Dr Isabella Gollini

Probability and Statistics

Solution Extra Examples – Lab – Goodness of fit

1 Birthdays – Adapted from 2013 Exam

The birthdays of a random sample of 200 students in a college were found to fall in the quarters of the year as follows:

Quarter	1	2	3	4
Frequency	62	48	44	46

(a) Plot the data.

(b) Write down the expected frequencies under the null hypothesis that the students are drawn from a population in which birthdays are uniformly distributed throughout the year.

(c) State the approximate distribution of the test statistic under the null hypothesis.

(d) Carry out the goodness of fit test and draw conclusions.

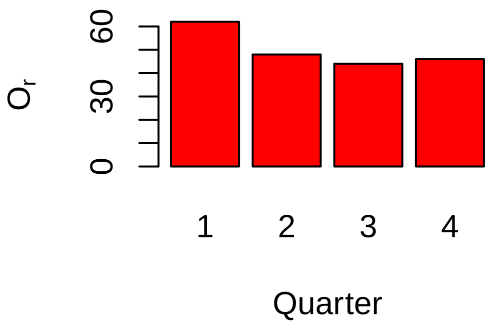
(e) Interpret your conclusion in the context of the application.

Solution:

(a)

```r
Quarter <- 1:4
Or <- c(62, 48, 44, 46)
n <- sum(Or)

barplot(Or,
  names = Quarter,
  xlab = "Quarter",
  ylab = expression(O[r]),
  col = "red")
```



(b) The expected frequencies under the null hypothesis are 50 for each quarter.

```
pr <- rep(1 / 4, 4)
Er <- pr * n

Er

## [1] 50 50 50 50
```

(c) The chi-square test statistic is

$$X^2 = \frac{(62-50)^2}{50} + \frac{(48-50)^2}{50} + \frac{(44-50)^2}{50} + \frac{(46-50)^2}{50} = 4.00.$$

Under the null hypothesis it has the chi-square distribution with 3 degrees of freedom.

(d) Using Table 7 of Lindley and Scott, $p = 1-F(4) = 1-0.7385 = 0.2615$. The p-value is not significant at the 5% level. Hence, there is no strong evidence to reject the null hypothesis that birthdays are uniformly distributed throughout the year.

Or with `R`:

```
test <- chisq.test(Or)
test

##
##  Chi-squared test for given probabilities
##
## data:  Or
## X-squared = 4, df = 3, p-value = 0.2615
```

or

```
test <- chisq.test(Or,
  p = pr)
test

##
##  Chi-squared test for given probabilities
##
## data:  Or
## X-squared = 4, df = 3, p-value = 0.2615
```

2 Homicides in London

Spiegelhalter and Barnett (2009)[1] in a paper in the statistics magazine *Significance* (`https://www.statslife.org.uk/significance`) examined data on the daily numbers of homicides in London for the 1095 days over the three year period from April 2004 to March 2007. The frequency distribution of the daily numbers of homicides is given in the table below.

Number of homicides	Frequency
0	713
1	299
2	66
3	16
4	1
≥ 5	0

Frequency distribution of daily numbers of homicides

(a) Calculate the sample mean and sample variance of the daily number of homicides.

(b) Carry out a chi-square goodness of fit test to test the hypothesis that the data may be regarded as a random sample from a Poisson distribution. Draw conclusions. [Pay attention to the number of degrees of freedom of the approximate distribution of the test statistic under the null hypothesis].

(c) As an alternative to the method of part (b), carry out a dispersion test to test the same hypothesis. Draw conclusions.

(d) Plot the observed and the expected frequencies.

(e) Interpret your conclusion in the context of the application.

Solution:

(a) $n = 1095$ days.

The sample mean is:

$$\begin{aligned}\bar{x} &= \frac{\sum_{r=0}^{4} r \times O_r}{1095} \\ &= \frac{0 \times 713 + 1 \times 299 + 2 \times 66 + 3 \times 16 + 4 \times 1}{1095} \\ &= 0.441\end{aligned}$$

homicides per day.

