Data Science and Statistical Computing

Tutorial 3, Week 6 Solutions

Q1

(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, $\mathbf{x}^{\star b}, b = 1, \dots, B$, 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,

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

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 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.

Q2

(a)

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

(b)

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

Therefore, we note:

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

Hence, just as with the π problem in class, we could do many simulations to estimate $\mathbb{P}(X \le \sqrt{2})$, 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 $i \in \{1, \dots, n\}$.
2. Estimate $\hat{p} = \frac{\sum_{i=1}^n \mathbb{1}\{x_i^2 \le 2\}}{n} \approx \mathbb{P}(X^2 \le 2) = \mathbb{P}(X \le \sqrt{2})$
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,

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

Consequently, the distribution of the indicator will be Bernoulli with this probability and our 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!

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

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