In This Section

Level C

Level C is designed to build on the foundation developed in Levels A and B. In particular, Levels A and B introduced students to statistics as an investigatory process, the importance of using data to answer appropriately framed questions, types of variables (categorical versus numerical), graphical displays (including bar graph, dotplot, stemplot, histogram, boxplot, and scatterplot), tabular displays (including two-way frequency tables for categorical data and both ungrouped and grouped frequency/relative frequency tables for numerical data), and numerical summaries (including counts, proportions, mean, median, range, quartiles, interquartile range, MAD, and QCR).

Additionally, Levels A and B covered common study designs (including census, simple random sample, and randomized designs for experiments), the process of drawing conclusions from data, and the role of probability in statistical investigations.

At Level C, all of these ideas are revisited, but the types of studies emphasized are of a deeper statistical nature. Statistical studies at this level require students to draw on basic concepts from earlier work, extend the concepts to cover a wider scope of investigatory issues, and develop a deeper understanding of inferential reasoning and its connection to probability. Students also should have increased ability to explain statistical reasoning to others.

At Level C, students develop additional strategies for producing, interpreting, and analyzing data to help answer questions of interest. In general, students should be able to formulate questions that can be answered with data; devise a reasonable plan for collecting appropriate data through observation, sampling, or experimentation; draw conclusions and use data to support these conclusions; and understand the role random variation plays in the inference process.

Specifically, Level C recommendations include:

I. Formulate Questions

→   Students should be able to formulate questions and determine how data can be collected and analyzed to provide an answer.

II. Collect Data

→   Students should understand what constitutes good practice in conducting a sample survey.

→   Students should understand what constitutes good practice in conducting an experiment.

→   Students should understand what constitutes good practice in conducting an observational study.

→   Students should be able to design and implement a data collection plan for statistical studies, including observational studies, sample surveys, and simple comparative experiments.

III. Analyze Data

→   Students should be able to identify appropriate ways to summarize numerical or categorical data using tables, graphical displays, and numerical summary statistics.

→   Students should understand how sampling distributions (developed through simulation) are used to describe the sample-to-sample variability of sample statistics.

→   Students should be able to recognize association between two categorical variables.

→   Students should be able to recognize when the relationship between two numerical variables is reasonably linear, know that Pearson's correlation coefficient is a measure of the strength of the linear relationship between two numerical variables, and understand the least squares criterion in line fitting.

IV. Interpret Results

→   Students should understand the meaning of statistical significance and the difference between statistical significance and practical significance.

→   Students should understand the role of p-values in determining statistical significance.

→   Students should be able to interpret the margin of error associated with an estimate of a population characteristic.

An Introductory Example–Obesity in America

Data and the stories that surround the data must be of interest to students! It is important to remember this when teaching data analysis. It is also important to choose data and stories that have enough depth to demonstrate the need for statistical thinking. The following example illustrates this.

Students are interested in issues that affect their lives, and issues of health often fall into that category. News items are an excellent place to look for stories of current interest, including items on health. One health-related topic making lots of news lately is obesity. The following paragraph relates to a news story that is rich enough to provide a context for many of the statistical topics to be covered at Level C.

A newspaper article that appeared in 2004 begins with the following lines: "Ask anyone: Americans are getting fatter and fatter. Advertising campaigns say they are. So do federal officials and the scientists they rely on. … In 1991, 23% of Americans fell into the obese category; now 31% do, a more than 30% increase. But Dr. Jeffrey Friedman, an obesity researcher at Rockefeller University, argues that contrary to popular opinion, national data do not show Americans growing uniformly fatter. Instead, he says, the statistics demonstrate clearly that while the very fat are getting fatter, thinner people have remained pretty much the same. …The average weight of the population has in-

creased by just seven to 10 pounds." The discussion in the article refers to adults.

The following are suggested questions to explore with students who have a Level B background in statistics, but are moving on to Level C.

→ Sketch a histogram showing what you think a distribution of weights of American adults might have looked like in 1991. Adjust the sketch to show what the distribution of weights might have looked like in 2002, the year of the reported study. Before making your sketches, think about the shape, center, and spread of your distributions. Will the distribution be skewed or symmetric? Will the median be smaller than, larger than, or about the same size as the mean? Will the spread increase as you move from the 1991 distribution to the 2002 distribution?

→ Which sounds more newsworthy: "Obesity has increased by more than 30%" or "On the average, the weight of Americans has increased by fewer than 10 pounds"? Explain your reasoning.

→ The title of the article is *The Fat Epidemic: He Says It's an Illusion.* [See *New York Times*, June 8, 2004, or *CHANCE*, Vol. 17., No. 4, Fall 2004, p. 3 for the complete article.] Do you think this is a fair title? Explain your reasoning.

→ The data on which the percentages are based come from the National Center for Health Statistics, National Health and Nutrition Examination Survey 2002. This is a survey of approximately 5,800 residents of the United States. Although the survey design is more complicated than a simple random sample, the margin of error calculated as if it were a simple random sample is a reasonable approximation. What is an approximate margin of error associated with the 31% estimate of obesity for 2004? Interpret this margin of error for a newspaper reader who never studied statistics.

For the curious, information about how obesity is defined can be found at *www.amstat.org/education/gaise/3.*

In answering these questions, students at Level C should realize that a distribution of weights is going to be skewed toward the larger values. This generally produces a situation in which the mean is larger than the median. Because 8% shifted over the obesity line between 1991 and 2002, but the average weight (or center) did not shift very much, the upper tail of the distribution must have gotten "fatter," indicating a larger spread for the 2002 data. Students will have a variety of interesting answers for the second and third questions. The role of the teacher is to help students understand whether their answers are supported by the facts. The last question gets students thinking about an important estimation concept studied at Level C.

The Investigatory Process at Level C

Because Level C revisits many of the same topics addressed at Levels A and B, but at a deeper and more sophisticated level, we begin by describing how the investigatory process looks at Level C. This general discussion is followed by several examples.

Formulating Questions

As stated at the beginning of Level A, data are more than just numbers. Students need to understand the types of questions that can be answered with data. For example, the question "Is the overall health of high-school students declining in this country?" is too big a question to answer with a statistical investigation (or even many statistical investigations). Certain aspects of the health of students, however, can be investigated by formulating more specific questions, such as "What is the rate of obesity among high-school students?"; "What is the average daily caloric intake for high-school seniors?"; "Is a three-day-a-week exercise regimen enough to maintain heart rate and weight within acceptable limits?" Question formulation, then, becomes the starting point for a statistical investigation.

Collecting Data—Types of Statistical Studies

Most questions that can be answered through data collection and interpretation require data from a designed study, either a *sample survey* or an *experiment*.

These two types of statistical investigations have some common elements—each requires randomization for both purposes of reducing bias and building a foundation for statistical inference and each makes use of the common inference mechanisms of margin of error in estimation and p-value in hypothesis testing (both to be explained later). But these two types of investigations have very different objectives and requirements. Sample surveys are used to estimate or make decisions about characteristics (parameters) of populations. A well-defined, fixed population is the main ingredient of such a study. Experiments are used to estimate or compare the effects of different experimental conditions (treatments), and require well-defined treatments and experimental units on which to study those treatments.

Estimating the proportion of residents of a city that would support an increase in taxes for education requires a sample survey. If the selection of residents is random, then the results from the sample can be extended to represent the population from which the sample was selected. A measure of sampling error (margin of error) can be calculated to ascertain how far the estimate is likely to be from the true value.

