The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 5,000 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Page 15: Broder, John M. and Megan Thee, "The 2006 Campaign; In Battleground State, Alarm Bells for Bush and G.O.P. in Poll Results," *New York Times,* April 18, 2006. http://www.nytimes.com/. From The New York Times on the Web © The New York Times Company. Reprinted with Permission.

The College Board wishes to acknowledge all the third party sources and content that have been included in these materials. Sources not included in the captions or body of the text are listed here. We have made every effort to identify each source and to trace the copyright holders of all materials. However, if we have incorrectly attributed a source or overlooked a publisher, please contact us and we will make the necessary corrections.

© 2007 The College Board. All rights reserved. College Board, Advanced Placement Program, AP, AP Central, AP Vertical Teams, Pre-AP, SAT, and the acorn logo are registered trademarks of the College Board. AP Potential and connect to college success are trademarks owned by the College Board. All other products and services may be trademarks of their respective owners. Visit the College Board on the Web: www.collegeboard.com.

Special Focus: Sampling Distributions

Important Notes

The materials in the following section are organized around a particular theme that reflects important topics in AP® Statistics. The materials are intended to provide teachers with professional development ideas and resources relating to that theme. However, the chosen theme cannot, and should not, be taken as any indication that a particular topic will appear on the AP Exam.

Within these materials, references to particular brands of calculators reflect the individual preferences of the respective authors; mention should not be interpreted as the College Board's endorsement or recommendation of a brand.

Why Sampling Distributions?

Chris Olsen
Thomas Jefferson High School
Cedar Rapids, Iowa

The outline of the AP Statistics course as it appears in the Course Description presents four basic topics: exploring data, sampling and experimentation, probability, and statistical inference. Each of the first three topics supports the "larger" idea of statistical inference.

The sampling distribution is the basis for inferential statistics, whether one is doing estimation or testing a hypothesis. It is our understanding of the behavior of sample statistics that logically forms the basis for making inferences. Without an understanding of sampling distributions, the process of making inferences is mechanical: What statistic? What table? Reject or not? Next case.

AP Statistics is a concept course, not a course in mere mechanics. For a student to be able to generalize what he or she learns in the first statistics course, the mechanics are not particularly helpful. The first step to the second course begins with an exposure to probability, random variables, and that preeminent random variable: the sample statistic. The probability distribution of a statistic—its sampling distribution—is the primordial source of the *p*-values and confidence interval lengths. This is not merely true for the statistics we encounter in the AP Statistics course—it is true of all inferential statistics.

In our statistics textbooks the processes of inference may be thought of as an *n*-act play, Act I: "Assumptions" and Act N: "Conclusion/Confidence Interval." Our textbooks will have a section or two prior to formal inference explaining sampling distributions but in our instruction they might sometimes recede into the background. To slim these sections would be as if the three witches in *Macbeth* did their bubbling, toiling and troubling while the initial credits rolled, and Macbeth—oblivious to their prattling—just grabbed a cup of soup and rode on without listening. Macbeth, of course, did not just ride off after his encounter with the witches, thank goodness. Without recurring consideration of the witches there is no drama in *Macbeth*; and without a recurring consideration of sampling distributions, there is little understandable basis for inference in statistics!

Though the witches actually appear in only four scenes in *Macbeth*, without comprehending their role and Macbeth's fascination with them we cannot properly interpret Macbeth's decisions and actions. Similarly, consideration of sampling distributions is what guides actions and decisions during the course of statistical inference. A familiarity and appreciation of the place of sampling distributions in the great N-Act play of inference will bring rewards to your students in the AP Statistics course and beyond in their next statistics course.

In these Special Focus Materials, Roxy Peck, former Chief Reader in AP Statistics, sketches the motivation for sampling distributions. Then two high school teachers, Corey Andreasen and Floyd Bullard, provide a wealth of ideas for teaching about them. AP Statistics students' mathematical knowledge of statistics can be improved, and our high school authors can choose the dynamism of simulation as a vehicle for teaching about sampling distributions. Indeed, one might argue that an experience with simulation before a mathematical presentation would improve those mathematical statistics courses!

Our "theme analogy" throughout is that sampling distributions are what-if scenarios, describing not the actual sample statistic we have but the perspective of all those sample statistics that might have been. It is this might-have-been that gives the sampling distribution its abstract quality; these classroom activities will translate the abstract into a more tactile and visual reality.

Sampling Distributions: Motivating the What-Ifs

Roxy Peck
California Polytechnic State University
San Luis Obispo, California

Sampling distributions. The topic that strikes fear into the hearts of introductory statistics teachers everywhere. Clearly this is the most abstract concept that we ask our students to come to terms with in the AP Statistics course. Nonetheless it is critical that students develop an understanding of sampling distributions if they are to comprehend the logic of statistical inference.

While the topic of sampling distributions is difficult for students because of its abstract nature, the basic idea of a sampling distribution is actually relatively simple. To illustrate the idea, let's begin with what may at first seem like a silly example. But please, do read on—the intention is to give a simple, concrete, intuitive example of what a sampling distribution is and how it is used to reach a conclusion in a hypotheses test.

I have a dog named Kirby. He is an adult dog and weighs 25 pounds. Suppose I ask you to decide if Kirby is a golden retriever.

An adult golden retriever.

If you are like most people knowledgeable about dogs, you probably would say that Kirby was not a golden retriever and that you were fairly certain that you were correct in your judgment. How would you reach such a conclusion? Informally, you would probably use what you know about the behavior of the random variable X = weight for adult golden retrievers. There is, of course, variability in the weights of golden retrievers—not all adult golden retrievers weigh exactly the same amount. But, even taking this variability into account, 25 pounds would be an extremely unusual weight for an adult golden retriever. In fact, it would be so unusual that you would probably be quite confident in saying that my dog is not a golden retriever.

In an analogy to a test of hypotheses, you could say that given the choice between

H_0: Kirby is a golden retriever

and

H_a: Kirby is not a golden retriever,

you felt that the information given ($x = 25$ lbs.) provided convincing evidence that enabled you to reject the null hypothesis. Can you be positive that your conclusion is correct? Probably not positive—Kirby might just be the smallest, skinniest golden retriever ever—but you are probably still convinced that the choice to reject the "golden retriever" hypothesis is the correct one. (And, in this instance you would indeed be correct—Kirby is a Welsh corgi.)

Let's think about the informal reasoning that led to the conclusion that Kirby was not a golden retriever. To put it in statistical language, you based your conclusion on the observed value of the random variable X = weight. The key to your being able to reach a decision depended on knowing something about the behavior of (i.e., the distribution of) the variable X = weight when the null hypothesis "golden retriever" is true. You relied on intuition and previous knowledge of golden retriever weights to make your assessment that 25 pounds would be a very unusual weight for a golden retriever. Had you not possessed the knowledge needed to make this judgment, it would have been possible to obtain the information necessary to approximate the weight distribution of adult golden retrievers by observing a large number of dogs known to be golden retrievers and then constructing a histogram of the observed weights. For example, if I had asked you if you thought that Kirby was a lesser southern ridge dog, some observation would probably be in order—your experience would be unlikely to come to your aid.

So what does all this have to do with statistical inference and sampling distributions? I would argue that exactly the same logic underlies the formal hypothesis testing procedures of the AP statistics course. In a test of hypotheses, we use data from a sample to reach a conclusion about a population characteristic (often called a parameter). For example, we might be interested in testing the claim that 70 percent of the students at a particular high school carry a cell phone against the alternative that this percentage is greater than 70 percent. A random sample of 100 students from the school will be selected and each student in the sample will be asked if he or she carries a cell phone. The sample proportion, p, will then be used as the basis for making a decision to either fail to reject $H_0 : p = 0.70$ or to reject $H_0 : p = 0.70$ in favor of the alternative $H_0 : p > 0.70$. How can we make this decision? Just as knowing something about the distribution of the random variable x = weight when the hypothesis "golden retriever" is true in the dog example led us to a conclusion. What is needed in the cell phone hypothesis test is information about the behavior of the sample proportion (i.e., the distribution of the sample proportion) when the null hypothesis of $p = 0.70$ is true.

Consider the following: the sample proportion from a random sample of size 100 is a random variable. How so? A random variable associates a value with each outcome in the sample space for some chance experiment. Here, think of the experiment as selecting a random sample of size 100 from the population of students at the high school. The sample space (set of all possible outcomes for this experiment) consists of all the different possible samples of size 100. The random variable $\hat{p}$ associates a value with each different sample (which is the proportion who carry a cell phone for that particular sample), and so $\hat{p}$ (or in fact any other sample statistic) can be regarded as a random variable.

Since a sample statistic is a random variable, then just like all random variables it has a probability distribution that describes its behavior. When the random variable of interest is a sample statistic, its probability distribution is called a sampling distribution.

So, if we knew the distribution of p when $H_0 : p = 0.70$ is true, we would know a lot about the behavior of $\hat{p}$ when samples of size 100 are selected from the population. In particular, we would be able to distinguish "usual" values from extreme values, and this provides what is needed to make a decision in a hypothesis test.

For example, if we knew that $\hat{p} = 0.80$ would be unlikely to occur when $p = 0.70$, we would be able to reject the null hypothesis $H_0 : p = 0.70$ with confidence if we observed a sample proportion of .80. On the other hand, if $\hat{p} = 0.73$ is a "usual" value for the sample proportion when $p = 0.70$, we would not be able to reject the hypothesis $H_0 : p = 0.70$.

What makes this scenario more difficult than the "golden retriever hypothesis" example is that most people can't rely on intuition and prior knowledge to make the assessment of what are usual values and what are unlikely values for the sample proportion random variable. It is here where simulation and statistical theory can help.

The general results about the sampling distributions of sample statistics (e.g., a sample mean, a sample proportion, the difference between two means or two proportions), provide the information that enables us to make the necessary distinction between usual and unusual values under the null hypothesis.

As you will see in the accompanying articles, simulation is a great way to approximate sampling distributions and to motivate theoretical results about the sampling distribution of sample statistics in many situations. But ultimately we rely on statistical theory (e.g., proven results such as "the distribution of the sample mean for a random sample of size n from a population with mean μ and standard deviation σ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ when the sample size is large") to tell us what we should expect to see when a particular null hypothesis is true.

So, let's compare the two scenarios considered here—the first, obvious and intuitive dog scenario and the second, more realistic cell phone scenario.

Dog Scenario	Cell Phone Scenario
H_0 : Kirby is a golden retriever H_a: Kirby is a not a golden retriever **Random variable:** X = weight **Observed value:** $x = 25$ **Question of interest:** Would the observed value $x = 25$ lbs. be unusual if Kirby is a golden retriever?	$H_0 : p = 0.70$ $H_a : p > 0.70$ **Random variable:** $\hat{p}$ = sample proportion **Observed value:** $\hat{p} = 0.80$ **Question of interest:** Would the observed value $\hat{p} = 0.80$ be unusual if $p = .70$?
Assessment: Based on what we know about the distribution of X = weight when H_0 is true, 25 is an unusual value. We reject the hypothesis that Kirby is a golden retriever in favor of the alternative hypothesis that Kirby is not a golden retriever.	**Assessment:** If H_0 is true, theory tells us that because the sample size is large—($np = 70$ and $n[1 - p] = 30$), $\hat{p}$ has a distribution that is approximately normal with mean .70 and standard deviation $\sqrt{\frac{p(1-p)}{n}} = .046$. The observed value of $\hat{p} = 0.80$ is an unusual value when H_0 is true because it is more than 2 standard deviations above the mean, which is unusual for a normal distribution. We reject the hypothesis that the proportion who carry a cell phone is 0.70 in favor of the alternative hypothesis that the proportion is greater than 0.70.

In my experience, students understand the dog example and find the reasoning intuitive. The only difference in the cell phone scenario is that we needed a little help when it came to making the "likely versus unlikely" assessment. Knowledge of the sampling distribution came to the rescue, providing the necessary information.

Consider trying an approach like this to motivate the study of sampling distributions. One reason that students have difficulty with the concept is that it is often introduced in the abstract and students don't see why they would need to know the information that sampling distributions provide. Once students understand this, it is much easier to introduce the formal concepts of sampling distributions.

Sampling Distributions: The What-Ifs with Hands-On Simulation

Floyd Bullard
The North Carolina School of Science and Mathematics
Durham, North Carolina

Sampling distributions are difficult for many students to understand. When students first learn about distributions, they do so in the context of population or sample data. A common graphical representation of such data is a dot plot; each dot in such a dot plot corresponds to a real element of the population or sample. But a sampling distribution is more abstract. If one imagines a dot plot of a sampling distribution, then each dot corresponds to a particular *possible* random sample, most of which in all likelihood never was and never will be collected or observed. What's more, the correspondence isn't a direct measure of a characteristic of an actual object as it is with population or sample data—each dot corresponds to some *function* of everything in the sample and may not have any meaning in the context of a single individual.

The sampling distribution of a statistic is a distribution of imaginary outcomes, each one possible in a hypothetical sense. Only one of them is actually realized and observed. For our students to understand such a distribution, they must take a rather large step from the concrete world of measurable, tangible things into the world of alternate realities—they must learn to play *what if.*

This article is about classroom practice. You will find here seven classroom activities that all involve using simulations to approximate sampling distributions. They are arranged in the order that I use them when I teach AP Statistics. I do not myself suggest teaching a single "unit" on simulations. Rather, I use simulations throughout the year to help teach many different concepts. By sowing seeds of understanding of sampling distributions early and often during the year, the concept—before it is a crucial element of the course—becomes more natural to students than would be the case if the only distributions they saw were of raw data.

My class size is typically around 20. I believe the activities in this article will work well for classes of between 12 and 24 students, although they may need to be modified slightly for class sizes toward the small side of that range. Some of the activities may be modified for classes with fewer than 12 students, although generally they will take longer, since the modification will often take the form of one student doing what would usually be a task for two.

The activities described in this article are:

1. *Capture/Recapture.* It can be completed in a single 50-minute class period and requires no previous knowledge of statistics. It introduces students to a number

of ideas that are important in the AP Statistics syllabus, including point estimates, simulations, assumptions, and graphical representations of data, sampling distributions, and inference.

2. *Polls (Sample Proportions).* I use this activity fairly early in the year, around October, to give students an intuitive introduction to inference for a single proportion. This is well before such inference is more formally covered in the syllabus. I like this early introduction because during election years it is always highly relevant. In addition, I find that foreshadowing a topic early makes its later and more formal discussion easier for students to grasp, almost as if the students had been unconsciously digesting it during the interim.
3. *The German Tank Problem.* This popular activity particularly stresses the concept of a sampling distribution and may be used to introduce that idea. It also introduces the ideas of bias and sampling variability, and how estimators can be evaluated.
4. *Baseball Players' Salaries (The Central Limit Theorem).* This activity introduces students to the Central Limit Theorem by having them sample baseball players from a known population and average their salaries. Students will see that although the population of salaries is highly skewed, the distribution of sample means is approximately normal when the sample size is fairly large.
5. *Standardized Mean Heights (the t-Distribution Family).* This 20-minute activity uses only a calculator and introduces students to the *t*-distribution family without entangling it with inference. The distributions are not plotted; rather, students call out loud simulated sample statistics and the heavy-tailed *t*-distribution is perceived *aurally*. With just five additional minutes, the activity may be extended to show students that as the sample size grows larger, the *t*-distribution has lighter and lighter tails, becoming more like the normal distribution.
6. *Baseball Players' Height/Weight Relationship (Regression Line Slopes).* With this activity we return to the list of Major League Baseball players. This time, multiple samples are taken and used to construct regression lines of weight predicted from height. The slopes vary from sample to sample and by plotting a distribution of the slopes, students will understand the slope as a sample statistic with a distribution—a fact that often eludes them when their experience of bivariate data is limited to single samples.
7. *Worm Species (the Chi-square Distribution, Sort-of).* This activity is meant to precede a lesson on the chi-square test of goodness-of-fit. The chi-square statistic is never actually used in the activity itself. Instead, the activity permits students to create their own measure of "discrepancy" between a claimed categorical distribution and a set of categorical data—and then to simulate the sampling distribution of the measure they devised. From there it is only a short step from the goodness-of-fit test *concept* to the actual chi-square test.

1. Capture/Recapture

The following "Capture/Recapture" simulation is my preferred "first day of class" activity; it is accessible to students on Day One and in addition foreshadows much of what will come later in the year—point estimates, sampling distributions, simulations, graphical representations, inference, and assumptions.

For this activity, I like to use plastic frogs and beads, but M&Ms or any colored tokens work just as well.

Put at least 100 frogs in a container such as a bag or tub. Show the students your container of frogs and then carry out the capture/recapture scenario, well-known to biologists and statisticians but probably unfamiliar to your students. That is, we will "capture" a certain number of frogs and "tag" them (here, by replacing the captured frogs with frogs of a different color), then release them back "into the wild." We will then capture another set of frogs, this time not tagging them but simply counting how many among those captured are already tagged. For teaching purposes, it is helpful if the sample sizes in the two phases are different from one another (so students won't confuse them). If you use about 100 frogs, then you should try to have around 20–30 frogs in both stages' samples, though you need not count them out exactly—indeed, *not* counting them out exactly more closely mimics the way such studies are actually carried out. If you use more than 200 frogs in your population, you might want to capture 35–50 frogs in each of your stages' samples.

After you have carried out both stages of sampling, ask your students to estimate the population size. Let us suppose that you captured and tagged 25 frogs in stage one, and then captured 29 frogs in stage two, finding 7 of them already tagged. Your students will likely set up the following proportion:

$$\frac{7}{29} = \frac{25}{N}$$

where N is the unknown population size. Solving for N in this case, we estimate the population to have about 104 frogs.

If time permits, you might want to lead your students in a discussion about what assumptions are being made when we compute that point estimate. One of the most important is that both capture stages involved a *simple random sample* of the frogs in the population. (This assumption is credible because the proportion stated above is based on a well-mixed frog population.) In practice, how should that impact how the actual study would be carried out? Among other things, since it is probably not reasonable to assume that the frogs are randomly shuffling themselves about at all times, it means that both capture stages must involve sampling from random locations. Additionally, we assume that between the two capture stages the population stays the same size, and the number of tagged frogs remains the same. This means that we should not wait too long between the two capture stages, since that would allow the population to change sizes, perhaps substantially. Also, the tags must not make the frogs any more or less likely to be captured the second time than the nontagged frogs. In particular, the tags must be harmless to the frogs, since a dead frog is one that is unlikely to be recaptured.