The sample variance is:

[1] `http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2009.00334.x/abstract`

$$
\begin{aligned}
s^2 &= \frac{\sum_{r=0}^{4} r^2 \times O_r - 1095 \times \bar{x}^2}{1094} \\
&= \frac{0^2 \times 713 + 1^2 \times 299 + 2^2 \times 66 + 3^2 \times 16 + 4^2 \times 1 - 1095 \times (0.441)^2}{1094} \\
&= \frac{723 - 212.96}{1094} \\
&= 0.466
\end{aligned}
$$

In R:

```
Homicides <- 0:5
Or <- c(713, 299, 66, 16, 1, 0)
n <- sum(Or)

cbind(Homicides, Or)

##      Homicides  Or
## [1,]         0 713
## [2,]         1 299
## [3,]         2  66
## [4,]         3  16
## [5,]         4   1
## [6,]         5   0
```

```
xbar <- sum(Homicides * Or) / n
xbar

## [1] 0.4410959

s2 <- (sum(Homicides^2 * Or)  - n * xbar^2) / (n - 1)
s2

## [1] 0.4661341
```

(b) In order to carry-out the goodness of fit test we have to obtain the probabilities and expected frequencies under the null hypothesis that the random sample is drawn from a Poisson distribution with parameter equal to the sample mean $\bar{x}$.

```
pr <- numeric(6)
pr[1:5] <- dpois(Homicides[-6], xbar)
pr[6] <- 1 - sum(pr[1:5])
names(pr) <- c(0:4, ">=5")
```

```
Er <- pr * n

cbind(Or, Er)

##       Or          Er
## 0    713 704.4474607
## 1    299 310.7288799
## 2     66  68.5306160
## 3     16  10.0761910
## 4      1   1.1111416
## >=5    0   0.1057107
```

Since for the chi-square test to be valid, the expected frequencies should (almost) all be greater than 5. In the present case this is not so, as the last two expected frequencies are both less than 5. The natural way of dealing with this issue is to amalgamate the last two cells to produce a cell for the number of homicides ≥ 4, so that the total number of cells is reduced by one to $k = 4$.

```
Or2 <- c(Or[1:3], sum(Or[4:6]))
pr2 <- c(pr[1:3], sum(pr[4:6]))
Er2 <- n * pr2

cbind(Or2, Er2)

##   Or2       Er2
## 0 713 704.44746
## 1 299 310.72888
## 2  66  68.53062
##     17  11.29304
```

Now we can perform the test:

The chi-square test statistic is

$$X^2 = \frac{(713 - 704.45)^2}{704.45} + \frac{(299 - 310.73)^2}{310.73} + \frac{(66 - 68.53)^2}{68.53} + \frac{(17 - 11.29)^2}{11.29} = 3.527.$$

Under the null hypothesis it has the chi-square distribution with $4 - 1 - 1 = 2$ degrees of freedom.

Using Table 7 of *Lindley and Scott* for the χ^2 distribution function with 2 degrees of freedom, we find that the corresponding p-value is given by $p = 1 - F_2(3.5) = 1 - 0.8262 = 0.1738$.

The p-value is not significant at the 5% level. Hence there is no significant evidence to reject the Poisson hypothesis.

In R

```
test <- chisq.test(Or2,
                   p = pr2)
test

##
##  Chi-squared test for given probabilities
##
## data:  Or2
## X-squared = 3.524, df = 3, p-value = 0.3177
```

Since the parameter for the Poisson distribution has been estimated, we have to correct the number of degrees of freedom, and find the correct p-value:

```
1 - pchisq(test$statistic, df = length(Or2) - 2)

## X-squared
## 0.1716991
```

The p-value is not significant at the 5% level. Hence there is no significant evidence to reject the Poisson hypothesis.

(c) The index of dispersion is:

$$\begin{aligned} I &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}} \\ &= \frac{(n-1) \times s^2}{\bar{x}} \\ &= \frac{1094 \times 0.466}{0.441} \\ &= 1156.018 \end{aligned}$$

which, under our Poisson hypothesis, has approximately the χ^2_{n-1} distribution.

Since the number of degrees of freedom is very large we can not use the Statistical tables to get the exact p-value, but we can use R:

```
pvalue <- 1 - pchisq(1156.018, df = 1094)
pvalue

## [1] 0.09412407
```

The p-value is not significant at the 5% level. Hence, again, there is no significant evidence to reject the Poisson hypothesis.