Testing to see if a new medication to improve breathing for asthma patients produces greater lung capacity than a standard medication requires an experiment in which a group of patients who have consented to participate in the study are randomly assigned to either

the new or the standard medication. With this type of randomized comparative design, an investigator can determine, with a measured degree of uncertainty, whether the new medication caused an improvement in lung capacity. Randomized experiments are, in fact, the only type of statistical study capable of establishing cause and effect relationships. Any generalization extends only to the types of units used in the experiment, however, as the experimental units are not usually randomly sampled from a larger population. To generalize to a larger class of experimental units, more experiments would have to be conducted. That is one reason why replication is a hallmark of good science.

Studies that have no random selection of sampling units or random assignment of treatments to experimental units are called *observational studies* in this document. A study of how many students in your high school have asthma and how this breaks down among gender and age groups would be of this type. Observational studies are not amenable to statistical inference in the usual sense of the term, but they can provide valuable insight into the distribution of measured values and the types of associations among variables that might be expected.

At Level C, students should understand the key features of both sample surveys and experimental designs, including how to set up simple versions of both types of investigations, how to analyze the data appropriately (as the correct analysis is related to the design), and how to clearly and precisely state conclusions for these designed studies. Key elements of the design and implementation of data collection plans for these types of studies follow.

Sample Surveys

Students should understand that obtaining good results from a sample survey depends on four basic features: the population, the sample, the randomization process that connects the two, and the accuracy of the measurements made on the sampled elements. For example, to investigate a question on health of students, a survey might be planned for a high school. What is the population to be investigated? Is it all the students in the school (which changes on a daily basis)? Perhaps the questions of interest involve only juniors and seniors. Once the population is defined as precisely as possible, one must determine an appropriate sample size and a method for randomly selecting a sample of that size. Is there, for example, a list of students who can then be numbered for random selection? Once the sampled students are found, what questions will be asked? Are the questions fair and unbiased (as far as possible)? Can or will the students actually answer them accurately?

When a sample of the population is utilized, errors may occur for several reasons, including:

→ the sampling procedure is biased
→ the sample was selected from the wrong population

→   some of the units selected to be in the sample were unable (or unwilling) to participate

→   the questions were poorly written

→   the responses were ambiguous

These types of errors should be considered carefully before the study begins so plans can be made to reduce their chance of occurring as much as possible. One way to resolve the bias in the sampling procedure is to incorporate randomness into the selection process.

Two samples of size 50 from the same population of students will most likely not give the same result on, say, the proportion of students who eat a healthy breakfast. This variation from sample to sample is called *sampling variability*. When randomness is incorporated into the sampling procedure, probability provides a way to describe the "long-run" behavior of this sampling variability.

Experiments

At Level C, students should understand that obtaining good results from an experiment depends upon four basic features: well-defined treatments, appropriate experimental units to which these treatments can be assigned, a sound randomization process for assigning treatments to experimental units, and accurate measurements of the results of the experiment. Experimental units generally are not randomly selected from a population of possible units. Rather, they are the ones that happen to be available for the study. In

experiments with human subjects, the people involved are often volunteers who have to sign an agreement stating they are willing to participate in the experimental study. In experiments with agricultural crops, the experimental units are the field plots that happen to be available. In an industrial experiment on process improvement, the units may be the production lines in operation during a given week.

As in a sample survey, replicating an experiment will produce different results. Once again, random assignment of experimental units to treatments (or vice versa) allows the use of probability to predict the behavior in the resulting values of summary statistics from a large number of replications of the experiment. Randomization in experiments is important for another reason. Suppose a researcher decides to assign treatment A only to patients over the age of 60 and treatment B only to patients under the age of 50. If the treatment responses differ, it is impossible to tell whether the difference is due to the treatments or the ages of the patients. (This kind of bias in experiments and other statistical studies is called *confounding*.) The randomization process, if properly done, will usually balance treatment groups so this type of bias is minimized.

Observational Studies

At Level C, students should understand that observational studies are useful for suggesting patterns in data and relationships between variables, but do not provide a strong foundation for estimating population parameters

or establishing differences among treatments. Asking the students in one classroom whether they eat a healthy breakfast is not going to help you establish the proportion of healthy breakfast-eaters in the school, as the students in one particular classroom may not be representative of the students in the school. Random sampling is the only way to be confident of a representative sample for statistical purposes. Similarly, feeding your cats Diet A and your neighbor's cats Diet B is not going to allow you to claim that one diet is better than the other in terms of weight control, because there was no random assignment of experimental units (cats) to treatments (diets). As a consequence, confounding may result. Studies of the type suggested above are merely observational; they may suggest patterns and relationships, but they are not a reliable basis for statistical inference.

Analyzing Data

When analyzing data from well-designed sample surveys, students at Level C should understand that an appropriate analysis is one that can lead to justifiable inferential statements about population parameters based on estimates from sample data. The ability to draw conclusions about the population using information from a sample depends on information provided by the sampling distribution of the sample statistic being used to summarize the sample data. At Level C, the two most common parameters of interest are the population proportion for categorical data and the population mean for numerical data. The appropriate sample statistics used to estimate these parameters are the sample proportion and the sample mean, respectively. At Level C, the sample-to-sample variability, as described by the sampling distribution for each of these two statistics, is addressed in more depth.

Exploring how the information provided by a sampling distribution is used for generalizing from a sample to the larger population enables students at Level C to draw more sophisticated conclusions from statistical studies. At Level C, it is recommended that the sampling distributions of a sample proportion and of a sample mean be developed through simulation. More formal treatment of sampling distributions can be left to AP Statistics and college-level introductory statistics courses.

Because the sampling distribution of a sample statistic is a topic with which many teachers may not be familiar, several examples are included here to show how simulation can be used to obtain an approximate sampling distribution for a sample proportion and for a sample mean.

Example 1: The Sampling Distribution of a Sample Proportion

Properties of the sampling distribution for a sample proportion can be illustrated by simulating the process of selecting a random sample from a population using random digits as a device to model various populations.

For example, suppose a population is assumed to have 60% "successes" ($p = .6$) and we are to take a random sample of $n = 40$ cases from this population. How far can we expect the sample proportion of successes to deviate from the true population value of .60? This can be answered by determining an empirical sampling distribution for the sample proportion.

One way to model a population with 60% successes (and 40% failures) is to utilize the 10 digits 0, 1,…, 9. Label six of the 10 digits as "success" and the other four as "failures." To simulate selecting a sample of size 40 from this population, randomly select 40 random digits (with replacement). Record the number of successes out of the 40 digits selected and convert this count to the proportion of successes in the sample. Note that:

Proportion of Successes in the Sample

$$= \frac{\text{Number of Successes in the Sample}}{\text{Sample Size}}$$

Repeating this process a large number of times, and determining the proportion of successes for each sample, illustrates the idea of the sample-to-sample variability in the sample proportion.