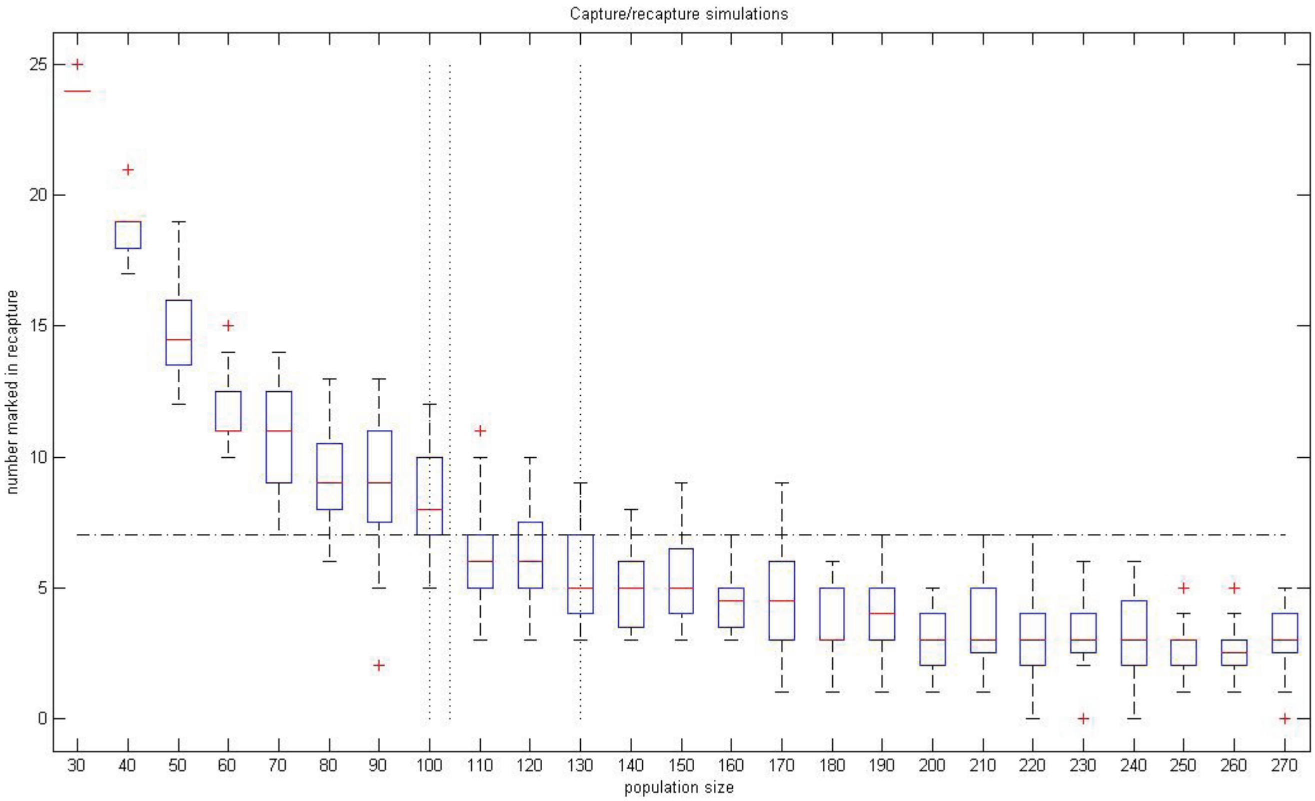
A point-estimate alone does not require a simulation, and indeed this activity is not helpful for middle school students if all that is desired is a point estimate. But a crucial part of inference is attaching to a point estimate some margin of error. Although the theoretical variability of the point estimate in this activity—the estimated population size—is not within the AP Statistics curriculum, students can estimate its variability through simulation. That's what this activity is primarily about: using simulations to assess variability and uncertainty—variability in the sample, and the consequent uncertainty about the parameter (population size).

When we design our simulation to estimate the variability in our statistic, we have to choose a population size to work with. But in reality you wouldn't know the population size either before or after doing your study. So how can you assess the variability of your statistic accurately?

One way is to see how the statistic behaves for a variety of different values of the true parameter—in this case population size. In order to help distinguish between actual data and simulations, I have the students use beads instead of frogs; my frogs are the "real" data, while their beads are simulations of possible other outcomes. Any colored tokens will do. One color represents untagged frogs and another color tagged frogs. Students simply switch one color for another to "tag the frogs."

Have students work in groups and assign them different population sizes ranging from 30 to 300 or so by 10s. Give them the beads or tokens they need to conduct the simulation themselves. The student group with population size $N = 30$ will require 30 beads, and so on. They are to tag as many beads as you did earlier with the "real" frogs by replacing that number of beads with another color, then mix their beads well, and then sample as many beads as you did in the second capture stage, thus replicating the earlier study exactly, except for the population size. They should then repeat the second-stage capture process a total of 20 times, counting each time how many "tagged frogs" they found in their sample. (They do not need to repeat the tagging process each time.)

On the board, draw a pair of perpendicular axes, one (the horizontal is better) marked "population size" and the other marked "number marked in recapture." After each student group has performed 20 simulations, have the group come to the board and draw a boxplot of their estimates over their actual frog population size. The end result should be a series of parallel boxplots such as the one shown in the picture below.



In the graph above, a horizontal dashed line is drawn at 7, the observed number of tagged frogs that we saw in our second capture stage. We now address the question "How many frogs are there in the population?" We already came up with a point estimate (about 104 frogs) using a proportion; that is where the middle of the three vertical dotted lines are drawn. But we know that the point estimate of 104 frogs is not necessarily exactly right. We really want to know *what other possible population sizes are consistent with our observation of 7 tagged frogs in the recapture stage*. For this, we look at which boxplots contain "7" as a "typical" value. Let us suppose that we define "typical" to be the middle 50% of the values—those represented by the center box in each boxplot. The graph suggests to us that populations ranging from $N = 100$ to $N = 130$ might very typically have resulted in 7 tagged frogs in the second capture stage. That is where the other two vertical dotted lines are drawn. Thus, under this definition of consistency between population and observation (i.e., observation falls in the middle 50 percent of its sampling distribution under a given population size), we estimate that there are between 100 and 130 frogs in the population. We now have not only a point estimate, but a range of other plausible values as well.

If your students do not know what boxplots are on the first day of class, you may use this activity a week or two later, as an application of that topic. Or you may use this activity on the first day of class but modified in the following way. Instead of constructing a boxplot of their 20 sampled values, students are to order their 20 sampled values from smallest to largest and keep only the middle 18 as “typical” (discarding the highest or lowest value if it occurs only once and is therefore not “typical”) and then draw a vertical line segment from their lowest typical value to their highest typical value in lieu of a boxplot. The rest of the activity works the same way.

2. Polls (Sample Proportions)

Some years are more interesting than others with respect to preelection polls, but every year around October you can easily find lots of polls about how people feel about different candidates for office. That’s a little early in the AP Statistics year to be teaching about confidence intervals, but it’s not too early to plant the seed of understanding sampling distributions, which is key to so much of inference. The following is an activity that can be done with students using any poll, not just a political one. My recommendation is to use one conducted by a reputable organization, such as Gallup or the *New York Times*. The latter is very good about printing with their polls a statement about how the poll was conducted and what its margin of error means. Discovering that meaning is what this activity is about. For example, the following statement from the *New York Times*, April 18, 2006, accompanied a poll of Ohio residents. A key statement is printed here in boldface.

> The latest New York Times/CBS News poll of Ohio is based on telephone interviews conducted Oct. 11 to Oct. 15 with 1,164 adults throughout the state. Of these, 1,020 said they were registered to vote.
>
> The sample of telephone exchanges called was selected by a computer from a complete list of Ohio exchanges. The exchanges were chosen so as to ensure that each area of the state was represented in proportion to its population. For each exchange, the telephone numbers were formed by random digits, thus permitting access to listed and unlisted numbers alike.
>
> Within each household, one adult was designated by a random procedure to be the respondent for the survey.
>
> The results have been weighted to take account of household size and number of telephone lines into the residence, and to adjust for variations in the sample relating to geographic region, race, sex, age, education and marital status.
>
> **In theory, in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult residents of Ohio.**
>
> For smaller subgroups the potential sampling error is larger. Shifts in results between polls over time also have a larger sampling error.
>
> In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll. Differences in the wording and order of questions, for example, can lead to somewhat varying results.

For the purpose of the present activity, we will use one of the results of the *Times*’ poll published October 18, 2006: When asked “Compared with previous congressional elections, this year are you more enthusiastic about voting or less enthusiastic?”

Forty-two percent of registered voters said "More." Let us suppose that we have just shared this result with our students. We now call to the students' attention the bold statement above: "In theory, in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult residents of Ohio." How, we ask, can they know that?

This activity requires a calculator such as the TI-83 that is capable of simulating binomial random variables. Tell your students that you are going to *simulate* a poll of 1,020 randomly selected registered Ohio voters. (At present, they are just to observe what you do, not conduct a simulation themselves.) The syntax on the TI-83 for simulating a binomial random variable with parameters *n* and *p* is `RandBin(n,p)`. Simulating a poll of 1,020 randomly selected registered Ohio voters therefore requires entering `RandBin(1020, p),` where *p* is the proportion of *all* registered voters in Ohio who are more enthusiastic about voting this year than in previous congressional election years. Unfortunately, *p* is not known to us.

Once you are sure that your students understand how you will simulate the poll, and the problem of not knowing *p*, ask them for suggestions. Many will want to use 0.42 for *p*; since that was in fact the actual sample proportion, it is our best guess as to what *p* really is. That's fine, but it is very important that the students understand that the 0.42 we are entering is just a guess as to the actual population proportion. We don't really know that $p = 0.42$.

For our sample of 1,020 Ohio voters, we enter `RandBin(1020, 0.42).` Let's suppose that after repeated trials, your class reported 433 "more enthusiastics." Then ask the students what they'd like to do with that number; hopefully someone will say, "Let's compute the sample proportion." That value would be $433/1020 = 0.4245$, which we round off to 42 percent.

If you happened to get 42 percent, ask your students, "I got 42 percent. If I simulate a new random sample, will I get 42 percent again?" Or, if you happened to get something other than 42 percent, ask your students, "I used 0.42 for *p*, but I got [let's say] 45 percent from my simulation. Why are they different?" The point of these questions is to guide students to seeing that the simulated sample proportion need not match the presumed population proportion, and that if a new sample is taken, you may get a sample proportion that not only differs from the population proportion but may also differ from the first result.

Once the students understand that, repeat the simulation. `RandBin(1020, 0.42).` Let's suppose that this time you get 414, and $414/1020 = 41\%$. (We are now playing the *what if* game. *What if* the sample had been *this* particular group of 1,020 people?) You want to be sure before continuing that the students understand what is happening each time you simulate a sample. You are simulating a new random sample of 1,020 registered Ohio voters, asking them the question about voting enthusiasm, and counting how many people in *that* random sample respond "more enthusiastic," still assuming that in the whole population, the true proportion who feel that way is 42 percent.

Once they understand that, generate a few more random samples. Let's say we now get a 44 percent and a 39 percent. We have now accumulated four samples, and therefore four sample proportions: 42 percent, 41 percent, 44 percent, and 39 percent. At this point draw a horizontal line on the board, put numbers under it for integer percents ranging from about 35 percent to 50 percent, and begin constructing a histogram by drawing an "X" over each of the four sample percentages you've obtained. Underneath the line, label the axis "% saying *more* in random sample, supposing $p = 0.42$ in population."

When you are confident that the students understand the simulation so far, then instruct them to all do the same thing you just did on your calculator, and write down the sample proportion they got. Take a quick survey in the class. "How many of you got 42%? How about 45%? Anyone higher than 50%? No? How about lower than 35%? No one?" You would like students to realize, even if it is at this point unconsciously, that while there is variability in their sample proportions, it is not dramatic. Few students, in fact, will have results greater than 45% or less than 39%.

Ask your students to do five or so more simulations each[1], and then to come to the board and continue constructing the histogram.

In my classes I do so many simulation activities during the year that my students are quite familiar with doing this by mid-October and I hardly need to instruct them at all. If I draw an axis on the board and put an "X" over it somewhere, they know that they'll shortly be at the board doing the same thing. This seems to me a good thing. We are constructing simulated sampling distributions so early in the school year that by the time we get to formally talking about what they are and giving them the name *sampling distribution*, my students already really know what they are: a sampling distribution is a histogram[2] of sample statistics you would get from many different possible random samples.

You should see on the board a more or less normal-shaped distribution centered on 42 percent. Remind the students once again of the newspaper statement: "In theory, in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all adult residents of Ohio." Does our simulation bear that out?

The students at that point will hopefully think to count how many of their simulated samples were off from the population proportion of 42 percent by more than three percentage points. And hopefully it will be only about 5 percent of your simulated samples. So while you haven't *proven* anything, you have at least seen what is meant by the newspaper statement.

[1] The number of simulations you ask them to do depends on how many students are in your class. A hundred or so simulations for the whole class is a nice target number, so for a class of 20 you might ask them each to do 4 or 5 simulated samples.

[2] I am aware of using sloppy language here: a distribution is not a histogram. The latter is a graphical representation of the former. But conceptually, students who associate the board histogram with the repeated-sample simulation likely have a correct understanding of what a sampling distribution is.

And you also have given students a healthy introduction to sampling distributions, even if you never use that term.

But the activity isn't over. Ask your students: "Are you convinced? Do you believe what the newspaper says about 19 out of 20 cases being within three percentage points?" It is not an easy question to answer, but with some guidance (it might help to point to what you wrote under the histogram), they may realize that your activity up to this point relied upon a supposition that may in fact not be true: that the real proportion of all Ohio registered voters who are more enthusiastic this year is 42 percent. The poll suggested that, but we don't really know. What if in fact the real value of p was something different? Suppose it's *very* different! Let's see what happens if we repeat the activity but this time use p = 0.80.

On another part of the board, draw a new axis and numbers ranging from 75 percent to 85 percent. Write under it "percentage saying *more* in random sample, supposing $p = 0.80$ in population." Get the ball rolling by doing one or two simulations yourself, entering `RandBin(1020, 0.80),` then ask them to do several simulations each (about as many as they did before), and put them on the histogram.

The result, not surprisingly, is that even with a dramatically different value of p, it is unusual for the sample proportion to differ by more than three percentage points from the population proportion. It is still the case that in about 19 out of 20 cases, we are within three percentage points of the population proportion.

And now the activity really is over. There are two concepts that have been addressed, both of them planting seeds of topics that will be covered more thoroughly later in the AP Statistics course: sampling distributions and confidence intervals.

Before moving on to the next activity, I will make three comments. First, you may notice that we are here playing *what if* on two levels. We are taking repeated samples and addressing the question, "what if *this* had been our random sample?" This is the *what if* that is being referred to in the title of this article, and it is the basis of sampling distributions. But we are also looking at what the entire sampling distribution would have looked like under two different values of p. What if p were 0.42? What if p were 0.80? That is conceptually a different matter.[3] Help students be aware that the reason we look at different possible samples is not the same as the reason we look at different possible parameter values. We do the former because we want to understand the behavior of a sample statistic over many repeated samples. We do the latter because we want to see whether, and how, that behavior depends upon the parameter value.

My second comment is that the calculator will actually permit the creation of many outcomes of a binomial random variable at a time, by adding an additional argument after n and p: `RandBin(n, p, N)` will create N binomial outcomes, each with parameters n

[3] This was also addressed in the Capture/Recapture activity, by having students see what the sampling distribution of recaptured frogs would look like under *many different possible population sizes*.

and *p*, and return them as a list. So you can actually *practice the activity yourself* before you do it with the class, like this: `RandBin(1020, 0.42,100)/1020→L1.` Then make a histogram of list L1 to see what to expect on the board when you do the activity in class.

I do *not* recommend that you actually do multiple simulations this way in class, however. There are two good pedagogical reasons for not doing this in class. First, done this way the activity would become a mystifying "black box" for many students. They push a button and they get a histogram, but they don't know what it means and they're no closer to understanding sampling distributions than they were before. Students need to see samples simulated[4], and a statistic computed for *each* sample in order to appreciate what's going into the sampling distribution. The second pedagogical reason is that *n* and *N* are two completely different things, and there's no need to invite students to confuse them.

Finally, my third comment is that at some point during this activity it may be worth pointing out to students, perhaps by way of asking leading questions, that the number or registered voters in Ohio—i.e., the population size—is irrelevant to the inference. You could, for example, at the conclusion of the activity ask the students whether the margin of error would be any larger if you sampled the same number of voters from the entire United States rather than just from Ohio. Very likely, some students will think that the margin of error should be larger since the samples would then represent a much smaller fraction of the population. But if you then press them to explain what would be different about the simulation activity, they may realize that nothing in the activity requires knowing or using the population size at all. For many students this is troubling because it is so counter to their intuition. Yet it is, of course, a fact, so exposing students to this fact about *N* while conducting this activity early in the year will serve them well later.

3. The German Tank Problem

Teachers tend to fall into three groups with respect to the German Tank Problem. There are some teachers who have never heard of it, a group which I happily find to be diminishing from year to year; there are those who have heard of it but not tried it in their own classrooms; and there are those who have tried it and love it.

Those who fall in the second group often have chosen not to use the activity because they fear that taking a class day—or even worse, two class days—for a single activity is too great a price to pay. They assume that it is time lost, that the rest of the syllabus will still take the same amount of time, and they will therefore be obliged to cut or crunch at some point in the future. I believe they are mistaken. An understanding of sampling distributions is very important for students in AP Statistics, and the German Tank activity is very good for introducing students to the concept. Time spent on the activity introducing sampling distributions early will actually save time in the long run. Future discussions about bias, the Central Limit Theorem, confidence intervals, *p*-values, significance levels, and other

[4] Admittedly, even the sampling procedure is a little bit of a "black box" in this activity. But I have found it is sufficiently accessible to students.

concepts associated with sampling distributions will go much more smoothly later in the course if students have a solid grasp of sampling distributions earlier. And I have found no better activity than this one for giving students that solid grasp.

The version of the German Tank problem I present here is a variation on an activity I learned about at an NCTM meeting, one that I have found works well for my students and can be done in one or two 50-minute class periods. My teaching colleague Dan Teague has written an excellent paper about a variation on this activity that does not have a war context, which he calls The Taxi Problem[5]. I prefer to preserve the war context, first because it is historic (it was a real mathematical problem during World War II whose solution had strategic implications), and also because I think it is good for students to see the full breadth of the real-world applications of mathematical problems.

The history behind this famous problem is (more or less) as follows. During WWII, Allied spies were asked to estimate the numbers of tanks the Germans had of various types. At about the same time, the Allies were able to capture a number of German tanks, and it was discovered that part numbers on the tanks had coded information that almost certainly indicated serial numbers from the same factories. The part numbers were decoded, and British mathematicians were given the serial numbers and asked to estimate the number of tanks. The mathematicians came up with estimates quite a bit lower than those given by the spies. Long after the war, it was discovered that the spies had been deceived by the Germans repainting their tanks to increase their apparent numbers. The mathematicians were much closer to getting the number of tanks right.[6]

For the classroom activity we simplify the problem by considering a population $\{1, 2, 3, \ldots N\}$ with an unknown parameter, the population size N, to be estimated. In advance of doing this activity you should prepare bags of numbered tags, such as squares of cardstock paper. The bags should all be identical, containing chits going from 1 up to the same number N. There should be one bag for every 3 or 4 students in your class. Let's suppose N is 342.

On the day of the activity, put your students in groups of 3 or 4 and give each group a bag. Tell your students the historic context and tell them that they are going to play the role of the British mathematicians. Each group shuffles up the chits in their bag and then draws 7 numbers at random. Each group's task is then to come up with (1) an estimate of N, and (2) a description of the process they used to come up with their estimate. The latter, they are instructed, must be sufficiently clear that it may be applied to any sample of 7 numbers.

If they finish early, they are asked to come up with another method. Students often come up with lots of good ideas, including things like "double the sample mean," "double the sample

[5] This paper can be found at http://courses.ncssm.edu/math/Talks/index.htm.

[6] This problem was first introduced to the world in 1947, shortly after many documents concerning WWII became declassified. The original article was *An Empirical Approach to Economic Intelligence in World War II* by Richard Ruggles and Henry Brodie, published in *the Journal of the American Statistical Association*, Vol. 42, No. 237. (Mar., 1947), pp. 72–91. Much has been published about it since then, and information can readily be found on the Web by searching for "German Tank Problem."

median," "six times the sample standard deviation," "the mean plus two standard deviations," and more. Occasionally a group comes up with "the smallest number in the sample plus the largest number in the sample" and I even once had a group say "8/7 times the largest number in the sample." Each of these may have a rational justification.

