

**Chapter 4 review exercises:**

- # 1 (a) The average ( $\bar{x}$ ) of the  $N = 6$  numbers can be found by plugging them in the formula

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

The standard deviation (SD) is equal to

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

- (b)  $0.5 \text{ SD} = 2.5$  and  $1.5 \text{ SD} = 7.5$ . Thus, the set of numbers within 0.5 SDs of the average ( $\bar{x}$ ) is  $\{48, 50, 50\}$ , since 0.5 SDs from  $\bar{x}$  covers the interval  $(50-2.5, 50+2.5) = (47.5, 52.5)$ . The set of numbers within 1.5 SDs of  $\bar{x}$  is  $\{48, 50, 50, 54, 57\}$ , since 1.5 SDs from  $\bar{x}$  covers the interval  $(50-7.5, 50+7.5) = (42.5, 57.5)$ .

- # 2 (a) (ii) has the smaller standard deviation (SD). Note that in (ii) the three additional values are equal to the average and that adding values close to the mean to a set of numbers that is overall more spread out will always 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. Intuitively, think of the last two values of (ii) as (extreme) outliers (they are both 49 away from the mean, while the other values are at most 25 away). Thus, we add two “outliers” and therefore obtain a higher SD in (ii) than in (i), even though the denominator in (ii) is larger than the one for (i).

- # 4 In the U.S., the average income is higher than the median income, since the U.S. income distribution (as in most other countries) is right skewed.<sup>1</sup> Those who have trouble following the argument might want to recall the example of baseball players’s salaries made in class. In their case, the graph was skewed to the right, indicating that the average income (salary)

<sup>1</sup>For a graphical representation of the U.S. income distribution for the bottom 98% of all U.S. households in 2005 (i.e., extreme outliers like Bill Gates or Warren Buffett were excluded) see <http://www.visualizingeconomics.com/2006/11/05/2005-us-income-distribution/>. On the horizontal axis of the graph you will see that the median household income in 2005 was \$ 46,326, while the mean household income was \$ 63,344.

is higher than the median income (salary). For the U.S. population as a whole, the situation is comparable.

For years of schooling completed, the average is again higher than the median. Intuitively, it will do to recall that almost everyone in this country has a high-school education (giving us a median value of 12 years), but less than half the population goes on to college-level study. As for the average, even though relatively few people have advanced degrees, even fewer drop out of high school. (You may also want to refer back to the data presented in chapter 3 in FPP, pp. 38-39. On page 38, you'll find data for 1991 that at least seems to indicate that the median is 12 years of schooling (i.e., a typical high school career), while the average is (slightly) higher.)

# 5 <sup>2</sup>

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

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

Now look at the various blood pressures. 80mm is unusually low, since it is more than 3 SDs (i.e.,  $\frac{124-80}{14} = 3.14$  to be exact) below the average. Less than 0.15% of all men age 18-24 in HANES had a blood pressure that low or lower. 115mm and 135mm are about average, since they are less than 1 SD (i.e.,  $\frac{124-115}{14} = 0.61$  and  $\frac{135-124}{14} = 0.79$  to be exact) below or above the average. Recall that according to the empirical rule 68% of all men age 18-24 in HANES have a blood pressure between 110mm and 138mm. Finally, 210mm is unusually high, since it is more than 6 SDs (i.e.,  $\frac{210-124}{14} = 6.14$  to be exact) above the average. This is an extreme outlier, far less than 0.0001% of all men age 18-24 in the HANES had blood pressure that high or higher.

