Samples and Populations: Homework Examples from ACE
Investigation 1: #19, 26, 33.
Investigation 2: #9, 27.
Investigation 3: #5.
Investigation 4: #7.

ACE Question	Possible Answer
ACE Investigation 1	
19. How much taller is a student in grades 6 – 8 than a student in grades K – 2? Explain. (See student text for histograms.)	Note: The two histograms shown have organized information about two groups of students. Each "bar" on the histogram shows how frequently a particular height was observed. For example, the furthest left bar on the histogram for K-2 heights indicates that only one student had a height between 105 and 110 centimeters. We cannot read information about a particular student in this group, but we have a general picture of heights of all the students in this group. ACE question 19 can only be answered if we think of comparing "typical" students in the two groups. 19. A typical student in grades 6 – 8 has a height between 150 and 175 centimeters. Very few students are outside this cluster of students. If we want to narrow this further we might use the mean or median, which are both in the 165 to 170 centimeter interval, closer to 165 centimeters. Meanwhile a typical student in grades K – 2 has height 115 to 135 centimeters, or, narrowing this further, about 125 centimeters. Thus, we might compare means and say that a typical student in grade 6 – 8 is about 40 centimeters (165 – 125) taller than a typical student in grades K – 2. (If we compared two actual students the difference might be more or less than this.)
26. Tim says that TastiSnak raisins are a better deal than Harvest Time raisins because there are more raisins in each box. Kadisha says that, because a box of either type contains half an ounce, both brands give you the same amount for your money.	Note: Box plots are another way to organize information to show the big picture. In histograms the data are organized in intervals and shown as bars. In box plots the information is put in order and divided into 4 equal-sized groups. Values called *quartiles* are used to divide the data. 25% of the data lies at or below the first quartile, 25%

The students found the number of raisins and the mass for 50 boxes of each type. They made the plots shown. Based on this information, which brand is a better deal? Explain. (See student text for graphs.)	between the first quartile and the second quartile (median), 25% between the median and the third quartile, and 25% at or above the third quartile. The "box" in the middle indicates the range of values for the middle 50% of the data, so is a way of talking about both how spread out or clustered the data are, and what is a typical value. 26. TastiSnak raisin boxes typically contain about 37 – 40 raisins per box, while Harvest Time boxes typically contain fewer raisins, about 28 – 31 raisins (see middle 50% of data). Comparing the medians, TastiSnak boxes have 10 more raisins. The typical weights per box of the two brands are not so different as the typical number per box. TastiSnak boxes typically contain between about 16.5 and 17.5 grams, while Harvest Time boxes typically contain about 16.2 to 16.8 grams. Comparing the medians, TastiSnak boxes weigh 0.5 grams more than Harvest Time boxes. So, whether you want more raisins or more weight you are more likely to find this in a TastiSnak box. (However, it is possible that two individual boxes will contradict this statement. The maximum number of raisins found in a Harvest Time box is 35. There are TastiSnak boxes with fewer raisins. Likewise we can find a box of Harvest Time that weighs more than a box of TastiSnak.)
33. Bill and Joe are interested in baseball. The histogram below shows data they collected about the duration of professional baseball games. The title and axes are missing. (See student text for graph.) a. What title and axis labels are appropriate fro this graph? b. B. What does the shape of the graph tell you about the length of a typical baseball game? c. About how many games are represented in the graph? d. Estimate the lower quartile, median, and upper quartile for these data. What do	33. a. Not answered here. b. Not answered here. c. The furthest left bar on the histogram indicates that 7 games lasted between 120 and 130 minutes. Continuing across the graph from the left we can total the number of games represented: 7 + 12 + 27 + 25 + 18 + 28 + 15 + 6 + 3 + 1 + 1 + 1 + 1 + 1 + 1 = 147 games. d. With 147 games to divide into 4 groups we have 36 shortest games, then the first quartile, then another 36 games, then the median game length, then another 36 games between the median and the third quartile, and another 36 game lengths above the third quartile. (36 + 1 + 36 + 1 + 36 + 1 + 36 = 147)

