Data Science and Statistical Computing

Tutorial 2, Week 4

Q1

The following data have been collected (sorted after collection for your convenience):

$$\begin{array}{cccccccccc} 3.353, & 3.454, & 3.549, & 3.556, & 3.750, & 3.766, & 3.812, & 3.835, & 3.916, & 3.921, \\ 3.984, & 3.988, & 4.092, & 4.138, & 4.161, & 4.226, & 4.369, & 5.372, & 5.808, & 5.889. \end{array}$$

We are interested in the 0.75 quantile: that is, the value x such that $\mathbb{P}(X \leq x) = 0.75$.

a) What is the 0.75 quantile in this data?

b) The sample is small, a histogram shows the data is not at all Normal looking, and we are not looking at the mean anyway, so we cannot use a standard Normal confidence interval. Write down in detail the steps you would take to create an empirical estimate of the 0.75 quantile and the uncertainty in that estimate.

Your friend follows the procedure you have described and produces the following 1000 estimates of the quantile (sorted for your convenience):

$$\begin{array}{cccccccccc} 3.835, & 3.916, & 3.916, & 3.916, & 3.916, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, \\ 3.921, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, & 3.921, \\ & & & & \vdots & & & & & \\ & & & & \text{(96 rows)} & & & & & \\ & & & & \vdots & & & & & \\ 5.372, & 5.372, & 5.372, & 5.372, & 5.372, & 5.372, & 5.372, & 5.372, & 5.372, & 5.372, \\ 5.808, & 5.808, & 5.808, & 5.808, & 5.808, & 5.808, & 5.808, & 5.808, & 5.808, & 5.808. \end{array}$$

c) You want to identify as large a value as possible which you are still confident is less than or equal to the 0.75 quantile. What value would you recommend to be 99% confident?

Q2

Consider a sample $\mathbf{x} = (x_1, \ldots, x_n)$ from some unknown distribution and for simplicity assume $x_i \neq x_j, \forall\, i \neq j$.

In lectures, we constructed an *empirical cumulative distribution function* based on this sample. This defines a new discrete random variable, which we will call Y. We saw that Y then has probability mass function

$$p(y) = \mathbb{P}(Y = y) = \begin{cases} \frac{1}{n} & \text{if } y \in \{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases}$$

and that $\mathbb{E}[Y] = \bar{x}$, $\text{Var}[Y] = \frac{n-1}{n} s_x^2$, where s_x^2 is the sample variance of $\mathbf{x}$.

Prove that for the mean of an iid sample of size m, $Y_1, \ldots, Y_m$, we have

$$\mathbb{E}[\bar{Y}] = \bar{x} \quad \text{and} \quad \text{Var}[\bar{Y}] = \frac{n-1}{n} \frac{s_x^2}{m}$$

Q3

Consider taking bootstrap resamples from a dataset of size n.

a. How many possible resamples are there in total? (Do not worry about uniqueness)
b. What is the probability of taking a bootstrap resample and discovering there are no repetitions in the sample?
c. (harder!) How many possible *unique* resamples of the data are there in total? (assume the original data contains no ties and the order should not matter eg the resamples (x_1, x_2, x_1) and (x_2, x_1, x_1) are the same)

If stuck, click for hint 1

Think of representing a bootstrap resample as a vector of counts of how many times each x_i is chosen, where the counts always sums to n. eg For a dataset (x_1, x_2, x_3), we could represent a bootstrap resample of (x_3, x_1, x_1) as the vector of counts $(2, 0, 1)$ where the sum is clearly 3.

If stuck, click for hint 2

Now, with the setup from hint 1, think of putting n balls in n urns.

Q4

Remember the small mouse data set from lectures. We decide to use the bootstrap to estimate the uncertainty in the *median* of the treatment group.

