Data Science and Statistical Computing

Tutorial 2, Week 4 Solutions

Q1

(a)

The 0.75 quantile (thought of in terms of the ecdf) will be the $0.75 \times 20 = 15$-th sample. Thus, 4.161.

(b)

Let the data be $\mathbf{x} = (x_1, \ldots, x_{20})$. Then, we would take B bootstrap resamples, $\mathbf{x}^{\star 1}, \ldots, \mathbf{x}^{\star B}$, where $\mathbf{x}^{\star b} = (x_1^{\star b}, \ldots, x_{20}^{\star b})$ 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 (the 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:

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

Q5

(a)

$$S(\mathbf{x}) = \left(\frac{1}{4}\right)^2 = 0.0625$$

(b)

We can calculate explicitly the probability of all possible values for $S(\mathbf{x})$. Therefore, we can calculate the quantile that we would end up reporting under the percentile method without actually doing the simulations!

So the number of zeros in a bootstrap resample of size $n = 4$ will be Binomially distributed, with probability $p = \frac{1}{4}$.

Hence we have,

k zeros	$S(\mathbf{x}_{k \text{ zeros}})$	$\mathbb{P}(k \text{ zeros})$
0	$\left(\frac{0}{4}\right)^2 = 0$	$\binom{4}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^4 = 0.3164$
1	$\left(\frac{1}{4}\right)^2 = 0.0625$	$\binom{4}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^3 = 0.4219$
2	$\left(\frac{2}{4}\right)^2 = 0.25$	$\binom{4}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^2 = 0.2109$
3	$\left(\frac{3}{4}\right)^2 = 0.5625$	$\binom{4}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^1 = 0.0469$
4	$\left(\frac{4}{4}\right)^2 = 1$	$\binom{4}{4}\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^0 = 0.0039$

We can see that $\sum_{i=0}^{2} \mathbb{P}(i \text{ zeros}) = 0.9492$ and $\sum_{i=0}^{3} \mathbb{P}(i \text{ zeros}) = 0.9961$. For $k = 3$ we have *over* 95% confidence, therefore we would take $k = 2$ and identify the probability of no calls in the next 2 minutes is at most 0.25 with 94.92% confidence.

(c)

Again, we can compute the theoretical $\mathbb{E}[\bar{S}^\star]$, we don't need to do B resamples and compute an empirical mean.

$$
\begin{aligned}
\mathbb{E}[\bar{S}^\star] &= \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} S(\mathbf{X}_i)\right] \\
&= \frac{1}{\cancel{B}}\cancel{B}\mathbb{E}[S(\mathbf{X})] \\
&= \sum_{k=0}^{4} S(\mathbf{x}_{k \text{ zeros}})\, \mathbb{P}(k \text{ zeros}) \\
&= \sum_{k=0}^{4} \left(\frac{k}{4}\right)^2 \binom{4}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{4-k} \\
&= 0 \times 0.3164 + 0.0625 \times 0.4219 + 0.25 \times 0.2109 + 0.5625 \times 0.0469 + 1 \times 0.0039 \\
&= 0.1094
\end{aligned}
$$

Therefore, the approximate bias (estimated via bootstrap) is:

$$
\widehat{\text{bias}}(\theta, \hat{\theta}) = \bar{S}^\star - S(\mathbf{x}) = 0.109375 - 0.0625 = 0.0469
$$

This does appear to be a biased estimator, especially in light of the magnitude of the bias relative to the magnitude of the estimator. The bootstrap bias estimator method shown in lectures will rarely give exactly zero even when an estimator *is* unbiased, but when the relative size of the bias is this large compared to the estimator we can see it is clearly biased.

(d)

Step 1

Compute the MLE, $\hat{\lambda}$, for the Poisson distribution.

$$
\begin{aligned}
0 = \frac{\partial \mathcal{L}}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left[\sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)\right] \\
&= \frac{\partial}{\partial \lambda}\left[-n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log\lambda - \sum_{i=1}^{n}\log(x_i!)\right] \\
\implies 0 &= -n + \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\lambda} \\
\implies \hat{\lambda} &= \frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}
$$

So we have $\hat{\lambda} = \frac{0+3+1+2}{4} = 1.5$

Step 2

Simulate many data sets of size 4, $\mathbf{x}^{\star b}, b = 1, \ldots, B$, from a Poisson distribution with $\lambda = 1.5$, and each time compute $S(\mathbf{x}^{\star b})$ by counting the number of simulated zeros, divided by 4, squared.

Step 3

Compute $\bar{S}^\star$ and $\widehat{\text{Var}}(S(\mathbf{x}))$ in the usual way.