Simulating the selection of 200 random samples of size 40 from a population with 60% successes and determining the proportion of success for each sample

resulted in the empirical distribution shown in Figure 25. This empirical distribution is an approximation to the true sampling distribution of the sample proportion for samples of size 40 from a population in which the actual proportion is .60.
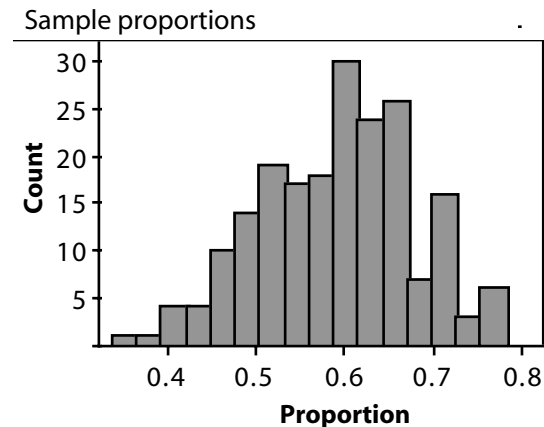

Figure 25: Histogram of sample proportions

Summarizing the above distribution based on its shape, center, and spread, one can state that this empirical sampling distribution has a mound shape (approximately normal). Because the mean and standard deviation of the 200 sample proportions are .59 and .08, respectively, the empirical distribution shown in Figure 25 has a mean of .59 and a standard deviation of .08.

By studying this empirical sampling distribution, and others that can be generated in the same way, students will see patterns emerge. For example, students will observe that, when the sample size is reasonably large

(and the population proportion of successes is not too near the extremes of 0 or 1), the shapes of the resulting empirical sampling distributions are approximately normal. Each of the empirical sampling distributions should be centered near the value of p, the population proportion of successes, and the standard deviation for each distribution should be close to:

$$\sqrt{\frac{p(1-p)}{n}}$$

Note that in Example 1, the mean of the empirical distribution is .59, which is close to .6, and the standard deviation is .08, which is close to:

$$\sqrt{\frac{.6(.4)}{40}} \approx .0775$$

A follow-up analysis of these empirical sampling distributions can show students that about 95% of the sample proportions lie within a distance of:

$$2\sqrt{\frac{.6(.4)}{40}} \approx 0.155$$

from the true value of p. This distance is called the *margin of error.*

Example 2: The Sampling Distribution of a Sample Mean

Properties of the sampling distribution for a sample mean can be illustrated in a way similar to that used for proportions in Example 1. Figure 26 shows the distribution of the sample mean when 200 samples of

30 random digits are selected (with replacement) and the sample mean is computed. This simulates sampling from a population that has a uniform distribution with equal numbers of 0s, 1s, 2s,…, 9s. Note that this population of numerical values has a mean, μ, of 4.5 and a standard deviation, σ, of 2.9.
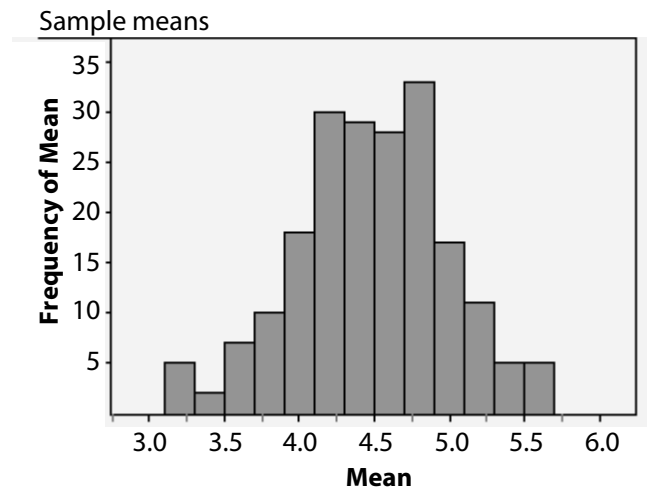
Figure 26: Histogram of sample means

The empirical sampling distribution shown in Figure 26 can be described as approximately normal with a mean of 4.46 (the mean of the 200 sample means from the simulation) and a standard deviation of 0.5 (the standard deviation of the 200 sample means).

By studying this empirical sampling distribution, and others that can be generated in similar ways, students will see patterns emerge. For example, students will

observe that, when the sample size is reasonably large, the shapes of the empirical sampling distributions are approximately normal. Each of the empirical sampling distributions should be centered near the value of μ, the population mean, and the standard deviation for each distribution should be close to:

$$\frac{\sigma}{\sqrt{n}}$$

Note that in Example 2, the mean of the empirical sampling distribution is 4.46, which is close to $\mu = 4.5$, and the standard deviation (0.5) is close to:

$$\sigma \big/ \sqrt{n} = 2.9 \big/ \sqrt{30} = 0.53$$

The margin of error in estimating a population mean using the sample mean from a single random sample is approximately:

$$2\frac{\sigma}{\sqrt{n}}$$

The sample mean should be within this distance of the true population mean about 95% of the time in repeated random sampling.

Interpreting Results

Generalizing from Samples

The key to statistical inference is the sampling distribution of the sample statistic, which provides information about the population parameter being estimated. As described in the previous section, knowledge of the sampling distribution for a statistic, like a sample proportion or sample mean, leads to a margin of error that provides information about the maximum likely distance between a sample estimate and the population parameter being estimated. Another way to state this key concept of inference is that an estimator plus or minus the margin of error produces an interval of plausible values for the population parameter. Any one of these plausible values could have produced the observed sample result as a reasonably likely outcome.

Generalizing from Experiments

Do the effects of the treatments differ? In analyzing experimental data, this is one of the first questions asked. This question of difference is generally posed in terms of differences between the centers of the data distributions (although it could be posed as a difference between the 90th percentiles or any other measure of location in a distribution). Because the mean is the most commonly used statistic for measuring the center of a distribution, this question of differences is generally posed as a question about a difference in means. The analysis of experimental data, then, usually involves a comparison of means.

Unlike sample surveys, experiments do not depend on random samples from a fixed population. Instead, they require random assignment of treatments to preselected experimental units. The key question, then,

is: "Could the observed difference in treatment means be due to the random assignment (chance) alone, or can it be attributed to the treatments administered?"

The following examples are designed to illustrate and further illuminate the important concepts at Level C by carefully considering the four phases of a statistical analysis—question, design, analysis, interpretation—in a variety of contexts.

Example 3: A Survey of Music Preferences

A survey of student music preferences was introduced at Level A, where the analysis consisted of making counts of student responses and displaying the data in a bar graph. At Level B, the analysis was expanded to consider relative frequencies of preferences and cross-classified responses for two types of music displayed in a two-way table. Suppose the survey included the following questions:

1. *What kinds of music do you like?*

   Do you like country music?

   Yes or No

   Do you like rap music?

   Yes or No

   Do you like rock music?

   Yes or No

2. *Which of the following types of music do you like most? Select only one.*

   Country   Rap/Hip Hop   Rock

In order to be able to generalize to all students at the school, a representative sample of students from the school is needed. This could be accomplished by selecting a simple random sample of 50 students from the school. The results can then be generalized to the school (but not beyond), and the Level C discussion will center on basic principles of generalization—or statistical inference.

A Level C analysis begins with a two-way table of counts that summarizes the data on two of the questions: "Do you like rock music?" and "Do you like rap music?" The table provides a way to separately examine the responses to each question and to explore possible connections (association) between the two categorical variables. Suppose the survey of 50 students resulted in the data summarized in Table 11.

As demonstrated at Level B, there are a variety of ways to interpret data summarized in a two-way table, such as Table 11. Some examples based on all 50 students in the survey include:

→   25 of the 50 students (50%) liked both rap and rock music.

