MAST20005/MAST90058: Week 10 Solutions

From the lectures, we know that the pdf of the kth order statistic $X_{(k)}$ is:

$$g_k(x) = k\binom{n}{k} F(x)^{k-1} \left(1 - F(x)\right)^{n-k} f(x).$$

1. Here $f(x) = \frac{1}{3}e^{-x/3}$ and $F(x) = 1 - e^{-x/3}$.

   (a) Using the above result,

   $$\begin{aligned} g_3(x) &= 3\binom{5}{3} \left(1 - e^{-x/3}\right)^{3-1} \left(e^{-x/3}\right)^{5-3} \frac{1}{3}e^{-x/3} \\ &= 10 \left(1 - e^{-x/3}\right)^2 e^{-x}, \quad x > 0. \end{aligned}$$

   (b) We need the probability that one or zero observations are larger than 5. The following derivation is very similar to the triangular distribution example from Module 9 (see the lecture notes).

   $$\begin{aligned} \Pr(X_{(4)} < 5) &= \binom{5}{4} F(x)^4 \left(1 - F(x)\right) + F(x)^5 \\ &= 5\left(1 - e^{-5/3}\right)^4 e^{-5/3} + \left(1 - e^{-5/3}\right)^5 \\ &= 0.7599 \end{aligned}$$

   (c) For $1 < X_{(1)}$ we need each observation to be larger than 1. In other words,

   $$\Pr(1 < X_{(1)}) = \left(1 - F(1)\right)^5 = \left(e^{-1/3}\right)^5 = e^{-5/3} = 0.1889$$

2. (a) The likelihood is

   $$L(\theta) = \begin{cases} e^{-\sum_i (x_i - \theta)} & \theta \leqslant \min(x_i), \\ 0 & \text{otherwise.} \end{cases}$$

   This is maximised when each $(x_i - \theta)$ is minimised, and this happens when θ is as large as possible but still satisifies the constraint given by the inequality. Hence $\hat{\theta} = \min(X_i) = X_{(1)} = Y$.

   (b) Firstly,

   $$F(x) = \int_\theta^x e^{-(t-\theta)}\, dt = 1 - e^{-(x-\theta)}, \quad x \geqslant \theta.$$

   Then,

   $$\begin{aligned} g_1(y) &= n\left(1 - F(y)\right)^{n-1} f(y) = 10 \left(e^{-(y-\theta)}\right)^9 e^{-(y-\theta)} \\ &= 10e^{-10(y-\theta)}, \quad y \geqslant \theta. \end{aligned}$$

   (c) Firstly,

   $$\mathbb{E}(Y) = \int_\theta^\infty y 10e^{-10(y-\theta)}\, dy.$$

   Substitute $z = y - \theta$,

   $$\begin{aligned} \mathbb{E}(Y) &= \int_0^\infty (z+\theta) 10e^{-10z}\, dz \\ &= \theta \int_0^\infty 10e^{-10z}\, dz + \int_0^\infty z 10e^{-10z}\, dz \\ &= \theta + \frac{1}{10} \end{aligned}$$

because the left integral evaluates to 1 since it integrates the pdf of an exponential distribution, and the right integral is the expected value of the same exponential distribution (so we know what its value is). Therefore, $\mathbb{E}(Y - \frac{1}{10}) = \theta$, which means $Y - \frac{1}{10}$ is an unbiased estimator of θ.

(d) Firstly (and substituting $z = y - \theta$ again),

$$\Pr(\theta < Y < \theta + c) = \int_{\theta}^{\theta+c} 10e^{-10(y-\theta)}\, dy = \int_0^c 10e^{-10z}\, dz = 1 - e^{-10c}.$$

Hence we need to solve $1 - e^{-10c} = 0.95$, which results in $c = 0.1\ln(20) = 0.300$. Now, simple rearranging gives

$$\Pr(\theta < Y < \theta + c) = \Pr(Y - c < \theta < Y)$$

so that a 95% confidence interval is $[y - 0.3, y]$.

(e) This was the 'boundary problem' example shown in the lectures as part of Module 2.

3. (a) $f(x) = 1$ and $F(x) = x$, so we have $g_1(x) = n(1-x)^{n-1}$, $0 < x < 1$.