(e)

This will follow closely the solution to part (c), except we have a different probability of zero which is induced by the parametric Poisson family we are simulating our data sets of size 4 from. Now, instead of the probability of a zero being in a bootstrap resample being $\frac{1}{4}$, we have that it is,

$$p(k = 0 \,|\, \hat{\lambda} = 1.5) = e^{-1.5}$$

Thus,

$$\begin{aligned}
\mathbb{E}[\bar{S}^\star] &= \sum_{k=0}^{4} S(\mathbf{x}_{k \text{ zeros}})\, \mathbb{P}(k \text{ zeros}) \\
&= \sum_{k=0}^{4} \left(\frac{k}{4}\right)^2 \binom{4}{k} \left(e^{-1.5}\right)^k \left(1 - e^{-1.5}\right)^{4-k} \\
&= 0 \times 0.3642 + 0.0625 \times 0.4185 + 0.25 \times 0.1803 + 0.5625 \times 0.0345 + 1 \times 0.0025 \\
&= 0.0931
\end{aligned}$$

The estimator $S(\mathbf{x})$ is of course unchanged, so our new estimate of bias is:

$$\widehat{\text{bias}}(\theta, \hat{\theta}) = \bar{S}^\star - S(\mathbf{x}) = 0.0931228 - 0.0625 = 0.0306$$

Even supplying a parametric family in this case we find that the empirical estimator based on this statistic is biased.

Q6

(a)

$$f(x) = \begin{cases} \frac{1}{2} & \text{if } x \in [0, 2] \\ 0 & \text{otherwise} \end{cases}$$

$$\implies F(x) = \int_{-\infty}^{x} f(z)\, dz = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{2} & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

(b)

$$\mathbb{P}(X \leq \sqrt{2}) = F(\sqrt{2}) = \frac{\sqrt{2}}{2}$$

Therefore, we note:

- $\sqrt{2} = 2\,\mathbb{P}(X \leq \sqrt{2})$, and;
- $\mathbb{P}(X \leq \sqrt{2}) \equiv \mathbb{P}(X^2 \leq 2)$

Hence, just as with the π problem in class, we could do many simulations to estimate $\mathbb{P}(X \leq \sqrt{2})$, in this case simulating a uniform, squaring and comparing to 2 (thus not needing to know how to take square roots).

Therefore, our algorithm is:

1. Simulate $x_i \sim \text{Uniform}(0, 2)$, for $i \in \{1, \ldots, n\}$.
2. Estimate $\hat{p} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_i^2 \leq 2\}}{n} \approx \mathbb{P}(X^2 \leq 2) = \mathbb{P}(X \leq \sqrt{2})$
3. Estimate $\widehat{\sqrt{2}} \approx 2\hat{p}$.

(c)

The distribution of the indicator for whether a sample is smaller than $\sqrt{2}$ will be Bernoulli with true probability $p = \frac{\sqrt{2}}{2}$. Hence, we can use a Normal approximation to the Binomial confidence interval (Stats I) to determine that 95% of the time our Monte Carlo estimate for p based on n samples will lie within:

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.707 \pm 1.96\sqrt{\frac{0.707(1-0.707)}{n}} = 0.707 \pm 1.96\sqrt{\frac{0.207}{n}}$$

Therefore, our confidence interval for the estimator of $\sqrt{2}$ will lie within:

$$1.414 \pm 2 \times 1.96\sqrt{\frac{0.207}{n}}$$

(d)

If we construct an estimator based on simulating X from a Uniform on $[0, 1.5]$, then

$$\mathbb{P}(X \leq \sqrt{2}) = F(\sqrt{2}) = \frac{\sqrt{2}}{1.5} = \frac{2\sqrt{2}}{3}$$

Consequently, the distribution of the indicator will be Bernoulli with this probability, and our estimator becomes $\widehat{\sqrt{2}} = 1.5\hat{p}$. The confidence intervals will become:

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.943 \pm 1.96\sqrt{\frac{0.943(1-0.943)}{n}} = 0.943 \pm 1.96\sqrt{\frac{0.054}{n}}$$

$$1.414 \pm \frac{3}{2} \times 1.96\sqrt{\frac{0.054}{n}}$$

Therefore, the width of the interval in which the estimator should lie has shrunk:

$$\frac{3}{2} \times 1.96\sqrt{\frac{0.054}{n}} < 2 \times 1.96\sqrt{\frac{0.207}{n}}$$

(e)

In fact, the Uniform [0,1.5] interval is 0.3827 times the width of the interval when estimating using Uniform simulations on [0,2].

This is really important! It means that I can get an equally accurate estimate with far fewer simulations!

