Probability and Statistics for JMC
Solutions 1 — Numerical Summaries

For the first 3 questions, let $(x_1, x_2, \ldots, x_n)$ be a sample of n real numbers and $m \in \mathbb{R}$ some measure of location for these data.

1. Show that $m = \overline{x}$, the sample mean, is the minimizer of the sum of squared deviations

$$\sum_{i=1}^{n} (x_i - m)^2 .$$

Take the derivative with respect to m and set it equal to 0 to find the value of m which extremizes the dispersion:

$$\begin{aligned}
\frac{d}{dm} \sum_{i=1}^{n} (x_i - m)^2 &= \sum_{i=1}^{n} \frac{d}{dm} (x_i - m)^2 \\
&= \sum_{i=1}^{n} 2 (x_i - m) (-1) \\
&= -2 \sum_{i=1}^{n} x_i + 2 \sum_{i=1}^{n} m \\
&= -2n\overline{x} + 2nm.
\end{aligned}$$

Setting the derivative equal to zero, $0 = -2n\overline{x} + 2nm$, shows the function is either a local maximum or minimum when $m = \overline{x}$.
To show this is unique and also a minimum notice that the dispersion $\sum_{i=1}^{n} (x_i - m)^2$ is a quadratic in m with a positive coefficient for m^2. So it's a parabola opening upward and therefore the derivative is zero at the a unique minimum.

2. Show by induction that $m = x_{(n+1)/2}$, the sample median, minimizes the sum of absolute deviations

$$\sum_{i=1}^{n} |x_i - m| .$$

(Note that the median is not necessarily the unique minimzer.)
[Hint: Assume the samples are ordered. Consider the base cases of $n = 1$ and $n = 2$ first. Then for the induction step prove the case for n assuming result holds for $n - 2$.]

Without loss of generality (w.l.o.g.) we can put the samples in increasing order (i.e. $x_1 \leq x_2 \leq \ldots \leq x_n$).
The base cases: For $n = 1$ the sum is just $|x_1 - m|$, which is minimized when $m = x_1$ and x_1 is the median. For $n = 2$, notice that when m is between x_1 and x_2, $|x_1 - m| + |x_2 - m|$ is just the distance from x_1 to m plus the distance from m to x_2, which is just the total length of the interval, $x_2 - x_1$. When m is outside $[x_1, x_2]$ then $|x_1 - m| + |x_2 - m| > x_2 - x_1$. Therefore, any m between x_1 and x_2 is a minimizer. In particular, the median, $(x_1 + x_2)/2$, is a minimizer.

Now suppose the result holds for $n-2$ and prove it for n. Break up the sum as follows:

$$\sum_{i=1}^{n} |x_i - m| = |x_1 - m| + |x_n - m| + \sum_{i=2}^{n-1} |x_i - m| .$$

Consider the x_1 and x_n terms. Like in the $n = 2$ case, as long as $x_1 \leq m \leq x_n$ these two terms add up to the constant $x_n - x_1$. Therefore, the whole sum is minimized when the $\sum_{i=2}^{n-1}$ term is minimized, which, by the inductive hypothesis, occurs at the median of the subsample $(x_2, \ldots, x_{n-1})$. But the median of the full sample is equal to the the median of the subsample because we just added two points x_1 and x_n at opposite ends. (If m is outside $[x_1, x_n]$ then clearly both $(|x_1 - m| + |x_2 - m|)$ and the $\sum_{i=1}^{n-1}$ term will be larger than when m is inside $[x_1, x_n]$.)

3. If we want m to be the mode of the sample, construct your own measure of dispersion for which m would be the minimizer. Describe how the equation you give acts as a (crude) measure of dispersion.

One possible measure of dispersion would be

$$\sum_{i=1}^{n} I(x_i \neq m),$$

where I is the indicator function ($I(\cdot) = 1$ when its argument is true and 0 when it's false).

If m is our measure of location of the data, then this measure of dispersion counts how many elements of the sample take some value other than m value. This would be minimized by the mode.

4. The blood plasma beta endorphin concentration levels for 11 runners who collapsed towards the end of the Great North Run were

66 72 79 84 102 110 123 144 162 169 414

Calculate the median and mean of this sample. Why might one have predicted beforehand that the mean would be larger than the median? Why might the standard deviation not be a very good measure of dispersion?

Median = 110, mean = 138.6.
Because of the right skew.
Because standard deviation will be sensitive to outlier 414 (so will the mean).

