 50 random pages from each book and estimate the parameter(
      lambda)
6 set.seed(987654)
7
8 BK1 = sample(1:403,50)
```

```
sample1<-data1[BK1,]
lambda1<-mean(sample1$numOfOccur)

BK2 = sample(1:307,50)
sample2<-data2[BK2,]
lambda2<-mean(sample2$numOfOccur)

BK3 = sample(1:269,50)
sample3<-data3$numOfOccur[BK3]
lambda3<-mean(sample3)
```


Code 3: confint.R source code.

```
confInt<-function(x,n)
{
  q<-qnorm(0.975)
  right<-x+q^2/(2*n)+sqrt(q^4/(4*n^2)+q^2*x/n)
  left<- x+q^2/(2*n)-sqrt(q^4/(4*n^2)+q^2*x/n)
  CI<-c(left,right)
  CI
}

#95% CI of parameter(lambda) of book 1
confInt(lambda1,50)

#95% CI of parameter(lambda) of book 2
confInt(lambda2,50)

#95% CI of parameter(lambda) of book 3
confInt(lambda3,50)
```

Code 4: bootstrap.R source code.

```
Book1 <- data1$numOfOccur
Book2 <- data2$numOfOccur
Book3 <- data3$numOfOccur

N <- 10^4

#Difference in means between Book 1 and Book 2
word.diff.mean <- numeric(N)

for(i in 1:N)
{
  Book1.sample <- sample(Book1, 50, replace = TRUE)
```

```
13    Book2.sample <- sample(Book2, 50, replace = TRUE)
14    word.diff.mean[i] <- mean(Book2.sample)-mean(Book1.sample)
15  }
16
17  hist(word.diff.mean)
18  abline(v=mean(Book2)-mean(Book1), col ="red") #observed main
        difference
19  diff.mean <- mean(Book2.sample)-mean(Book1.sample)
20  diff.mean
21  CI.95 <- quantile(word.diff.mean, c(0.025, 0.975))
22  CI.95
23
24  #Difference in means between Book 1 and Book 3
25  word.diff.mean2 <- numeric(N)
26
27  for(i in 1:N)
28  {
29    Book1.sample <- sample(Book1, 50, replace = TRUE)
30    Book3.sample <- sample(Book3, 50, replace = TRUE)
31    word.diff.mean2[i] <- mean(Book3.sample)-mean(Book1.sample)
32  }
33
34  hist(word.diff.mean)
35  abline(v=mean(Book3)-mean(Book1), col ="red") #observed mean
        difference
36  diff.mean <- mean(Book3.sample)-mean(Book1.sample)
37  diff.mean
38  CI.95 <- quantile(word.diff.mean2, c(0.025, 0.975))
39  CI.95
40
41  #Difference in means between Book 2 and Book 3
42  word.diff.mean3 <- numeric(N)
43
44  for(i in 1:N)
45  {
46    Book2.sample <- sample(Book2, 50, replace = TRUE)
47    Book3.sample <- sample(Book3, 50, replace = TRUE)
48    word.diff.mean3[i] <- mean(Book3.sample)-mean(Book2.sample)
49  }
50
51  hist(word.diff.mean)
52  abline(v=mean(Book3)-mean(Book2), col ="red") #observed mean
        difference
53  diff.mean <- mean(Book3.sample)-mean(Book2.sample)
54  diff.mean
```

```
CI.95 <- quantile(word.diff.mean3, c(0.025, 0.975))
CI.95
```

Code 5: goodnessOfFit.R source code.