→   29 of the 50 students (58%) liked rap music.

→   19 of the 50 students (38%) did not like rock music.

Table 11: Two-Way Frequency Table

		Like Rock Music?		
		Yes	No	**Row Totals**
Like Rap Music?	Yes	25	4	29
	No	6	15	21
Column Totals		31	19	50

One type of statistical inference relates to conjectures (hypotheses) made before the data were collected. Suppose a student says "I think **more than 50%** of the students in the school like rap music." Because 58% of the students in the sample liked rap music (which is more than 50%), there is evidence to support the student's claim. However, because we have only a sample of 50 students, it is possible that 50% of all students like rap (in which case, the student's claim is not correct), but the variation due to random sampling might produce 58% (or even more) who like rap. The statistical question, then, is whether the sample result of 58% is reasonable from the variation we expect to occur when selecting a random sample from a population with 50% successes.

One way to arrive at an answer is to set up a hypothetical population that has 50% successes (such as even and odd digits produced by a random number generator) and repeatedly take samples of size 50 from it, each time recording the proportion of even digits.

The sampling distribution of proportions so generated will be similar to the one below.

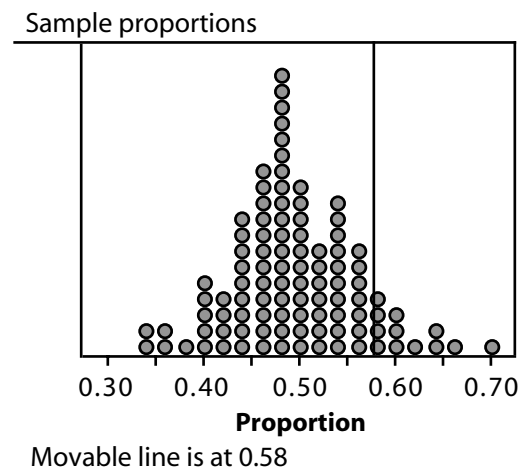Sample proportions

Movable line is at 0.58

Figure 27: Dotplot of sample proportions from a hypothetical population in which 50% like rap music

Based on this simulation, a sample proportion greater than or equal to the observed .58 occurred 12 times out of 100 just by chance variation alone when the actual population proportion is .50. This suggests the result of .58 is not a very unusual occurrence when sampling from a population with .50 as the "true" proportion of students who like rap music. So a population value of .50 is plausible based on what was observed in the sample, and the evidence in support of the student's claim is not very strong. The fraction of times the observed result is matched or exceeded (.12 in this investigation) is called the approximate

p-value. The p-value represents the chance of observing the result observed in the sample, or a result more extreme, when the hypothesized value is in fact correct. A small p-value would have supported the student's claim, because this would have indicated that if the population proportion was .50, it would have been very unlikely that a sample proportion of .58 would have been observed.

Suppose another student hypothesized that **more than 40%** of the students in the school like rap music. To test this student's claim, samples of size 50 must now be repeatedly selected from a population that has 40% successes. Figure 28 shows the results of one such simulation. The observed result of .58 was reached only one time out of 100, and no samples produced a proportion greater than .58. Thus, the approximate

Sample proportions
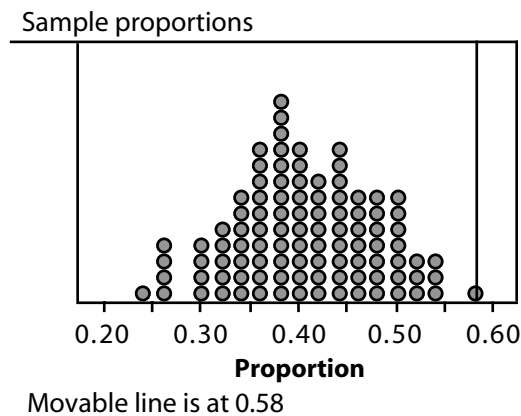

Movable line is at 0.58

Figure 28: Dotplot of sample proportions from a hypothetical population in which 40% like rap music

p-value is .01, and it is not likely that a population in which 40% of the students like rap music would have produced a sample proportion of 58% in a random sample of size 50. This p-value provides very strong evidence in support of the student's claim that more than 40% of the students in the entire school like rap music.

Another way of stating the above is that .5 is a plausible value for the true population proportion, based on the sample evidence, but .4 is not. A set of plausible values can be found by using the margin of error introduced in Example 1. As explained previously, the margin of error for a sample proportion is approximately:

$$2\sqrt{\frac{p(1-p)}{n}}$$

However, in this problem, the true value of p is unknown. Our sample proportion ($\hat{p} = .58$) is our "best estimate" for what p might be, so the margin of error can be estimated to be:

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2\sqrt{\frac{.58(.42)}{50}} \approx .14$$

Thus, any proportion between .58 − .14 = .44 and .58 + .14 = .72 can be considered a plausible value for the true proportion of students at the school who like rap music. Notice that .5 is well within this interval, but .4 is not.

Another type of question that could be asked about the students' music preferences is of the form "Do those who like rock music also tend to like rap music?" In other words, is there an association between liking rock music and liking rap music? The same data from the random sample of 50 students can be used to answer this question.

According to Table 11, a total of 31 students in the survey like rock music. Among those students, the proportion who also like rap music is $25/31 = .81$. Among the 19 students who do not like rock music, $4/19 = .21$ is the proportion who like rap music. The large difference between these two proportions (.60) suggests there may be a strong association between liking rock music and liking rap music. But could this association simply be due to chance (a consequence only of the random sampling)?

If there were no association between the two groups, then the 31 students who like rock would behave as a random selection from the 50 in the sample. We would expect the proportion who like rap among these 31 students to be close to the proportion who like rap among the 19 students who don't like rock. Essentially, this means that if there is no association, we expect the difference between these two proportions to be approximately 0. Because the difference in our survey is .6, this suggests that there is an association. Can the difference, .6, be explained by the random variation we expect when selecting a random sample?

To simulate this situation, we create a population of 29 1s (those who like rap) and 21 0s (those who do not like rap) and mix them together. Then, we select 31 (representing those who like rock) at random and see how many 1s (those who like rap) we get. It is this entry that goes into the (yes, yes) cell of the table, and from that data the difference in proportions can be calculated. Repeating the process 100 times produces a simulated sampling distribution for the difference between the two proportions, as shown in Figure 29.

Differences between proportions
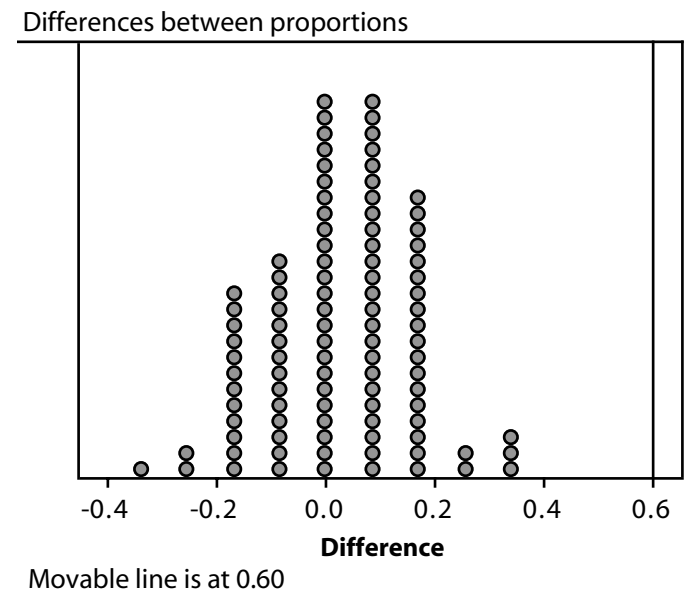

Movable line is at 0.60

Figure 29: Dotplot showing simulated sampling distribution

The observed difference in proportions from the sample data, .6, was never reached in 100 trials, indicating that the observed difference cannot be attributed to chance alone. Thus, there is convincing evidence of a real association between liking rock music and liking rap music.

