

- #1 (a) You can find many sets of four numbers that satisfy the two given conditions. Examples include $\{14, 15, 19, 20\}$, $\{13, 16, 18, 21\}$, $\{12, 17, 19, 20\}$, and etc. You need to verify that four numbers satisfy the two conditions. For example, for $\{14, 15, 19, 20\}$,

$$\bar{x} = \frac{14 + 15 + 19 + 20}{4} = 17,$$

and

$$IQR = Q_3 - Q_1 = \frac{19 + 20}{2} - \frac{14 + 15}{2} = \frac{19 + 20 - (14 + 15)}{2} = 5,$$

where the 1st quartile (or 25% percentile), i.e. Q_1 , is the median of the lower half of the data and the 3rd quartile (or 75% percentile), i.e. Q_3 , is the median of the upper half of the data.

- (b) Likewise, there are many correct answers, including $\{26, 27, 27, 28\}$, $\{25, 27, 27, 29\}$, $\{25, 27, 27, 27\}$, and etc. You need to verify that four numbers satisfy the two conditions. For example, the numbers $\{26, 27, 27, 28\}$ have mode = 27, since the number 27 appears the most, and have median = 27 since the two middle numbers in the data are both 27.
- (c) The numbers $\{3, 3, 3, 3\}$ have mean = 3 since

$$\bar{x} = \frac{3 + 3 + 3 + 3}{4} = 3,$$

and the largest number (i.e. maximum) = 3.

- (d) There are many correct answers, such as $\{-1, 0, 0, 1\}$, $\{-2, 0, 0, 2\}$, $\{-3, 0, 0, 3\}$, and etc. You need to verify that four numbers satisfy the three conditions. For example, the numbers $\{-1, 0, 0, 1\}$ have mode = 0 (since 0 appears most in the data), median = 0 (since the middle two numbers are 0), and mean = 0 since

$$\bar{x} = \frac{-1 + 0 + 0 + 1}{4} = 0$$

Chapter 4 review exercises:

- # 1 (a) The average (\bar{x}) of the 6 numbers can be obtained as follows:

$$\bar{x} = \frac{\sum_{i=1}^6 x_i}{6} = \frac{41 + 48 + 50 + 50 + 54 + 57}{6} = 50$$

Given the average, \bar{x} , is 50, we can find the standard deviation (SD) as follows:

$$\begin{aligned}
 \text{SD} &= \sqrt{\frac{\sum_{i=1}^6 (x_i - \bar{x})^2}{6}} \\
 &= \sqrt{\frac{(41 - 50)^2 + (48 - 50)^2 + (50 - 50)^2 + (50 - 50)^2 + (54 - 50)^2 + (57 - 50)^2}{6}} \\
 &= \sqrt{\frac{(-9)^2 + (-2)^2 + (0)^2 + (0)^2 + (4)^2 + (7)^2}{6}} \\
 &= 5
 \end{aligned}$$

(b) First, note that $0.5 \text{ SD} = 2.5$. Hence, for numbers within 0.5 SD of the average, we are looking for x that satisfies $47.5 \leq x \leq 52.5$. The numbers in this range are 48 and 50.

Second, note that $1.5 \text{ SD} = 7.5$. For numbers within 1.5 SD of the average, we are looking for x that satisfies $42.5 \leq x \leq 57.5$. The numbers in this range are 48, 50, 54, and 57.

2 (a) (ii) has a smaller standard deviation (SD) than (i). Note that in (ii) the three additional values are equal to the average. Recall that adding values close to the mean to a set of numbers will result in a lower SD. Also, if we computed the SD for both data sets, we would find that the numerator is the same for both (i) and (ii), yet the denominator is larger for (ii) (namely 10 vs. 7).

(b) (i) has the smaller SD. Adding these two observations (e.g. 99 and 1) increases the spread of the distribution, and as the spread of the distribution increases the SD increases.

4 In the U.S., more people earn a small amount of money, and few people are rich. Thus, the U.S. income distribution is right skewed. When a distribution is skewed to the right, the average is higher than the median. Hence, the average income would be higher than the median income.

For years of schooling completed, the average is again higher than the median. Recall that most people in the US have only a high-school education, and fewer people go on to college-level study or above. The distribution is again right skewed, so the average years of education would be higher than the median.

5 In order to evaluate if a blood pressure is unusually high, unusually low, or about average recall the empirical rule(s):

- (a) Approximately 68% of all observations are within ± 1 standard deviation (SD) of the average.
- (b) Approximately 95% of all observations are within ± 2 SDs of the average.
- (c) Approximately 99.7% of all observations are within ± 3 SDs of the average.

Now look at the various blood pressures. 80mm is unusually low, since it is more than 3 SDs (i.e., $\frac{116-80}{11} = 3.2727$) below the average. Less than 0.15% of all men age 18-24 in HANES had a blood pressure that is lower than 80mm.