In R:

Calculate the corrected sum of squares:

```
ss <- sum((Or * Homicides^2)) - n * xbar^2
ss

## [1] 509.9507
```

or

```
ss <- s2 * (n - 1)
ss

## [1] 509.9507
```

The index of dispersion

```
index <- ss / xbar
index

## [1] 1156.099
```

The p-value

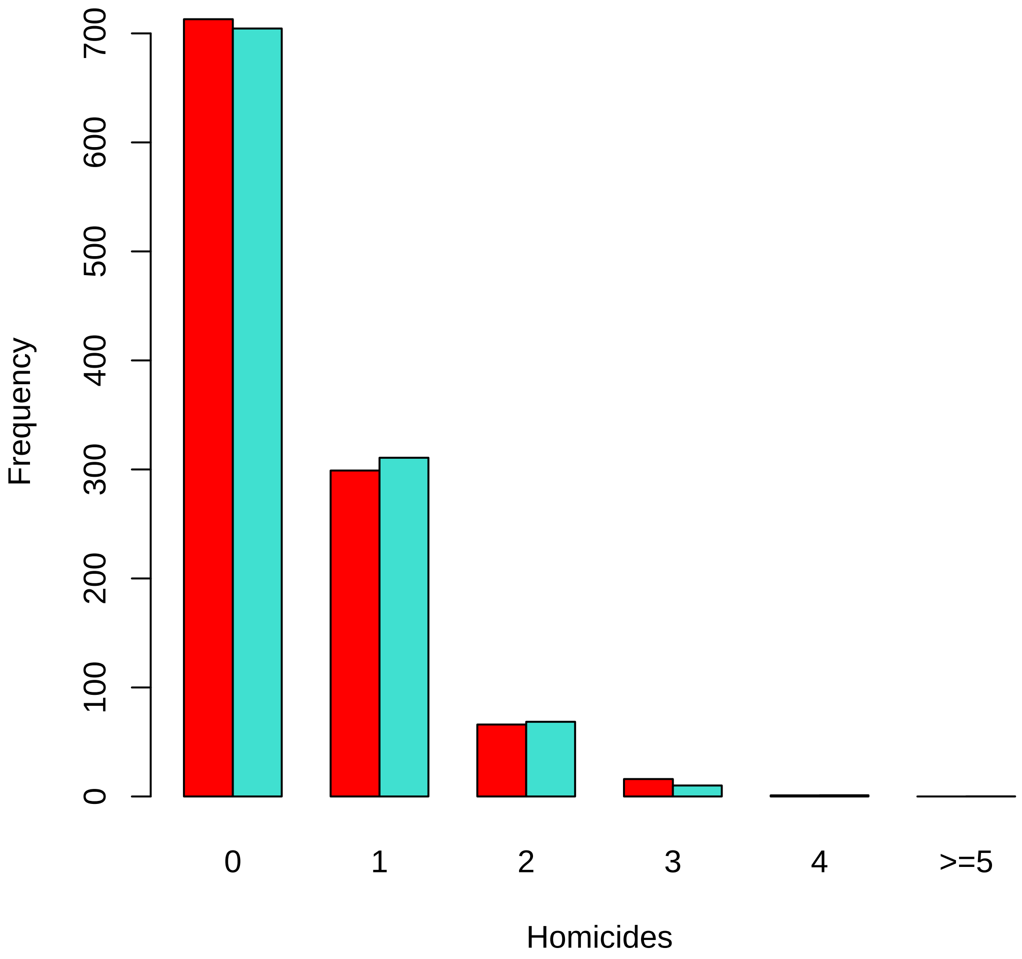
```
pvalue <- 1 - pchisq(index, df = n - 1)
pvalue

## [1] 0.09384287
```

The p-value is not significant at the 5% level. Hence, again, there is no significant evidence to reject the Poisson hypothesis.

(d)

```
barplot(rbind(Or, Er),
  names = c(0:4, ">=5"),
  xlab = "Homicides",
  ylab = "Frequency",
  beside = TRUE,
  col = c("red", "turquoise"))
```



(e) The p-value is not significant at the 5% level. Hence there is no significant evidence to reject the hypothesis that the daily numbers of homicides follows a Poisson distribution.