these numbers tell you about the length of a typical baseball game?	So counting from the left side of the graph we find that the 1st quartile must be in the interval 140 – 150 minutes. We have no way of knowing exactly what value this has, so 145 minutes is a fair approximation for the 1st quartile. Likewise, the median is approximately 165 minutes, and the third quartile is approximately 175 minutes. This tells us that typically a baseball game lasts from 145 to 175 minutes.
ACE Investigation 2	
9. A radio host asked her listeners to call in to express their opinions about a local election. What kind of sampling method is she using? Do you think the results of this survey could be used to describe the opinions of all the show's listeners? Explain.	9. This is a *voluntary-response sample*. The decision about who is to be included in the sample is left entirely in the hands of the responders. This sample is likely to be biased for several reasons. Only those who feel strongly are likely to take the time to call in. In fact it may be that those listeners who are very unhappy about the recent election will be disproportionately represented in the sample. Next we don't know if the survey had to be answered right away; if it did then only those listeners with access to a phone at that moment, which might be during the work-day, can answer. Note: Not only can the survey in this problem not be used to represent the opinions of all listeners to the show, it definitely must not be used to represent the opinions of the entire population in the listening area, since people choose shows according to their own tastes and prejudices. If the radio station wanted to draw conclusions about the opinions of all listeners they would have to devise a way to choose a *random sample* of listeners, and then follow up by making sure that whether to respond or not was not left entirely in the hands of the listener.
27. There are 350 students in a school. Ms. Cabral's class surveys two random samples of students to find out how many went to camp last summer. Here are the results: Sample 1: 8 of 25 attended camp Sample 2: 7 of 28 attended camp a. Based on the results from Sample 1, what	27. a. Not answered here. b. Not answered here. c. Not answered here. d. This question goes to the heart of the issue of "randomness." We use random sampling to eliminate any bias, *and* because only with random samples can we make any probabilistic

fraction of the students in the school would you predict attended camp? How many students is this?

b. Based on the results from Sample 2, what fraction of the students in the school would you predict attended camp? How many students is this?

c. Which sample predicts that the greater fraction of students attended camp?

d. One of Ms. Cabral's students says, "We were careful to choose our samples randomly. Why did the two samples give us different predictions?" How would you answer the student's question?

statements about the outcomes. (Such as "We are 95% certain that …") But the point about random sampling is that each sample of a given size has the same probability of occurring. Therefore, in the case of Sample 1, for example, if we ask 25 students, "Did you attend camp?" we are as likely to have picked one group of 25 as any other group of 25, and the 25 students we picked *may* all say "yes" to our question. We know sample results vary but that *most* random samples will result in proportions of "yes" that cluster around the proportion of "yes" in the population they are drawn from. So the proportion of "yes" answers in any one random sample is *probably* close to the proportion of "yes" answers in the population. If we apply this reasoning to the two samples collected we would have to deduce that Sample 1's proportion, 32%, is probably close to the fraction of students in the entire school, but so is Sample 2's proportion, 25%. (In fact, students will learn in later Statistics classes to make a statement like, "Based on Sample 1, we can predict with 95% confidence that the percentage of students at the school who attended camp is 32% plus or minus …" and then they would give a margin of error that builds in the idea that samples do vary and so do the predictions based on these samples.)

ACE Investigation 3

5.

Yung-nan wants to estimate the number of beans in a large jar. She takes out 150 beans, marks each with a red dot, returns then to the jar, and mixes them with the unmarked beans. She then takes four samples from the jar.

Sample	Total beans	Beans with Red Dots
1	25	3
2	150	23
3	75	15
4	250	25

5.

a. Not answered here.

b. The reasoning in the diagram assumes that the proportion of "red" in Sample 3 is the same as the proportion of "red" in the population. That is, 20% of Sample 3 (first shaded bar) is "red," so we believe that 20% of the entire jar (second shaded bar) is "red." Now we know that we only marked 150 beans "red" in the entire population, so we have to complete the second shaded bar by reasoning that each of the sections in the picture would be another 20% or 150 beans. This would result in 5(150) = 750 beans.

Note: students may have other ways to reason

a. Which sample has the greatest percent of beans that are marked with a red dot? Use this sample to predict the number of beans in the jar.

b. The shaded bars below are a visual way to think about making a prediction from Sample 3. Explain what the bars show and how they can be used to estimate the number of beans in the whole jar.

Sample 3
Beans in sample: 75

15, or 20% marked				

Whole jar
Beans in entire jar:?

150, or 20% marked				

c. Which sample has the least percent of beans marked with a red dot? Use this sample to predict the number of beans in the jar.

d. What is your best guess for the total number of beans in the jar?

proportionally about this.

c. Not answered here.

d. Students may reason logically about this in different ways, based on what they know about sampling and proportions. They might, for example, say that the largest sample can be trusted more than the other samples, because they have observed that statistical results from large samples are more likely to cluster around the statistic of the underlying population, than statistical results from small samples. If they choose this line of reasoning they will reason that 10% of the entire jar is "red" and that since we know that 150 beans in all are "red" then the entire jar contains 1500 beans.