After 10 minutes or so, have the students reveal their methods and their estimates (speaking technically, the methods are "estimat**ors**" and the actual numbers are estimat**es**) and I write them in two columns on the board. The same method often comes up multiple times. When this happens, I write the different groups' estimates in a row next to it. On the board you may see something like the following:

Method	*Estimate of population size*
Double the mean	358, 480, 404
Six times the standard deviation	874
Four times the standard deviation	515, 353
Sample max plus sample min	320
Third quartile plus one standard deviation	499
Double the median	408, 644, 212

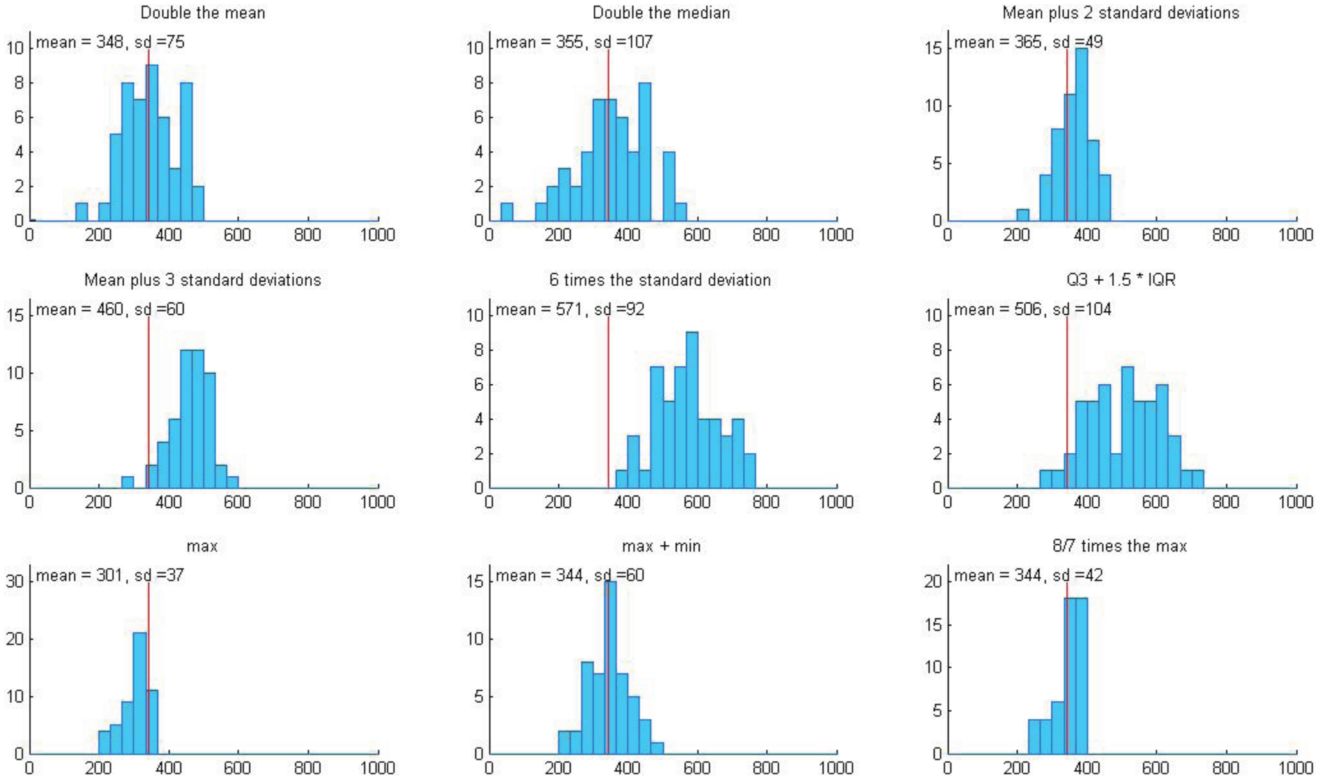
The students usually want to know the "true" answer, and at this point it could be revealed. I then find the estimate on the board that comes closest to that number and point to it and say, "This estimate is closest. Therefore, this method [whichever it is] must be the best method for estimating, right?" A few students will say "Yes," but most will see that the goodness of the estimator cannot be judged by a single estimate based on one random sample. Is the estimate good because the method is good or because the sample was "lucky"?

A discussion then begins with students about how one might judge estimators. If you can't judge an estimator based on how it did in practice with a sample (after all, in practice, we usually get only one sample), then how are we to judge it? One answer is: We judge it on how "well" it would perform over many possible random samples—and this brings in the idea of simulations.

The students should then simulate their own random samples of size 7 from a population whose size N is known. It is probably a good idea here to use the number that you actually used for the bag of numbers. Let us suppose it is $N = 342$. A sample of size 7 may be simulated on the TI-83 thus: `randInt(1,342,7)→L1`, where "→" is the "store" function. (Occasionally students will get a duplicate in a list—have them replace the sample with another.) Instruct your student groups to create 50 or so simulated samples from the population and apply their method (estimator) to each sample, recording the estimate that each sample produces. (Note that this is the moment when the *what if* game is being played. "What if this had been our actual sample? . . . What if *this* had been our actual sample?") When they're finished, they should make a histogram of their estimates (this is the estimated *sampling distribution of their statistic*) on their calculator and then on the board.

It is also helpful to have them report the mean and standard deviation of their sampling distribution.

Below is a set of nine histograms, each based on a different estimator, showing the sorts of histograms that are typical. In these graphs, the population size $N = 342$ was used, and each histogram reflects 50 simulated samples. A vertical line is drawn at $N = 342$ to make it easier to see where the true population parameter lies. Additionally, the same horizontal scale is used for all graphs to make it easier to compare the distributions' spreads. Finally, the mean and standard deviation for each sampling distribution is given.



After the histograms are drawn on the board, the discussion resumes once more: Which estimator (method) is "best"? We've made it clearer now what is meant by "best" in that we've specified that it must be "good over many random samples," but we still haven't defined "good." Do we want to choose the estimator that is exactly right most often? Perhaps, but a method that generally comes very close but never actually gets it exactly right may still be a good estimator. What then?

This is an excellent time to discuss bias and variability. All other things being equal, lack of bias is a good thing, and so is low variability. Put together, they make a good estimator; an estimator with low bias and low variability results in estimates that are pretty close most of the time. If a student group comes up with "six times the sample standard deviation," a simulation based on 50 random samples will show a clear bias, leaning towards overestimation of N. Likewise "the mean plus three standard deviations."[7] "Two times the mean" can be seen to have lower variability than "two times the median," even though both are unbiased. You have to pay attention to the scales on the histograms to compare these appropriately.

"Sample minimum plus sample maximum" does surprisingly well. Students are always impressed by that one. They eventually will probably want to know what the British mathematicians did. Although the real-world problem involved an unknown upper *and lower* bound to the population, the mathematicians chose as their estimator the equivalent of what, for this activity, would be 8/7 times the sample maximum. This happens to be (though you need not share this with students) the estimator having smallest variance among all unbiased estimators of *N*. But interestingly, the distribution of this statistic is skewed, not symmetric. Students don't like that. They think something must be wrong with a statistic if its distribution is skewed. But that is, of course, not so. There is no inherent reason to prefer a symmetric distribution over a skewed one.

A few students have pointed out that in the context of the German tanks, bias in one direction may be worse than bias in the other. It may, for example, be much more dangerous to underestimate your enemy's strength than to overestimate it. This is an excellent point. Although *unbiasedness* and low variability are good things, there is in fact no single gold standard by which to compare all estimators. It depends on what you want the estimator to do.

Here is a final comment on an issue that I do not suggest should be a focus in class, but which, if your students raise it, may warrant a brief discussion, such as the following: Joe Student: "We performed all of our simulations using $N = 342$ because we knew that was the right answer. But in real life you wouldn't know what the right answer was. So how could you perform the simulations?" You: "Good point, Joe! One thing we could do is perform the simulations for a variety of different plausible values of *N*. Remember how we did that with the Capture/Recapture problem on the first day of class? Or it might turn out that the way an estimator performs for one value of *N* is about the same as the way it performs for any value of *N*. For this problem, for example, we can observe the following: If we change *N*, we really only change the *scale* of values in our sample, and therefore, for these estimators, we also only change the *scale* of the sampling distributions. How the estimators perform relative to other estimators would be the same, even if *N* were larger or smaller."[8] Joe: "Thanks, O Wise and Sage Instructor!" (Well, OK, we might be reaching for that last comment.)

4. Baseball Players' Salaries (The Central Limit Theorem)

I have found that a good way to simulate samples from an actual population is to create a complete list of a population whose properties can be determined and to index them with

[7] Both of these methods are justified by students using properties of the normal distribution, but this population is not normal.

[8] It very rarely happens—I have never had it happen—but it is *possible* for students to concoct an estimator for which this is not true. If an estimator of *N* involves *adding a constant*, then that will not "stretch" as *N* changes.

consecutive integers ("ID numbers").[9] Students can then use their calculators to generate random integers and refer to the list to see which population member has that ID number. In this way they can fairly quickly construct random samples from a population, and the sampling process is transparent, not hidden by the technology.

One such complete list that I have found useful in the classroom is all Major League Baseball (MLB) players. Posted on AP Central® is such a roster, including players' names, teams, jersey numbers, positions, ages, heights, weights, and salaries.[10] Although the activities described in this article only involve sampling salaries, (this activity and heights and weights, a later activity), other items are included because they make the data more accessible to students, and they may be of use to teachers in other sampling activities they may devise on their own.

The purpose of this activity, is to demonstrate the Central Limit Theorem. The salaries of MLB players are highly skewed, but the sampling distribution of sample means is fairly normal when the sample size is around 20.

For this activity students should each have a list ($N = 866$) in their hands, or at least one list per pair of students. They are to randomly choose a player from the list by entering on their calculators:

```
    randInt(1,866)
```

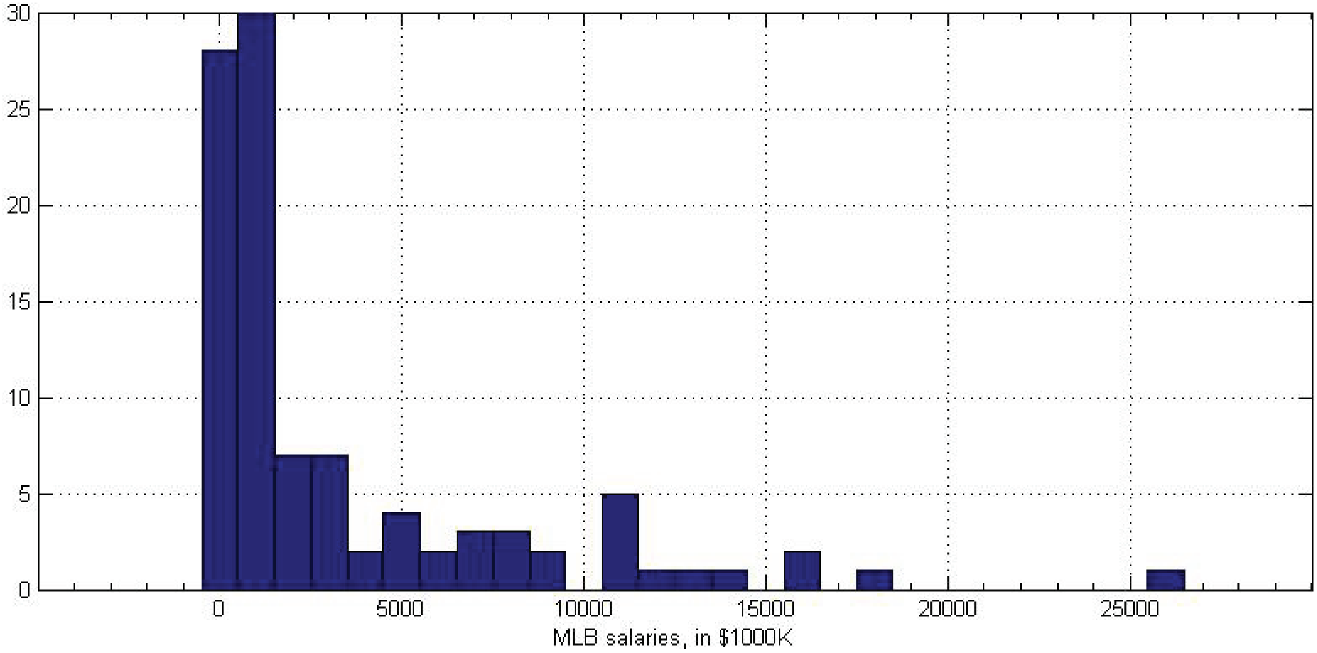
Then they look in the list to find the salary of the player whose ID number is the one they just found. They should write down that salary on a piece of paper, sample again, get a new salary, etc. Ask the students to do this several times each, enough to have a total of about 100 samples among all your students.[11] As an example, if you have 20 students working in pairs, then each pair should sample about 10 MLB players. I would recommend that you explicitly say "I want each pair of students to get *about* 10 *or so* randomly sampled players' salaries." This performs the valuable function of deemphasizing the number 10, because it is not crucial in this activity.

While they are sampling baseball players' salaries, draw a horizontal axis on the board, on top of which a histogram of salaries will be constructed. As your students finish sampling their players, they should come to the board and put X's over the salaries they sampled. Ask them to round to the nearest $1,000K (i.e., million). Below is a histogram of what 100 sampled salaries might look like.

[9] For this I thank my teaching colleague Gloria Barrett.

[10] It is possible that the list posted online will be more up-to-date than the list which is referred to in this activity description. Naturally, the activity and the concepts are the same—only the particular players and their salaries will have changed.

[11] I have found from experience that about 100 X's in a board histogram are usually sufficient to show the important characteristics of the distribution.



Several things are obvious from the histogram. First of all, the salaries are very right-skewed. Second, a typical salary is around $1,000K, a million dollars. It is not clear what the mean salary is, since the skew makes it difficult to tell, but it would appear to be around $3,000K ($3 million) for this set of 100 draws.

The next phase of this activity is to have students sample again, but this time they are to take samples of size $n = 5$ and average the five salaries together. The easiest way to take a sample of size $n = 5$ is probably to have students enter this on their calculators:

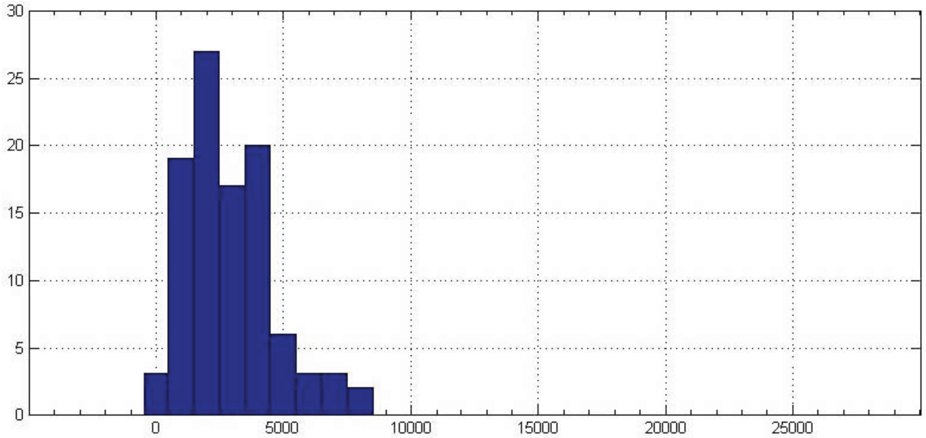
```
Sort(randInt(1,866,5))→L1
```

They then use the list editor mode to see the five ID numbers, and they fill in the corresponding salaries in list L2. The sorting done above just makes it easier to flip through the MLB list to find the five players. If a student or student pair gets a list of five players that includes a duplicate, then they should replace that player or the whole sample.[12] Once again have them repeat this process over and over until the entire class has about 100 sample means. As before, do not mention the number 100 to the class, or put any special importance on the number of samples they are each to collect. Indeed, what I sometimes find is helpful is to monitor the students' samples, and after I sense that there are about 100 sample means in the room, I begin instructing the groups to go to the board individually, regardless of how many they've completed. That way, the class isn't waiting for the slowest group to finish, no one feels pressured, and no one attaches any importance to the number of samples that were collected. This is very important, because if given the chance, students will confuse

[12] Mathematically, it makes little difference whether such a sample is kept or replaced. In fact, the Central Limit Theorem is more correctly demonstrated if you sample *with* replacement, permitting such duplicate players in a sample. But pedagogically, this is hard to explain. I find it better just to go with what we do in actual practice, which is sampling without replacement.

the sample size n with the number of samples collected. It is helpful if the new histogram of sample means is drawn parallel to the population histogram, with the axes matching up, but it isn't necessary if there isn't room for both histograms.

After about 100 sample means are marked as X's on the board, the new histogram might look something like this:



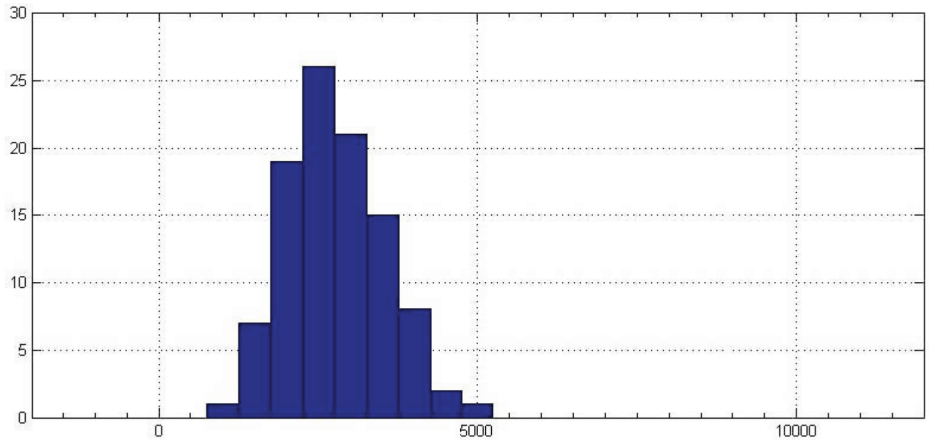
Mean salaries in samples with n=5, in $1000K

We see that the distribution is much less skewed than before, but with $n = 5$, your students will probably still be able to detect some skew. The center of the distribution is about the same as before, but the spread is clearly less. That reflects two facts that students may already know: The mean of sample means equals the population mean[13], and the standard deviation of sample means equals the population standard deviation divided by the square root of the sample size n.

In the final phase of this simulation, students should repeat as before, only this time using samples of size $n = 20$. Duplicate players in a sample of size 20 will be fairly common. When such a sample appears, students should just toss it out (since in practice we sample without replacement) and get a new sample. Alternately, they could just replace the duplicate players in the sample. (This is slightly less convenient, though.)

If students are taking 10 samples each, you might want to reduce that number so as to save some class time. Additionally, you may want this time to have students round up to the nearest $500K instead of $1,000K. There will be fewer bins this time, and the sample means will be much more concentrated around the center of the sampling distribution. The resulting histogram may look something like the one below, which reflects the means from 100 random samples.

[13] Notice that the word "mean" has been used three times in this phrase, each time referring to something different! The mean (over many random samples) of sample means (over the five MLB players' salaries in each sample) equals the population mean (of all 866 players' salaries). This is equivalent to saying that the sample mean is an unbiased estimator of the population mean.



Mean salaries in samples with n=5, in $1000K

Now the distribution is clearly less skewed and less spread out, but its center remains in about the same place, about $2,500K or $3,000K.

It is possible that students will ask you what the true population mean is, and even if they don't, it may be a good idea to tell them. (It is $2,761K.) We can then see that all three of the distributions are indeed centered on about that value.

There are three lessons in this activity. First, for any sample size, the sample mean is an unbiased estimator of the population mean. That is to say, although any particular sample may have a mean that is higher or lower than the actual population mean, over many repeated samples, the *mean of the sample means* will equal the population mean. Students will not generally grasp the meaning of "unbiased" unless they understand completely what a sampling distribution is. Hopefully by the time they do this activity they will have already become comfortable with the concept. If not, there's no time like the present!

The second lesson of this activity is that the distribution of sample means becomes less spread out as the sample size increases. The practical importance of this is that you can estimate a population mean with greater precision if you use a larger sample size.

