Tutorial 3: Solutions

January 31, 2018

Question 5.4.54, Page 237

Suppose the sediment density (g/cm) of a randomly selected specimen from a certain region is normally distributed with mean 2.65 and standard deviation 0.85.

(a) If a random sample of 25 specimens is selected, what is the probability that the sample average sediment density is at most 3.00? Between 2.65 and 3.00?

Let X be a specimen where $X \sim N(\mu = 2.65, \sigma^2 = 0.85^2)$. Then $\overline{X} \sim N(\mu = 2.65, \sigma^2/n = 0.85^2/25)$. From this, we have that $Z := \dfrac{\overline{X} - 2.65}{0.85/\sqrt{25}} \sim N(0,1)$.

$$\begin{aligned}
\mathbf{P}\left(\overline{X} \leq 3.00\right) &= \mathbf{P}\left(\frac{\overline{X} - 2.65}{0.85/5} \leq \frac{3.00 - 2.65}{0.85/5}\right) \\
&= \mathbf{P}\left(Z \leq 2.059\right) \\
&= 0.98025
\end{aligned}$$

$$\begin{aligned}
\mathbf{P}\left(2.65 \leq \overline{X} \leq 3.00\right) &= \mathbf{P}\left(\frac{2.65 - 2.65}{0.85/5} \leq \frac{\overline{X} - 2.65}{0.85/5} \leq \frac{3.00 - 2.65}{0.85/5}\right) \\
&= \mathbf{P}\left(0 \leq Z \leq 2.059\right) \\
&= \mathbf{P}\left(Z \leq 2.059\right) - \mathbf{P}\left(Z \leq 0\right) \\
&= 0.98025 - 0.50000 \\
&= 0.48025
\end{aligned}$$

(b) How large a sample size would be required to ensure that the first probability in part (a) is at least .99?

$$\begin{aligned}
\mathbf{P}\left(\overline{X} \leq 3.00\right) &\geq 0.99 \\
\mathbf{P}\left(\frac{\overline{X} - 2.65}{0.85/\sqrt{n}} \leq \frac{3.00 - 2.65}{0.85/\sqrt{n}}\right) &\geq 0.99 \\
\mathbf{P}\left(Z \leq \sqrt{n}\left(\frac{3.00 - 2.65}{0.85}\right)\right) &\geq 0.99 \\
\sqrt{n}\left(\frac{3.00 - 2.65}{0.85}\right) &\geq 2.3263 \\
n &\geq 31.918
\end{aligned}$$

n must be an integer so we round it up to 32. Therefore, we need a sample size of at least 32 in order to ensure the probability in (a) is at least 0.99.

Question 5.5.72, Page 243

I plan to leave my office at precisely 10:00 a.m. and wish to post a note on my door that reads, "I will return by t a.m." What time t should I write down if I want the probability of my arriving after t to be 0.01?

Let T be the total time it takes to run all three errands, including travel time: $T = X_1+X_2+X_3+X_4 = \sum_{i=1}^{4} X_i$. Then we have that

$$\mathbf{E}(T) = \mathbf{E}\left(\sum_{i=1}^{4} X_i\right) = \sum_{i=1}^{4} \mathbf{E}(X_i) = 15+5+8+12 = 40$$

$$\mathbf{Var}(T) = \mathbf{Var}\left(\sum_{i=1}^{4} X_i\right) = \sum_{i=1}^{4} \mathbf{Var}(X_i) = 4^2+1^2+2^2+3^2 = 30$$

Since each X_i was from a normal distribution, we conclude that $T \sim N(\mu = 40, \sigma^2 = 30)$. Converting to standard normal, $Z := \frac{T-40}{\sqrt{30}} \sim N(0,1)$. We want to find $\mathbf{P}(T \geq t) = 0.01$.

$$\begin{aligned}
\mathbf{P}\left(\frac{T-40}{\sqrt{30}} \geq \frac{t-40}{\sqrt{30}}\right) &= 0.01 \\
\mathbf{P}\left(Z \geq \frac{t-40}{\sqrt{30}}\right) &= 0.01 \\
1-\mathbf{P}\left(Z \geq \frac{t-40}{\sqrt{30}}\right) &= 1-0.01 \\
\mathbf{P}\left(Z \leq \frac{t-40}{\sqrt{30}}\right) &= 0.99 \\
\frac{t-40}{\sqrt{30}} &= 2.3263 \\
t &= 52.741
\end{aligned}$$

Since $t = 52.741$ guarantees that $\mathbf{P}(T \geq t) = 0.01$, we must round upwards to the nearest minute. Therefore, the sign that we post should read "I will return by 10:53 a.m.".