Or, they may calculate the estimated total using each sample, as in part b and then average their results.

Or they may add all of their samples together as if they had in fact drawn 500 beans and found 66 "red."
66 "red" out of 500 = 150 "red" out of how many total?

Note: Each of these ways of reasoning is logical and based on what students know so far about sampling. In fact, given more knowledge about data analysis they would make predictions in terms of intervals based on these samples, rather than single numerical answers.)

ACE Investigation 4

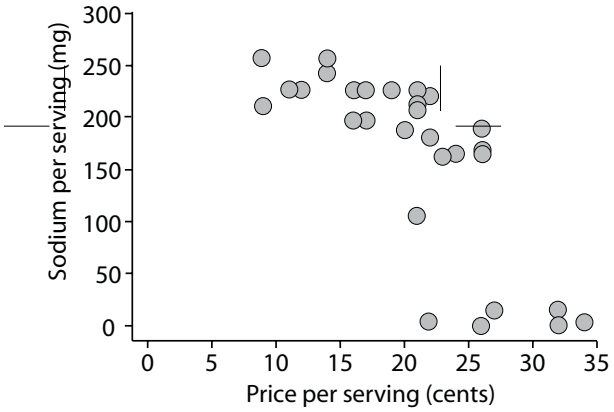
7.

A different type of scatter plot is shown below.

Note: In Investigation 1 students learned to make histograms and box plots; the purpose of both of these is to focus on what is a typical value in the data set, and on how much variability is present in the data set. In Investigation 4 students use a different kind of graph, a scatterplot, to compare two values for each observation; the purpose is now to see if there is a relationship, and, if so, to make predictions.

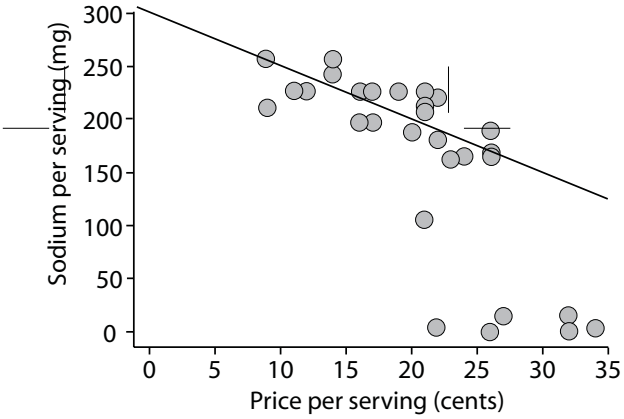
7.

There does appear to be a trend, or relationship,

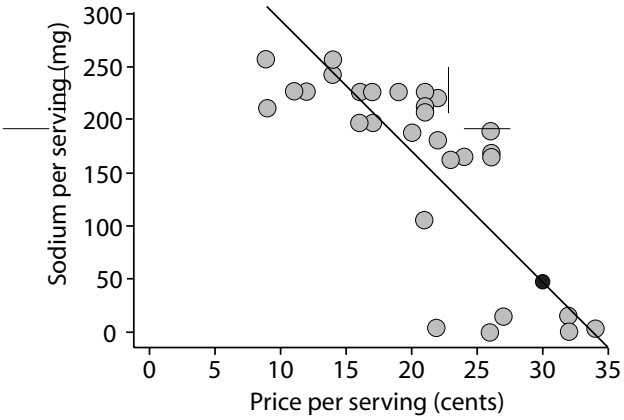


Some of the peanut butters have the same price preserving and sodium content. When this happens, the "dots" slightly overlap. Suppose you know the price per serving for a peanut butter. Can you predict the amount of sodium in a serving? Explain.

shown on the graph. As the price per serving increases the amount of sodium per serving decreases. We might try to model this trend by drawing a line that passes through the middle of the cloud of data. (See below.)



However, there are data points at the lower right of the graph, where the sodium values are near zero, that may not fit this line very well. Students may adjust the line to try to minimize the distances between all points and the line. A sensible position for the linear model is shown below.



	Based on the linear model drawn we can make predictions, though we would not want to assert that our prediction is exact. Students may want to make their predictions as intervals, rather than as a single numerical value. For example, one might predict that for a price per serving of 30 cents the amount of sodium present is about 100 mg.