Dr Isabella Gollini

Probability and Statistics

13 Goodness of Fit for the Poisson Distribution

13.1 Example — ant-hill data

A large area is divided up into 4×4 metre squares and a random sample of 50 such squares is selected. The number of ant-hills in each sample square is counted, which yields the following frequency distribution:

Number of hills, r	0	1	2	3	4	5	total
Observed frequency, O_r	13	8	12	10	5	2	50

Table 1: Frequency distribution of numbers of ant-hills in 4×4 metre squares

It is suggested that the numbers of ant-hills vary randomly from square to square according to a Poisson distribution. Are the data consistent with this suggestion? Formally, we test the null hypothesis that the data are a random sample from a Poisson distribution.

We are going to carry out a chi-square goodness of fit test using the same form of test statistic as in Section 12,

$$X^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r} \ , \tag{1}$$

where k is the number of cells in the table of frequencies, in this case 6, with a more natural labelling of the cells as $0, 1, \ldots, 5$. We have the observed frequencies O_r as given in Table 1, but we shall have to calculate the expected frequencies $E_r = np_r$, where n is the sample size, $n = 50$ in the present example, and the p_r are the cell probabilities. This requires calculation of the Poisson probabilities as specified in Section 5.5,

$$p_r = e^{-\mu} \frac{\mu^r}{r!} \qquad (r = 0, 1, 2, \ldots). \tag{2}$$

We are not given the parameter μ of the Poisson distribution, but we have to calculate an estimate of μ from the observed data. We saw in Section 5.5 that the mean of the Poisson distribution is identical with its parameter μ. Hence, as discussed in Sections 7 and 8, an unbiased estimate of μ is given by the sample mean $\bar{x}$, which in this case, using the data of Table 1, is

$$\bar{x} = \frac{1}{n} \sum_{r=0}^{5} rO_r = \frac{(0 \times 13) + (1 \times 8) + (2 \times 12) + (3 \times 10) + (4 \times 5) + (5 \times 2)}{50} = 1.84 \ ,$$

the mean number of ant-hills per square. We then use the estimated value, $\mu = 1.84$, to calculate the probabilities p_r.

We might use Table 2 of *Lindley and Scott* for the Poisson distribution function, but, given our μ value, interpolation is required, and it is simpler to calculate the p_r directly, using

$$p_0 = e^{-\mu}$$

and the recurrence relation noted in Section 5.6,

$$p_r = \frac{\mu}{r}\, p_{r-1} \qquad (r = 1, 2, 3, \ldots).$$

Because we want the cell probabilities to sum to 1 and the expected frequencies to sum to 50, we modify the description of the final cell to be number of hills "≥ 5" and the corresponding probability to be $1 - F_4 = 1 - \sum_{r=0}^{4} p_r$. We then obtain the probabilities and expected frequencies as shown in Table 2.

Number of hills, r	0	1	2	3	4	≥ 5	total
Observed frequency, O_r	13	8	12	10	5	2	50
probability, p_r	0.1588	0.2922	0.2688	0.1649	0.0759	0.0394	1
Expected frequency, E_r	7.94	14.61	13.44	8.24	3.79	1.97	50

Table 2: Observed and expected frequencies of ant-hills

As stated in Section 12.2, it is often suggested that, 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 hills "≥ 4", so that the total number of cells is reduced by one to $k = 5$, as shown in Table 3, in which the calculation of the chi-square statistic is also presented, yielding the value 7.012.

Number of hills, r	0	1	2	3	≥ 4	total
Observed frequency, O_r	13	8	12	10	7	50
probability, p_r	0.1588	0.2922	0.2688	0.1649	0.1152	1
Expected frequency, E_r	7.94	14.61	13.44	8.24	5.76	50
$(O_r - E_r)^2/E_r$	3.225	2.991	0.154	0.376	0.267	7.012

Table 3: Calculation of chi-square statistic after amalgamation of cells

When, in calculating the probabilities associated with each cell, estimates of the parameters of the distribution being fitted (in our case the Poisson distribution) have been made, based upon the sample data being analyzed, the degrees of freedom are reduced by one for each estimated parameter. In general, then, the degrees of freedom ν of the chi-square statistic are given by

$$\nu = k - 1 - d, \tag{3}$$

where k is the number of cells and d is the number of fitted parameters. In the present case there is one estimated parameter, the Poisson parameter $\mu = 1.84$, and hence

$$5 - 1 - 1 = 3$$

degrees of freedom. Using Table 7 of *Lindley and Scott* for the χ^2 distribution function with 3 degrees of freedom, we find that the corresponding p-value is given by

$$p = 1 - F(7.012) = 1 - 0.928 = 0.072,$$

which is not significant at the 5% level. We conclude that there is no strong evidence to reject the null hypothesis that the numbers of ant-hills vary from square to square according to a Poisson distribution.

As always, we can use instead Table 8 of the percentage points of the χ^2 distribution. We see that $\chi^2_3(5) = 7.815$. As our calculated value of the chi-square statistic is smaller than this, we do not reject the null hypothesis at the 5% significance level.

