Statistics
Chapter 1

Section 1.2 Exercises

Suggested Problems 1. Variables and Data; Types of Variables

1. The administrators of a hospital want to know how many hours each patient stayed in the hospital after being admitted last year. Two thousand patient records are randomly chosen from last year's admissions and the length of stay is examined for each patient.
a) State the population
b) State the sample

2. A manager at a manufacturing plant wants to monitor the masses of drill bits coming off the manufacturing line today. One hundred drill bits are randomly selected throughout the day and the mass of each is recorded.
a) State the population
b) State the sample

For Questions 3-5 read the paragraph below:
Cars are observed entering a campus parking lot. The following data is recorded for each car: make, model, number of occupants, number of seatbelts in vehicle, distance displayed on odometer, diameter of front left tire.

3. List the variables.

4. State the type of each variable.

5. State the experimental unit.

For Questions 6-8 read the paragraph below:
Seventeen soil samples are taken from Quadrants A, B, C and D of a drilling site. For each sample the following data is obtained and recorded: depth at which sample was drawn, mass of sample to the nearest gram, quadrant sample was drawn from, and primary element by mass.

6. List the variables.

7. State the type of each variable.

8. State the experimental unit.

9. State the type of each variable below:
a) Time to complete a math test
b) Country of birth
c) Total cost of textbooks for this term
d) Mass of student's wallet

10. State the type of each variable below:
a) Student's height, rounded to the nearest inch
b) Employer from current or last job
c) Number of college or university course credits achieved
d) Last three symbols of postal code

1. a) The lengths of stay for all patients the hospital admitted last year.
   b) The lengths of stay for the 2000 patients whose records are chosen.

2. a) The masses of all drill bits manufactured today.
   b) The masses of the 100 selected drill bits.

3. make
   model
   # of occupants
   # of seatbelts
   distance on odometer
   diameter of front left tire

4. Make : qualitative
   Model : qualitative
   # of occupants : (quantitative) discrete
   # of seatbelts : (quantitative) discrete
   distance on odometer : (quantitative) discrete
   diameter of front left tire : (quantitative) Continuous

5. The car

(6) depth of sample
mass of sample (nearest g)
quadrant
primary element by mass

(7) Depth of Sample : (quantitative) Continuous
Mass of Sample (nearest g) : (quantitative) discrete
Quadrant : qualitative
Primary Element by Mass : qualitative
(for example Carbon, Nitrogen etc.)

(8) The soil sample

(9) a) (Quantitative) Continuous
b) Qualitative
c) (Quantitative) Discrete
d) (Quantitative) Continuous

(10) a) (Quantitative) Discrete
b) Qualitative
c) (Quantitative) Discrete
d) Qualitative

Section 1.4: Measures of Centre and of Variability

Exercises

1. The top ten movies (Skyfall, The Hobbit, etc.) and their profits (in millions of dollars) from last weekend are reported in Monday Magazine. Calculate the mean and median for this data.

   profits: 1.4, 4.1, 1.2, 1.3, 5.8, 5.0, 2.6, 1.8, 2.9, 5.9, 2.5, 5.3

2. Calculate the mean and median for the data set: 35, 47, 29, 42, 38, 39, 42.

3. Pat finds the mean height of all twelve students in her physics class to be 68.0 inches. Just as she's finished that calculation, one more student walks in late. If that student is 63.0 inches tall, what is the mean height of all thirteen students?

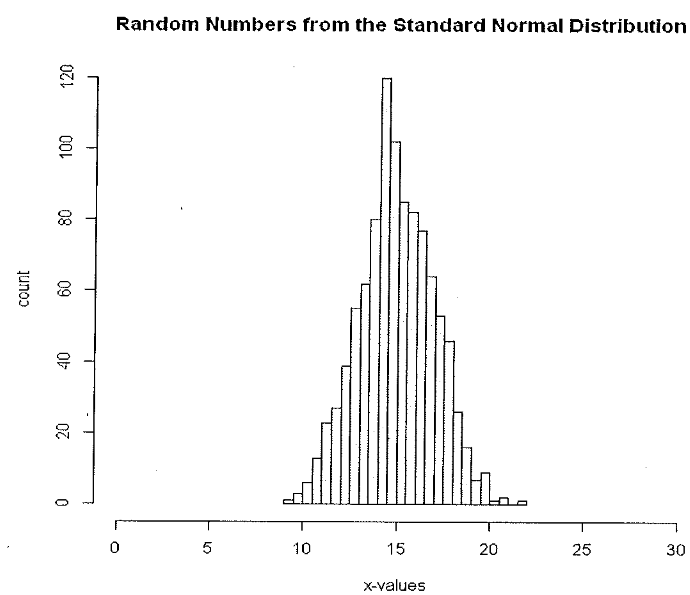
4. The Victoria Real Estate Board claims that in October of 2012, the average cost of a single-family home in Greater Victoria was $592,000, while the median was $527,000. Why is the mean greater than the median for housing prices? Explain.