Example 4: An Experiment on the Effects of Light on the Growth of Radish Seedlings

What is the effect of different durations of light and dark on the growth of radish seedlings? This question was posed to a class of biology students who then set about designing and carrying out an experiment to investigate the question. All possible relative durations of light to dark cannot possibly be investigated in one experiment, so the students decided to focus the question on three treatments: 24 hours of light, 12 hours of light and 12 hours of darkness, and 24 hours of darkness. This covers the extreme cases and one in the middle.

With the help of a teacher, the class decided to use plastic bags as growth chambers. The plastic bags would permit the students to observe and measure the germination of the seeds without disturbing them. Two layers of moist paper towel were put into a disposable plastic bag, with a line stapled about 1/3 of the way from the bottom of the bag (see Figure 30) to hold the paper towel in place and to provide a seam to hold the radish seeds.
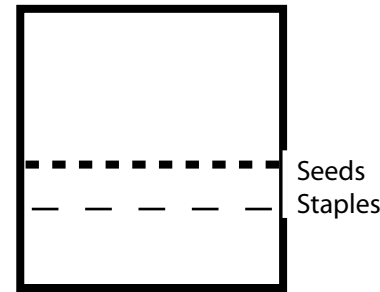

Figure 30: Seed experiment

Although three growth chambers would be sufficient to examine the three treatments, this class made four growth chambers, with one designated for the 24 hours of light treatment, one for the 12 hours of light and 12 hours of darkness treatment, and two for the 24 hours of darkness treatment. One hundred twenty seeds were available for the study. Thirty of the seeds were chosen at random and placed along the stapled seam of the 24 hours of light bag. Thirty seeds were then chosen at random from the remaining 90 seeds and placed in the 12 hours of light and 12 hours of darkness bag. Finally, 30 of the remaining 60 seeds were chosen at random and placed in one of the 24 hours of darkness bags. The final 30 seeds were placed in the other 24 hours of darkness bag. After three days, the lengths of radish seedlings for the germinating seeds were measured and recorded. These data are provided in Table 12; the measurements are in milli-

Table 12: Lengths of Radish Seedlings

Treatment 1 24 light	Treatment 2 12 light, 12 dark	Treatment 3 24 dark		Treatment 1 24 light	Treatment 2 12 light, 12 dark	Treatment 3 24 dark	
2	3	5	20	10	17	15	30
3	4	5	20	10	20	15	30
5	5	8	22	10	20	15	30
5	9	8	24	10	20	15	31
5	10	8	25	10	20	15	33
5	10	8	25	10	20	15	35
5	10	10	25	10	21	16	35
7	10	10	25	10	21	20	35
7	10	10	25	14	22	20	35
7	11	10	26	15	22	20	35
8	13	10	29	15	23	20	35
8	15	11	30	20	25	20	36
8	15	14	30	21	25	20	37
9	15	14	30	21	27	20	38
						20	40

meters. Notice that not all of the seeds in each group germinated.

A good first step in the analyses of numerical data such as these is to make graphs to look for patterns and any unusual departures from the patterns. Box-plots are ideal for comparing data from more than one treatment, as you can see in Figure 31. Both the centers and the spreads increase as the amount of darkness increases. There are three outliers (one at 20

mm and two at 21 mm) in the Treatment 1 (24 hours of light) data. Otherwise, the distributions are fairly symmetric, which is good for statistical inference.

In Figure 31, Treatment 1 is 24 hours of light; treatment 2 is 12 hours of light and 12 of darkness; treatment 3 is 24 hours of darkness.

The summary statistics for these data are shown in Table 13.
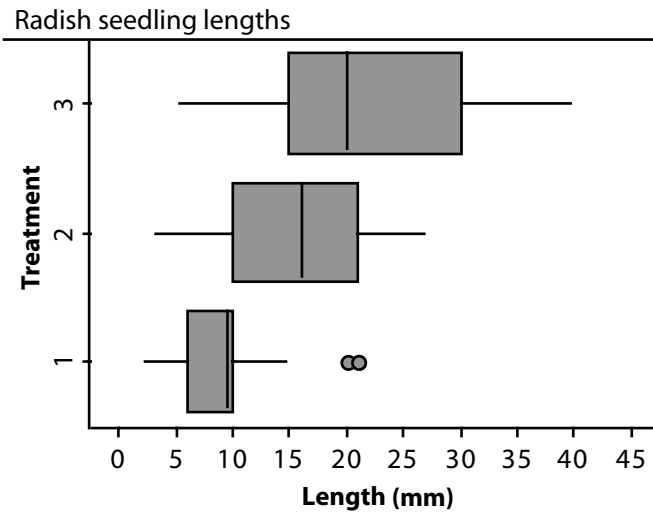
Radish seedling lengths

Figure 31: Boxplot showing growth under different conditions

Experiments are designed to compare treatment effects, usually by comparing means. The original question on the effect of different periods of light and dark on the growth of radish seedlings might be turned into two questions about treatment means. Is there evidence that the 12 hours of light and 12 hours of dark (Treatment 2) group has a significantly higher mean than the 24 hours of light (Treatment 1) group? Is there evidence that the 24 hours of dark (Treatment 3) group has a significantly higher mean than the 12 hours of light and 12 hours of dark (Treatment 2) group? Based on the boxplots and the summary statistics, it is clear that the sample means differ. *Are these*

Table 13: Treatment Summary Statistics

Treat-ment	n	Mean	Median	Std. Dev.
1	28	9.64	9.5	5.03
2	28	15.82	16.0	6.76
3	58	21.86	20.0	9.75

differences large enough to rule out chance variation as a possible explanation for the observed difference?

The Treatment 2 mean is 6.2 mm larger than the Treatment 1 mean. If there is no real difference between the two treatments in terms of their effect on seedling growth, then the observed difference must be due to the random assignment of seeds to the bags; that is, one bag was simply lucky enough to get a preponderance of good and lively seeds. But, if a difference this large (6.2 mm) is likely to be the result of randomization alone, then we should see differences of this magnitude quite often if we repeatedly re-randomize the measurements and calculate a new difference in observed means. This, however, is not the case, as one can see from Figure 32. This dotplot was produced by mixing the growth measurements from Treatments 1 and 2 together, randomly splitting them into two groups of 28 measurements, recording the difference in means for the two groups, and repeating the process 200 times.

