Question 1

Suppose $Z_1, Z_2, \ldots, Z_n$ are independent and identically distributed random variables following an unknown distribution F_Z. The mean μ of the distribution F_Z is unknown, but the variance of F_Z is known to be $\sigma^2 = 7$. Suppose we observe $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ as $\mathbf{z} = (z_1, z_2, \ldots, z_n)$. Given that the sample mean is $\overline{z} = 6$ and $n = 12$, construct a 95% confidence interval for μ.

Solution to Question 1

If the distribution is unknown, but the variance is known, we can use Chebyshev's inequality. For any X and any $k > 0$,

$$\mathrm{P}\left(|X - \mathrm{E}(X)| < k\sqrt{\mathrm{Var}(X)}\right) \geq 1 - \frac{1}{k^2}.$$

Since the Z_i are independent, we know from Proposition 1.2.6 that

$$\mathrm{E}\left(\overline{Z}\right) = \mu, \qquad \mathrm{Var}\left(\overline{Z}\right) = \frac{\sigma^2}{n}$$

Then, applying Chebyshev's inequality to $X = \overline{Z}$,

$$\begin{aligned}
&\mathrm{P}\left(\left|\overline{Z} - \mu\right| < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\
\Rightarrow &\mathrm{P}\left(\left|\mu - \overline{Z}\right| < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\
\Rightarrow &\mathrm{P}\left(-k\frac{\sigma}{\sqrt{n}} < \mu - \overline{Z} < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}. \\
\Rightarrow &\mathrm{P}\left(\overline{Z} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{Z} + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}.
\end{aligned}$$

To find the value of k,

$$\begin{aligned}
&1 - \frac{1}{k^2} = 0.95 \\
\Rightarrow &\frac{1}{k^2} = 0.05 = \frac{1}{20} \\
\Rightarrow &k^2 = 20 \\
\Rightarrow &k = \sqrt{20} = 2\sqrt{5}
\end{aligned}$$

So, if $1 - \frac{1}{k^2} = 0.95$, then $k = 2\sqrt{5}$. Since $\overline{z} = 6$ and $n = 12$ and $\sigma^2 = 7$, the 95% confidence interval is

$$\left(\overline{z} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{z} + k\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(6 - 2\sqrt{5} \cdot \frac{\sqrt{7}}{\sqrt{12}}, 6 + 2\sqrt{5} \cdot \frac{\sqrt{7}}{\sqrt{12}}\right)$$

$$= \left(6 - 2\sqrt{5} \cdot \frac{\sqrt{7}}{2\sqrt{3}}, 6 + 2\sqrt{5} \cdot \frac{\sqrt{7}}{2\sqrt{3}}\right)$$

$$= \left(6 - \frac{\sqrt{5}\sqrt{7}}{\sqrt{3}}, 6 + \frac{\sqrt{5}\sqrt{7}}{\sqrt{3}}\right).$$

Question 2

Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and each follows the same distribution which has mean μ and variance σ^2. Recall the definitions of the sample mean and sample variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

where $\overline{X}$ is an estimator of μ and S^2 is an estimator of σ^2. Suppose it is known that for this distribution,

$$\mathrm{Var}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = 2(n-1)\sigma^4.$$

Stating any results used from the notes:

(a) Show that $b_{\sigma^2}(S^2) = 0$, where $b_{\sigma^2}(S^2)$ is the bias of S^2.

(b) Prove that the mean squared error of S^2 is $\dfrac{2\sigma^4}{n-1}$.

(c) Suppose that one defines $W = \dfrac{1}{n+1}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ as an alternative estimator of σ^2. Compute $b_{\sigma^2}(W)$, the bias of W.

(d) Compute $\mathrm{Var}\,(W)$.

(e) Compute the mean squared error of W, and show that it is less than the mean squared error of S^2.

(f) Which estimator would you prefer to use to estimate σ^2? Justify your answer, stating the advantages and disadvantages of both estimators.