5. Tom is running a small business with five employees, including himself. The salaries of the five people (in thousands of dollars) are 30, 45, 50, 55, and 75, with Tom making the highest salary.

   a) calculate the mean and median of these salaries
   b) if Tom gives everyone a $2000 bonus, what happens to the mean and median?
   c) if Tom gives everyone a 5% raise, what happens to the mean and median?
   d) if Tom decides to keep everyone else's salary the same, but raise his own salary by $10,000, what happens to the mean and the median?

6. Consider the following histogram. Is the standard deviation equal to

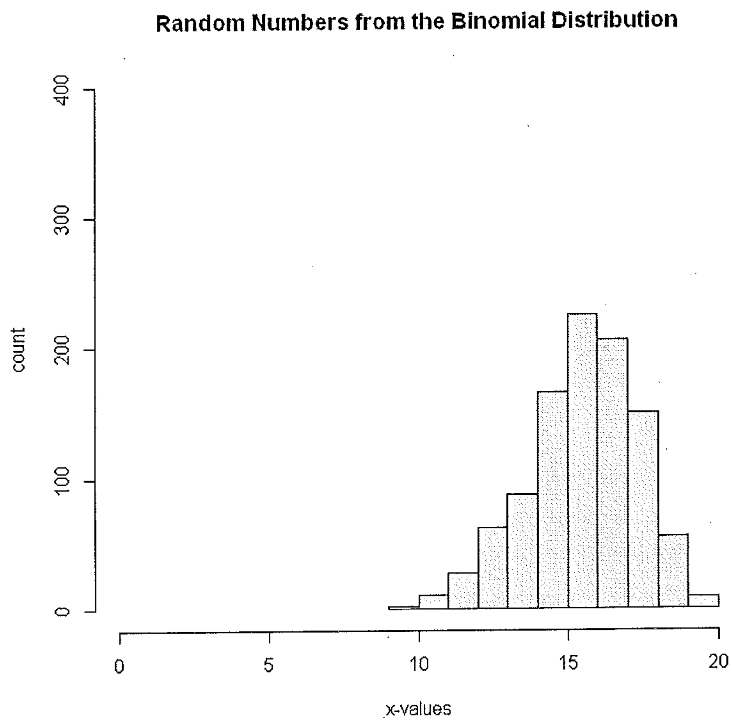
   a) 0.5
   b) 2
   c) 15
   d) 20



7. Consider the following data set: 7, 7, 7, 7, 7, and 7. What is the mean and the median? What is the range? Without calculating it, what would be the standard deviation?

8. Consider the following histogram. Is the standard deviation equal to

a) 1
b) 2
c) 5
d) 10
e) 15
f) 20



9. Pat, when entering quiz scores into her spreadsheet, accidentally put an extra zero on the end of one student's score (making it 380/40 instead of 38/40), and then calculated the mean, median, range, and standard deviation for the section. She then noticed her mistake and recalculated all of the quantities. For the following quantities, state whether the corrected value will be higher, lower, or the same as the value calculated with the incorrect quiz score:

a) mean
b) median
c) range
d) standard deviation

10. Consider the following sets of data. Without calculating any values, state which set will have the higher standard deviation (or will they be the same?).

a) Set 1: 2, 3, 9, 16, 17 Set 2: 2, 8, 9, 10, 17
b) Set 1: 2, 3, 9, 16, 17 Set 2: 3, 4, 10, 17, 18

Section 1.4: Statistical Quantities

Solutions

1. The mean is 3.31667, or just 3.3. There are twelve points, so the median is the 12/2+1/2=6.5th point, which means the average of the 6th and 7th points. Therefore, the median is (2.6+2.9)/2 = 2.75.