- # 6 (a) Have a look at the shape of the three histograms first. Histogram (i) is left skewed, histogram (ii) is approximately symmetric and centered at 50 and histogram (iii) is right skewed. Now, if each of the three

---

<sup>2</sup>The response in the main body is only for the 3rd edition. The logic of the problem as it shows up in the 4th edition is identical. Here, the numbers we are given are 80, 115, 120, and 210. The mean is 116 and the SD is 11. The corresponding z-scores, then, are -3.27, -0.09, 0.36, and 8.54. The first and last of these are unlikely (unusually low and unusually high, respectively) if we assume the normal distribution is a good approximation of the distribution of blood pressures.

averages above belong to one (and only one) of the histograms, than 50 clearly belongs to (ii). To see this imagine a set of 9 numbers between 0 and 100, distributed symmetrically around 50 according to (ii) (e.g., 20, 40, 40, 50, 50, 50, 60, 60, 80). You can now easily check by plugging the numbers into the formula for the average that it is 50.

To match the remaining averages to the remaining histograms, recall that the mean is “outlier sensitive”. That is, adding a clear outlier to the left (right) of a set of numbers pulls the average of that set of numbers to the left (right). To see this assume you had a set of 4 equal numbers, for example 5, 5, 5, 5. Obviously, the average of those 4 numbers is 5. Now add the number 10 to the set. The new average is 5.83, which is greater than 5. Now add 0 instead of 10 to the set and you will see that the new average is 4.16, which is clearly smaller than 5. Thus, 40 belongs to histogram (i) and 60 to histogram (iii), since right (left) skewed distributions indicate that there are some outliers on the right (left) end of the horizontal scale.

- (b) Recall that the median is defined as the number that splits the sorted set of numbers in half (i.e., 50% of all observations are below and the other 50% of them are above the median.). Also, recall that the average is “outlier sensitive” as illustrated in (a). Based on those two facts it follows immediately that the statement “the median is less than the average” belongs to histogram (iii), the statement “the median is about equal to the average” belongs to histogram (ii) and the statement “the median is bigger than the average” belongs to histogram (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% (i.e., the distribution covers all observations). 50 can then be ruled out, since  $\pm 1$  SD of 60 (= the average by (a)) is equal to (10, 110), which covers almost the whole curve. However, by the empirical rule, it should only cover approximately 68%. By the same argument 5 can also be ruled out, since the area between 55 and 65 contains clearly less than 68% or all observations.
- (d) False. While histogram (i) is left and histogram (iii) is right skewed, they are skewed about to the same degree. That is, histogram (i) mirrors histogram (iii) across a vertical line through 50. The standard deviation is defined as  $SD = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$ . Note that because histogram (i) mirrors histogram (iii) on a vertical line through 50 the numerator of the fraction under the square root is identical. If N is the same in (i) and (iii) as the figures suggest, then the SDs of (i) and (iii) are equal and the statement therefore false.

- # 7 (a) To find the values in pounds, simply multiply each average kilogram value and each SD by 2.2 (male: average 145.2, SD 19.8 / female:

average 121, SD 19.8).

- (b) Applying the empirical rule, we know that roughly 68% of all observations are within  $\pm 1$  SD ( $57 = 66 - 9$ ,  $75 = 66 + 9$ ) from the mean.
- (c) If we took the men and women together, the new “joint” SD would go up. To see why, note that the observed values will be spread out more than before (i.e., the two subgroups were each more homogeneous than the new, larger group).

For some intuition, consider two groups that both have a SD of 0 before pooling. One group consists of only one observation, namely 0. The other group has four elements, all of which are 100. As should be clear, a new group with 0, 100, 100, 100, 100 will have a SD  $> 0$  (we obtain, in fact, SD = 40), so simply adding up the old SD is misleading.

- # 9 (a) Yes, this accident will affect the average ( $\bar{x}$ ). In fact, it will lead to an increase in  $\bar{x}$ . To see by how much, we need to look at the formula of  $\bar{x}$ , i.e.,

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N},$$

where  $N=1000$ . Now consider the formula including the mistake

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

Thus, due to the mistake  $\bar{x}$  increases by \$887.4.

- (b) No, the median is not affected. To see this, recall that the median is the 50th percentile (i.e., 50% of all observations are smaller than and the other 50% of them are greater than the median.). Since the accident only increased the maximum number, which did not change the order around the median, the median remains unaffected.