Solution to Question 2

Part (a) From Proposition 1.2.6 in the notes, $\mathrm{E}[S^2] = \sigma^2$. Therefore,

$$b_{\sigma^2}(S^2) = \mathrm{E}[S^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$

Part (b)

Method 1: First use the assumption above to compute the variance of S^2 as

$$\begin{aligned}
\mathrm{Var}\left(S^2\right) &= \mathrm{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{(n-1)^2}\mathrm{Var}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{(n-1)^2}\left(2(n-1)\sigma^4\right) \\
&= \frac{2\sigma^4}{n-1}.
\end{aligned}$$

Now, using Theorem 1.5.24 from lectures, and the fact given in (i) that $b_{\sigma^2}(S^2) = 0$, the mean squared error of S^2 is computed as

$$\left(b_{\sigma^2}(S^2)\right)^2 + \mathrm{Var}\left(S^2\right) = (0)^2 + \frac{2\sigma^4}{n-1} = \frac{2\sigma^4}{n-1}.$$

Method 2: Recalling that $\mathrm{E}[S^2] = \sigma^2$, the mean squared error is computed directly as

$$\mathrm{E}[(S^2 - \sigma^2)^2] = \mathrm{Var}\left(S^2\right),$$

and one computes $\mathrm{Var}\left(S^2\right) = \frac{2\sigma^4}{n-1}$ as in Method 1, and so the mean squared error is

$$\mathrm{E}[(S^2-\sigma^2)^2] = \frac{2\sigma^4}{n-1}.$$

Method 3: One expands the definition of the mean squared error (using the linearity of expectation)

$$\begin{aligned}\mathrm{E}[(S^2-\sigma^2)^2] &= \mathrm{E}[(S^2)^2 - 2\sigma^2 S^2 + \sigma^4] = \mathrm{E}[(S^2)^2] - 2\sigma^2\mathrm{E}[S^2] + \mathrm{E}[\sigma^4] \\ &= \mathrm{E}[(S^2)^2] - \sigma^4\end{aligned}$$

and then one computes $\mathrm{E}[(S^2)^2]$ using

$$\mathrm{E}[(S^2)^2] = \mathrm{Var}\left(S^2\right) + (\mathrm{E}[S^2])^2 = \mathrm{Var}\left(S^2\right) + (\sigma^2)^2,$$

and then this becomes the same as Method 2.

Part (c)

Recalling from Proposition 1.2.6 in the notes that $\mathrm{E}[S^2] = \sigma^2$, and noticing that $W = \dfrac{n-1}{n+1}S^2$,

$$\begin{aligned}b_{\sigma^2}(W) &= \mathrm{E}[W] - \sigma^2 = \mathrm{E}[\tfrac{n-1}{n+1}S^2] - \sigma^2 = \tfrac{n-1}{n+1}\mathrm{E}[S^2] - \sigma^2 = \tfrac{n-1}{n+1}\sigma^2 - \sigma^2 \\ &= \left(\tfrac{n-1}{n+1} - 1\right)\sigma^2 \\ &= \left(\tfrac{n-1-(n+1)}{n+1}\right)\sigma^2 \\ \Rightarrow b_{\sigma^2}(W) &= \frac{-2}{n+1}\sigma^2\end{aligned}$$

Part (d)

Method 1:

$$\begin{aligned}\mathrm{Var}\,(W) &= \mathrm{Var}\left(\frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{(n+1)^2}\mathrm{Var}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\ &= \frac{1}{(n+1)^2}2(n-1)\sigma^4 \\ &= \frac{2(n-1)}{(n+1)^2}\sigma^4\end{aligned}$$

Method 2 (very similar to Method 1):

$$\begin{aligned}\mathrm{Var}\,(W) &= \mathrm{Var}\left(\tfrac{n-1}{n+1}S^2\right) \\ &= \left(\frac{n-1}{n+1}\right)^2\mathrm{Var}\left(S^2\right) \\ &= \frac{(n-1)^2}{(n+1)^2}\frac{2\sigma^4}{n-1} \\ &= \frac{2(n-1)}{(n+1)^2}\sigma^4\end{aligned}$$

Part (e)

The mean squared error can be computed using Theorem 1.5.24 and the bias and variance from Parts (c) and (d):

$$\begin{aligned}
\left(b_{\sigma^2}(W)\right)^2 + \operatorname{Var}(W) &= \left(\frac{-2}{n+1}\sigma^2\right)^2 + \frac{2(n-1)}{(n+1)^2}\sigma^4 \\
&= \frac{4}{(n+1)^2}\sigma^4 + \frac{2(n-1)}{(n+1)^2}\sigma^4 \\
&= \frac{4+2n-2}{(n+1)^2}\sigma^4 \\
&= \frac{2(n+1)}{(n+1)^2}\sigma^4 \\
&= \frac{2}{n+1}\sigma^4
\end{aligned}$$

Since for any value of n the inequality $\frac{2}{n+1} < \frac{2}{n-1}$ is true, then

$$\operatorname{MSE}(W) = \frac{2\sigma^4}{n+1} < \frac{2\sigma^4}{n-1} = \operatorname{MSE}(S^2)$$

and so the mean squared error of W is less than the mean squared error of S^2.