The third lesson is of course the Central Limit Theorem: The sampling distribution of the sample means becomes more nearly normal as the sample size gets larger, going from 1 to 5 to 20. (Again: The number of simulations that students performed is irrelevant!) Although we've only seen this for one population, the CLT is in fact true for any finite population. In this activity, the population of MLB players' salaries is quite skewed, but the means of samples of size $n = 20$ is approximately normal. The practical importance of this is substantial: With large enough samples, we can know the shape of the distribution of sample means even if we don't know the shape of the distribution of the population. It is the CLT that allows us to perform inference on sample means by invoking properties of the normal distribution.

I have two additional comments on this activity. First, I want to comment that this activity is similar to the popular classroom activity involving sampling pennies and averaging their ages.[14] Both of them are excellent in suggesting to students the behavior of sampling distributions, the nature of the sampling process, and especially the Central Limit Theorem. Both begin with nonnormal populations and result in fairly normal distributions for sample means with $n = 20$ or $n = 25$ or so.

Second comment: This activity is good for demonstrating that the CLT "works" even when the population distribution is skewed. We could have done the activity using baseball players' ages or heights or weights and the same thing would have resulted, but it would be less dramatic since those populations are all pretty normal to begin with. Indeed, such an activity might not convince students of the power of the CLT. However, it should be pointed out that for data this skewed, means are often not the best way to summarize the population. The median baseball player salary would be a better representative of a "typical" salary than the mean, which is very influenced by outliers. In this case, the median salary is $950K, about a fourth the size of the mean of $2,761K. Both of these are, of course, very large salaries by most people's standards. But there is no need to exaggerate the value further. In fact, about 70 percent of MLB players earn salaries lower than the mean.

5. Standardized Mean Heights (The *t*-Distribution Family)[15]

Note: This activity is already published by the College Board on AP Central under "Teaching Resource Materials" in an article titled "Three Calculator Simulation Activities." (http://apcentral.collegeboard.com/apc/members/courses/teachers_corner/49152.html)

This activity introduces students to the *t* distribution family and unfolds in several steps.[16]

First, we will simulate heights of adult American males, assuming the population to be normal with mean 70 inches and standard deviation 2.6 inches, which is pretty accurate.

```
randNorm(70,2.6)
```

Then we simulate three at a time:

```
randNorm(70,2.6,3)
```

[14] For example, see a description in *Activity-Based Statistics* by Scheaffer, Watkins, Witmer, Gnanadesikan, and Erickson. 2nd Edition, Key College Press, 2004. This activity will be discussed in the following article by Corey Andreasen.

[15] Thanks to my teaching colleague Julie Graves for helping me develop this activity.

[16] The syntax throughout this activity is that of the TI-83/84, the calculator models that are probably the most widely used in statistics classrooms. Of course, the activities may be done with any calculator or computer having basic random-number-generating functions. On the TI-8x calculators, the random-number-generating functions are located under the math → prb menu. The notation $X \sim N(\mu,\sigma)$ used in this document indicates that X is a random variable having a normal distribution with mean μ and standard deviation σ.

On the TI you have to scroll to the right after doing this in order to see all three heights in the list. Now it gets a little bit tricky. The colon (same button as the decimal) can be used to separate commands that are entered on a single line. The output you see is the result of the last command. (For example, 1→X:X + 1 would report back "2".) Therefore, use the following command to (1) simulate three men's heights, and then (2) compute the standardized *z*-score for the sample mean, given that the population mean is 70 and the population standard deviation is 2.6. The function mean( ) on the TI-83 is located under the 2nd-list-math menu.

```
    randNorm(70,2.6,3)→L1:(mean(L1)-270)/(2.6/sqrt(3))
```

The reason the commands are separated by a colon rather than entered separately is to allow students to repeat the simulation quickly and easily simply by pressing the ENTER button repeatedly. The ENTER button, when pressed after no new commands are entered, reexecutes the last instruction line.

Have your students press ENTER a few times to get a feel for the sort of numbers it produces. Despite having studied the topic, many students do not immediately see that the numbers produced by this simulation should have a standard normal distribution. It helps to write on the board the same computation in correct notation, with μ and σ, and then substitute in 70 and 2.6:

$$
\begin{aligned}
z &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\
&= \frac{\bar{x} - 70}{2.6/\sqrt{3}}
\end{aligned}
$$

Now ask your students to press ENTER repeatedly, which will repeat the entire simulation many times. Anytime they see a number that is larger in magnitude than 3, they are to say it out loud. You are likely to hear an occasional "3.1" or "−3.3," but numbers much farther from zero will be quite rare. No students are likely to see any numbers greater in magnitude than 4.

The last step is to replace 2.6 in the expression above with stdDev(L1), a quantity that depends upon the "data." This is equivalent to replacing σ of the population with its estimate *s*, computed from a sample. The function stdDev( ) on the TI-83 is located under the 2nd-list-math menu.

```
    randNorm(70,2.6,3)→L1:(mean(L1)-270)/(stdDev(L1)/sqrt(3))
```

Once again, ask your students to say out loud any numbers they see that are greater in magnitude than 3. You will almost certainly hear several 4s, a 6, perhaps even a 12—numbers that would be unheard of from a standard normal distribution. (Numbers greater than 12 are rare even in a *t* distribution with 2 degrees of freedom, but not with 20 students simulating this 50 times or so each.)

Now do this yourself on the overhead calculator several times until you get a pretty large number, say 7 or larger in magnitude, and then stop. Ask your students what they expect to see in list `L1`. What makes the value of the standardized mean so big? There are two possible explanations: the sample mean is pretty far from the true mean of 70 or else the sample standard deviation is pretty small (or, more likely, both). After they've thought about it and perhaps given those answers, look in list `L1`. Very likely, you will find three numbers that are all at least an inch away from 70 inches, and all in the same direction, and they will likely be fairly close to one another, making the sample standard deviation, *s*, relatively small.

The point of this is to convince students that the distribution of the *t* statistic is more spread out than the normal distribution, and the reason is that you're dividing by a random quantity *s* that may vary a lot when the sample size is small. The variability in *s* is what creates the heavy tails in the *t* distribution. The reason the distribution begins to look more normal when the sample size gets larger is that the variability in *s* decreases.

If you repeat this activity with larger and larger sample sizes, you should have fewer and fewer students saying large numbers out loud.

6. Baseball Players' Height/Weight Relationship (Regression Line Slopes)

Students often fail to understand that when they construct a regression line on bivariate data, the slope of the regression line is in fact a sample statistic—and that it therefore has a sampling distribution. The reason of course is that the sample of data was only one sample that happened to be observed, and had the sample been different, the regression line would have been different too. This is true both when measuring bivariate data on a random sample of a population (as is done in this activity), as well as when measuring the response variable in a controlled experiment in which the explanatory variable is determined by experimental design.

I am sympathetic with the students' difficulty, and I have seen it every year I have taught AP Statistics. Students who are comfortable with the idea of a sampling distribution of sample means are less comfortable with the idea of a sampling distribution of regression line slopes. This activity is meant to make the latter more accessible.

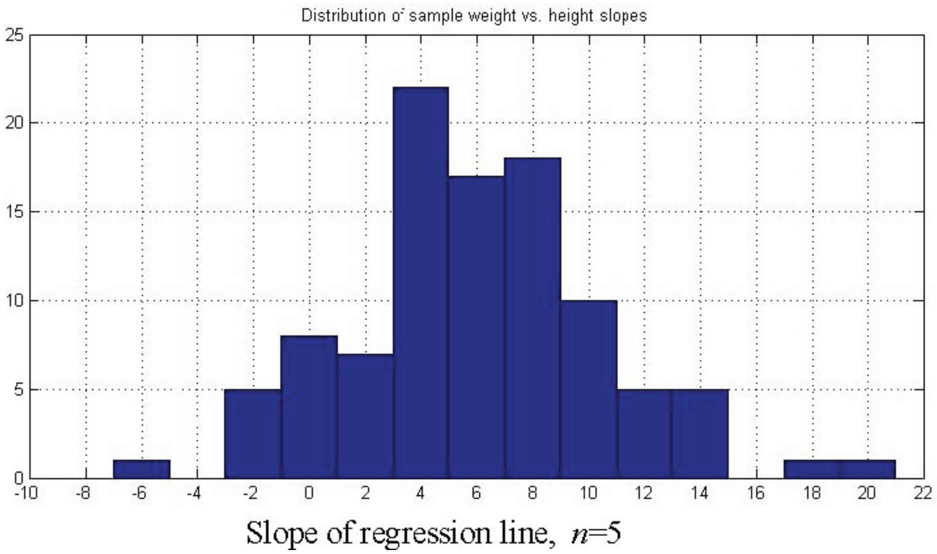
An earlier activity involved sampling baseball players from a list of all MLB players (a complete population) and recording their salaries. This activity is very similar, and so it will be described in slightly less detail. The point is for students to recognize that the slope of a regression line depends on what random sample was observed, and so it has a distribution. Furthermore, the distribution of regression line slopes is (under certain circumstances) approximately normally distributed.

First have your students sample five MLB players at random from the list:

```
SortA(randInt(1,866,5))→L1
```

As before, if a player is duplicated in their sample, they should get a new player or a new random sample.

Next, they record in lists L2 and L3 the heights and weights, respectively, of the players. Then they plot those heights and weights on their calculator and compute the regression line, where weight is the response variable and height is the explanatory variable. Then they note the slope of the regression line, writing it down somewhere. Finally, they repeat this several times each (as before, it would be nice to have about 100 sample slopes among all your students to get a nice-looking histogram). As the students finish, they come to the board and put their sample slopes on a class histogram, as X's. For this activity, it makes a nice histogram if they round their slopes off to the nearest even integer. Below is a typical histogram of 100 simulated samples' slopes.



It is particularly instructive if any students happen to get negative slopes. (This is quite likely to happen if you do 100 simulations.) Do students' faces light up or show puzzlement as it dawns on them what is strange about this? If not, that is a red flag for you: The student may not appreciate the strangeness of the negative numbers in this distribution. The explanation for the negative numbers is that with samples of just $n = 5$ MLB players, it is possible to get the occasional sample of five such that the heights and weights have a *negative* correlation. (This is much less likely with larger samples.) You might like to have a student who gets a negative slope stop and share a scatterplot with the whole class of that particular sample's heights and weights.

The main point of this activity is simply to show that slopes *have* a sampling distribution. You could stop the activity right here. But if you want to carry it further, you could repeat the activity with samples of size $n = 12$, and observe that the distribution of slopes is still centered on about 6 lbs/inch but is narrower. You are unlikely to see any negative slopes among 100 samples of size 12.

You also may remark that the distribution of slopes is approximately normal. This happens here because the heights and weights together have a bivariate distribution that is approximately "bivariate normal." This topic is beyond the AP syllabus, but among the consequences of bivariate normality are the three main assumptions we do require students to know are necessary for inference: (1) a linear relationship between the conditional mean of Y given X, and X itself, (2) normally distributed Y values around those means, and (3) constant standard deviation around those means conditioned on any particular X. And incidentally, it is the case that if those conditions are met, then the sample regression line slope is an unbiased estimator of the population regression line slope. In the case of the baseball players' heights and weights, the population regression line has a slope of 5.5 lbs./inch.

When we look at output from a computer that has performed regression based on a sample, we usually get, in addition to the slope of the regression line, an estimate of its standard error. It is probably not best to discuss this with your students until they are more comfortable with this topic, but a brief discussion might proceed as follows: That estimate of standard error is *not* equal to the standard deviation that you see in the histogram of slopes. The histogram of slopes that your class produced has a standard deviation that is unknowable when all you have is a single sample of data. You can *estimate* that spread using the data in your sample, and that's what the computer output gives you. But it is only an estimate, and it is variable—just as with univariate data, you can only estimate σ with *s* from a particular random sample. And just as inference in the context of univariate data required the *t*-distribution family, so inference in the context of bivariate data requires the *t*-distribution family, and for exactly the same reason: The standard error of the slope is estimated, not known.

7. Worm Species (The Goodness-of-Fit Test)

It is assumed at the beginning of this lesson that students are already familiar with the structure of hypothesis tests. (It is not actually required for the lesson, but if the students are familiar with hypothesis tests, then the class may discuss results with terms and phrases such as "p-value" and "reject the null hypothesis.")

Begin by preparing a bag containing four colors of gummy worms: 15 yellow worms, 35 blue worms, 15 green worms, and 35 red worms. Then tell your students—without showing them what is in the bag—that you claim the bag contains 35 percent yellow worms, 15 percent blue worms, 35 percent green worms, and 15 percent red worms. Write that claim on the board.

Next, tell your students that you are going to draw 10 worms at random from the bag. Do so or have a student do so, and write the results on the board. Hopefully, there will be between the claim and the data a discrepancy that is dispersed over all four colors. One goal of this activity is for students to understand that sometimes observed data may not differ from a claimed distribution in any single category sufficiently to convincingly reject the claim; but that over many categories, the cumulative discrepancies may yet convince.

Now ask the students whether they believe your claim. Some will say they do not. Ask them why. They will likely respond that the data don't "fit" the claimed distribution, or that the data "deviate too much" from the claimed distribution, or with some similar statement. You should then ask, "Oh? By how much do these data deviate from the claim?" They will have to think for a moment, and they may come up with different answers. This is important, for you want them to be thinking about how to quantify the discrepancy between a claimed distribution and an observed distribution within a sample. It is not quite as straightforward as quantifying the discrepancy between a claimed proportion and a sample proportion, for then the most natural and obvious measure is the difference between the two (or, with more sophistication, the standardized difference between the two). But now there are four proportions to reckon with.

Allow the students to discuss freely their different measures of discrepancy. Do not suggest to them the chi-square statistic, and if any student knows of it already, suggest that student remain silent so as not to stifle the creative and constructive discourse of the others. (For example, you might draw a χ^2 on the board and ask any students who have seen this to "keep the secret.") Often students will propose the sum of the absolute values of the differences in the proportions. Or, having studied variances, students may suggest the sum of the squares of the differences in the proportions. Or they may suggest the average of the absolute deviations. Occasionally they come up with something a little more exotic, but they will almost certainly not come up with the chi-square statistic, and that's fine. That isn't the point of this activity.

Have the class discuss the different proposed discrepancy measures and agree upon one of them to continue with. It doesn't really matter what they pick, so long as it is a quantity that is larger when the distributions have a larger discrepancy and smaller when they have a smaller discrepancy. You might write on the board, "D =" and their measure, either in words or in symbols.

Next, have the students calculate the value of the discrepancy measure for the claim and the data with which you began class. For example, if you observed 1 yellow, 4 blues, 2 greens, and 3 reds, the measure of discrepancy might be:

D = sum of absolute values of differences between claimed proportions and actual proportions

$$D = |0.35 - 0.1| + |0.15 - 0.4| + |0.35 - 0.2| + |0.15 - 0.3| = .25 + .25 + .15 + .15 = 0.8$$

Next ask your students, “You said earlier that you didn’t think my claim was true, and the reason you gave was that the discrepancy between the claimed distribution and the distribution of the data was ‘too large.’ Now we’ve measured the discrepancy and found it to be 0.8, according to your own measure of discrepancy. How do we know whether this is ‘too large’ or not? Just how large is ‘too large,’ anyway?”

If your students are accustomed to doing simulations, they may suggest one at this point. If not, you will have to suggest it to them. Either way, you should allow them to figure out how to do the simulation themselves. This isn’t hard. They need to be provided with bags and tokens (more gummy worms, or some other token, like plastic beads or kindergarten “counters”[17]) and allowed to put tokens in the bag according to what they think the simulation requires. Before beginning the execution, however, you should verify that they have it right: The bags need to contain tokens according to the *claimed* distribution of colors. Additionally, they need to have enough tokens in their bag that their sample doesn’t represent a sizeable chunk (say, more than 10 percent) of it. One hundred tokens (as you used yourself) will do for samples of size 10. To execute the simulation, they then need to do as you did at the start of class, drawing 10 tokens from the bag and calculating the value of D (the “discrepancy score”) for the sample. Allow each student to do several simulations while you draw a number line on the board with tick marks for every tenth. (Or at other appropriate locations, depending on what discrepancy score your students came up with. It might be worth having a very brief discussion with the students about the range of possible values of their discrepancy score.)

As the students complete their simulations, they should come to the board and mark x’s over the scores they obtained, constructing a histogram of the distribution of their discrepancy score under the claimed distribution. It should then be evident just how “unusual” the actual observed value of 0.8 (or whatever) was, and a rejection of the claim can be more rationally justified. A p-value can also be estimated based on the proportion of simulations that resulted in a discrepancy score greater than the observed one.

For homework, students should then go home and read in their texts about the chi-square statistic and come to class prepared to teach it or discuss it. I like to pick a random student to teach a chi-square goodness-of-fit lesson for 5 or 10 minutes. Essentially, I want them to say that the process is exactly what they did in class, only with a different measure of discrepancy. The extent to which you discuss the rationale behind the chi-square statistic

[17] I prefer to use different objects for the real data (the 10 sampled worms at the beginning of class) and the data sets that they simulate through multiple draws from a known population they constructed. By making them different you emphasize the fact that what they are doing is a simulation, not collecting additional real data.

with your class is then up to you. This simulation activity gets them in the right frame of mind to understand the concept behind it.

You will note that although this activity is meant to introduce the chi-square goodness-of-fit test, a different and unconventional statistic is actually used for the activity. This is an example of a *constructivist* lesson. Although you guide the activity and can be pretty confident how it will turn out, students contribute the crucial elements themselves. The chi-square statistic is not a natural one, and indeed it would be confusing for many students, for not only is it strange in its construction, it isn't even obvious that it is a measure of discrepancy between the data and the claimed distribution at all. More to the point: this activity really isn't meant to teach about the chi-square statistic. It is meant to teach students about the goodness-of-fit test and about the sampling distribution of a statistic that measures the discrepancy between two different categorical distributions. The actual computation of the chi-square statistic is a burden and should be left to calculators and computers.

Some teachers worry that students will take away the wrong lesson from this activity; that they'll remember the statistic D and think they need to know it for their next test. I think exactly the opposite is true: if students are taught the chi-square statistic, they'll think its computation is the important part of the lesson. The important thing to focus on is *the sampling distribution of a statistic that measures a discrepancy between a claimed categorical distribution and some data*. If students are told explicitly that their formula for D was good for the activity because of its simplicity but now needs to be replaced with the formula that scientists really use (i.e., the chi-square statistic), they suffer little confusion.[18]

Conclusion

Sampling distributions are less accessible than distributions of data because they involve summary statistics rather than direct measurements, and in practice they invoke samples other than the one that was actually observed. But helping students to understand sampling distributions need not be difficult. Following these recommendations will help:

- Look at sampling distributions early and often. Have students make boxplots and histograms of sampling distributions early in the year, before you even name them, and continue looking at them occasionally throughout the rest of the year. Even after students have mastered the concept of a sampling distribution, they may need reinforcement with more activities to grasp that the chi-square statistic or the slope of a regression line has a sampling distribution.
- Use hands-on simulations. Students will grasp sampling distributions better if they get to draw samples themselves, rather than just looking at pictures of sampling distributions in a text or on a computer screen. Ideally, you would like them to progress

[18] One might very reasonably ask why anyone *does* use the chi-square statistic? What's wrong with the D formula that we presented here in this activity? There are two reasons. First of all, the chi-square statistic has a distribution that is not very dependent upon the actual distribution of the categories in the population, only upon the number of categories there are. "D" here does not have that property. Second, the chi-square statistic results in a more powerful test—power in the technical sense, i.e., more likely to reject a false null. The "D" statistic here gives too much weight to relatively infrequent categories, making it too easy for other distributions to "masquerade" as the one in the null if all you see is D.

toward being able to simulate a sampling distribution using applets or computer software such as Fathom. Then many more simulations can be done, and can be done faster. But it's better to start with something more transparent, in which each sample is simulated, and its corresponding statistic computed, by itself. (The next article in this collection addresses the use of technology in teaching sampling distributions.)