The observed difference of 6.2 mm was exceeded only one time in 200 trials, for an approximate p-value of

> " Experiments are designed to compare treatment effects, usually by comparing means. "
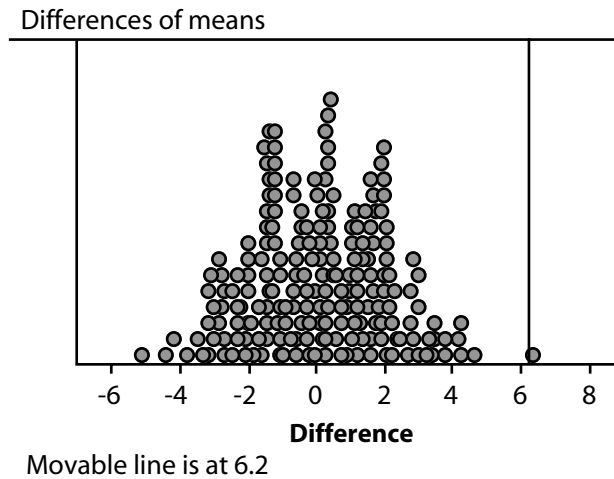
Differences of means

Movable line is at 6.2

Figure 32: Dotplot showing differences of means

1/200. This is very small, and gives extremely strong evidence to support the hypothesis that there is a statistically significant difference between the means for Treatments 1 and 2. The observed difference of 6.2 mm is very unlikely to be due simply to chance variation.

In a comparison of the means for Treatments 2 and 3, the same procedure is used, except that the combined measurements are split into groups of 28 and 58 each time. The observed difference of 6 mm was exceeded only one time out of 200 trials (see Figure 33), giving extremely strong evidence of a statistically significant difference between the means for Treatments 2 and 3. In summary, the three treatment groups show statistically significant differences in mean growth that cannot reasonably be explained by the random as-
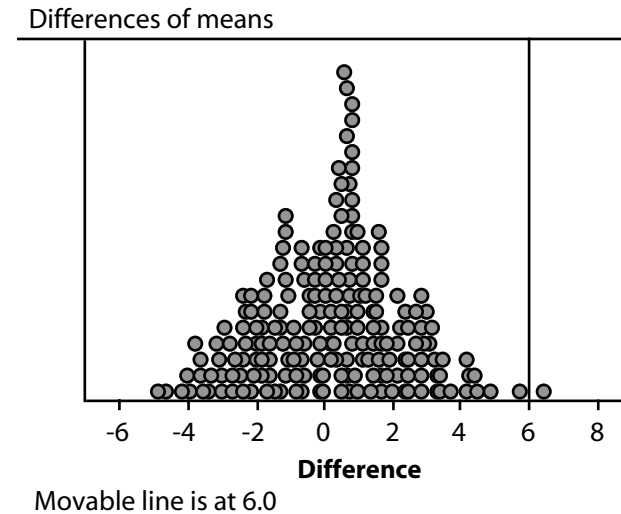

Differences of means

Movable line is at 6.0

Figure 33: Dotplot showing differences of means

signment of seeds to the bags. This gives us convincing evidence of a treatment effect—the more hours of darkness, the greater the growth of the seedling, at least for these three periods of light versus darkness.

Students should be encouraged to delve more deeply into the interpretation, relating it to what is known about the phenomenon or issue under study. Why do the seedlings grow faster in the dark? Here is an explanation from a biology teacher. It seems to be an adaptation of plants to get the seedlings from the dark (under ground) where they germinate into the light (above ground) as quickly as possible. Obviously, the seedling cannot photosynthesize in the dark and is using up the energy stored in the seed to power the

growth. Once the seedling is exposed to light, it shifts its energy away from growing in length to producing chlorophyll and increasing the size of its leaves. These changes allow the plant to become self-sufficient and begin producing its own food. Even though the growth in length of the stem slows, the growth in diameter of the stem increases and the size of the leaves increases. Seedlings that continue to grow in the dark are spindly and yellow, with small yellow leaves. Seedlings grown in the light are a rich, green color with large, thick leaves and short stems.

What is the density of the Earth? This is a question that intrigued the great scientist Henry Cavendish, who attempted to answer the question in 1798. Cavendish estimated the density of the Earth by using the crude tools available to him at the time. He did not literally take a random sample; he measured on different days and at different times, as he was able. But the density of the Earth does not change over time, so his measurements can be thought of as a random sample of all the measurements he could have taken on this constant. The variation in the measurements is due to his measurement error, not to changes in the Earth's density. The Earth's density is the constant that is being estimated.

This is a typical example of an estimation problem that occurs in science. There is no real "population" of measurements that can be sampled; rather, the sample data is assumed to be a random selection from the conceptual population of all measurements that could have been made. At this point, there may be some confusion between an "experiment" and a "sample survey" because Cavendish actually conducted a scientific investigation to get his measurements. The key, however, is that he conducted essentially the same investigation many times with a goal of estimating a constant, much like interviewing many people to estimate the proportion who favor a certain candidate for office. He did not randomly assign treatments to experimental units for the purpose of comparing treatment effects.

The famous Cavendish data set contains his 29 measurements of the density of the Earth, in grams per cubic centimeter. The data are shown below [Source: *http://lib.stat.cmu.edu/DASL*]:

5.50	5.57	5.42	5.61	5.53	5.47	4.88
5.62	5.63	4.07	5.29	5.34	5.26	5.44
5.46	5.55	5.34	5.30	5.36	5.79	5.75
5.29	5.10	5.86	5.58	5.27	5.85	5.65
5.39						

One should look at the data before proceeding with an analysis. The histogram in Figure 34 shows the data to be roughly symmetric, with one unusually small value. If Cavendish were alive, you could ask him if he had
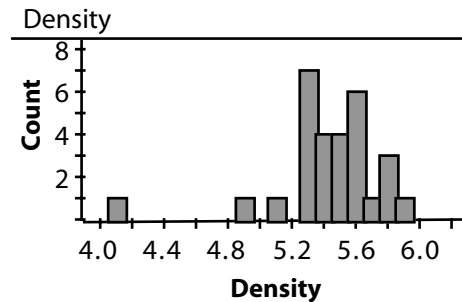
Density

Figure 34: Histogram of Earth density measurements

made a mistake (and that is certainly what you should do for a current data set).

The mean of the 29 measurements is 5.42 and the standard deviation is 0.339. Recall that the margin of error for the sample mean is:

$$2\frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation. In this problem, the population standard deviation is not known; however, the sample standard deviation provides an estimate for the population standard deviation. Consequently, the margin of error can be estimated to be:

$$2\frac{s}{\sqrt{n}} = 2\frac{0.339}{\sqrt{29}} = 0.126$$

The analysis shows that any value between 5.420 − 0.126 and 5.420 + 0.126, or in the interval (5.294, 5.546), is a plausible value of the density of the Earth. That is, any value in the interval is consistent with the data obtained by Cavendish. Now, the questionable low observation should be taken into account, as it will lower the mean and increase the standard deviation. If that measurement is regarded as a mistake and removed from the data set, the mean of the 28 remaining observations is 5.468 and the standard deviation is 0.222, producing a margin of error of 0.084 and an interval of plausible values of (5.384, 5.552).

Students now can check on how well Cavendish did; modern methods pretty much agree that the average density of the Earth is about 5.515 grams per cubic centimeter. The great 18th century scientist did well!