To carry out the analysis in `R`, you have to prepare the dataset by entering a vector named `r` that lists the number of hills and the frequency variable `Or` that lists the corresponding frequencies.

```
r <- c(0:5)
Or <- c(13, 8, 12, 10, 5, 2)
names(Or) <- r
Or

##  0  1  2  3  4  5
## 13  8 12 10  5  2
```

Then we have to calculate the vector containing the expected probabilities under the hypothesis that the data follows a Poisson distribution.

```
n <- sum(Or)
n

## [1] 50

xbar <- sum(r * Or) / n
xbar

## [1] 1.84

pr <- numeric(6)
pr[1:5] <- dpois(0:4, lambda = xbar)
pr[6] <- 1 - sum(pr[1:5])
round(pr, 3)

## [1] 0.159 0.292 0.269 0.165 0.076 0.039

sum(pr)

## [1] 1
```

Calculate the expected frequencies.

```
Er <- pr * n
Er

## [1]  7.940871 14.611203 13.442307  8.244615  3.792523  1.968481
```

Since there are expected frequencies lower than 5 we calculate the new values of `Or` and `pr`, like in Table 3.

```
Or2 <- c(Or[1:4], sum(Or[5:6]))
pr2 <- c(pr[1:4], sum(pr[5:6]))

cbind(Or2, pr2)

##   Or2       pr2
## 0  13 0.1588174
## 1   8 0.2922241
## 2  12 0.2688461
## 3  10 0.1648923
##     7 0.1152201
```

We can use the function `chisq.test` in order to calculate the test statistic. But you have to notice that the p-value is wrong:

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

##
##  Chi-squared test for given probabilities
##
## data:  Or2
## X-squared = 7.0095, df = 4, p-value = 0.1354
```

The function `chisq.test` calculates the p-value assuming the test statistic follows a χ^2 distribution with $(k-1)$ degrees of freedom, as it does not allow to include the information that the parameter of the Poisson distribution has been estimated by using the data, and so, the test statistic follows a χ^2 distribution with $(k-1-1)$ degrees of freedom. The correct p-value is obtained by using:

```r
1 - pchisq(test$statistic, df = 3)

##  X-squared
## 0.07159437
```

13.2 The dispersion test for the Poisson distribution

An alternative approach to testing the goodness of fit of a Poisson distribution to a sample of data is based upon the fact that, as shown in Section 5.5, for a Poisson distribution the variance is equal to the mean, i.e., $\sigma^2 = \mu$.

We saw in Section 8.3 that for a random sample of size n from a $N(\mu, \sigma^2)$ distribution

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1} \ , \tag{4}$$

or, equivalently,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \ .$$

Suppose instead that we have a random sample of size n from a Poisson distribution. Replacing $\sigma^2 \equiv \mu$ in Equation (4) by its sample estimate $\bar{x}$, we obtain what is known as the *index of dispersion*,

$$I = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}} \ , \tag{5}$$

which, under our Poisson hypothesis, has approximately the χ^2_{n-1} distribution. In what is known as the *dispersion test*, the statistic I of Equation (5), i.e., the index of dispersion, is used as a test statistic for testing the null hypothesis that the data are a random sample from a Poisson distribution.

If our data are not from a Poisson distribution then what usually (although not always) turns out to be the case is that the sample variance is greater than the sample mean — we have what is known as *overdispersion*, so that the value of the test statistic tends to be inflated. So we may take the alternative hypothesis to be a one-sided one, that the variance of the distribution from which we are sampling is greater than its mean, in which case we carry out a one-tail test based upon the χ^2_{n-1} distribution.

Example — ant-hill data (continued)

Recall that for the ant-hill data the sample size was $n = 50$ and the sample mean was $\bar{x} = 1.84$. The sum of squares of the data is

$$\sum_{i=1}^{n} x_i^2 = \sum_{r=0}^{5} r^2 O_r = (0^2 \times 13) + (1^2 \times 8) + (2^2 \times 12) + (3^2 \times 10) + (4^2 \times 5) + (5^2 \times 2) = 276.$$

The corrected sum of squares is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 276 - (50 \times 1.84^2) = 276 - 169.28 = 106.72 \ .$$

The sample variance is $s^2 = 106.72/49 = 2.178$. The sample variance is greater than the sample mean, but is it so much greater that it provides significant evidence to reject the hypothesis that the data are a random sample from a Poisson distribution?

The index of dispersion is $I = 106.72/1.84 = 58$ (exactly). From Table 8 of *Lindley and Scott* with some interpolation we find $\chi^2_{49}(5) \approx 66.3$, so our test statistic is not significant at the 5% level. There is no strong evidence to reject the hypothesis that the data are a random sample from a Poisson distribution.

We repeat the analysis in the `R` output below. The p-value of the test statistic is calculated to be 0.177, which is not significant at the 5% level.

Calculate the corrected sum of squares:

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

## [1] 106.72
```

The index of dispersion

```
index <- ss / xbar
index

## [1] 58
```

The p-value

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

## [1] 0.1774292
```