MAST20005/MAST90058: Week 2 Problems

1. Let $X_1, X_2, \ldots, X_9$ be a random sample. We are told that $\mathbb{E}(X_3) = 7$ and $\text{var}(X_4) = 4$.

   (a) What is $\text{sd}(X_2)$?

   (b) What is $\text{var}(X_7 + X_8)$?

   (c) What is $\text{cov}(X_3, X_4)$?

   (d) What is an approximate distribution of the sample mean, $\bar{X}$?

2. Let $Y = X_1 + \cdots + X_{15}$ be the sum of iid rvs, each with pdf $f(x) = (3/2)x^2$ where $-1 < x < 1$.

   (a) What is $\mathbb{E}(X_1)$?

   (b) Calculate $\mathbb{E}(Y)$ and $\text{var}(Y)$.

   (c) We would like to calculate $\Pr(-0.3 < Y < 1.5)$. Use the Central Limit Theorem to approximate this probability. *Hint*: $\Phi(-0.1) = 0.4602$ and $\Phi(0.5) = 0.6915$, where $\Phi(\cdot)$ is the standard normal cdf.

3. In each of the following scenarios, is the sample that is described a random sample? What assumptions are being made? Are they realistic? What is the 'population' in each case?

   (a) Tingjin runs a plant experiment. He creates ten pots, plants an identical seed in each one, and leaves them in the same spot in the sun. After 6 weeks he measures the height of each plant, giving measurements $x_1, x_2, \ldots, x_{10}$.

   (b) Damjan measures the height of all of his immediate family members, giving measurements y_1, y_2, y_3, y_4.

   (c) Every day, Robert counts the number of people sitting down on South Lawn. He does this for 100 days in a row, giving counts $z_1, z_2, \ldots, z_{100}$.

4. Consider the following realisations from X:

   ```
   43.1  48.9  42.6  43.7  41.0
   ```

   Note that: $\sum_{i=1}^5 x_i = 219.3$, $\sum_{i=1}^5 x_i^2 = 9654.27$.

   (a) Reminder from the lectures: the sample mean, $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, is a measure of location; the sample variance, $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, is a measure of spread; the sample standard deviation, $s = \sqrt{s^2}$, is another measure of spread.

      i. Show that $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $s^2 = (n-1)^{-1}(\sum_{i=1}^n x_i^2 - n\bar{x}^2)$.

      ii. Compute the mean and standard deviation for the dataset above.

   (b) How would your answers to part (a) change if we multiplied each data point by 2?

   (c) Reminder: the default ('Type 7') sample quantiles in R are defined as $\hat{\pi}_p = x_{(k)}$, where $k = 1 + (n-1)p$. The set of statistics {minimum, $\hat{q}_1$, $\hat{q}_2$, $\hat{q}_3$, maximum} is often called the 'five-number summary', where the $\hat{q}_i$ are the sample quartiles.

      i. Calculate the five-number summary of the above dataset.

      ii. The interquartile range (IQR) is $\hat{q}_3 - \hat{q}_1$. Calculate the IQR of the above dataset.

      iii. Draw a box plot for this dataset using these quantiles.

(d) An alternate definition of sample quantiles is given by: $\widetilde{\pi}_p = x_{(i)} + r \cdot (x_{(i+1)} - x_{(i)})$, where $(n+1)p = i + r$ such that i is an integer and $0 \leqslant r < 1$.

i. Re-compute the five-number summary using this definition.

ii. Show that $\widetilde{\pi}_p$ are the 'Type 6' quantiles (as defined in the lectures).

(e) Outliers are observations that don't seem to belong with the rest of the data. They can occur through data entry errors or problems with an experiment. Extreme observations are not necessarily errors but there is a crude convention for when to identify and label them as 'outliers': an observation x is an outlier if $x < \hat{\pi}_{0.25} - k \times \text{IQR}$ or $x > \hat{\pi}_{0.75} + k \times \text{IQR}$, where typically $k = 1.5$. These are often depicted graphically on a box plot, by only extending the whiskers up to $k \times \text{IQR}$ from each quartile and plotting each outlier as an individual point. (This is the default way that R will draw box plots.) According to this convention, are there any outliers in the sample above?

5. Create a sample of 4 numbers from $\{1, 2, 3, 4\}$, with repeats allowed, that maximises the sample variance.

6. The following are Prostaglandin-endoperoxide synthase 2 (COX2) measurements on tissue samples from 10 mice (COX2 is a protein involved in inflamatory processes related to cancer):

```
10.39  10.43  9.99  11.17  8.91  11.20  11.38  7.74  10.61  11.11
```

(a) Calculate the five-number summary (you may use either Type 6 or Type 7 quantiles).

(b) Are there any outliers in this sample (using R's default convention)?

(c) Draw a box plot for this dataset.

7. The following are observations on maximum rainfall (cm/day) in a year measured by a weather station in Tasmania (King Island Airport) in 10 consecutive years:

```
9.9  4.7  20.5  1.8  4.7  9.8  20.5  20.2  6.5  3.0
```

(a) Draw a histogram of these data.

(b) Suppose that the random variable Y follows the extreme value (EV) distribution which depends on a location parameter, θ, and a scale parameter, ξ. This distribution is obtained as maximum of a set of values (maximum wind speed, precipitation, peak flow, etc.). This distribution has the property that we can write Y as

$$Y = \theta + \xi Z,$$

where Z has the standard EV distribution with cdf $F(z) = e^{-e^{-z}}$. Thus, the inverse cdf function is $F^{-1}(p) = -\ln(-\ln p)$, so we expect that,

$$x_{(k)} \approx \theta + \xi F^{-1}\left(\frac{k}{n+1}\right).$$

Use a QQ plot to assess whether the EV model looks correct for these data.
How could you estimate the parameters θ and ξ based on your plot?
(Drawing the plot will be easier with a computer in the lab class, but you can discuss this problem in the tutorial beforehand.)