MAST20005/MAST90058: Assignment 1

Due date: 11am, Friday 4 September 2020

Instructions: Please submit your assignment via the LMS, ensuring that you follow the submission instructions provided online. **Remember to submit ON TIME, since late submission will receive zero points. Do not wait until the last minute!** We suggest that you submit your assignment promptly once you finish all questions. You can always re-submit your assignment before the deadline if you make any changes.

If for any reason you think you will not be able to submit on time, you need to notify the Subject coordinator Tingjin Chu in a timely manner (as soon as you become aware of any issue and preferably prior to the deadline). In general, a medical certificate is required. Note that extensions are only granted in exceptional circumstances and only for a very limited time period.

Questions labelled with '**(R)**' require use of R. Please provide appropriate R commands and their output, along with sufficient explanation and interpretation of the output to demonstrate your understanding. **Such R output should be presented in an integrated form together with your explanations.** All other questions should be completed without reference to any R commands or output, except for looking up quantiles of distributions where necessary. Make sure you give enough explanation so your tutor can follow your reasoning if you happen to make a mistake. Please also try to be as succinct as possible. Each assignment will include marks for good presentation.

Problems:

1. **(R)** Tingjin runs a plant experiment. He creates thirty pots, plants an identical seed in each one, and leaves them in the same spot in the sun. After six weeks he measures the height of each plant (in cm),

```
173.1  61.5 123.3 100.4  20.4  20.9
228.4   1.0   6.8  11.4   7.7  40.7
 15.8 422.4  58.2  19.9  38.8 121.0
118.6 174.9  87.2  14.0 204.7  81.9
 57.3 177.0  14.1 137.0  76.4 330.2
```

   (a) Give basic summary statistics for these data and produce a box plot. Briefly comment on centre, spread and shape of the distribution.

   (b) Assuming an exponential distribution, compute maximum likelihood estimate for the parameter.

   (c) Tom thinks that the data follow $X \sim \text{Exp}(100)$ with the density function

$$f(x) = \frac{1}{100} e^{-x/100}.$$

   Draw a QQ plot to compare the data against the exponential distribution Exp(100). Include a reference line. Comment on the fit of the model to the data. (Use 'Type 6' quantiles for the sample quantiles.)

   (d) Jen also thinks that the data follow $X \sim \text{Exp}(100)$. However, Jen draws a QQ plot to compare the data against the exponential distribution Exp(1). Will the above approach work? If Jen's hypothesis is correct, what will be the approximate slope and intercept of the best fitting line?

2. Let $X_1, \ldots, X_n$ be a random sample from the lognormal distribution, Lognormal(μ, λ), whose pdf is:

$$f(x \mid \mu, \lambda) = \frac{1}{x\sqrt{2\pi\lambda}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\lambda}\right\}, \quad x > 0.$$

   (a) Show that the MLE of μ and λ are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i$ and $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(\ln X_i - \hat{\mu})^2$.

   (b) It is known that $\ln X_i \sim \mathrm{N}(\mu, \lambda)$.

      i. Derive the standard deviation of $\hat{\lambda}$.

      ii. Derive a $100 \cdot (1 - \alpha)\%$ confidence interval for λ.

   (c) **(R)** Consider the following dataset:

```
 12.9   2.3   2.4  65.0   6.7   1.8   1.5   1.7
248.7   1.0   2.0   4.9   3.6   4.1   6.8
```

      i. Calculate the standard error of $\hat{\lambda}$.

      ii. Assuming a lognormal distribution is an appropriate model for these data, compute the maximum likelihood estimate of λ and give a 95% CI.

3. Let $X \sim \mathrm{Unif}(0, \theta)$, a uniform distribution with an unknown endpoint θ.

   (a) Suppose we have a single observation on X.

      i. Find the method of moments estimator (MME) for θ and derive its mean and variance.

      ii. Find the maximum likelihood estimator (MLE) for θ and derive its mean and variance.

   (b) The *mean square error* (MSE) of an estimator is defined as $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$.

      i. Let $\mathrm{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. Show that,

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2.$$

      ii. Compare the MME and MLE from above in terms of their mean square errors.

      iii. Find an estimator with smaller MSE than either of the above estimators.

   (c) Suppose we have a random sample of size n from X.

      i. Find the MME and derive its mean, variance and MSE.

      ii. Find the MLE and derive its mean, variance and MSE.

      iii. Consider the estimator $a\hat{\theta}$ where $\hat{\theta}$ is the MLE. Find a that minimises the MSE.

Some information that might be useful:

$$\mathbb{E}\left(X_{(1)}\right) = \frac{\theta}{n+1}, \quad \mathbb{E}\left(X_{(1)}^2\right) = \frac{2\theta^2}{(n+1)(n+2)}, \quad \mathbb{E}\left(X_{(n)}\right) = \frac{n\theta}{n+1}, \quad \mathbb{E}\left(X_{(n)}^2\right) = \frac{n\theta^2}{n+2}$$

4. **(R)** We have a random sample of size 20 from a normal distribution. We wish to estimate the population mean. Damjan suggests taking the average of the sample minimum and sample maximum. Allan thinks this will be a poor estimator and says we should use the sample mean instead. Do a simulation in R to compare these two estimators in terms of their bias and variance. Include a side-by-side boxplot that compares their sampling distributions.

5. Let X_1, X_2, X_3 be independent random variables with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = (\sigma/i)^2 > 0$, for $i = 1, 2, 3$. Consider the following four estimators of μ:

$$
\begin{aligned}
T_1 &= \frac{1}{4}(X_1 + X_2) + \frac{1}{2}X_3 \\
T_2 &= \frac{1}{3}(X_1 + 2X_2 + 3X_3) \\
T_3 &= \frac{1}{3}(X_1 + X_2 + X_3) \\
T_4 &= \frac{1}{2}(X_1 + X_2) + \frac{1}{4}X_3^2
\end{aligned}
$$

(a) Which of these estimates are unbiased? Show your working.

(b) Among the unbiased estimators, which one has the smallest variance?

(c) Can you suggest an unbiased estimator which has a smaller variance than above unbiased estimators?