Question 1

Recall from Section 8.3.1 in Prof. Veraart's notes that the p.d.f. of the uniform distribution on the interval (a, b) is

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \mathrm{U}(a, b)$ to indicate that the random variable X follows this distribution.

(a) If $X \sim \mathrm{U}(a, b)$, compute $\mathrm{E}(X)$.

(b) If $X \sim \mathrm{U}(a, b)$, compute $\mathrm{Var}(X)$.

Solution to Question 1

Part (a):

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x = \left[\frac{1}{2}\cdot\frac{x^2}{b-a}\right]_a^b = \frac{a+b}{2}$$

Part (b):

$$\mathrm{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \int_a^b \frac{x^2}{b-a}\mathrm{d}x = \left[\frac{1}{3}\cdot\frac{x^3}{b-a}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

$$\Rightarrow \mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - (\mathrm{E}(X))^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Question 2

In each part below, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{x}' = \{x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{n+n'}\}$.

(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) Given $\mathbf{x}$ with sample mean μ, for any other finite value $\mu' \neq \mu$, construct $\mathbf{x}'$ so that the sample mean of $\mathbf{x}'$ is μ', and furthermore $\mathbf{x}'$ has the smallest possible number of elements.

(e) Given $\mathbf{x}$ with sample median m, for any other finite value $m' \neq m$, construct $\mathbf{x}'$ so that the sample median of $\mathbf{x}'$ is m', and furthermore $\mathbf{x}'$ has the smallest possible number of elements.

Solution to Question 2

Part (a): Any sample with two elements will work, or any sample with $n \geq 3$ elements with all elements equal would also work.

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): 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): 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 $x_{(n+1)} = m'$. To show that constructing $\mathbf{x}'$ with $n' = n+1$ is the construction with the smallest possible value of n' (that will work for all cases), take $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, with $x_i = m' - i$ for $i = 1, 2, \ldots, n$. Then, all values in $\mathbf{x}$ are less than m', and adding any n values will result in the median still being less than m', since the median of a set of $2n$ elements is less than the $(n+1)$th largest element (in particular, it is any value in the open interval $(m'-1, m')$, usually taken to be the midpoint, since the set contains an even number of elements).

Question 3

Suppose you conduct an experiment and record the following measurements: $\mathbf{x} = \{6, 3, 10, 3, 10, 8, 8, 7, 2\}$.

(a) Compute the sample mean of $\mathbf{x}$.

(b) Compute the sample variance of $\mathbf{x}$.

(c) Compute the sample median of $\mathbf{x}$.

(d) Compute the interquartile range of $\mathbf{x}$.

(e) Which do you think, in general, is more computationally intensive for a large number of observations: computing the sample mean or computing the sample median?

(f) Which do you think, in general, is more computationally intensive for a small number of observations when doing the calculation by hand: computing the variance or computing the interquartile range?

Solution to Question 3

Part (a): There are nine values, and so

$$\overline{x} = \frac{1}{9}(6+3+10+3+10+8+8+7+2) = \frac{57}{9} = \frac{19}{3}$$

Part (b):

$$s^2 = \frac{1}{9-1}\left(\sum_{i=1}^{9} x_i^2 - (9)\overline{x}^2\right) = \frac{1}{8}\left(435 - (9)\frac{361}{9}\right) = \frac{1}{8}(74) = \frac{37}{4}$$

Part (c): Sorting the values:

$$\mathbf{x} = \{2, 3, 3, 6, 7, 8, 8, 10, 10\} \tag{1}$$

Since there are nine values, the median is the fifth largest, i.e. $m = 7$.

Part (d): The difficult part here is finding the indices that correspond to $q_{0.25}$ and $q_{0.75}$. One can use the formula for the index $i_{0.25}$ (the index for the element that is $q_{0.25}$):

$$i_{0.25} = \frac{1}{2}\left\lfloor \frac{n+1}{2} + 1 \right\rfloor = \frac{1}{2}\left\lfloor \frac{n+3}{2} \right\rfloor.$$

Think of this calculation as follows: if the index for the median is $\frac{n+1}{2}$, then $i_{0.25}$ is the midpoint between the median index and the smallest index (i.e. 1). In the case that $n = 9$, $i_{0.25} = \frac{1}{2}\left\lfloor \frac{9+3}{2} \right\rfloor = 3$, so $q_{0.25}$ is the third smallest observation, namely, 3.