115mm and 120mm are about average, since they are less than 1 SD (i.e., $\frac{116-115}{11} = 0.0909$ and $\frac{120-116}{11} = 0.3636$, respectively) below and above the average. This is consistent with the empirical rule according to which 68% of all men age 18-24 in HANES have a blood pressure between 105mm (i.e. $116-11=105$) and 127mm (i.e. $116+11=127$).

Finally, 210mm is unusually high, since it is much more than 3 SDs (i.e., $\frac{210-116}{11} = 8.5455$) above the average. This is an extreme outlier, and much less than 0.15% of all men age 18-24 in the HANES had blood pressure that high or higher.

- # 6 (a) Histogram (i) is left skewed, histogram (ii) is approximately symmetric and centered at 50 and histogram (iii) is right skewed.

It is straightforward to see that (ii) has an average of 50. Note that (ii) is a symmetric distribution around 50, which is its mode. In a symmetric unimodal distribution, the mode, the median, and the average (or mean) are about the same.

(i) is left-skewed, meaning the average is smaller than the median, and (iii) is right-skewed, where the average is larger than the median. Hence, given the shape of the distribution, the average of (i) is smaller than the average of (ii) (i.e. 50), which is smaller than the average of (iii). This indicates (i) has the average of 40, (ii) 50, and (iii) 60.

(b) The median is less than the average, when the distribution is right skewed, i.e. (iii). The median is about equal to the average when the distribution is symmetric, i.e. (ii). The median is bigger than the average when distribution is left skewed, i.e. (i)

(c) The standard deviation of histogram (iii) is about 15. To see this, recall the empirical rule(s) stated above and that the total area under the curve is equal to 100%. Also note that the average of (iii) according to part (a) is 60.

50 can then be ruled out, since that would mean ± 1 SD of 60 is equal to (10,110), which covers almost the whole curve, while according to

the empirical rule, ± 1 SD should only cover approximately 68%. By the same argument 5 can also be ruled out, since the area between 55 and 65 contains less than 68% of all observations.

(d) False. While histogram (i) is left skewed and histogram (iii) is right skewed, they seem to be skewed about to the same degree. Note that the standard deviation is defined as $SD = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$. Since histogram (i) almost mirrors histogram (iii) on a vertical line through 50 the numerator of the fraction under the square root is almost identical. If N is the same in (i) and (iii) as the figures suggest, then the SDs of (i) and (iii) are almost equal and the statement therefore is false.

- # 7 (a) Let \bar{x}_m and $\bar{\sigma}_m$ be the average and SD of the men's weight, while \bar{x}_w and $\bar{\sigma}_w$ be the average and SD of the women's weight. To get the values in pounds, multiply each average kilogram value and each SD by 2.2.

Therefore, $\bar{x}_m = 145.2lb$ and $\bar{\sigma}_m = 19.8lb$, and $\bar{x}_w = 121lb$ and $\bar{\sigma}_w = 19.8lb$.

(b) Note that 57 kg is 1 SD below the average weight of the men (i.e. $57 = 66 - 9$), while 75 kg is 1 SD above the average (i.e. $75 = 66 + 9$). Recall the empirical rule, saying roughly 68% of all observations are within ± 1 SD of the average. Therefore, we know that approximately 68% of men weigh between 57kg and 75kg.

(c) The spread the distribution of the combined (i.e. men and women) group's weights is now bigger than that of only men's or only women's weights. Recall that the SD increases as the spread of the distribution increases. Therefore, the SD of the combined group is bigger than 9 kg.

- # 9 (a) Yes. This accident will increase the average (\bar{x}). To see by how much \bar{x} increases, let us consider \bar{x}_{old} , which refers to the average when the highest income (i.e. x_{100}) in the file was \$98,600, and \bar{x}_{new} , which refers to the average when the highest income (i.e. x_{100}) is changed to \$986,000.

Note that $\bar{x}_{old} = \frac{\sum_{i=1}^{1000} x_i}{1000}$, where $x_{100} = 98,600$.

Now consider the average of the income when $x_{100} = 986,000$

$$\begin{aligned}\bar{x}_{\text{new}} &= \frac{\sum_{i=1}^{999} x_i + 986,000}{1000} \\ &= \frac{\sum_{i=1}^{999} x_i + 98,600 - 98,600 + 986,000}{1000} \\ &= \frac{\sum_{i=1}^{999} x_i + 98,600}{1000} + \frac{887,400}{1000} \\ &= \frac{\sum_{i=1}^{1000} x_i}{1000} + \frac{887,400}{1000} \\ &= \bar{x}_{\text{old}} + 887.4\end{aligned}$$

Hence, the average when the highest income is changed to \$986,000 is bigger than the old average by \$887.40.

(b) No, the median is not affected. Note that the number of observations is still the same (i.e. $N=1000$), implying that the two middle numbers are the 500th and 501st observations as before. Therefore, the median (i.e. $\frac{x_{500} + x_{501}}{2}$ where x_{500} is the 500th observation and x_{501} is the 501st observation), stays the same since the 500th and 501st observations remain the same.