- # 12 The data reported in the question can be seen as *consistent* with a claim that there is a permanent underclass, although they are not likely to be sufficient to establish the claim. However, the extent to which the data is consistent with or able to establish a claim depends on both one’s definition of “permanent underclass” and on the measure (or operational definition) used to identify those who are in poverty.

One possible definition of a permanent underclass would be a block of the population which spend its entire life in poverty, roughly speaking. If this is to what is referred by “permanent underclass”, then it is not sufficient to show that the size of the block of the population in poverty is constant.

Instead, one would also need to know that those within the block are consistently identified as poor. Of course, there are alternative definitions of “permanent underclass” and each would place different demands on data in order to establish the claim. Next, consider the way poverty was operationalized. Here, we can imagine two classes of measures: relative and absolute. Depending on which kind of measure was used, identifying individuals as impoverished would comport with our sense of an underclass to various degrees.

## Chapter 5 Review Exercises

# 2 The printout does not look reasonable. Since these test scores are converted to standard units (i.e., z scores from a normal table), we know from the empirical rule that over 99 percent of the data should lie between -3 and 3 under a standard normal curve. Out of these ten scores, there are seven that lie outside of this interval, which is unlikely to happen (at least 7 percent of the data lie outside of three standard deviations, contradicting our empirical rule).

# 3 <sup>3</sup>

- (a) Since  $z = \frac{x-\mu}{\sigma}$ , we have  $z = \frac{600-466}{110} \approx 1.22$  in this case. Table 1 from the handout shows this z score corresponds to 0.1112, which means about 11.12 percent of students scored over 600 in 1967.
- (b) Use the same formula we have  $z = \frac{600-423}{110} \approx 1.61$  for the year 1994. The same table shows 0.0537 corresponds to this z score. Therefore, 5.37 percent of students scored over 600 in 1994.

# 4 <sup>4</sup>

- (a) In this case,  $z = \frac{600-500}{120} \approx 0.83$ . The normal table shows 0.2033 corresponds to  $z = 0.83$ . As a result, 20.33 percent of men got over 600 in 1994.
- (b) For women,  $z = \frac{600-460}{120} \approx 1.17$ . The normal table shows 0.1210 corresponds to this z score, which means 12.1 percent of women got over 600 in 1994.

<sup>3</sup>The solutions to the corresponding problems in the 4th edition use the same approach but rely on different numbers. For the 4th edition, the percentage of students scoring over 700 if the mean was 543 and the sd was 110 is about 8%. The percentage of students scoring over 700 if the mean is 499 and the sd is unchanged is roughly 3.5%.

<sup>4</sup>The solutions to the corresponding problems in the 4th edition use the same approach as is provided above, but the numbers are different. Assuming the mean for men is 538 and for women is 504 where both SDs are 120, then the percentage of each that scores over 700 is about 9% and 5%, respectively.

# 7 <sup>5</sup>

- (a) A student who scored 350 on the Math SAT was at the *6.68* percentile of the score distribution.

In this question,  $z = \frac{350-500}{100} \approx -1.5$  and the normal table shows 0.0668 corresponds to  $z = 1.5$  (recall that normal distribution is symmetric).

- (b) To be at the 75th percentile of the distribution, a student needed a score of about *567* points on the Math SAT.

In this question, the 75th percentile corresponds to  $z = 0.67$  or  $z = 0.68$  (find the  $z$  score that corresponds to 0.25 in the normal table). Use  $z = 0.67$ , we have  $0.67 = \frac{x-500}{100}$ . Solve for  $x$ , we get  $x = 567$ .

- # 8 (a) True.  
(b) False.  
(c) True.  
(d) True.  
(e) True.  
(f) False.

See the lecture on Change of Scale for explanations.

---

<sup>5</sup>The solutions to the corresponding problems in the 4th edition use the same approach but have different numbers in the problem. Here, a student who scored a 400 if the mean is 550 and the SD is 100 would be at the 7th percentile. A student at the 75th percentile would have to score a 617.