Given the data $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$, we use the notation $x_{(i)}$ to denote the ith smallest observation in the sample, so:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq x_{(5)} \leq x_{(6)} \leq x_{(7)}$$

For example, for the mouse data $x_{(1)} = x_3 = 16$, $x_{(2)} = x_7 = 23$, etc.

a. If I take a bootstrap resample $\mathbf{x}^\star$, show that the median of $\mathbf{x}^\star$ must equal one of the observations in the data, $x_{(i)}$.
b. (harder!) Prove that the probability that the median of a bootstrap sample is $x_{(i)}$ is given by:

$$\sum_{j=0}^{3} \left[\binom{7}{j} \left(\frac{i-1}{n} \right)^j \left(1 - \frac{i-1}{n} \right)^{7-j} - \binom{7}{j} \left(\frac{i}{n} \right)^j \left(1 - \frac{i}{n} \right)^{7-j} \right]$$

Q5

A telephone switchboard receives the following number of calls in a 4-minute period:

$$\mathbf{x} = (0, 3, 1, 2)$$

We assume that the number of calls are independent and identically distributed in any 1-minute period and want to estimate the probability that there are no calls for the next 2 minutes. We propose estimating this non-parametrically by the following statistic:

$$S(\mathbf{x}) = \left(\frac{\sum_{i=1}^{4} \mathbb{1}\{x_i = 0\}}{4} \right)^2$$

(a) What is the estimate for the probability of no calls in the next 2 minutes for this data?

You decide to use bootstrap methodology to study this estimator, and notice that you don't even need to do any simulation in this case! Can you see why? Discuss in your group before proceeding.

Hint for discussion if stuck

How many possible values can $S(\mathbf{x})$ take here?

Could you exactly compute the probability of observing those values when you resample $\mathbf{x}$?

Without doing any simulation:

(b) Using the bootstrap percentile confidence interval method for your estimator, $S(\mathbf{x})$, find the largest value $\eta \in [0, 1]$ for which you are at most 95% confident that $\mathbb{P}(\text{no calls in next 2 mins}) \leq \eta$.

(c) Determine $\mathbb{E}[\bar{S}^\star]$ and hence determine the bootstrap estimate of bias (if any) in the estimator.

You read that the Poisson distribution, with probability mass function,

$$p(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

is often used to model the number of arrivals in a queue.

(d) List the detailed steps to perform a parametric bootstrap estimate of the uncertainty in the probability that there are no calls for the next 2 minutes under the assumption of a Poisson number of calls each minute.

(e) Without performing any simulation, again determine $\mathbb{E}[\bar{S}^\star]$ and hence determine the parametric bootstrap estimate of bias (if any) in the estimator.

Q6

In the lecture we saw a method of estimating π by looking at the proportion of randomly sampled uniform values on a square falling inside the circle contained within. We will use similar ideas to estimate $\sqrt{2}$ in this question.

We certainly know $0 \leq \sqrt{2} \leq 2$, so:

(a) Write down the probability density function for the random variable X having uniform distribution between 0 and 2, and calculate the cumulative distribution function.

(b) Write down $\mathbb{P}(X \leq \sqrt{2})$ and hence propose an algorithm to estimate $\sqrt{2}$ *without* having to know how to take square roots.

(c) What is the distribution of $\mathbb{1}\{X \leq \sqrt{2}\}$? Given that we know this probability (we can in fact compute $\sqrt{2}$), write down a 95% confidence interval for the value of your Monte Carlo estimate of $\sqrt{2}$ based on your algorithm in (b), when the number of Monte Carlo samples n is large.

From a pilot run of simulations, you notice that $1.5^2 > 2$.

(d) What is the effect on your estimator if you now construct a new estimator based on simulations from a uniform distribution between 0 and 1.5?

(e) If I take 1000 samples using the Uniform [0,2] approach, how many samples must I take under the Uniform [0,1.5] approach to be equally accurate in terms of the confidence interval for the estimators?

Q7

We are interested in evaluating the integral

$$\int_0^1 xe^{-x}\,dx$$

a) By noting that $f_X(x) = \lambda e^{-\lambda x}$ is the probability density function of an Exponential random variable with rate parameter λ, write the above integral as an expectation with respect to an Exponential distribution.

b) Compute the expectation exactly by doing the integration required (*Hint:* integration by parts)

Because we can compute it exactly, we wouldn't actually use simulation to approximate it, but this makes it a good example to understand the behaviour of Monte Carlo integration, since we can analytically compute the accuracy too!

c) Write down a Monte Carlo integration algorithm that estimates the integral using simulations from an Exponential distribution.

d) Compute the variance of your Monte Carlo estimator. (*Hint:* use $\text{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$, and you already know $\mathbb{E}[Y]$ from (b))