Example 6: Linear Regression Analysis—Height vs. Forearm Length

Regression analysis refers to the study of relationships between variables. If the "cloud" of points in a scatterplot of paired numerical data has a linear shape, a straight line may be a realistic model of the relationship between the variables under study. The least squares line runs through the center (in some sense) of the cloud of points. Residuals are defined to be the deviations in the y direction between the points in the scatterplot and the least squares line; spread is now the variation around the least squares line, as

Table 14: Heights vs. Forearm Lengths

Forearm (cm)	Height (cm)	Forearm (cm)	Height (cm)
45.0	180.0	41.0	163.0
44.5	173.2	39.5	155.0
39.5	155.0	43.5	166.0
43.9	168.0	41.0	158.0
47.0	170.0	42.0	165.0
49.1	185.2	45.5	167.0
48.0	181.1	46.0	162.0
47.9	181.9	42.0	161.0
40.6	156.8	46.0	181.0
45.5	171.0	45.6	156.0
46.5	175.5	43.9	172.0
43.0	158.5	44.1	167.0

Height vs. forearm length

$Height = 2.76 Forearm + 45.8$ $r^2 = 0.64$

Figure 35: Scatterplot and residual plot

measured by the standard deviation of the residuals. When using a fitted model to predict a value of y from x, the associated margin of error depends on the standard deviation of the residuals.

Relationships among various physical features, such as height versus arm span and neck size versus shoe size, can be the basis of many interesting questions for student investigation. If I were painting a picture of a person, how could I get the relative sizes of the body parts correct? This question prompted students to carry out an investigation of one of the possible relationships, that between forearm length and height.
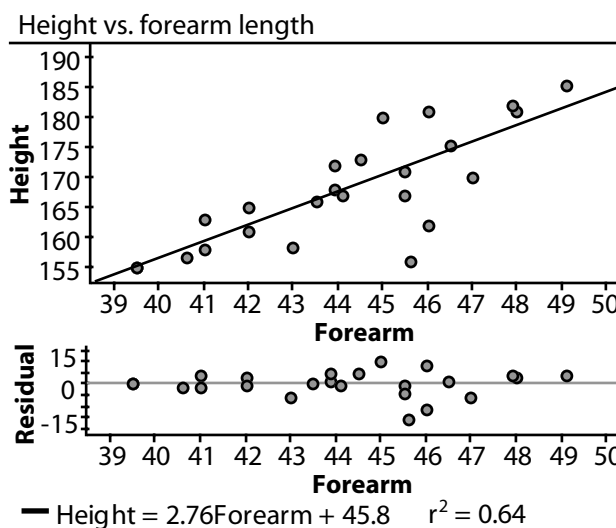
The students responsible for the study sampled other students on which to make forearm and height measurements. Although the details of how the sample actually was selected are not clear, we will suppose that it is representative of students at the school and has the characteristics of a random sample. An important consideration here is to agree on the definition of "forearm" before beginning to take measurements. The data obtained by the students (in centimeters) are provided in Table 14.

A good first step in any analysis is to plot the data, as we have done in Figure 35. The linear trend in the plot is fairly strong. The scatterplot, together with Pearson's correlation coefficient of .8, indicate that a

line would be a reasonable model for summarizing the relationship between height and forearm length.

The scatterplot includes a graph of the least squares line:

Predicted Height = 45.8 + 2.76(Forearm Length).

The plot below the scatterplot shows the residuals. There are a few large residuals but no unusual pattern in the residual plot. The slope (about 2.8) can be interpreted as an estimate of the average difference in heights for two persons whose forearms are 1 cm different in length. The intercept of 45.8 centimeters cannot be interpreted as the expected height of a person with a forearm zero centimeters long! However, the regression line can reasonably be used to predict the height of a person for whom the forearm length is known, as long as the known forearm length is in the range of the data used to develop the prediction equation (39 to 50 cm for these data). The margin of error for this type of prediction is approximately 2(standard deviation of the residuals). For these data, the standard deviation of the residuals is 5.8 (not shown here, but provided as part of the computer output), so the margin of error is 2(5.8) = 11.6 cm. The predicted height of someone with a forearm length of 42 cm would be:

Predicted Height = 45.8 + 2.76(42) = 161.7 cm

With 95% confidence, we would predict the height of people with forearm length 42 cm to be between 150.1 cm and 173.3 cm (161.7 ± 11.6).

Is the slope of 2.8 "real," or simply a result of chance variation from the random selection process? This question can be investigated using simulation. A description of this simulation is included in the Appendix to Level C.

Example 7: Comparing Mathematics Scores— An Observational Study

Data often are presented to us in a form that does not call for much analysis, but does require some insight into statistical principles for correct interpretation. Standardized test scores often fall into this category. Table 15 gives information about the state mean scores on the National Assessment of Educational Progress (NAEP) 2000 Grade 4 mathematics scores for Louisiana and Kentucky. Even though these scores are based on a sample of students, these are the scores assigned to the states, and consequently, they can be considered observational data from that point of view.

Table 15: NAEP 2000 Scores in Mathematics

	Overall Mean	Mean for Whites	Mean for Non-whites	% White
Louisiana	217.96	229.51	204.94	
Kentucky	220.99	224.17		87

To see if students understand the table, it is informative to ask them to fill in a few omitted entries.

→ Fill in the two missing entries in the table (53% and 199.71).

More substantive questions involve the seeming contradictions that may occur in data of this type. They might be phrased as follows.

→ For the two states, compare the overall means. Compare the means for whites. Compare the means for nonwhites. What do you observe?

→ Explain why the reversals in direction take place once the means are separated into racial groups.

It is genuinely surprising to students that data summaries (means in this case) can go in one direction in the aggregate but can go in the opposite direction for each subcategory when disaggregated. This phenomenon is called Simpson's Paradox.

Example 8: Observational Study—Toward Establishing Causation

Observational studies are the only option for situations in which it is impossible or unethical to randomly assign treatments to subjects. Such situations are a common occurrence in the study of causes of diseases. A classical example from this field is the relationship between smoking and lung cancer, which prompted heated debates during the 1950s and 1960s. Society will not condone the notion of assigning some people to be smokers and others to be nonsmokers in an experiment to see if smoking causes lung cancer. So the evidence has to be gathered from observing the

Table 16: Cigarette Smoking and Lung Cancer

	Lung Cancer Cases	Controls	Totals
Smokers	647	622	1,269
Non-smokers	2	27	29

world as it is. The data collection process still can be designed in clever ways to obtain as much information as possible.

Here is an example from the smoking versus lung cancer debates. A group of 649 men with lung cancer was identified from a certain population in England. A control group of the same size was established by matching these patients with other men from the same population who did not have lung cancer. The matching was on background variables such as ethnicity, age, and socioeconomic status. (This is called a case-control study.) The objective, then, is to compare the rate of smoking among those with lung cancer to the rate for those without cancer.

First, make sure students understand the nature of the data in Table 16. Does this show, for example, that there was a very high percentage of smokers in England around 1950? The rate of smoking in these groups was $(647/649) = .997$ for the cancer patients and $(622/649) = .958$ for the controls. If these data had resulted from a random assignment or selection, the difference of about 4 percentage points would be

statistically significant (by methods discussed earlier), which gives the researcher reason to suspect there is an association here that cannot be attributed to chance alone. Another way to look at these data is to think about randomly selecting one person from among the smokers and one person from among the nonmokers. The smoker has a chance of 647/1269 = .51 of being in the lung cancer column, while the nonsmoker has only a 2/29 = .07 chance of being there. This is evidence of strong association between smoking and lung cancer, but it is not conclusive evidence that smoking is, in fact, the cause of the lung cancer. (This is a good place to have students speculate about other possible causes that could have resulted in data like these.)

