Data Science and Statistical Computing

Tutorial 2, Week 4 Solutions

Q1

(a)

The 0.75 quantile will be the $0.75 \times 20 = 15$-th sample. Thus, 4.161.

(b)

Let the data be $\mathbf{x} = (x_1, \dots, x_{20})$. Then, we would take B bootstrap resamples, $\mathbf{x}^{\star 1}, \dots, \mathbf{x}^{\star B}$, where $\mathbf{x}^{\star b} = (x_1^{\star b}, \dots, x_{20}^{\star b})$ and each $x_i^{\star b}$ is a sample (with replacement) from $\mathbf{x}$.

Then, we would define the statistic $S(\mathbf{x}^{\star b})$ to be the function which sorts resampled times from smallest to largest, and takes the 15th value. Or, more formally:

$$S(\mathbf{x}^{\star b}) = x_{(15)}^{\star b}$$

where the subscript in brackets denotes that this is the 15th ordered value (these are called *order statistics*, which you might encounter in later years of study).

Then, we would compute

$$\widehat{\text{Var}}(S(\mathbf{x})) = \frac{1}{B-1} \sum_{b=1}^{B} \left(S(\mathbf{x}^{\star b}) - \bar{S}^\star \right)^2$$

where $\bar{S}^\star = \frac{1}{B} \sum_{b=1}^{B} S(\mathbf{x}^{\star b})$

Note: The estimate we report is still $S(\mathbf{x}) = 4.161$, and *not* $\bar{S}^\star$. We only use $\bar{S}^\star$ in order to estimate the uncertainty (we will see in lectures soon one other use is to estimate the bias).

(c)

There are 1000 bootstrap sampled values of $S(\mathbf{x}^{\star b})$ here. Therefore, if we want as large a value as possible, yet be 99% certain it is less than or equal to the 0.75 quantile, then we can only be larger than 1% (=10) of the bootstrap samples. Therefore, we choose the value 3.921 (10th value in the ordered list).

Hence, we are 99% confident that the 0.75 quantile is greater than or equal to 3.921.

Q2

$$\begin{aligned}
\mathbb{E}[\bar{Y}] &= \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} Y_i \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[Y_i \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \bar{x} \\
&= \frac{m\bar{x}}{m} \\
&= \bar{x}
\end{aligned}$$

$$
\begin{aligned}
\text{Var}[\bar{Y}] &= \text{Var}\left[\frac{1}{m}\sum_{i=1}^{m} Y_i\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\text{Var}[Y_i] \\
&= \frac{m}{m^2}\text{Var}[Y] \\
&= \frac{n-1}{n}\frac{s_x^2}{m}
\end{aligned}
$$

Q3

(a)

In total, because we sample with replacement we can choose any of the n observations in any of the n resampling places, so in total there are n^n resamples (many not unique).

(b)

$$
\mathbb{P}(\text{no repeats}) = \prod_{j=0}^{n-1}\left(1 - \frac{j}{n}\right) = \frac{n!}{n^n}
$$

(c)

Making use of the two hints, we can imagine n balls going into n urns as represented by n `o`'s for the balls and $n-1$ `|`'s for the division between urns. So with $n = 7$, the resample leading to vector of counts $(2, 1, 0, 2, 1, 0, 1)$ (sum 7) can be visualised as:

`oo|o||oo|o||o`

It is then clear that we have a total of $2n-1$ symbols (`o` or `|`) and we need to choose the location of $n-1$ bars (or n balls). Each arrangement corresponds to a unique resample, so the total number of unique resamples is

$$
\binom{2n-1}{n-1} \equiv \binom{2n-1}{n}
$$

Q4

(a)

The median is the middle value in an ordered data set, so for odd sized data sets this will be a particular observation. Since this data is size 7, $(x_1, \ldots, x_7)$, every resample will also be size 7, $(x_1^\star, \ldots, x_7^\star)$, where $x_i^\star = x_j$ for some j.

The ordered resample is then $(x_{(1)}^\star, \ldots, x_{(7)}^\star)$ and the median will be $x_{(4)}^\star$, which will be equal to one of the original observations.

(b)

We know from part (a) that the median must be equal to one of the values. Therefore, we can do the following:

$$
\begin{aligned}
\mathbb{P}(\text{median} = x_{(i)}) &= \mathbb{P}(\text{median} > x_{(i-1)}) - \mathbb{P}(\text{median} > x_{(i)}) \\
&= \mathbb{P}(\text{at most 3 resamples} \leq x_{(i-1)}) - \mathbb{P}(\text{at most 3 resamples} \leq x_{(i)}) \\
&= \sum_{j=0}^{3} \binom{7}{j} \left(\frac{i-1}{n}\right)^{j} \left(1 - \frac{i-1}{n}\right)^{7-j} - \binom{7}{j} \left(\frac{i}{n}\right)^{j} \left(1 - \frac{i}{n}\right)^{7-j}
\end{aligned}
$$