If $i_{0.25}$ is not an integer, but a half-integer, then one takes $q_{0.25}$ to be the average of the two observations at indices $\lfloor i_{0.25} \rfloor$ and $\lceil i_{0.25} \rceil$ (similar to the median).

To get $i_{0.75}$, one can use $i_{0.25}$, since if $i_{0.25}$ is k, then $i_{0.75}$ is $n+1-k$:

$$i_{0.75} = n + 1 - i_{0.25}.$$

So, $i_{0.75} = 9 + 1 - 3 = 7$ and $q_{0.75} = 8$. Then, IQR $= q_{0.75} - q_{0.25} = 8 - 3 = 5$.

Part (e): Computing the median requires sorting the data, which is intuitively more work (for large n) than adding up numbers, and adding up squares of numbers, which is all that is required for computing the mean.

This can be described mathematically using notation for computational complexity, which you may not have seen yet. In case you are interested, the most efficient known sorting algorithm for real-valued numbers is $\Theta(n \log_2 n)$, while adding up numbers is simply $\Theta(n)$; this can loosely be interpreted to mean that "about $n \log_2 n$ steps" are needed to compute the median, but only "about n steps" are needed to compute the mean.

Part (f): 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. This is especially true if the numbers have several decimal places, e.g. $1.425, 52.321, 6.234$....

Question 4

Suppose that a population is taking part in a vote and an unknown proportion p of the voters supports a particular option, labelled A. Suppose it is possible to interview a sample of n randomly selected voters and record $\widehat{p}$, the proportion of that sample that supports option A. What value of n should be chosen so that with high confidence (probability at least 95%) $\widehat{p}$ is within 0.01 of p?

Solution to Question 4

One will notice the similarity to Exercise 1.3.4 in the notes, and we start the same way. Let us label our sample of n voters from 1 to n, and let X_i be the random variable with value $x_i = 1$ if voter i supports option A, and $x_i = 0$ otherwise. By this construction, each $X_i \sim \text{Bernoulli}(p)$, where p is the unknown parameter we wish to estimate, and $\widehat{p} = \overline{x}$. Since each X_i has mean $\mathrm{E}(X_i) = p$ and variance $\mathrm{Var}(X_i) = p(1-p)$, using Proposition 1.2.6, $\mathrm{E}\left(\overline{X}\right) = p$ and $\mathrm{Var}\left(\overline{X}\right) = p(1-p)/n$. Therefore, for any $\epsilon > 0$, Chebyshev's Inequality in Theorem 1.3.2 gives us:

$$\mathrm{P}\left(\left|\overline{X} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}.$$

Furthermore, using Corollary 1.1.14, one can remove the unknown p on the right-hand side to obtain

$$\mathrm{P}\left(\left|\overline{X} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}.$$

Now, we want to find the value of n so that (when $\epsilon = 0.01$)

$$\mathrm{P}\left(\left|\overline{X} - p\right| \geq 0.01\right) \leq 1 - 0.95 = 0.05.$$

Instead of trying to directly bound $\mathrm{P}\left(\left|\overline{X} - p\right| \geq 0.01\right)$, we instead can bound $\frac{1}{4n\epsilon^2}$. We solve:

$$\begin{aligned}
&\frac{1}{4n\epsilon^2} \leq \frac{5}{100} \\
\Rightarrow &4n\epsilon^2 \geq \frac{100}{5} \\
\Rightarrow &4n(0.01)^2 \geq 20 \\
\Rightarrow &n \geq \left(\frac{20}{4}\right)(100)^2 = 50000
\end{aligned}$$

Therefore, taking a sample of at least 50,000 voters will give us an estimate of p to within $\epsilon = 0.01$ with probability 95%.

The next page contains the statements of Proposition 1.2.6, Theorem 1.3.2 and Corollary 1.1.14 from the notes.

Proposition 1.2.6. Suppose that the sample $X_1, X_2, \ldots, X_n$ are independently sampled from a distribution F_X that has mean μ and finite variance σ^2. Then

1. $\mathrm{E}\left(\overline{X}\right) = \mu$,
2. $\mathrm{Var}\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$,
3. $\mathrm{E}\left(S^2\right) = \sigma^2$.

Theorem 1.3.2 (Chebyshev's Inequality)**.** If X is a random variable with mean μ and variance σ^2, then for all $c > 0$,

$$\mathrm{P}\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2}.$$

Corollary 1.1.14. Suppose $X \sim \text{Bernoulli}(p)$, for some $p \in [0, 1]$. Then $\mathrm{Var}\left(X\right) = p(1-p) \leq \frac{1}{4}$.