(b) Using integration by parts,

$$\mathbb{E}(X_{(1)}) = \int_0^1 xn(1-x)^{n-1}\, dx = \left[-x(1-x)^n - \frac{1}{n+1}(1-x)^{n+1}\right]_0^1 = \frac{1}{n+1}.$$

4. (a) The pdf is symmetric about θ, with the function on either side being an exponential function that can be though of as two exponential distributions put 'back-to-back' (hence the nickname *double exponential distribution*). The expectation of X can be split into two integrals, one of each side of θ, and because of symmetry they will cancel out.

In more detail, let $Z = X - \theta$. This means Z has a symmetric pdf around 0, $f(z) = \frac{1}{2}e^{-|z|}$. Therefore,

$$\begin{aligned}
\mathbb{E}(Z) &= \int_{-\infty}^{\infty} z\frac{1}{2}e^{-|z|}\, dz = \int_{-\infty}^{0} \frac{z}{2}e^{-|z|}\, dz + \int_0^{\infty} \frac{z}{2}e^{-|z|}\, dz = \int_{-\infty}^{0} \frac{z}{2}e^{z}\, dz + \int_0^{\infty} \frac{z}{2}e^{-z}\, dz \\
&= \int_0^{\infty} \frac{-z}{2}e^{-z}\, dz + \int_0^{\infty} \frac{z}{2}e^{-z}\, dz = -\int_0^{\infty} \frac{z}{2}e^{-z}\, dz + \int_0^{\infty} \frac{z}{2}e^{-z}\, dz = 0.
\end{aligned}$$

This then implies $\mathbb{E}(X - \theta) = 0$, so we have $\mathbb{E}(X) = \theta$.

Using the hint, we can also exploit the symmetry to derive the variance,

$$\begin{aligned}
\operatorname{var}(Z) = \mathbb{E}(Z^2) &= \int_{-\infty}^{\infty} z^2\frac{1}{2}e^{|z|}\, dz = \int_{-\infty}^{0} \frac{z^2}{2}e^{-|z|}\, dz + \int_0^{\infty} \frac{z^2}{2}e^{-|z|}\, dz \\
&= \int_{-\infty}^{0} \frac{z^2}{2}e^{z}\, dz + \int_0^{\infty} \frac{z^2}{2}e^{-z}\, dz = \int_0^{\infty} \frac{z^2}{2}e^{-z}\, dz + \int_0^{\infty} \frac{z^2}{2}e^{-z}\, dz \\
&= \int_0^{\infty} z^2e^{-z}\, dz = 2,
\end{aligned}$$

and from this it follows that $\operatorname{var}(X) = 2$.

(b) For a sample mean we have $\mathbb{E}(\bar{X}) = \mathbb{E}(X) = \theta$ and $\operatorname{var}(\bar{X}) = \frac{1}{n}\operatorname{var}(X) = \frac{2}{n}$.

(c) Using the asymptotic distribution of the sample median, we have $\mathbb{E}(\hat{M}) \approx m$ and $\text{var}(\hat{M}) \approx (4nf(m)^2)^{-1}$. Due to symmetry, we know that $m = \theta$ (the population median is the same as the population mean), which means we have $f(m) = f(\theta) = \frac{1}{2}$ and thus: $\mathbb{E}(\hat{M}) \approx \theta$ and $\text{var}(\hat{M}) \approx \frac{1}{n}$.

(d) $\hat{M}$ is better. Both estimators are (approximately) unbiased but $\hat{M}$ has a smaller variance, so is more likely to be closer to the true value of θ. Note that this is the reverse of the situation of sampling from a normal distribution, where the sample mean is the better estimator.

(e) We did this already in the past! See the solution for question 1(c)iii from week 3. The MLE is the sample median, $\hat{M}$.

5. We use a confidence interval based on the order statistics. Since we are interested in the median, we would like a 'symmetric' interval formed by taking the ith lowest and ith largest order statistics, we just need to determine the most appropriate value of i. For $i = 1$ we have the interval $(x_{(1)}, x_{(14)})$, for $i = 2$ we have $(x_{(2)}, x_{(13)})$, and so on. Calculating the confidence levels for each of these leads us to using $(x_{(4)}, x_{(11)}) = (1.8, 6.26)$ as the best choice since it has a confidence level of 94.26%, which is very close to the desired 95%. This particular confidence level can be calculated in R using:

```
pbinom(10, size = 14, prob = 0.5) - pbinom(3, size = 14, prob = 0.5)
```