Question 6.1.10, Page 263

(a) Show that $\overline{X}^2$ is not an unbiased estimator for μ^2. [Hint: For any rv Y, $\mathbf{E}\left(Y^2\right) = \mathbf{Var}\left(Y\right) + \mathbf{E}\left(Y\right)^2$. Apply this with $Y = \overline{X}$.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\mathbf{Var}\left(\overline{X}\right) = \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}\left(X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$\mathbf{E}\left(\overline{X}\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i\right) = \frac{n\mu}{n} = \mu$

$\mathbf{E}\left(\overline{X}^2\right) = \mathbf{Var}\left(\overline{X}\right) + \left(\mathbf{E}\left(\overline{X}\right)\right)^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2$

Therefore, $\overline{X}^2$ is not an unbiased estimator of μ^2.

(b) For what value of k is the estimator $\overline{X}^2 - kS^2$ unbiased for μ^2? [Hint: Compute $\mathbf{E}\left(\overline{X}^2 - kS^2\right)$.]

We want to find k such that $\mathbf{E}\left(\overline{X}^2 - kS^2\right) = \mu^2$.

$\mathbf{E}\left(\overline{X}^2 - kS^2\right) = \mathbf{E}\left(\overline{X}^2\right) - k\mathbf{E}\left(S^2\right) = \frac{\sigma^2}{n} + \mu^2 - k\sigma^2$

In order for $\frac{\sigma^2}{n} + \mu^2 - k\sigma^2 = \mu^2$ to be true, we need $k = \frac{1}{n}$.

Therefore, an unbiased estimator of μ^2 is $\overline{X}^2 - \frac{S^2}{n}$.

Finding a MLE: Exponential

Suppose we are given a sample $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\theta)$, where θ is an unknown parameter ($\theta > 0$). Find a Maximum Likelihood Estimate for θ. Is the maximum likelihood estimator that you obtain also an unbiased estimator?

The likelihood function is simply the product of the probability density functions of each X_i.

$$\begin{aligned}
\mathcal{L}(\theta \mid \vec{x}) &= \prod_{i=1}^{n} \frac{1}{\theta}\exp\left\{\frac{-x_i}{\theta}\right\} \\
&= \theta^{-n}\exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right\}
\end{aligned}$$

We take the log of both sides since the log-likelihood will be easier to work with.

$$l(\theta \mid \vec{x}) := \ln(\mathcal{L}(\theta \mid \vec{x})) = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

$$\frac{dl}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i$$

We know that the maximum of $l(\theta)$ occurs when the first derivative equals 0. Thus we set the first derivative equal to 0 and we solve for $\widehat{\theta}$.

$$\begin{aligned}
0 &= -\frac{n}{\widehat{\theta}} + \frac{1}{\widehat{\theta}^2}\sum_{i=1}^{n} x_i \\
\frac{n}{\widehat{\theta}} &= \frac{1}{\widehat{\theta}^2}\sum_{i=1}^{n} x_i \\
\widehat{\theta} n &= \sum_{i=1}^{n} x_i \\
\widehat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}
\end{aligned}$$

We need to find the second derivative of our log-likelihood function and plug in our MLE to verify that this is indeed a maximum.

$$\begin{aligned}
\frac{d^2 l}{d\theta^2} &= \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i \\
\left.\frac{d^2 l}{d\theta^2}\right|_{\theta=\widehat{\theta}} &= \frac{n}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} - \frac{2\sum_{i=1}^{n} x_i}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^3} \\
&= \frac{n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} - \frac{2n^3}{\left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{-n^3}{\left(\sum_{i=1}^{n} x_i\right)^2}
\end{aligned}$$

Notice that n represents the number of elements in our sample (which is always non-negative) and therefore $-n^3$ is always negative. Our denominator is a quantity squared, and squared real values are always positive. Therefore, our second derivative evaluated at our MLE is less than 0. Recall from first year calculus that this means that $\widehat{\theta}$ is in fact a maximum and we conclude that $\widehat{\theta} = \overline{X}$ is a MLE for θ.

We now want to check whether $\widehat{\theta}$ is an unbiased estimator of θ. We do this by taking the expected value of $\widehat{\theta}$ and checking whether it equals θ. Recall that for a random variable $X \sim \text{Exp}(\theta)$, $\mathbf{E}(X) = \theta$.

$$\begin{aligned}
\mathbf{E}\left(\widehat{\theta}\right) &= \mathbf{E}\left(\overline{X}\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(X_i) \\
&= \frac{n\theta}{n} \\
&= \theta
\end{aligned}$$

Therefore we conclude that our MLE $\widehat{\theta} = \overline{X}$ is an unbiased estimator of θ. It is important to note that a Maximum Likelihood Estimator is **not** always an unbiased estimator!

Also note that being an unbiased estimator means that after several trials, the average of your estimates of the unknown parameter (expected value) will equal your unknown parameter. This does not guarantee that your estimator actually hit the true value dead-on, as illustrated in tutorial. You could be far off, you could be close. Sometimes we prefer a biased estimator if we know that it estimates a value very, very close to the true value but is consistently off by a certain amount.