Another step in establishing association in observational studies is to see if the increase in exposure to the risk factor produces an increase in incidence of the disease. This was done with the same case-control study by looking at the level of smoking for each person, producing Table 17.

Table 17: Level of Cigarette Smoking and Lung Cancer

Cigarettes/ Day	Lung Cancer Cases	Controls	Probability
0	2	27	0.07
1–14	283	346	0.45
15–24	196	190	0.51
25+	168	84	0.67

The term "probability" is used in the same sense as above. If a person is randomly selected from the 1–14 level, the chance that the person falls into the cancer column is .45, and so on for the other rows. The important result is that these "probabilities" increase with the level of smoking. This is evidence that an increase in the disease rate is associated with an increase in cigarette smoking.

Even with this additional evidence, students should understand that a cause and effect relationship cannot be established from an observational study. The main reason for this is that these observational studies are subject to bias in the selection of patients and controls. Another study of this type could have produced a different result. (As it turned out, many studies of this type produced remarkably similar results. That, coupled with laboratory experiments on animals that established a biological link between smoking and lung cancer, eventually settled the issue for most people.)

The Appendix to Level C contains more examples of the types discussed in this section.

The Role of Probability in Statistics

Teachers and students must understand that statistics and probability are not the same. Statistics uses probability, much as physics uses calculus, but only certain aspects of probability make their way into statistics. The concepts of probability needed for introductory statistics (with emphasis on data

analysis) include relative frequency interpretations of data, probability distributions as models of populations of measurements, an introduction to the normal distribution as a model for sampling distributions, and the basic ideas of expected value and random variation. Counting rules, most specialized distributions and the development of theorems on the mathematics of probability should be left to areas of discrete mathematics and/or calculus.

Understanding the reasoning of statistical inference requires a basic understanding of some important ideas in probability. Students should be able to:

→ Understand probability as a long-run relative frequency;

→ Understand the concept of independence; and

→ Understand how probability can be used in making decisions and drawing conclusions.

In addition, because so many of the standard inferential procedures are based on the normal distribution, students should be able to evaluate probabilities using the normal distribution (preferably with the aid of technology).

Probability is an attempt to quantify uncertainty. The fact that the long-run behavior of a random process is predictable leads to the long-run relative frequency interpretation of probability. Students should be able to interpret the probability of an outcome as the long-run proportion of the time the outcome should occur if the random experiment is repeated a large number of times. This long-run relative frequency interpretation of probability also provides the justification for using simulation to estimate probabilities. After observing a large number of chance outcomes, the observed proportion of occurrence for the outcome of interest can be used as an estimate of the relevant probability.

Students also need to understand the concept of independence. Two outcomes are independent if our assessment of the chance that one outcome occurs is not affected by knowledge that the other outcome has occurred. Particularly important to statistical inference is the notion of independence in sampling settings. Random selection (with replacement) from a population ensures the observations in a sample are independent. For example, knowing the value of the third observation does not provide any information about the value of the fifth (or any other) observation. Many of the methods used to draw conclusions about a population based on data from a sample require the observations in a sample to be independent.

Most importantly, the concepts of probability play a critical role in developing statistical methods that make it possible to make inferences based on sample data and to assess our confidence in such conclusions.

To clarify the connection between data analysis and probability, we will return to the key ideas presented in the inference section. Suppose an opinion poll shows 60% of sampled voters in favor of a proposed new law. A basic statistical question is, "How far

> " Probability is an attempt to quantify uncertainty. "

might this sample proportion be from the true population proportion?" That the difference between the estimate and the truth is less than the margin of error approximately 95% of the time is based on a probabilistic understanding of the sampling distribution of sample proportions. For large random samples, this relative frequency distribution of sample proportions is approximately normal. Thus, students should be familiar with how to use appropriate technology to find areas under the normal curve.

Suppose an experimenter divides subjects into two groups, with one group receiving a new treatment for a disease and the other receiving a placebo. If the treatment group does better than the placebo group, a basic statistical question is, "Could the difference have been a result of chance variation alone?" The randomization allows us to determine the probability of a difference being greater than that observed under the assumption of no treatment effect. In turn, this probability allows us to draw a meaningful conclusion from the data. (A proposed model is rejected as implausible, not primarily because the probability of an observed outcome is small, but rather because it is in the tail of a distribution.) An adequate answer to the above question also requires knowledge of the context in which the question was asked and a sound experimental design. This reliance on context and design is one of the basic differences between statistics and mathematics.

As demonstrated earlier, the sampling distribution of a sample mean will be approximately normal under random sampling, as long as the sample size is reasonably large. The mean and standard deviation of this distribution usually are unknown (introducing the need for inference), but sometimes these parameter values can be determined from basic information about the population being sampled. To compute these parameter values, students will need some knowledge of *expected values*, as demonstrated next.

According to the March 2000 Current Population Survey of the U.S. Census Bureau, the distribution of family size is as given by Table 18. (A family is defined as two or more related people living together. The number "7" really is the category "7 or more," but very few families are larger than 7.)

Table 18: Family Size Distribution

Family Size, x	Proportion, p(x)
2	0.437
3	0.223
4	0.201
5	0.091
6	0.031
7	0.017

Notice first the connection between data and probability: These proportions (really estimates from a very large sample survey) can be taken as approximate

probabilities for the next survey. In other words, if someone randomly selects a U.S. family for a new survey, the probability that it will have three members is about .223.

Second, note that we now can find the mean and standard deviation of a random variable (call it X), defined as the number of people in a randomly selected family. The mean, sometimes called the *expected value* of X and denoted by E(X), is found using the formula:

$$E(X) = \sum_{\substack{all\ possible \\ x\ values}} x \cdot p(x)$$

which turns out to be 3.11 for this distribution. If the next survey contains 100 randomly selected families, then the survey is expected to produce 3.11 members per family, on the average, for an estimated total of 311 people in the 100 families altogether.

The standard deviation of X, SD(X), is the square root of the variance of X, V(X), given by:

$$V(X) = \sum_{\substack{all\ possible \\ x\ values}} [x - E(X)]^2 \cdot p(x)$$

For the family size data, V(X) = 1.54 and SD(X) = 1.24.

Third, these facts can be assembled to describe the expected sampling distribution of the mean family size in a random sample of 100 families yet to be taken. That sampling distribution will be approximately normal in shape, centering at 3.11 with a standard deviation of $1.24/\sqrt{100} = 0.124$. This would be useful information for the person designing the next survey.

In short, the relative frequency definition of probability, the normal distribution, and the concept of expected value are the keys to understanding sampling distributions and statistical inference.

Summary of Level C

Students at Level C should become adept at using statistical tools as a natural part of the investigative process. Once an appropriate plan for collecting data has been implemented and the resulting data are in hand, the next step usually is to summarize the data using graphical displays and numerical summaries. At Level C, students should be able to select summary techniques appropriate for the type of data available, produce these summaries, and describe in context the important characteristics of the data. Students will use the graphical and numerical summaries learned at Levels A and B, but should be able to provide a more sophisticated interpretation that integrates the context and objectives of the study.

At Level C, students also should be able to draw conclusions from data and support these conclusions using statistical evidence. Students should see statistics as providing powerful tools that enable them to answer questions and to make informed decisions. Students also should understand the limitations of conclusions

based on data from sample surveys and experiments, and should be able to quantify uncertainty associated with these conclusions using margin of error and related properties of sampling distributions.