Probability and Statistics for JMC
Exercises 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

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

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 $n - 2$.]

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.

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?

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

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

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

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)?

Partial answers:

4. median=110, mean=138.6
5. median=20.5, mean=18.7, sd=7.9, cov=19.2, cor=0.51
6. median=29.025, LQ=26.9075, UQ=33.31
7. mean=286, sd=333, skewness=1.43, skewness of log transformed=0.26
8. 40 mph