2. The mean is 38.8571. (You can round to 38.9 if you like.)

3. To find the mean, we want the sum of all of the heights divided by the total number of students. Since the average of the twelve students is 68.0 inches, the total of all of those heights is just 68.0 times 12, which is 816.0 inches. Adding the height of the thirteenth student brings the total to 879.0 inches, then dividing by 13 gives a mean of 67.6 inches.

4. The histogram of Victoria housing prices will not be symmetrical: there is a lower limit for the price of single-family homes, while there can be house prices in the millions of dollars. Just a few very expensive homes will bring up the mean but not affect the median in any way, which is why the mean is greater than the median.

5. The means and medians are:

   a) mean = $51,000 and median = $50,000
   b) the mean and median will each increase by $2000: mean is now $53,000 and the median $52,000
   c) the mean and median will both increase by a factor of 1.05 (they are multiplied by 1.05): mean is now $53,550 and median is $52,500
   d) the mean will become $53,000 but the median will stay the same

6. Looking at the histogram, you can estimate the standard deviation by picking a "width" about the mean/average that most of the data points fall within. From this histogram, the standard deviation is about half of 5, since most of the data falls between approximately 12.5 and 17.5 (ish). And the closest value given that matches that is (b) 2.

7. The mean and median are both 7. The range is 0. The standard deviation is also 0, since all points lie exactly on the mean and $(x-\overline{x})$ is zero for each point.

8. Using the same reasoning as for question 6, most of the data seems to fall between 12.5 and 17.5, so the standard deviation is around 2.5 (ish). So the closest option given is (b) again.

9. New values:

   a) The corrected mean will be lower, since one value was lowered.

b) The median will remain unchanged (assuming that the 38/40 was in the upper half of the scores to begin with, so changing it to 380 and back won't affect that)
c) The corrected range will be lower, since the highest point has changed.
d) The standard deviation will be lower, since the corrected point's distance from the mean is lower than the uncorrected value.

10. a) Set 1's values are farther from the mean on average than Set 2's data points. So Set 1 will have a higher standard deviation.

b) Set 2's data points are just Set 1's points moved up by 1 unit. So each point's distance from the mean will be the same as Set 1, and the standard deviations will be the same also.

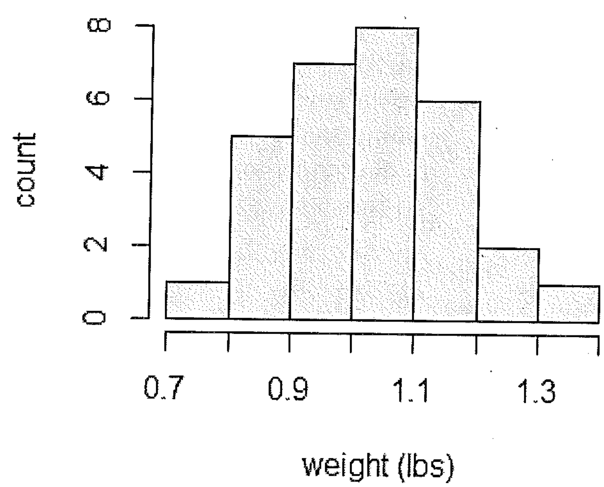
Section 1.5: The Empirical Rule

Exercises

1. A market researcher listed the weights (in pounds, sadly, considering that we are supposed to be a metric country) for 30 packages of ground beef in Thrifty's meat display.

   1.08, 0.99, 0.97, 1.18, 1.09, 1.28, 0.83, 1.06, 1.14, 1.38, 0.75, 0.96, 1.08, 0.87, 0.89, 0.89, 0.96, 1.01, 1.12, 1.06, 0.93, 1.24, 0.89, 0.98, 1.14, 0.92, 1.18, 1.17, 1.02, 1.03

   The mean and standard deviation of this sample are 1.04 and 0.14 lbs, respectively, and the data is graphed in the plot below.



   a) Describe the shape of the resulting distribution. Is the distribution mound-shaped?
   b) Find the percentage of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. In your table, also state what percentages you expect to see in these intervals using either Tchebysheff or the Empirical Rule.
   c) Do the percentages obtained in part b) agree with those given by the Empirical Rule? By Tchebysheff? Should they?