- Have your students construct class histograms, in which you draw an axis on the board and they, individually or in pairs or small groups, contribute "Xs" stacked on top of one another. Don't worry about the Xs being perfect or all exactly the same size, and don't worry about a vertical scale. You're mainly after the fact that samples produce a sampling distribution. You can discern its shape, center, and spread without a vertical scale.
- Don't let your students get hung up on how many simulations they are performing. This number is not important. (That's another reason not to use a vertical scale on your board histograms.) The only thing you need to do is make sure there are enough simulated values to make a clear histogram. Usually 100 is about right.
- Don't worry that these activities are taking too much class time. Of course you should have a plan for the year, but the inclusion of these activities in that plan will not only help students understand the concepts better than lectures, they will actually save you time as well, because they will find future topics involving sampling distributions so much more straightforward.
- If you are not careful, some students will draw an incorrect conclusion from many classroom simulation activities: they will think that taking many different random samples is what is done in actual practice. You should be on the alert for signs that students are thinking this way. Fortunately, it is not hard to lead them away from that notion. Just a clear, straightforward statement may do the trick: "Remember that what we did was only a *simulation*. We wanted to see what sorts of sample statistics *might* have resulted from different random samples. In practice, how many samples do you get?" (Hopefully, students respond with confidence, "One!") "Right. One. We are playing the 'what if' game to see what sorts of results are 'typical' by taking lots of different 'pretend' samples."

One of the delights of teaching AP Statistics is that it is so hands-on. I hope this set of activities has given you ideas about new things you can do with your students that are both engaging and enlightening.

Sampling Distributions: The What-Ifs with Technology

Corey Andreasen
Sheboygan North High School
Sheboygan, Wisconsin

Two pedagogical reasons one might have for having students generate simulated sampling distributions in class are:

- So that students can better understand what the individual values in a sampling distribution represent. Because sampling distributions are the basis for inference procedures, understanding them is imperative. Simulated sampling distributions generated by the students themselves can be seen to agree with theoretical results, helping students trust and have a greater understanding of those theoretical results.
- In situations where a theoretical sampling distribution is not known to the students, the simulated sampling distribution can be used for inference directly.

While it is crucial that students experience concrete simulations of sampling distributions, given the usual class time constraints, it is difficult to gather many values of a sample statistic if one is limited to rolling dice and drawing colored counters out of bags. Doing simulations by hand provides students with an irreplaceable understanding of what a sampling distribution is, but for a sampling distribution to be useful in teaching practice, and to get precise information from simulated sampling distributions, many samples are needed—perhaps thousands or even tens of thousands. Computers and calculators provide a means for gathering many samples, calculating the sample statistics quickly and efficiently, and displaying the results graphically.

Below, activities will demonstrate ways to use technology to provide a deeper understanding of sampling distributions and their importance in statistics. While these activities use computer technology to generate many samples and their statistics quickly, it is instructive to begin by having students use the concrete materials as described in the previous article. Then, before jumping to a large number of samples, the technology can be used to reproduce what the students did by hand. In doing so, students can see that the results from the technology do, indeed, match the results they got by hand. This helps students trust that the technology is mimicking the same process. If not used carefully, this same technology can obscure processes, presenting an appearance much like a mysterious "magic black box" to many students, not unlike the look and feel of mathematical theory. It is also instructive, when possible and appropriate, to have students program the simulations themselves to make the entire process more transparent.

Using Sampling Distributions to Detect Evidence of Discrimination

Because hypothesis testing requires a fair amount of background knowledge, it is usually among the later topics covered in the AP Statistics course. But a hypothesis test is really designed to ask a fairly simple question: If we were to take as many samples as we wished, assuming the null hypothesis is true, about what proportion of the sample statistics would be at least as extreme as the one we observed? To illustrate this idea, we will use a legal case as an example.

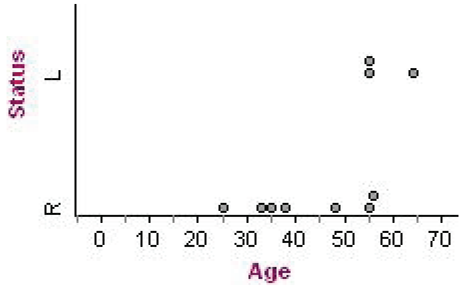
Westvaco Corporation, a producer of paper products, laid off a large part of the workforce in its envelope division in 1991. Robert Martin, a 55-year-old engineer who was one of those laid off, brought suit against the Westvaco Corporation for age discrimination (*Martin v. Envelope Division of Westvaco Corp*., CA No. 92-03121-MAP, 850 Fed. Supp. 83 (1994)). In most discrimination cases, statistical analysis can shed some light, and this case is no exception.

The original Westvaco data are:

Ages in Years of Laid off and Retained Westvaco Employees

Laid off	55	55	64				
Retained	25	33	35	38	48	55	56

The layoffs at Westvaco occurred over five rounds among both hourly and salaried employees. In this paper, we will look at only one round of layoffs: the second round of layoffs of the hourly workers. The result of this round is shown in the comparative dot plot in Figure 1. The mean age of the three workers laid off in the second round is 58 years, while the mean age of those retained is 41.4 years. Does this disparity support a claim of age discrimination?



At first students will simply want to compare the mean age of those laid off to the mean age of those retained. This is a reasonable place to start, but it does not answer the question of whether there is age discrimination—there is no standard with which to compare that difference in mean ages! After the mean difference in ages is calculated, we still don't know whether the difference is large enough to accuse Westvaco of discrimination.

The essential question is "How likely is a difference in mean ages this large (or larger) if the layoffs were made without regard to age?" A layoff of randomly selected workers could be appropriately modeled by randomly selecting ages of the workers. Of course, Westvaco did not randomly select the workers—possibly they used a nondiscriminatory scheme of some sort. Workers were selected for layoffs based on some criteria, and that criteria may be associated with age. For example, workers with obsolete skills may be laid off first, and these might be the older workers in general. In that case we would expect to see a higher mean age for those laid off. It is also possible that a difference in means this large or larger *could* reasonably be attributed to chance; Robert Martin, the plaintiff, has a very weak case at best. On the other hand, if this difference is too large to be attributed to chance, Westvaco will be asked to explain their actions.

How might we use simulation to help determine whether the results of the layoff can be reasonably attributed to chance? One approach would be to perform a simulation to analyze the results of chance layoffs. We could randomly select from the population three workers to be laid off, calculate the mean age from that sample, and compare that simulated mean to the mean of those actually laid off (58 years). Then we repeat this process many times and create a distribution of the sample means.

Evidence of discrimination would require a comparison between the mean age of those laid of to the mean age of those retained. However, because we know the age of all 10 people involved in this round of layoffs, we could calculate both means from the one sample. Rather than work this entire calculation into the simulation, we simplify it by considering only the mean age of those selected for layoffs.

We begin by having students design this simulation on their own and do a few trials. There are many different presentations of simulations, with a variety of steps and sequences in the literature. We will use a simple four-step simulation process: Assumptions, Model, Repetition, and Conclusion. At each step in the simulation, students will fill in the specifics of the problem at hand. Now let's use this model to tackle the Westvaco simulation!

Assumptions: The ages of the three workers are selected randomly and independently.

The next phase is the construction of a "model," or representation of a random selection process. The model that follows is one example. The randomization could be done in one of several different ways, including a random digit table, a random number generator on a calculator, rolling a 10-sided die, etc.

Model:

In this phase, we construct a model, or representation of a random selection process that mirrors our assumptions and the "structure" of the problem. The necessary randomizing could be done in one of several different ways, including a random digit table, a random number generator on a calculator, rolling a 10–sided die, etc.

1) Write the ages of the 10 workers {25, 33, 35, 38, 48, 55, 55, 55, 56, 64} on identical index cards.

2) Shuffle the cards and select three of them without replacement.
3) Record the mean of the ages on the three cards.

Note that the process above completely describes one *run* of the simulation, including the calculation of the sample statistic being recorded. The mention of whether or not replacement is allowed is important whenever sampling is done, and thus it is important that this be specified in every simulation model involving sampling.

Repetition: Repeat this procedure a large number (say 25) times. This set of 25 runs constitutes the simulation. Record the results and analyze the resulting data from the repetitions. With sample sizes of 25, a dotplot is particularly effective.

Conclusion: Analyze the data and determine the proportion of sample means that are 58 or over and decide whether this indicates that the three ages chosen for layoffs are unusual enough to ask Westvaco to explain itself.

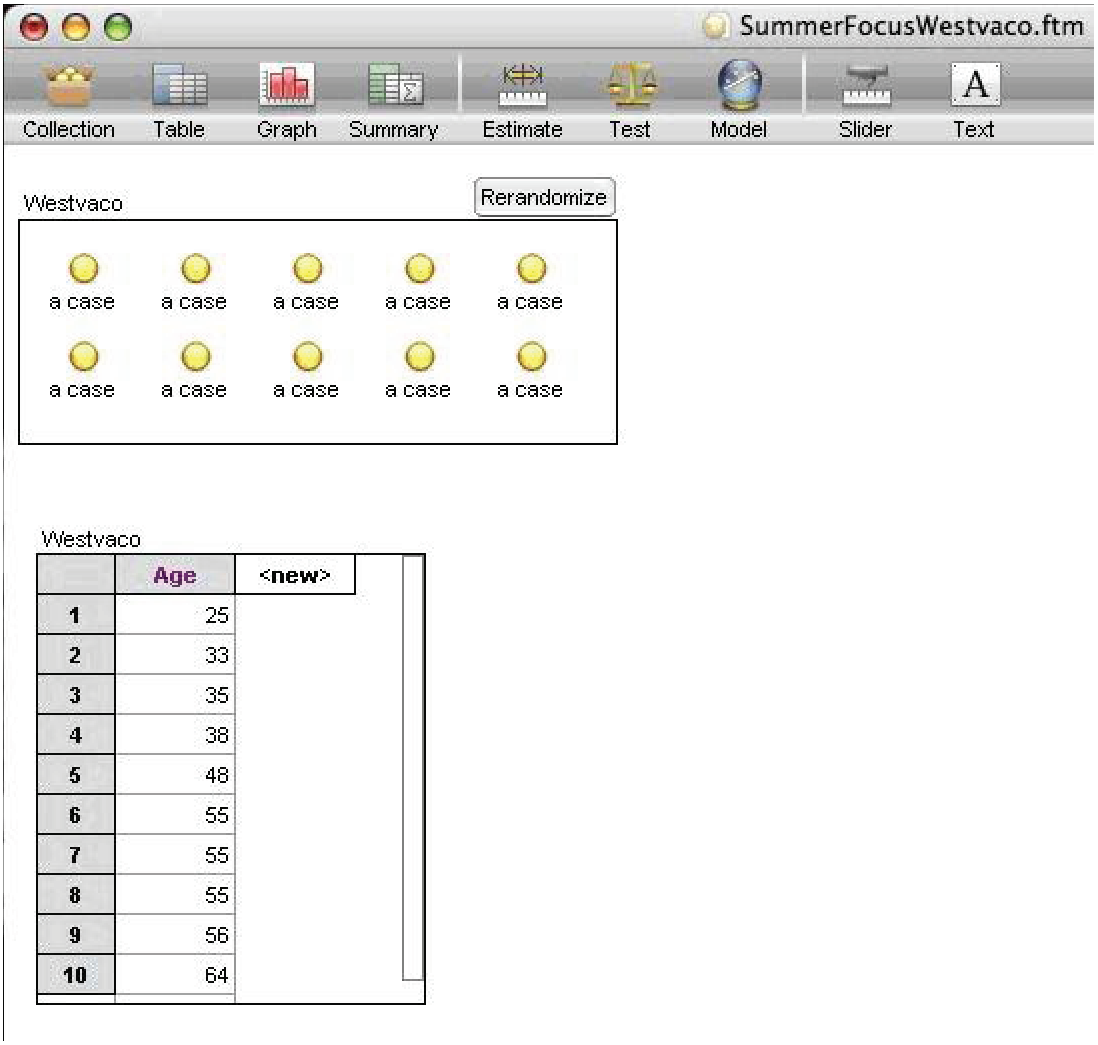
Having students articulate the four steps helps them to see what is essential to that simulation, since as many trials of this are required for students to understand what a plot of the sampling distribution communicates about the "right" number of replications. One possible strategy is to have the students do simulations in small groups until each group has accumulated some number of means, then have the group come to the board and contribute their sample means to a single class histogram that you collectively build on the board. Then you can lead the class in a brief discussion with such questions as "What's a typical number here? What does that mean? Does it appear that 58 is a typical result?" (Your focus here should be on their development and communication of what they believe a "typical" age might be if the layoff process was, in fact, unrelated to age.) After you are confident that they understand what the sampling distribution shows, then lead them to the recognition that a simulation involving hundreds or thousands of random samples would give a much clearer picture of the sampling distribution of the mean for this population.

Because effective simulation requires a large number of samples, we recommend the use of statistical software. Programs such as Minitab, DataDesk, and JMP will perform simulations effectively. For purposes of this article, we will use the program Fathom, by Key Curriculum Press (www.keypress.com). Fathom was designed from scratch to be a teaching tool, and it forces the simulation designer to be very careful and explicit in specifying the steps of a simulation. Some examples of simulations using other technologies will be provided below. Our discussion will provide a generic description of what we are doing and why in addition to presenting specific Fathom directions so that the reader can follow AND participate. (A free trial version of Fathom is available at their Web site.) Our first activity will explain in detail how to set up and run each step of the simulation using Fathom.

One last bit of advice before proceeding: The computer screen can get rather cluttered when setting up a simulation. Some forethought into how objects will be placed on the screen can help students gain insight into the structure of the simulation. But a poor or cluttered screen can make it difficult for students to decipher.

When we change to using technology for simulation, we can see some changes in the design: The random selection will be done on the computer rather than using slips of paper, and we will do many more than 25 trials.

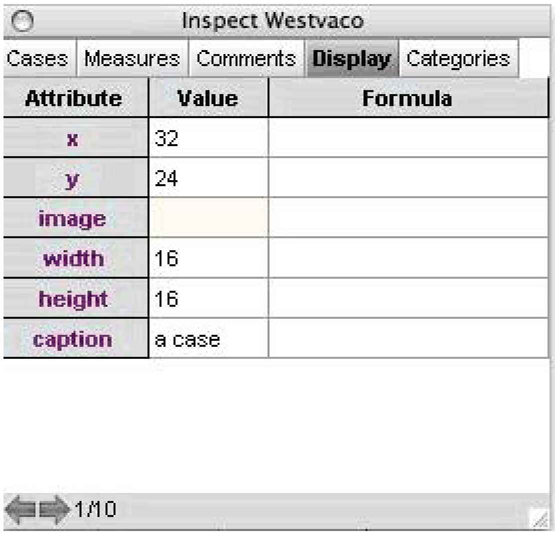
In Fathom there is a row of icons above the document workspace called the "tool shelf." You begin setting up the model by dragging a new case table from the tool shelf to the document workspace. While not actually necessary, you may wish to drag the windows to the positions shown in the diagrams for the sake of clarity in following our presentation. Type the attribute name "Age" where it says **<New>** in the table and hit **Return**, then enter the 10 ages into the table. A gold box called the **Collection** will appear. To help students understand the upcoming steps in the Fathom execution of the simulation, drag the lower right corner of the box as shown in Figure 2 so the icons for each worker can be seen.



Westvaco

	Age	<new>
1	25	
2	33	
3	35	
4	38	
5	48	
6	55	
7	55	
8	55	
9	56	
10	64	

The next steps will change the caption under the icons to show the workers' age rather than the generic "a case" that appears as a default. This is not needed for the simulation to work but is less abstract for students and thus pedagogically helpful when the sample is selected. Double-click inside the frame of the collection to get the **inspector** for the collection. Select the **Display** tab as shown in Figure 3.



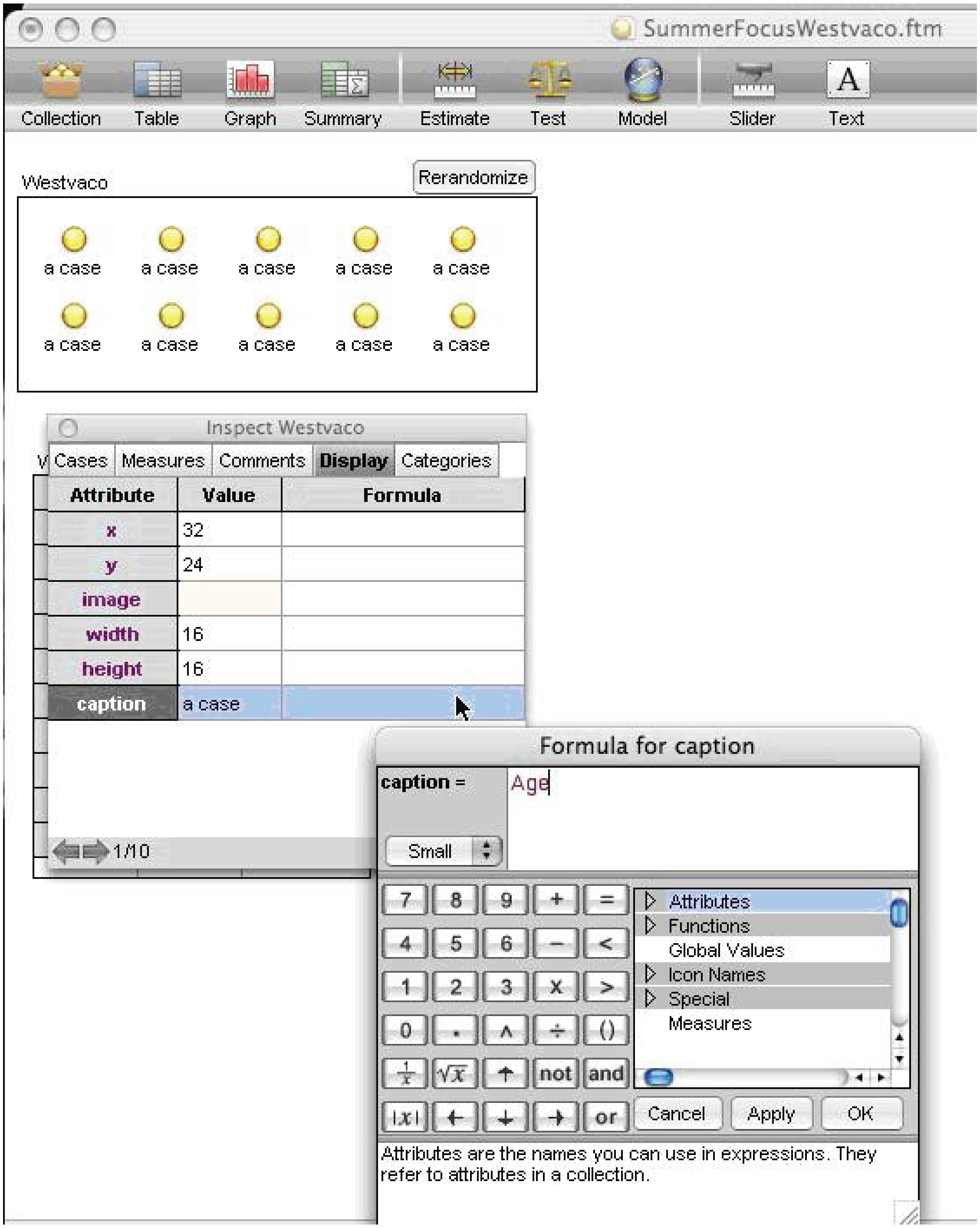
Attribute	Value	Formula
x	32	
y	24	
image		
width	16	
height	16	
caption	a case	

To the right of **Caption**, under **Formula**, double-click to open up the formula editor, and type "Age" as in Figure 4 on page 43. Then click OK and close the inspector.