To find the number of samples n_2 for Uniform[0,1.5] to achieve the same accuracy as n_1 samples from Uniform[0,2]:

$$\begin{aligned}
\frac{\frac{3}{2} \times 1.96\sqrt{\frac{0.054}{n_2}}}{2 \times 1.96\sqrt{\frac{0.207}{n_1}}} &= 1 \\
\implies \frac{\sqrt{n_1}}{\sqrt{n_2}} &= \frac{1}{0.3827} \\
\implies n_2 &= \left(0.3827\sqrt{n_1}\right)^2
\end{aligned}$$

So if I do 1000 simulations with a Uniform [0,2], I can achieve the same accuracy using only $\left(0.3827\sqrt{1000}\right)^2 \approx 147$ samples from a Uniform [0,1.5]!

Q7

(a)

If we have that $X \sim \text{Exp}(\lambda = 1)$, then using the hint we can see that the integral is simply:

$$\begin{aligned}\mu = \int_0^1 x\lambda e^{-\lambda x}\,dx &= \int_0^1 x f_X(x)\,dx \\ &= \int_{\mathbb{R}} x\ \mathbb{1}\{x \in [0,1]\} f_X(x)\,dx \\ &= \mathbb{E}\left[X\ \mathbb{1}\{X \in [0,1]\}\right]\end{aligned}$$

(b)

Using integration by parts,

$$\begin{aligned} u &= x, & dv &= e^{-x} \\ du &= 1, & v &= -e^{-x}\end{aligned}$$

$$\begin{aligned}\int xe^{-x}\,dx &= -xe^{-x} + \int e^{-x}\,dx \\ &= -e^{-x}(1+x) \\ \implies \int_0^1 xe^{-x}\,dx &= \left[-e^{-x}(1+x)\right]_0^1 \\ &= -e^{-1}(1+1) + e^{-0}(1+0) \\ &= 1 - 2e^{-1} \\ &\approx 0.2642\end{aligned}$$

Therefore, $\mathbb{E}\left[X\ \mathbb{1}\{X \in [0,1]\}\right] = 1 - 2e^{-1} \approx 0.2642$.

(c)

1. Simulate $x_1, \ldots, x_n$ from an Exponential($\lambda = 1$) distribution
2. Calculate
$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\ \mathbb{1}\{x_i \in [0,1]\}$$

(d)

We know from lectures that

$$\text{Var}\left(\hat{\mu}_n\right) = \mathbb{E}\left[(\hat{\mu}_n - \mu)^2\right] = \frac{\sigma^2}{n}$$

where here we have

$$\sigma^2 = \text{Var}\left(X\ \mathbb{1}\{X \in [0,1]\}\right)$$

with $X \sim \text{Exp}(\lambda = 1)$.

Based on the hint, we want to find:

$$\text{Var}\left(X\ \mathbb{1}\{X \in [0,1]\}\right) = \mathbb{E}\left[\left(X\ \mathbb{1}\{X \in [0,1]\}\right)^2\right] - \mathbb{E}\left[X\ \mathbb{1}\{X \in [0,1]\}\right]^2$$

We already know from (b) that $\mathbb{E}\left[X\ \mathbb{1}\{X \in [0,1]\}\right] = 1 - 2e^{-1}$, so we just need:

$$\begin{aligned}\mathbb{E}\left[\left(X\ \mathbb{1}\{X \in [0,1]\}\right)^2\right] &= \int_{\mathbb{R}} \left(x\ \mathbb{1}\{x \in [0,1]\}\right)^2 f_X(x)\,dx \\ &= \int_0^1 x^2 e^{-x}\,dx\end{aligned}$$

Again, a simple application of integration by parts gives:

$$\begin{aligned}\int x^2 e^{-x}\,dx &= -x^2 e^{-x} - \int -2xe^{-x}\,dx \\ &= -x^2 e^{-x} + 2\int xe^{-x}\,dx\end{aligned}$$

We know the remaining integral from (b), so we get:

$$\begin{aligned}\int x^2 e^{-x}\,dx &= -x^2 e^{-x} + 2\left(-e^{-x}(1+x)\right) \\ \implies \int_0^1 x^2 e^{-x}\,dx &= \left[-e^{-x}(2+2x+x^2)\right]_0^1 \\ &= -e^{-1}(2+2+1) + e^0(2+0+0) \\ &= 2 - 5e^{-1} \\ &\approx 0.1606\end{aligned}$$

So,

$$\begin{aligned}\text{Var}\left(X\ \mathbb{1}\{X \in [0,1]\}\right) &= \underbrace{(2-5e^{-1})}_{\mathbb{E}\left[\left(X\ \mathbb{1}\{X\in[0,1]\}\right)^2\right]} - \underbrace{(1-2e^{-1})^2}_{\mathbb{E}[X\ \mathbb{1}\{X\in[0,1]\}]^2} \\ &= 1 - e^{-1} - 4e^{-2} \\ &\approx 0.09078\end{aligned}$$

Thus,

$$\text{Var}\left(\hat{\mu}_n\right) \approx \frac{0.09078}{n}$$