2. A set of data has a mean of 75 and a standard deviation of 5. The histogram shows a more-or-less symmetric, mound-shaped distribution.

   a) What can you estimate about the proportion of measurements that fall between 70 and 80? Between 65 and 85?
   b) What can you say with certainty about the proportion of measurements that fall between 70 and 80? Between 65 and 85?

3. The top ten movies (Skyfall, Men in Black III, etc.) and their profits (in millions of dollars) from last weeked are reported in Monday Magazine. Consider this data to be a sample of movie profits for weekends in the year 2012.

profits: 1.1, 1.1, 1.2, 1.3, 5.1, 6.0, 9.6, 9.8, 9.9, 9.9

The mean and standard deviation of this data set are 5.5 and 4.1 million dollars, respectively.

a) By looking at the data, do you think that this distribution relatively mound-shaped?
b) Find the percentage of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$.
c) How do the percentages obtained in part c compare with those given by the Empirical Rule and Tchebysheff's Theorem? Do they agree? Do you expect them to?

4. In the 2000 Olympic Games in Sydney, the mean time on the 800 m event was 137 seconds with a standard deviation of 4 seconds. Gertrud Bacher of Italy finished in 129 seconds. Assuming that these running times have a mound-shaped distribution and that there were 38 competitors in total, on average how many of the competitors would you expect to finish ahead of the Italian? (She won the event, by the way.)

Section 1.5: The Empirical Rule

Solutions

1. Thrifty Foods:

   a) The distribution is roughly symmetrical and only has one peak (it is unimodal). Yes, it is mound-shaped.
   b) Consider the following table.

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$	0.90 – 1.18	21	70.0%	~68%	---
$\bar{x} \pm 2s$	0.76 – 1.32	28	93.3%	~95%	$\geq 75\%$
$\bar{x} \pm 3s$	0.62 – 1.46	30	100%	~99.7%	$\geq 89\%$

   c) See the table above.
   d) The agreement between the actual numbers and the Empirical Rule is pretty good (considering that there are only 30 points in the distribution, so each point represents 3% of the measurements), which is to be expected since the distribution is mound-shaped. As always, Tchebysheff is accurate since the predictions are true for all distributions.

2. Since the distribution is symmetrical, the Empirical Rule applies which allows you to estimate percentages. And Tchebysheff's theorem, which applies to all distributions, allows you to calculate with certainty lower limits on percentages.

   a) Between 70 and 80 is one standard deviation away from the mean, so by the Empirical Rule we can estimate that ~68% of measurements lie in this range. Between 65 and 85 is two standard deviations away from the mean, so ~95% of measurements will lie in this range.
   b) Tchebysheff's theorem has nothing to say about the percentage of measurements within one standard deviation of the mean, so we can't say anything with certainty about those that lie between 70 and 80. But *at least* 75% will fall between 65 and 85.

3. Movies:

   a) There's a group of measurements down around 1.1, then a couple in the middle, then a big group up at 9.9. So this distribution has more than one peak, and is not mound-shaped.

b) Here's a table:

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$	1.4 – 9.6	3	30.0%	~68%	---
$\bar{x} \pm 2s$	(-2.7) – 13.7	10	100%	~95%	$\geq 75\%$
$\bar{x} \pm 3s$	(-6.8) – 17.8	10	100%	~99.7%	$\geq 89\%$

c) The actual percentages don't agree with the Empirical Rule predications at all, which is not surprising since the distribution is not mound-shaped and the Empirical Rule does not apply. As always, Tchebysheff applies and is accurate.

4. Bacher finished 8 seconds ahead of the mean, which is two standard deviations. Since the distribution is mound-shaped, the Empirical Rule applies, and ~95% of the competitors' times should fall within two standard deviations of the mean. So, ~5% of the times will lie outside of this range, and since the distribution is symmetrical, we can assume that ~2.5% will lie below Bacher's time. Multiplying this by the number of competitors gives 0.95 competitors who will finish ahead of Bacher, which rounds to 1 person on average finishing ahead of her.