In the next few steps, we will explain how to tell Fathom to select a sample of three workers without replacement. Click once on the collection so it is outlined in a blue frame, and from the **Collection** menu, select **Sample Cases**. A second box appears, this one containing blue spheres and labeled **Sample of Collection 1**. Drag this box open as you did for the original collection. You will see 10 icons of workers sampled from the collection. To change this to a sample of three workers, double-click in the frame of this sample collection to open its inspector as in Figure 5 on page 44.

Click the check box to turn off **With Replacement** and change the number of cases to three as shown. Then click **Sample More Cases**. You should see three icons representing three workers in the sample collection frame. Click **Sample More Cases** again, and you will see another sample of three workers.

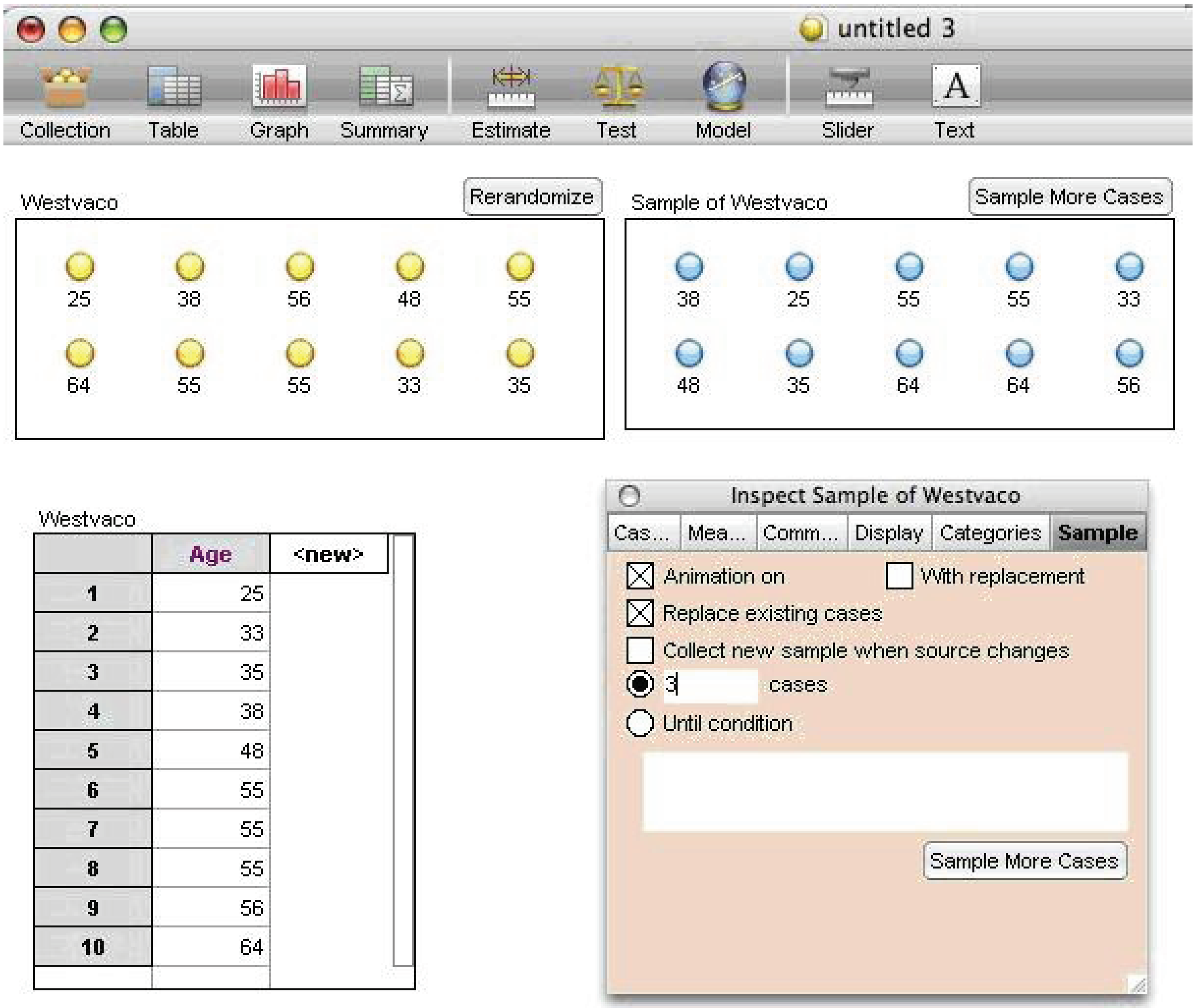
We will now tell Fathom which statistic to calculate—in this case, the sample mean. Students could calculate the mean age of the sample manually since the three ages are displayed, and it is worthwhile to have them do this for one sample. It encourages them to trust that Fathom does the same thing they did with index cards. To have the software calculate the mean, you will define a **Measure**. Double-click within the frame of the sample collection to open the



Attribute	Value	Formula
x	32	
y	24	
image		
width	16	
height	16	
caption	a case	

caption = Age

Attributes are the names you can use in expressions. They refer to attributes in a collection.



Westvaco

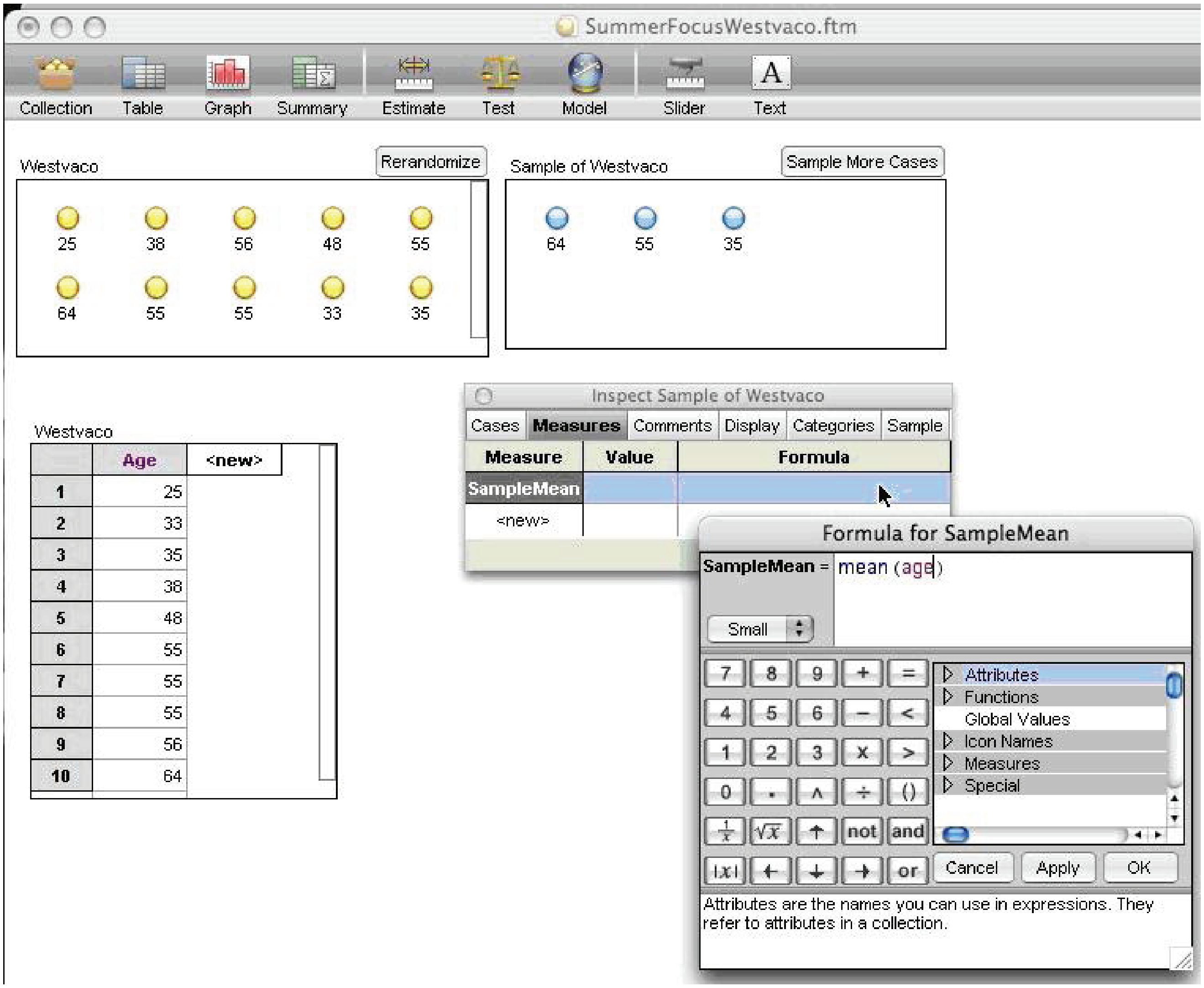
	Age	<new>
1	25	
2	33	
3	35	
4	38	
5	48	
6	55	
7	55	
8	55	
9	56	
10	64	

☒ Animation on □ With replacement
☒ Replace existing cases
□ Collect new sample when source changes
● 3 cases
○ Until condition

inspector, then select the **Measures** tab. Name the measure **SampleMean** and double-click in the formula box to open the formula editor. Type "mean(age)" into the editor as in Figure 6 on page 45 and click OK.

The sample mean displayed in the inspector should match the hand calculation.

Now we have defined the model and moved to the Repetition phase of the simulation. Close the inspector (the window where you just entered the measure **SampleMean**) and click once on the box of the **Sample of Westvaco** collection. Under the **Collection** menu, select **Collect Measures**. Now a third box will appear, a collection of the means of the samples. Double-clicking on **Measures of Samples of Collection 1** box will open the inspector, where you can determine the number of samples you want to collect. It will be instructive to begin with 1 sample mean. Change the number of measures from 5 to 1 and select **Replace Existing Cases**. Click **Collect More Measures**. Notice that the animation shows a blue sphere bouncing from **Collection 1** to the **Sample of Collection 1** showing the sample



Westvaco

	Age	<new>
1	25	
2	33	
3	35	
4	38	
5	48	
6	55	
7	55	
8	55	
9	56	
10	64	

Measure	Value	Formula
SampleMean		
<new>		

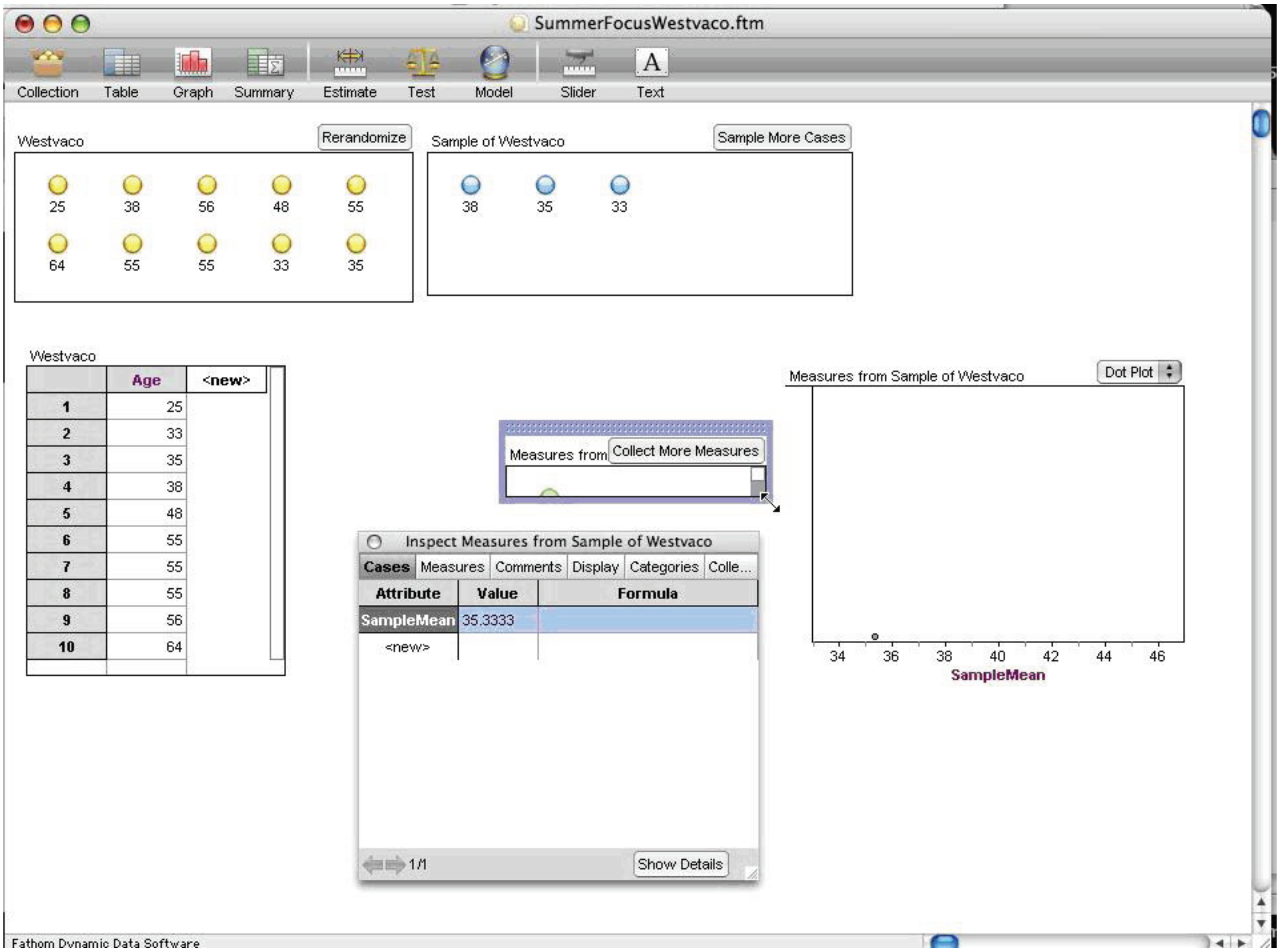
SampleMean = mean (age)

Attributes are the names you can use in expressions. They refer to attributes in a collection.

being taken. A green sphere then bounces to the **Measures from Sample of Collection 1** showing the mean of the sample being recorded. Placing the Measures box below the sample collection forces the spheres to go in different directions, making clear the two steps: *collect the sample* and *record the mean*.

To plot the sample mean on a dotplot, drag a new graph from the tool shelf to the document. In the inspector for the **Measures from Sample of Collection 1**, select the **Cases** tab and drag the attribute **SampleMean** to the horizontal axis of the graph. (Drag the word "SampleMean," not the number here.) Select the **Measures** box and drag the lower right corner just enough to display the "Collect More Measures" button. The screen should now be arranged as shown in Figure 7 on page 46.

Now, every time you click "Collect More Measures," you will see a new sample being collected, a mean being recorded, and the mean being plotted on the graph. Note that the



Westvaco

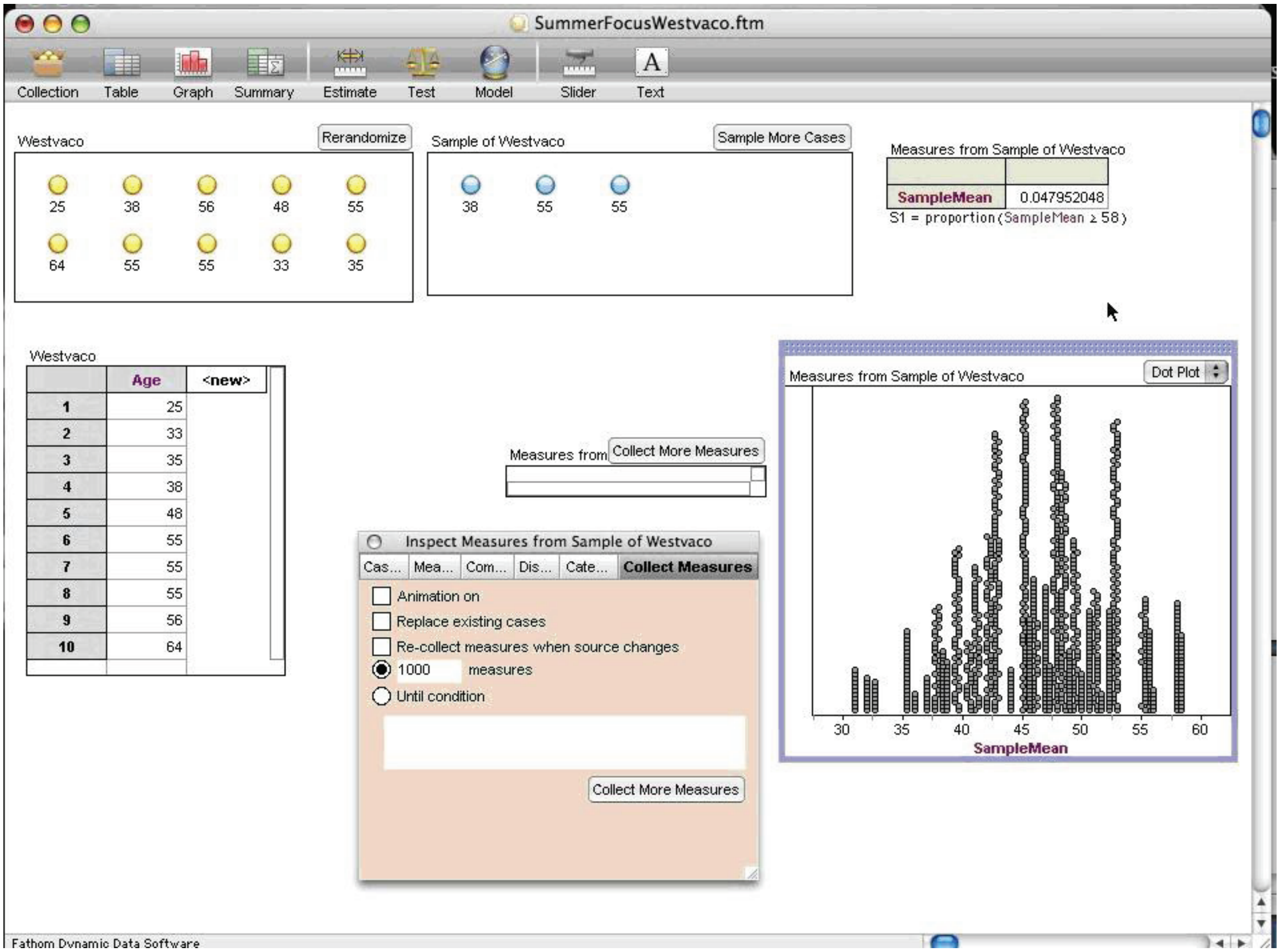
	Age	<new>
1	25	
2	33	
3	35	
4	38	
5	48	
6	55	
7	55	
8	55	
9	56	
10	64	

Attribute	Value	Formula
SampleMean	35.3333	
<new>		

mean shown in the inspector matches the position of the plotted point. Do this (or have students do this) a few times. Then, in the inspector, click on the Collect Measures tab and deselect “Replace existing cases.” Then go back to the Cases tab. Collect More Measures and see that the new means are now added to the old plot.

One benefit of having both steps in this process visible is that there is a clear distinction between the sample size (3) and the number of samples, which keeps going up as you Collect More Measures. Because this distinction is often unclear to students this merits a bit of classroom discussion. In the window for the sample, you see three blue spheres. That is the sample size. Each dot in the plot represents a different sample.

Now it’s time to take a bunch of samples quickly. In the inspector, click on the Collect Measures tab and change the number of measures to 1,000. Turn off the animation to speed things up and click Collect More Measures. (Students seem to enjoy setting the number of samples to 10,000 and watching the animation—for about ten seconds. It quickly gets pretty boring to watch—trust me!)



