STA 6384, Report 1.8

Carson Slater *Baylor University*

Problem: Work problem 1.10, p. 30 of Agresti.

Table 1.3 contains Ladislaus von Bortkiewicz's data on deaths of soldiers in the Prussian army from kicks by army mules (Fisher 1934, Quine and Seneta 1987). The data refer to 10 army corps, each observed for 20 years. In 109 corps-years of exposure, there were no deaths, in 65 corps-years there was one death, and so on. Estimate the mean and test whether probabilities of occurrences in these five categories follow a Poisson distribution (truncated for 4 and above).

Table 1.3 Data on Deaths by Mule Kicks, for Exercise 1.10

Number of Deaths	Number of Corps-Years
0	109
1	65
2	22
3	3
4	1
≥ 5	0

The problem asks for two things based on the Bortkiewicz data on deaths by mule kicks:

1. To estimate the mean number of deaths per corps-year.
2. To test the hypothesis that the data follow a Poisson distribution, using five categories where the last category is for 4 or more deaths.

We combine the original data for counts of 4 and ≥ 5. The total number of observations (corps-years) is $N = 109 + 65 + 22 + 3 + 1 = 200$.

The mean ($\bar{x}$) of the sample is the best estimate for the Poisson parameter λ. We calculate the sample mean using the original, un-binned data frequencies to get the most accurate estimate.

$$\bar{x} = \frac{\sum_i f_i x_i}{N}$$

Using the frequencies from the original data table (0 deaths: 109, 1 death: 65, 2 deaths: 22, 3 deaths: 3, 4 deaths: 1, ≥ 5 deaths: 0):

$$\bar{x} = \frac{(0 \times 109) + (1 \times 65) + (2 \times 22) + (3 \times 3) + (4 \times 1)}{200}$$

$$\bar{x} = \frac{0 + 65 + 44 + 9 + 4}{200} = \frac{122}{200} = 0.61$$

Thus, our estimate for the mean λ of the Poisson distribution is $\hat{\lambda} = 0.61$.

We will now test whether the data follow a Poisson distribution with $\lambda = 0.61$.

- **Null Hypothesis (H_0):** The number of deaths by mule kicks per corps-year follows a Poisson distribution.
- **Alternative Hypothesis (H_a):** The number of deaths does not follow a Poisson distribution.

The probability mass function (PMF) for a Poisson distribution is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

We use $\hat{\lambda} = 0.61$ to find the expected probability for each category, and then find the expected frequency $E_i = N \times P(k)$.

- $P(X = 0) = \frac{e^{-0.61}(0.61)^0}{0!} \approx 0.54335$
  $E_0 = 200 \times 0.54335 = 108.67$
- $P(X = 1) = \frac{e^{-0.61}(0.61)^1}{1!} \approx 0.33144$
  $E_1 = 200 \times 0.33144 = 66.29$
- $P(X = 2) = \frac{e^{-0.61}(0.61)^2}{2!} \approx 0.10109$
  $E_2 = 200 \times 0.10109 = 20.22$
- $P(X = 3) = \frac{e^{-0.61}(0.61)^3}{3!} \approx 0.02055$
  $E_3 = 200 \times 0.02055 = 4.11$
- $P(X \geq 4) = 1 - [P(0) + P(1) + P(2) + P(3)]$
  $P(X \geq 4) \approx 1 - (0.54335 + 0.33144 + 0.10109 + 0.02055) = 1 - 0.99643 = 0.00357$
  $E_{\geq 4} = 200 \times 0.00357 = 0.71$

Note: A condition for the χ^2 test is that all expected frequencies should be reasonably large, typically $E_i \geq 5$. Here, $E_3 = 4.11$ and $E_{\geq 4} = 0.71$ are below this threshold. For a more robust analysis, one might combine the last three categories ($k = 2, k = 3, k \geq 4$). However, proceeding as the problem directs (with five categories), we note this limitation.

The Chi-squared test statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)^2}{E_i}$$

We compare our calculated χ^2 value to a critical value from the χ^2 distribution. The degrees of freedom (df) are:

$$df = (\text{number of categories}) - 1 - (\text{number of estimated parameters})$$

Table 1: Calculation of the χ^2 Test Statistic

Category (k)	**Observed (O_i)**	**Expected (E_i)**	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
0	109	108.67	0.1089	0.0010
1	65	66.29	1.6641	0.0251
2	22	20.22	3.1684	0.1567
3	3	4.11	1.2321	0.2998
≥ 4	1	0.71	0.0841	0.1184
Total	**200**	**200**		$\chi^2 =$ **0.5910**

$$df = 5 - 1 - 1 = 3$$

We estimated one parameter, λ, from the data.

Let's set a significance level of $\alpha = 0.05$. The critical value for the χ^2 distribution with 3 degrees of freedom is:

$$\chi^2_{0.05,3} = 7.815$$

Our calculated test statistic is $\chi^2_{calc} = 0.591$.

Since $0.591 < 7.815$, our test statistic is not in the rejection region. We fail to reject the null hypothesis.