Data Science and Statistical Computing

Tutorial 1, Week 2

Q1

Let my data be:

$$x_1 = 1.49, x_2 = 1.51, x_3 = 1.48, x_4 = 1.54, x_5 = 1.50$$

$$x_6 = 1.46, x_7 = 1.44, x_8 = 1.56, x_9 = 1.45, x_{10} = 1.47$$

Recall, the Bootstrap takes *resamples* of the data (ie samples *with replacement* from the original data to create a psuedo data set of the same size). What is the probability that a Bootstrap resample of the above data has:

a. exactly three samples equal to 1.45?
b. at most two samples ≤ 1.48?
c. exactly two samples ≤ 1.48 and *all* other samples > 1.52?

Q2

Reminder: A one-sided Monte Carlo hypothesis test simulates N 'pseudo' data sets from the null hypothesis of the same size as the original data, and calculates the test statistic for each one, $\{t_i : i \in \{1, \dots, N\}\}$. We then take the empirical average of how many simulated test statistics are more extreme than the observed test statistic, t_{obs}, as our estimate of the p-value:

$$p \approx \hat{p} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{t_i \geq t_{\text{obs}}\}$$

Question: You have collected the following data of size $n = 3$:

2, 0, 1

and are interested in testing the hypothesis that it is Poisson(λ) distributed, with:

$$H_0 : \lambda = 1$$
$$H_1 : \lambda > 1$$

You recall that the expectation of a Poisson random variable is λ, so your friend suggests using the sum of the observations as the test statistic, $T = h(X_1, X_2, X_3) = \sum_{i=1}^{3} X_i$.

a. Is your friend's suggestion ok as a test statistic and why (or why not)?
b. What is the observed test statistic, t_{obs}, here?

You now proceed to simulate some data from a Poisson distribution with $\lambda = 1$ using the R programming language and get the following returned:

```
2, 1, 0        1, 1, 2
0, 2, 0        0, 2, 0
2, 1, 1        0, 3, 1
3, 1, 0        0, 0, 1
1, 4, 1        1, 1, 1
```

c. What is N here?
d. What is the test statistic, t_i, for each Monte Carlo simulated data set?
e. What would you estimate the p-value is based on this (admittedly tiny) Monte Carlo simulation?
f. At a significance level $\alpha = 0.1$, do you reject, or not reject, the null hypothesis?

Q3

Let p denote the true (exact) p-value (ie if we know the exact distribution of the test statistic),

$$p = \mathbb{P}(T \geq t_{\text{obs}} \mid H_0 \text{ true})$$

a. What is the exact p-value of the test in the previous question?
   Hint: You may use, without proof, that:

$$X_i \sim \text{Pois}(\lambda) \quad \Longrightarrow \quad \sum_{i=1}^{n} X_i \sim \text{Pois}(n\lambda)$$

b. What is the probability that a single Monte Carlo simulated test statistic will be greater than or equal to t_{obs}?

c. For N Monte Carlo simulated test statistics, what is the distribution of the number that exceed t_{obs}?

A problem for Monte Carlo testing is that the estimated p-value is random. The *resampling risk* is defined to be the probability that the Monte Carlo simulated p-value and the true p-value are on different sides of the significance threshold, α, because this is the situation when the Monte Carlo test will be wrong.

$$\text{resampling risk} = \begin{cases} \mathbb{P}(\hat{p} > \alpha) & \text{if } p \leq \alpha \\ \mathbb{P}(\hat{p} \leq \alpha) & \text{if } p > \alpha \end{cases}$$

d. What is the resampling risk of the Monte Carlo simulated hypothesis test in the last question?

Q4

In the lecture we saw the mouse data with lifetimes in days for the treatment group $(x_1, \ldots, x_7)$ and control group $(y_1, \ldots, y_9)$

a. For each group, what effect would there be on (i) the sample mean, and (ii) the standard error of the sample mean, if all the lifetimes were expressed in weeks?

b. How many standard errors from zero would the difference $\bar{x} - \bar{y}$ be now?

c. If we have a new dataset where each observation in the original data is repeated N times (ie, we get the value x_1 repeated N times, as well as the value x_2 repeated N times, etc), what would the effect be on the standard error of the sample mean? (is this roughly a factor of $\frac{1}{\sqrt{N}}$? We'll see this factor cropping up a lot later in the course!)
   Hint: First show the mean is unchanged, then write down the standard error of the new mean and put it in terms of the standard error of the original mean.