Tutorial 3

ST2137-2420

Material

This tutorial covers the topics and concepts from chapter 3 of the course textbook. The questions offer practice on understanding numerical data through plots, numerical summaries, comparisons with theoretical distributions, and associations.

Dataset: Student Performance

1. Create and plot a correlation matrix for the output variables G1, G2 and G3 for the student performance data.
2. Using the columns `G1`, `G2` and `G3`, create a new dataframe with 2 columns: `grade_type` and `grade_score`. There should be $395 \times 3 = 1185$ rows in the new dataframe. The unique values in `grade_type` column should be `final`, `second` and `first`. Here are some sample rows:

```
  grade_type grade_score
1      first           5
2      first           5
3      first           7
4      first          15
5      first           6
6      first          15
```

3. Create histograms for G1, G2, G3; summarise and compare them.
4. Create a boxplot of the single numeric variable `absences` using `boxplot()`, and also the 5-number summary using `summary()`. Extract the rows corresponding to the outliers and study them. What would you investigate next? *Hint: see the help page for information on the object returned by `boxplot`.*

Assessing "Poisson-ness" of a Dataset

The dataset in `er_arrivals.csv` contains the number of arrivals to an Emergency Room in U.K. over 13 months in the 1960s. It is commonly assumed (to begin with) that the number of arrivals on each day Y_j follows a Poisson pmf:

$$P(Y_j = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

To test/visualise this assumption, we can make a Poisson-ness plot. Here's how it works. First, suppose we observe $Y_1, Y_2, ..., Y_N$ and we wish to assess if it follows a Poisson distribution. We compute the count of counts:

$$X_k = \sum_{j=1}^{N} I(Y_j = k) \quad \text{for } k = 0, 1, 2, ...$$

where

$$I(Y_j = k) = \begin{cases} 1, & Y_j = k \\ 0, & Y_j \neq k \end{cases}$$

Thus X_k will be the number of times that Y_j took on the value of k. We let L be the maximum observed value of k, i.e. $\sum_{k=0}^{L} X_k = N$. We have observed our entire sample, so we consider N to be a fixed integer.

$$\begin{aligned} E(X_k) &= N \times \frac{e^{-\lambda}\lambda^k}{k!} \\ \text{(Taking logs on both sides) } \ln E(X_k) &= \ln N - \lambda + k\ln(\lambda) - \ln(k!) \end{aligned}$$

Thus, using X_k as an estimate of $E(X_k)$, a plot of $\phi_k = \ln(k!X_k/N)$ (on the y-axis) against k (on the x-axis) should yield a straight line with slope equals to $\ln(\lambda)$. Here is how this plot can be used:

- Systematic deviations, e.g. curvature from this line indicate that the Poisson distribution is unsuitable for this data.
- A regression can be fitted to estimate λ (slope will be $\ln(\lambda)$).

Complete the following questions in R and Python for this approach:

6. Read the data as a data frame `er_arrivals` in R/Python, ensuring that the date column is in a suitable format.
7. Create a Poisson-ness plot for this data.
8. Compute two different estimates of λ:
    1. Using the sample mean (this is also the MLE).
    2. Using the slope from the Poisson-ness plot