Measures from Sample of Westvaco

SampleMean	0.047952048

S1 = proportion(SampleMean ≥ 58)

Westvaco

	Age	<new>
1	25	
2	33	
3	35	
4	38	
5	48	
6	55	
7	55	
8	55	
9	56	
10	64	

Inspect Measures from Sample of Westvaco

Cas... | Mea... | Com... | Dis... | Cate... | **Collect Measures**

□ Animation on
□ Replace existing cases
□ Re-collect measures when source changes
● 1000 measures
○ Until condition

Collect More Measures

Finally we arrive at the *Conclusion* step, where we determine the typical proportion of sample means that are at least 58. Drag a **Summary Table** from the shelf to the document. In the **Measures** inspector, click the cases tab and drag the attribute name **SampleMean** to the row title cell in the summary table. The mean is automatically calculated, but you're interested in the proportion of sample means that are 58 or greater, not the mean. Double-click on the measure S1 = mean( ) and change the formula to **proportion(SampleMean ≥ 58)** (hold down **option** as you type > to get ≥). In the sample shown in Figure 8, 0.048, or about 4.8 percent, of the sample means were 58 or more.

After this long and detailed procedure it is important to review with students what has just been accomplished and what they are looking at. Taking many samples is playing that game of "What If?" In reality—that is, in the context of the case—we have only one sample to look at. As mentioned earlier, the actual "real life" sample was not selected randomly because people were selected for layoffs based on some criteria that may or may not be independent of age. But what if we *could* repeatedly select samples? Would an average age of 58 be reasonably likely if the selection process was independent of age? If the sample statistic would not be at all unusual with random sampling, Westvaco is off the hook. On the other hand, if this sample would have an unusually high average age compared with other random samples, we might reasonably ask Westvaco to explain what layoff selection procedure they used and let the court determine whether it was lawful.

According to the simulated sampling distribution shown here, there is a little less than a 5 percent chance that a result at least as extreme as this one would result if age was unrelated to the criteria for layoff. Such an extreme value would not usually be thought of as a common result and seems to provide more evidence for the plaintiff than for the defendant.

An activity like this can be done early in the AP Statistics course. An early look at sampling distributions allows students to explore the logical principles that underpin statistical inference and start learning about the big questions of statistics before the approximation techniques that depend on the Central Limit Theorem are introduced.

The German Tank Problem with Technology

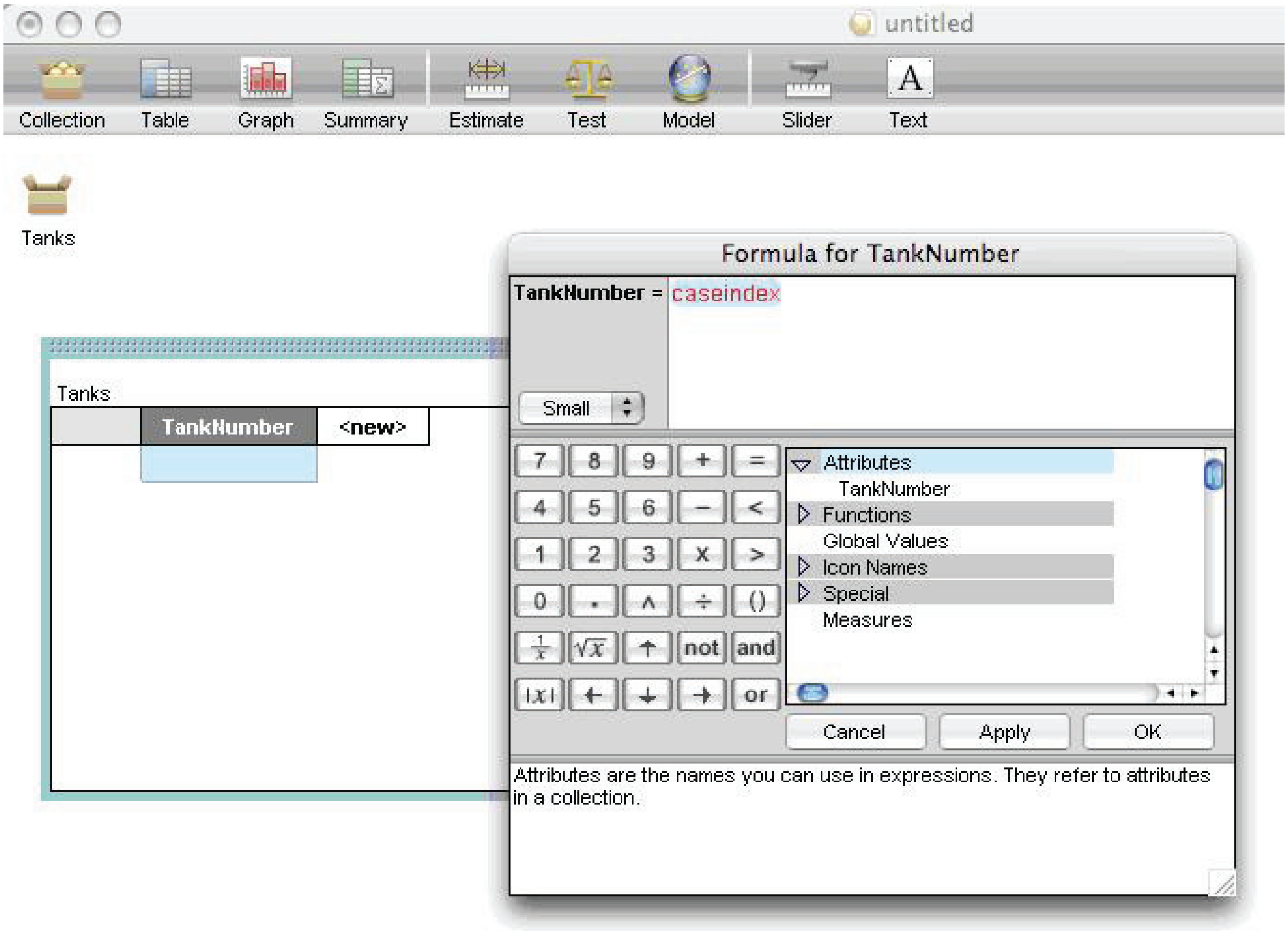
In the previous article we introduced the German Tank Problem. Here we will once again use Fathom, focusing intently on how students can use computer technology to define their own statistics and simulate sampling distributions of those statistics.

Earlier in this section we alluded to the problem of the “mysterious black box.” If students do not understand what is happening “behind the scenes” in a computer simulation, then they may follow your instructions correctly, click all the right buttons, and see exactly the sampling distribution they’re supposed to see and yet still not understand what it represents. A nearly failsafe way to avoid this problem is to have them construct the simulation process themselves. If students are working in groups, you should try to be sure that the students doing the construction on the computer are not those who already understand sampling distributions well but rather those students who do not.

When the students construct the simulation, they do so in a way that allows them to see the results of each step so they understand what the computer is doing. The procedure described below may not be the most efficient way to generate the sampling distribution, but it is very instructive.

To perform the German Tank simulation, we naturally need to create a population of tanks. Following along with your computer, open a new Fathom document and drag a **New Table** from the shelf to the document. Name the attribute “TankNumber.” Because each row in a case table is numbered with a caseindex, the formula TankNumber = caseindex will specify the creation of a list of integers, 1, 2, . . ., n. Single-click on the word TankNumber to select the attribute and from the **Edit** menu, select **Edit Formula.** Enter the formula “caseindex.”

The text of the word will turn red to indicate that Fathom recognizes the command as in Figure 9 below.

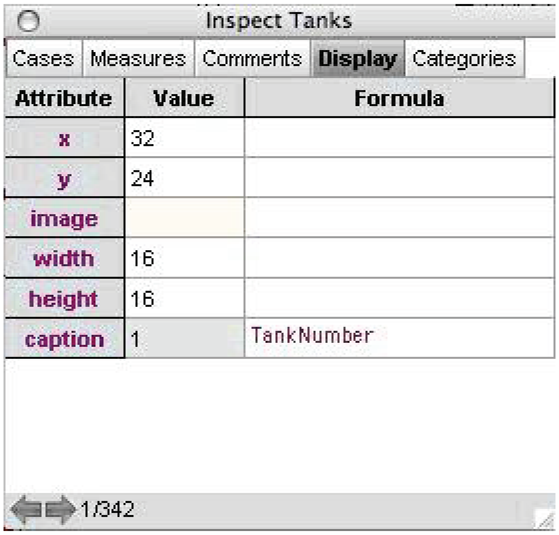


TankNumber = caseindex

Attributes are the names you can use in expressions. They refer to attributes in a collection.

Click **OK.** Notice that no tanks show up in the table. You created a formula to determine the values for any cases (tanks) you have, but you have not actually created any tanks. To do this, select **New Cases** under the **Collection** menu. In the dialog box that opens, type 342. This will create 342 tanks as in the previous section, and a gold box (the collection) filled with little spheres will appear on the screen.

The next couple of steps are for illustrative purposes and mimic the steps in our Westvaco simulation. They are not absolutely necessary for the simulation but are important pedagogically because they help students understand the simulation process. First, under the gold box, double-click on the text “Collection 1” and, in the dialog box that opens, rename the collection “Tanks.” Next, select the collection and drag the lower right corner of the blue frame so you can see the icons. Each ball represents one tank, and you should note that each tank is labeled “a case.” It is more instructive to have each case labeled with its number because it will be easy to see which tanks were selected for the sample. To change this label, double-click on the collection to open its inspector and select the **Display** tab. To the right of **Caption**, double-click in the **Formula** box, and in the formula editor, type in TankNumber as in Figure 10.



Attribute	Value	Formula
x	32	
y	24	
image		
width	16	
height	16	
caption	1	TankNumber

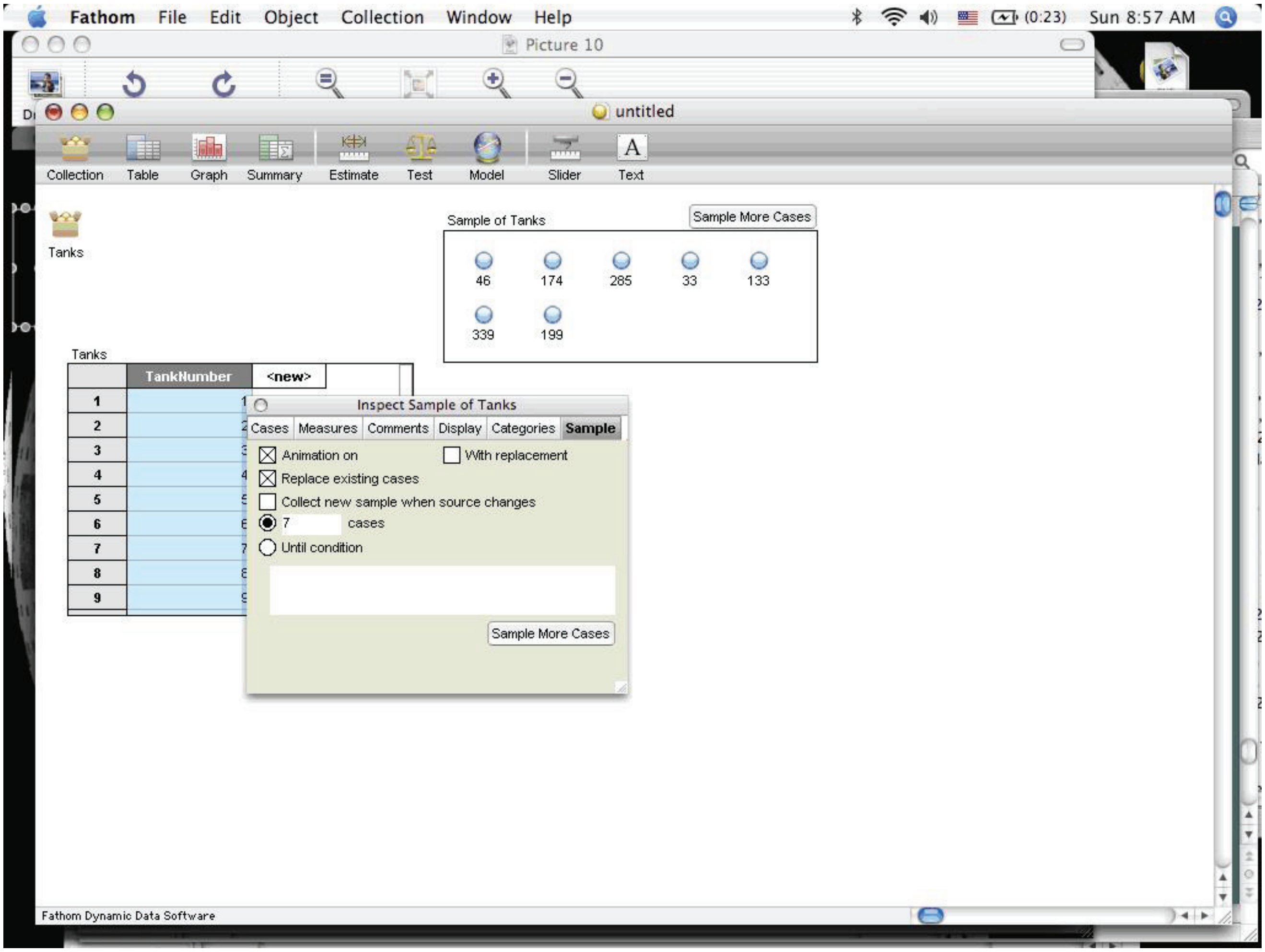
The caption for each icon in the collection will now show the number of the tank it represents. Close the inspector and minimize the collection to an icon by dragging the lower right corner.

To select a sample, single-click on the collection so the blue frame appears. Then, from the **Collection** menu, choose **Sample Cases**. A new blue box will appear, and you should see a blue ball pass from the collection box to the sample box to show that a sample is being selected. Single-click on the sample box and then drag the lower right corner of the frame until you can see all 10 balls in the sample. (Ten is the default sample size.) Now you can again see why we changed the captions: It is much more instructive to be able to see which tanks were selected when we do the simulation.

To change the size of the sample, double-click somewhere inside the sample frame to open the inspector for the sample. Change the number of cases to 7, turn off **With Replacement**, and click on **Sample More Cases** as in Figure 11 on page 51.

Now you can see the sample of seven tanks. Click the **Sample More Cases** button above the sample, and a new set of seven tanks will appear. I typically allow students to do this several times to see the repeated sampling.

In the classroom, students can get lost in the steps of this process so it is good to remind them the purpose of our activity. We want to estimate the number of tanks in the collection based on one sample. A sample statistic will be used to estimate the population maximum, and in this context the formula can technically be referred to as an **estimator**. Unlike the situation in World War II, we will have the opportunity to take many samples from the population. We will also be able to check our answers because we know the size of our population.



Tanks

	TankNumber	<new>
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	

Inspect Sample of Tanks

Cases | Measures | Comments | Display | Categories | **Sample**

☒ Animation on □ With replacement

☒ Replace existing cases

□ Collect new sample when source changes

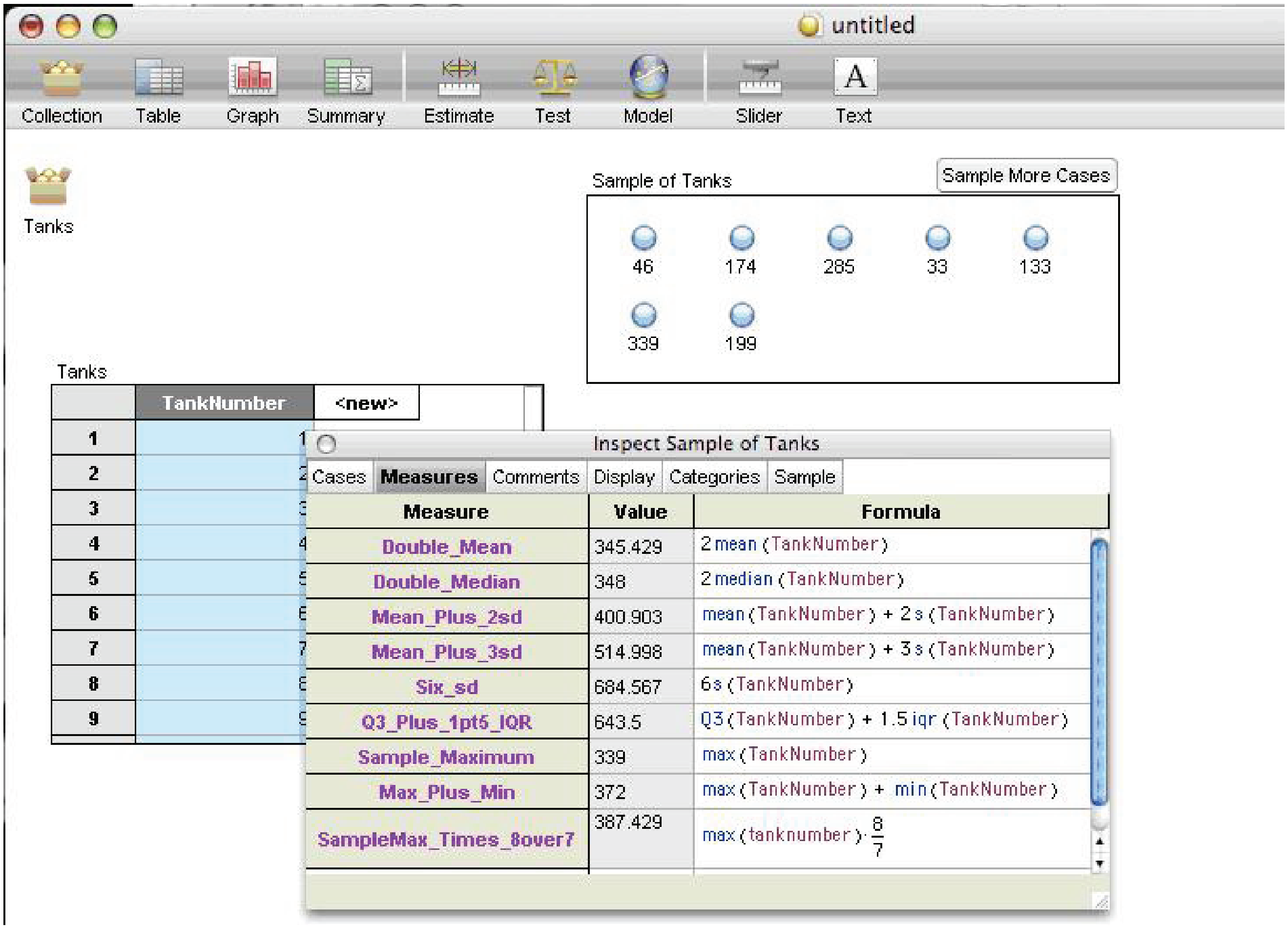
● 7 cases

○ Until condition

Sample More Cases

Some typical estimators were introduced in the earlier article of this Focus chapter. Students can calculate one or all of these estimators, or others they create, for each sample. To do multiple estimators, reopen the inspector for the sample (double-click somewhere in the sample), and click the **Measures** tab. Then enter the name of the estimator, such as "Double_Mean" for double the mean. (Fathom does not allow spaces in names of attributes or measures.) Double-click in the cell for the formula and type in 2*mean(TankNumber). Figure 12 on page 52 illustrates how to enter formulas for the nine estimators described earlier. Note that the name of the estimator can be different, but there are restrictions on the symbols that can be used. The formula is strict in its syntax.

Have students look at the numbers selected in the sample (the captions on the icons) and calculate their estimator with a calculator so they can compare to the result given by the software. This will reinforce trust in what the computer is doing as well as allow students to check for errors in the formula they entered.



