# 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 $m$, $Y_1, \ldots, Y_m$, has

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

## 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 $i$th 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 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:

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