Part (f)

Does not matter which estimator is preferred, both can be justified.

The advantage of S^2 is that it is unbiased, but the disadvantage is that it has a higher mean squared error.

The advantage of W is that it has a lower mean squared error but the disadvantage is that it is biased.

Question 3

Suppose you are conducting an experiment and record the following nine measurements:

$$\mathbf{x} = \{5.6,\ 3.2,\ 11.7,\ 3.2,\ 13.8,\ 8.4,\ 8.4,\ 7.5,\ 2.1\}$$

and you want to compute a measure of central tendency and dispersion for this data.

(a) Compute the sample mean and sample variance of $\mathbf{x}$ (if you like, you may use a calculator).

(b) Compute the sample median and the sample interquartile range of $\mathbf{x}$.

(c) Suppose that you did not have a computer or calculator with you, and you were asked to compute either (a) or (b) to two decimal places (or as a fraction). Which option would you choose, and why?

Solution to Question 3

(a) The sample mean and variance can be computed directly:

$$\begin{aligned}
\overline{x} &= \frac{1}{9}(5.6 + 3.2 + 11.7 + 3.2 + 13.8 + 8.4 + 8.4 + 7.5 + 2.1) = \frac{63.9}{9} = 7.1 = \frac{71}{10} \\
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \frac{1}{8}\left((-1.5)^2 + (-3.9)^2 + (4.6)^2 + (-3.9)^2 + (6.7)^2 + (1.3)^2 + (1.3)^2 + (0.4)^2 + (-5.0)^2\right) \\
&= \frac{1}{8}\left(2.25 + 15.21 + 21.16 + 15.21 + 44.89 + 1.69 + 1.69 + 0.16 + 25.00\right) \\
&= \frac{1}{8}\left(127.26\right) = \frac{1}{8}\left(\frac{12726}{100}\right) = \frac{12726}{800} = 15.9075 \approx 15.91
\end{aligned}$$

(b) To compute the sample median and sample interquartile range, one first sorts the data:

$$2.1, 3.2, 3.2, 5.6, 7.5, 8.4, 8.4, 11.7, 13.8$$

Since there are 9 values, the median is the 5th value, i.e. the median is 7.5.

For the interquartile range, since there are nine values, the $q_{0.25}$ value is the 3rd value, i.e. 3.2 and the $q_{0.75}$ value is the 7th value, i.e. 8.4. Therefore the interquartile range is $8.4 - 3.2 = 5.2$.

(c) Given this data, (b) is perhaps easier to compute by hand than (a). This is somewhat subjective but, when doing the calculation by hand, sorting a small number of values (less than 20?), and then finding the appropriate quantiles for the IQR seems to be less work than computing $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ for the variance. In this example the numbers only have one decimal place, but this is especially true if the numbers have several digits/decimal places, e.g. $1.425, 52.321, 6.234...$.

In case you are wondering, R gives the same solution:

```
x <- c(5.6,  3.2, 11.7,  3.2, 13.8,  8.4,  8.4,  7.5,  2.1)
cat("mean: ", mean(x), ", variance: ", var(x), "\n", sep="")
#> mean: 7.1, variance: 15.91
cat("median: ", median(x), ", interquartile range: ", IQR(x), "\n", sep="")
#> median: 7.5, interquartile range: 5.2
```

Question 4

Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, with $n > 1$.

(a) Find a sample $\mathbf{x}$ where its sample median equals its sample mean.

(b) Find a sample $\mathbf{x}$ where its sample median is greater than its sample mean.

(c) Find a sample $\mathbf{x}$ where its sample median is smaller than its sample mean.

(d) Suppose that the sample $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is known to have sample mean $\overline{x} = \mu$, but the precise values of the x_i in the sample $\mathbf{x}$ are unknown. Given **any** other finite value $\mu' \neq \mu$, we will add n' elements to $\mathbf{x}$ to construct $\mathbf{x}'$, i.e. $\mathbf{x}' = \{x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{n+n'}\}$, so that the sample mean of $\mathbf{x}'$ is μ'. What must the smallest value of n' be in order to ensure this, and furthermore what are the values of $x_{n+1}, \ldots, x_{n+n'}$?