5. The table below gives the blood plasma beta endorphin concentrations of 11 runners before and after the race. Find the median, the mean, and the standard deviation of the [after]-[before] differences. Also, calculate the covariance and correlation of the before and after concentration levels.

Before	4.3	4.6	5.2	5.2	6.6	7.2	8.4	9.0	10.4	14.0	17.8
After	29.6	25.1	15.5	29.6	24.1	37.8	20.2	21.9	14.2	34.6	46.2

Differences are `(25.3, 20.5, 10.3, 24.4, 17.5, 30.6, 11.8, 12.9, 3.8, 20.6, 28.4)`.
Median = 20.5, mean = 18.7, sd = 7.9, cov = 19.2, cor = 0.51.

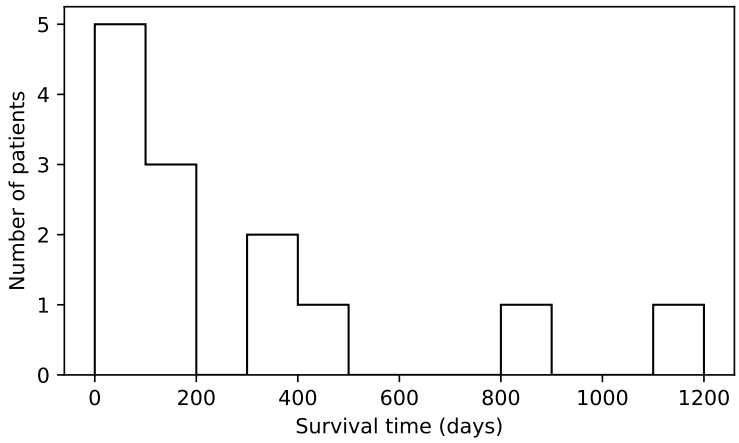
6. The data below give the percentage of silica found in each of 22 chondrites meteors. Find the median and the upper and lower quartiles of the data.

20.77, 22.56, 22.71, 22.99, 26.39, 27.08, 27.32, 27.33, 27.57, 27.81, 28.69, 29.36, 30.25, 31.89, 32.88, 33.23, 33.28, 33.40, 33.52, 33.83, 33.95, 34.82

Even number of data points so median is halfway between the 11th and 12th (when sorted into increasing order): median = (28.69, 29.36)/2 = 29.025. Lower quartile is 1/4th the way through, i.e. LQ = $x_{(n+1)/4} = x_{(23/4)} = x_{(5.75)}$ so three quarters of the way between $x_{(5)}$ and $x_{(6)}$: LQ = $26.39 + 0.75(27.08 - 26.39) = 26.9075$. Similarly, UQ = $x_{3(n+1)/4} = x_{(17.25)} = 0.75x_{(17)} + 0.25x_{(18)} = 0.75(33.28) + 0.25(33.40) = 33.31$.

7. The list below shows the survival time (in days) of patients undergoing treatment for stomach cancer. Using bins with edges at $0, 100, 200, 300, \ldots, 1200$ plot a histogram of the data. Compute the mean and standard deviation of the data. Why is the mean larger than the apparent mode of the data? Calculate the skewness of the data and of the log transformed data.

124, 42, 25, 45, 412, 51, 1112, 46, 103, 876, 146, 340, 396



Labeling the histogram is important. Mean = 286, sd = 333. Mean is larger than apparent mode because data is skewed toward large values. Whenever you have data that must be positive (like # days) but sd $\gtrsim$ mean that's an indication that the data are going to be skewed to the right.
Skewness = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})/s)^3 = 1.43$.
Log transformed data is where each sample is replaced with its log, i.e. $y_i = \log x_i$ and find skewness of the y_i's. Skewness of log transformed = 0.26 (much closer to zero, histogram of the y_i is more symmetric).

8. A car travels for 10 miles at 30 mph and then a further 10 miles at 60 mph. What was its average speed (i.e. total distance over total time)?

Total distance is 20 miles, total time is (10 miles / 30mph) + (10 miles / 60mph) = 0.5 hr, so average speed is 20 miles/0.5hr = 40 mph.

This is an example of the harmonic mean: average speed is the harmonic mean of the speeds during the two segments:

$$\text{Avg speed} = \frac{1}{\frac{1}{2}\frac{1}{30\text{mph}} + \frac{1}{2}\frac{1}{60\text{mph}}},$$

where the factors of $\frac{1}{2}$ are the fraction of the total distance covered in each segment.