```
#Book1
hist(data1$numOfOccur, prob=T, breaks=seq(-0.5,12.5,by=1),main = "Book1",xlab="number of Occurrences",ylab="percentage of counts")
#plot a probability density function of Poission of estimated lambda(1)
curve((1.02)^x/exp(1.02)/factorial(x),add=TRUE)

observed <- hist(data1$numOfOccur, breaks=seq(-0.5,12.5,by=1),plot=F)$counts

expected <- dpois(0:11, 1.02)
expected[12] <- 1-sum(expected[1:11])
counts.expected<-expected*sum(observed)
counts.expected
counts.expected[5]<-sum(counts.expected[5:length(counts.expected)])
counts.expected <- counts.expected[1:5]
counts.expected

observed[5] <- sum(observed[5:length(observed)])
observed <- observed[1:5]
observed

table1<-rbind(observed,counts.expected)
table1
#chi-square goodness of fit test for book 1
chisq(table1)
pchisq(chisq(table1), length(table1[1,])-2, lower.tail = FALSE) #df=k-l-1 since the parameter(lambda) is estimated from the data


#Book2
hist(data2$numOfOccur, prob=T, breaks=seq(-0.5,12.5,by=1),main = "Book2",xlab="number of Occurrences",ylab="percentage of counts")
#plot a probability density function of Poission of estimated lambda(1.7)
curve((1.64)^x/exp(1.64)/factorial(x),add=TRUE)

```

```
observed <- hist(data2$numOfOccur, breaks=seq(-0.5,12.5,by=1),plot=F)$counts
expected <- dpois(0:11, 1.64)
expected[12] <- 1-sum(expected[1:11])
counts.expected<-expected*sum(observed)
counts.expected
counts.expected[6]<-sum(counts.expected[6:length(counts.expected)])
counts.expected<-counts.expected[1:6]
counts.expected

observed[6] <- sum(observed[6:length(observed)])
observed <- observed[1:6]

table2<-rbind(observed,counts.expected)
table2
#chi-square goodness of fit test for book 2
chisq(table2)
pchisq(chisq(table2), length(table2[1,])-2, lower.tail = FALSE) # df=k-l-1 since the parameter(lambda) is estimated from the data


#Book3
hist(data3$numOfOccur, prob=T, breaks=seq(-0.5,12.5,by=1),main = "Book3",xlab="number of Occurrences",ylab="percentage of counts")
#plot a probability density function of Poission of estimated lambda(2.2)
curve((2.24)^x/exp(2.24)/factorial(x), add=TRUE)

#chi-square test for book 3
observed <- hist(data3$numOfOccur, breaks=seq(-0.5,12.5,by=1),plot=F)$counts
observed

expected <- dpois(0:11, 2.24)
expected[12] <- 1-sum(expected[1:11])
counts.expected<-expected*sum(observed)
counts.expected
counts.expected[7]<-sum(counts.expected[7:length(counts.expected)])
counts.expected<-counts.expected[1:7]
counts.expected

observed[7] <- sum(observed[7:length(observed)])
```

```
observed <- observed[1:7]

table3<-rbind(observed,counts.expected)
table3
#chi-square goodness of fit test for book 3
chisq(table3)
pchisq(chisq(table3), length(table3[1,])-2, lower.tail = FALSE) #
    df=k-l-1 since the parameter(lambda) is estimated from the data
```

C Data Collection Source Code - Java

Code 6: Fix.java source code.

```
import java.io.*;
import java.util.*;
import java.util.regex.*;

public class Fix
{
        public static void main(String[] args) throws IOException
        {
                //read in the txt file
                BufferedReader fin = new BufferedReader(new
                   FileReader("Return.txt"));
                //write out the new txt file
                PrintWriter fout = new PrintWriter(new FileWriter(
                   "Returnfixed1.txt"));

                String line;
                while ((line = fin.readLine()) != null)
                {
                        //if the line is empty, ignore it
                        if (line.equals(""))
                                continue;
                        //replace every other character besides
                           alphabet letters and space into blank
                        line = line.replaceAll("[^A-Za-z ]","");
                        //replace new line into blank
                        line = line.replaceAll("  ","");
                        //print out the replaced line
                        fout.println(line);
                }
                fout.close();
                fin.close();
```

```
29              }
30  }
```