(e) Suppose that $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is known to have sample **median** m, but the values of the x_i in the sample $\mathbf{x}$ are unknown . Given any other finite value $m' \neq m$, we will add n' elements to $\mathbf{x}$ to construct $\mathbf{x}'$ so that the sample median of $\mathbf{x}'$ is m'. What must the smallest value of n' be in order to ensure this, and furthermore what are the values of $x_{n+1}, \ldots, x_{n+n'}$? Choose the values of $x_{n+1}, \ldots, x_{n+n'}$ so that the sample median of $\mathbf{x}'$ will be m', no matter the values in the original sample $\mathbf{x}$.

Solution to Question 4

Part (a): Any sample with two elements will work, or any sample with $n \geq 3$ elements with all elements equal would also work, e.g. $\{2, 3\}$ has sample mean 2.5 which is also a median, while $\{7, 7, 7\}$ has sample mean and median equal to 7.

Part (b): Example: $x = \{0, 0, 5, 5, 5\}$ has mean 3 and median 5.

Part (c): Example: $x = \{0, 0, 0, 5, 5\}$ has mean 2 and median 0.

Part (d): Intuitively, adding a single element that is large/small enough will be able to shift the sample mean to any value of your choosing.

Take $\mathbf{x}' = \mathbf{x} \cup \{x_{n+1}\}$. Then, solving for x_{n+1}:

$$\begin{aligned}
\mu' &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}\sum_{i=1}^{n} x_i + \frac{1}{n+1}x_{n+1} = \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n+1}x_{n+1} \\
&= \frac{n}{n+1}\cdot\mu + \frac{1}{n+1}x_{n+1} \\
\Rightarrow (n+1)\,\mu' &= n\mu + x_{n+1} \\
\Rightarrow x_{n+1} &= n(\mu' - \mu) + \mu'
\end{aligned}$$

Therefore, only one observation x_{n+1} is needed to create the sample $\mathbf{x}'$ with any desired mean.

Part (e): One quickly realises that for, say, $n = 20$, adding a single value will not be enough to change the median to any value of one's choosing. Indeed, the median of a sample either needs to be an element of the sample, or 'between' two elements of a sample. The only way to change the median of a sample to **any** arbitrary value would be to add at least n elements that are large/small enough. Thinking about it a bit more, one realises that adding the value m' enough times to the sample will change the median to m'

Construct $\mathbf{x}'$ with $n' = n + 1$ and $x_{n+1} = x_{n+2} = \cdots = x_{2n+1} = m'$. Then $\mathbf{x}'$ consists of $2n + 1$ values, $n + 1$ of which are equal to m'. In particular the "middle" observation once the observations are sorted, i.e. the median, is $x_{(n+1)} = m'$.

Actually, we if we used $n' = n$ values, all of which were equal to m', then the median would also be m'; remember, for an even number of observations, any value in the closed interval $[x_{(n)}, x_{(n+1)}]$ would be a median, including the endpoints $x_{(n)}$ and $x_{(n+1)}$ (here the same has $2n$ values). Why must either $x_{(n)}$ or $x_{(n+1)}$ be equal to m'? Well, in this set of $2n$ values, at least n of them are equal to m', and if we sorted all $2n$ values, either $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are equal to m', or $x_{(n)}, x_{(n+1)} \ldots, x_{(2n)}$ are equal to m', or $x_{(i)}, x_{(i+1)}, \ldots, x_{(i+n-1)}$ are equal to m'; so in any case, either $x_{(n)}$ or $x_{(n+1)}$ is equal to m'.

To show that constructing $\mathbf{x}'$ with $n' = n$ is the construction with the smallest possible value of n' (that will work for all cases), consider the case $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ with $x_i = 0$ for $i = 1, 2, \ldots, n$ and suppose $m' = 1$. Then, no matter which $n - 1$ values you add, the median value is the element $x_{(n)}$ (since that is the middle element of the set of $2n - 1$ elements). Furthermore, no matter which $n - 1$ values you add, by a similar argument to that above, $x_{(n)} = 0$, since at least n out of the $2n - 1$ elements are 0.

The purpose of Parts (a), (b) and (c) are to show that the mean and median may be the same, or may be very different. The purpose of Part (d) and Part (e) are to show that the median is more robust than the mean to a single/a few extreme value(s).