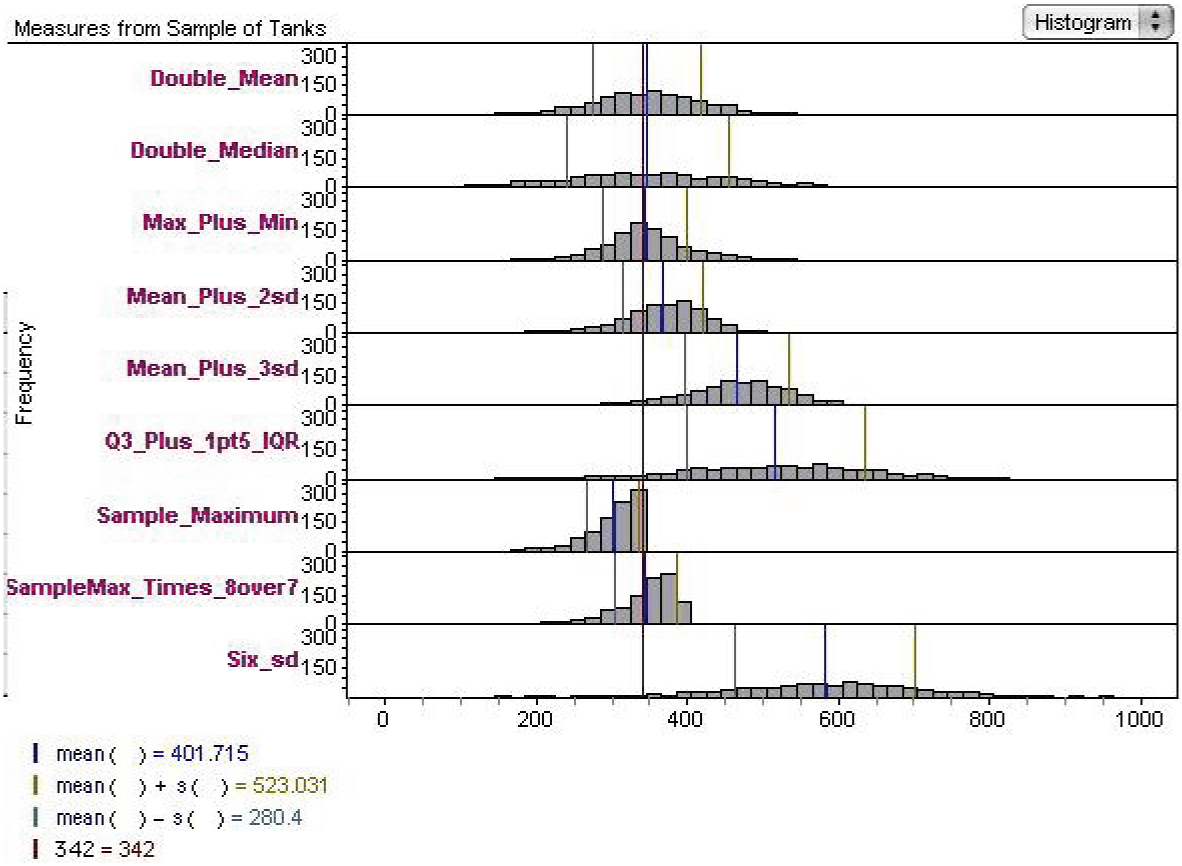
Measure	Value	Formula
Double_Mean	345.429	2 mean (TankNumber)
Double_Median	348	2 median (TankNumber)
Mean_Plus_2sd	400.903	mean (TankNumber) + 2 s (TankNumber)
Mean_Plus_3sd	514.998	mean (TankNumber) + 3 s (TankNumber)
Six_sd	684.567	6 s (TankNumber)
Q3_Plus_1pt5_IQR	643.5	Q3 (TankNumber) + 1.5 iqr (TankNumber)
Sample_Maximum	339	max (TankNumber)
Max_Plus_Min	372	max (TankNumber) + min (TankNumber)
SampleMax_Times_8over7	387.429	max (tanknumber) · $\frac{8}{7}$

Fathom provides a quick and easy way to compare the estimators for any given sample. Click **Sample More Cases,** and we can see the estimate each of these would provide for that sample. Repeatedly generating new samples allows students to see how the different estimators relate to each other, and they can get a feel for which tends to be closest to the population size of 342. They can play the *What if* game many times at once. "What if this were our sample? How would *my* estimator have performed? How would the others have performed?" Students can try to see if any of the estimators are consistently near the correct value, consistently too high or too low, etc. But we are interested in a more global way to look at the estimators.

We want to know which sample statistics would be the most reliable in predicting the population maximum. To assess this reliability, we need to see how the distribution of each of these statistics, used as estimators, behaves. (In the language of Fathom these estimators are called *measures.*) In order to understand the estimator's behavior, we need to collect some samples and calculate some estimates.

Single-click on the sample so the blue frame appears, and from the **Collection** menu, select **Collect Measures**. A blue sphere passes from the collection to the sample, showing a sample being collected, then a green sphere passes from the sample to the newly formed measures box to show each set of estimators being collected. Double-click on the green measures box to open its inspector, set the number of measures to 1000, turn off the animation, and click **Collect More Measures**.

Click the **Cases** tab to see the results of each sample. The arrows in the lower left corner allow you to scroll through them one at a time. To create a picture of the distribution of one of the estimators, drag a new graph from the shelf to the document and drag the name of the estimator from the inspector to the horizontal axis of the graph. To plot the mean on the graph, click once on the graph to select it, then from the **Graph** menu, select **Plot Value**. In the formula editor, type "mean( )." You could also display the mean plus and minus one standard deviation. Select **Plot Value** again and type "mean( ) + s( )" and again with "mean( ) − s( )." You can compare the different estimators by dragging each name to the horizontal axis and dropping it on the "plus" sign that appears as the axis is highlighted. Then you can **Plot Value** and type 342 to display the population maximum to see how each estimator performed. You can change the bin width and the scales by double-clicking on the graph and editing the **Properties**. Figure 13 has a bin width of 20.



Now is the time to discuss some qualities of a good estimator. *Double the mean*, *double the median*, and *maximum plus minimum* all seem to be centered at the population maximum, and students will see this as a desirable quality. Students also seem to recognize the benefit of a sampling distribution with a smaller spread. This is an ideal time to introduce the term *unbiased estimator* as an estimator whose average is the population parameter, and to confirm that this and a small spread are important qualities in an estimator.

The Central Limit Theorem

One of the most important and useful topics in an introductory statistics course is the Central Limit Theorem, which says that the sampling distribution of the sample mean becomes more nearly normal in shape as the sample size increases. This is the basis for the large sample inference procedures for means and proportions we teach in AP Statistics.

Helping students develop an intuitive feel for this is an important part of our teaching strategy. An activity developed by Ann Watkins and Richard Scheaffer in *Activity-Based Statistics* does this quite well, especially when technology is employed to allow many samples to be gathered quickly. What follows is a modified version of their activity.

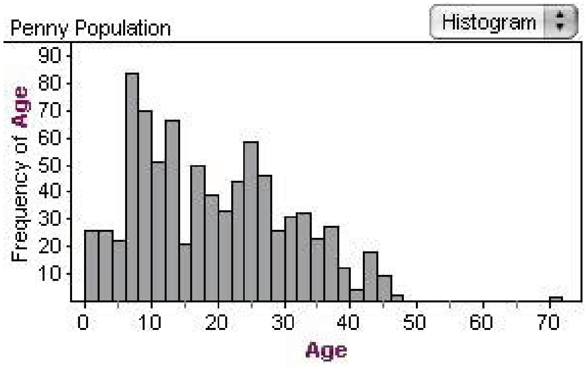
I have a bin of over 800 pennies in my classroom and also have their dates in a Fathom file. Early in the year I tell students that when they walk into the room, they are to reach into the bin, give it a good mix, and take out a penny. They then plot the age of the penny on a sheet of chart paper I have on my wall. Every student does this every day for several weeks, and we watch the dotplot develop.

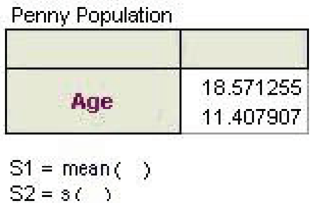
After three weeks or so I ask them to draw four pennies, calculate the mean of the ages, and record that on a second axis that I place below the first. Later we do samples of nine pennies. By the time we are ready to introduce the CLT, we have some nice dotplots ready to examine. Students see that the spread of the sampling distribution becomes smaller and the shape becomes more symmetric, but it takes many, many samples to really see the convergence to near normality. That's where the technology comes in.

Using statistics software, students create a histogram of the penny ages with the mean marked on the plot. They also create a summary table showing the mean and the standard deviation of the population. Then we duplicate the dotplots they did by hand to see that the plots created by the technology match what they did by hand. This helps students to understand what the technology is doing and to trust that it matches reality. The plots below show the population and the means of 200 samples each of size 1, 4, and 9. It is good to take a moment to remind students how these plots were created. A sample was selected, the mean age calculated and plotted. Use this to remind students of the difference between *n* (the sample size) and the number of samples. Also point out that the center of the distribution doesn't change much, though the shape becomes more symmetric and the distribution less spread out.

Then, to save time, pull up a document prepared ahead of time. The students have seen how to create these distributions and that they match reality, and are now ready to look at the patterns in the next set of plots and summary statistics. Figure 14 contains a histogram of the population of penny ages along with a summary table showing the mean and standard deviation, as well as plots and summary tables for 2,000 samples each of size 1, 4, 9, 16, 25, and 100. The reason for choosing sample sizes that are squares will be apparent in a moment.

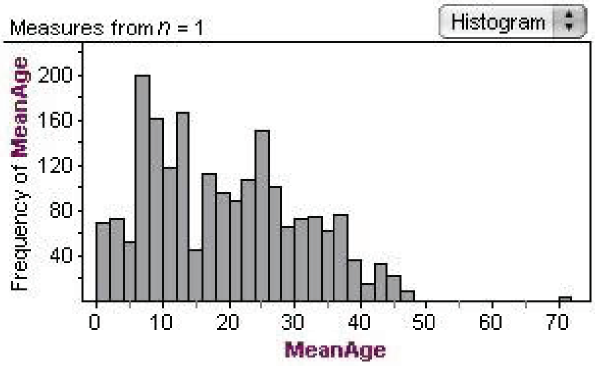
Students had already observed that the center doesn't change much. The numbers in the summaries make this clearer. There is virtually no change in the mean of the means, which we symbolize as, irrespective of the sample size. It was also clear, even from the plots they created by hand, that the distribution became more symmetric and mound-shaped—more nearly normal—with increasing *n*. This is the **Central Limit Theorem** in action! Notice that the population distribution was skewed to the right. When the sample size was one, the sampling distribution looked very much like the population distribution. As we used larger samples, the shape of the sampling distribution gradually changed from the skewed shape to a very symmetric shape.

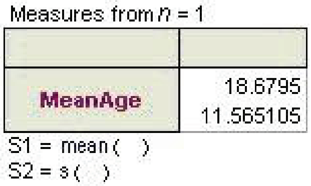



Penny Population

Age	18.571255 11.407907

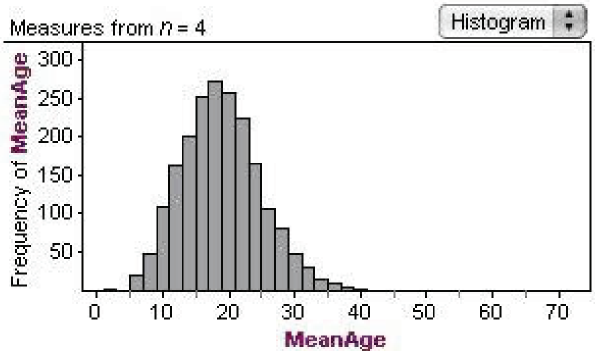
S1 = mean( )
S2 = s( )

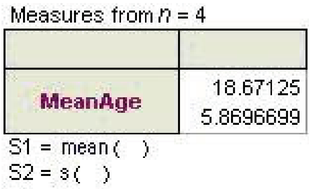



Measures from *n* = 1

MeanAge	18.6795 11.565105

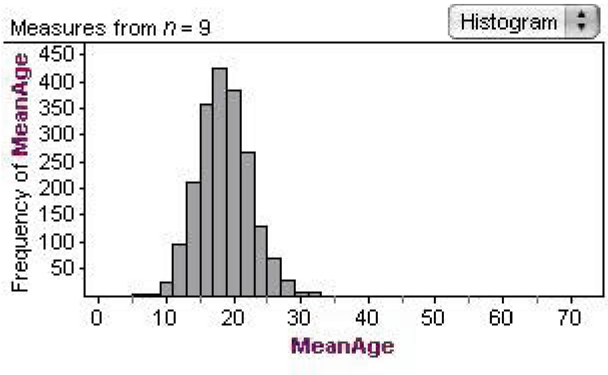
S1 = mean( )
S2 = s( )

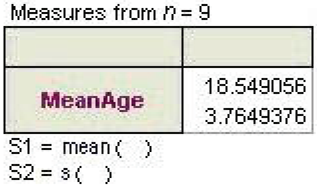



Measures from *n* = 4

MeanAge	18.67125 5.8696699

S1 = mean( )
S2 = s( )

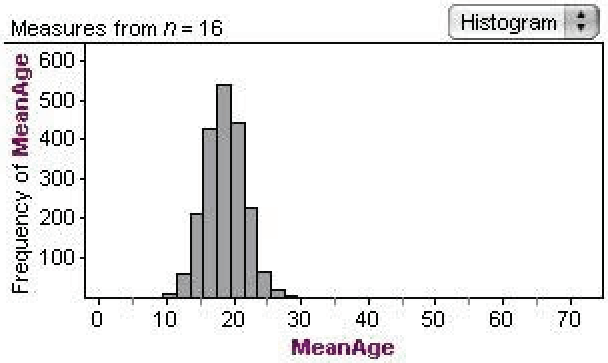


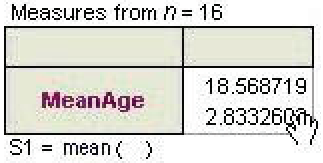


Measures from n = 9

MeanAge	18.549056 3.7649376

S1 = mean ( )
S2 = s ( )

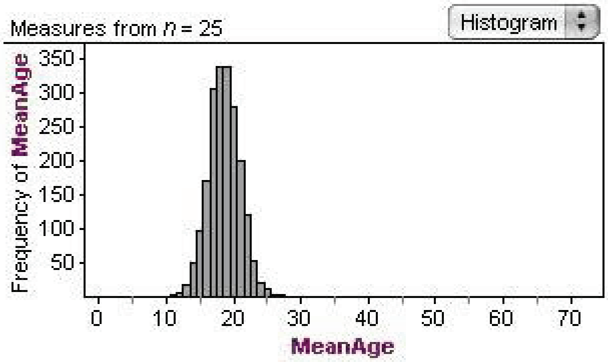


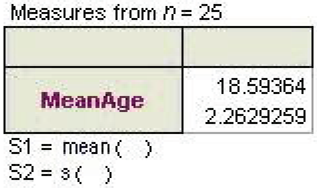


Measures from n = 16

MeanAge	18.568719 2.8332607

S1 = mean ( )
S2 = s ( )

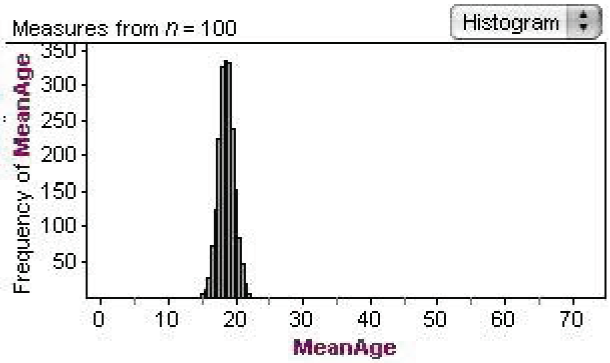


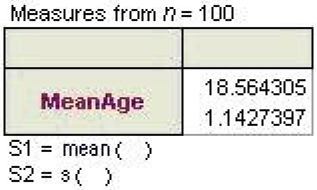


Measures from n = 25

MeanAge	18.59364 2.2629259

S1 = mean ( )
S2 = s ( )





Measures from n = 100

MeanAge	18.564305 1.1427397

S1 = mean ( )
S2 = s ( )

Now look at the spread. A quick observation shows that the standard deviation decreases as the sample size increases, but the summary tables allow a more detailed examination. Create a table of sample size and standard deviation of the means, $\sigma_{\bar{x}}$.

Sample Size	Standard Deviation	Observation
1	11.56	Very close to σ
4	5.87	Very close to $\sigma/2$
9	3.76	Very close to $\sigma/3$
16	2.83	Very close to $\sigma/4$
25	2.26	Very close to $\sigma/5$
100	1.14	Very close to $\sigma/10$

It doesn't take long to observe that the standard deviation of the means is approximately equal to the population standard deviation divided by the square root of the sample size. In symbols:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This can take some time to sink in, so you should save this document for the next time a student asks why you divide by the square root of the sample size. You may also wish to have this document available when the time comes to discuss the so-called "10 percent condition" that the sample size be no more than 10 perecent of the population size. If you have a second set of plots and summaries for samples taken without replacement, you can compare the standard deviations of the means with those from this activity. Almost no difference will be observed for small sample sizes, but when the sample size is 100, the standard deviation without replacement will be noticeably smaller than with replacement.

Body Fat: The Sampling Distribution of the Slope of a Regression Line

The sampling distribution of the slope of a regression line seems more difficult for students to visualize than those for sample means and proportions. The following activity allows students to see that the slope of a regression line based on a sample is simply another sample statistic and, like every sample statistic, it has a sampling distribution.

We shall begin with a context that meets the assumptions of the Simple Linear Regression Model:

- The relation between x and y is captured by the model, $y = \alpha + \beta x + Error$.
- The distribution of errors for each fixed x-value has a mean of 0.
- For each fixed x-value, the *Errors* in the population are approximately normally distributed.

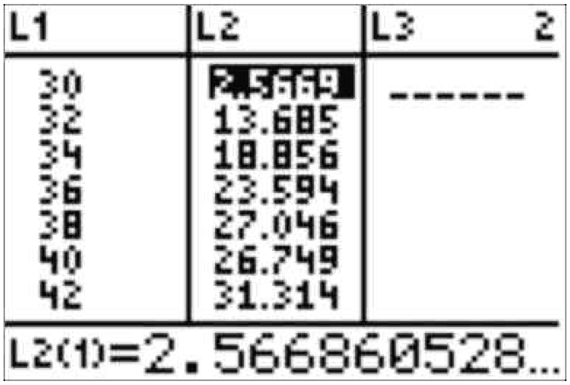
- For each fixed *x*-value, the *y*-values have the same standard deviation.
- The *Errors* associated with different observations are independent of one another.

In this activity, we will examine the relationship between the percent of body fat and waist size in men between the waist sizes of 30 and 45 inches. Of course, the percent of a person's body that is made up of fat depends on several factors, but there is a linear relationship between these two variables for men in this waist size range. We will also change technology gears and demonstrate the use of a graphing calculator (in our case a TI−84+).

For purposes of our simulation we will assume that men with a 30-inch waist average 8.3 percent body fat, and that each additional inch in waist size carries with it, on average, 1.7 percentage points more body fat. This means the regression line for the population would have the equation:

$$\%BodyFat = -42.7 + 1.7\ \ WaistSize$$

We will simulate sampling men from this age range, one from each of several waist sizes. In our simulation, we will assume a normally distributed body fat percentage at each of these waist sizes with a standard deviation of about 4.7 percentage points. We will use a TI−84+ graphing calculator for this simulation. In List #1, enter the waist sizes 30, 32, 34, 36, 38, 40, 42, and 44. We will now generate the body fat percentages in List #2. The predicted value of the body fat is given by the regression equation, and we want to introduce random error into each selection to simulate randomly selecting from all men with the given waist size. To do so, move the cursor to the title bar of L_2 and enter the formula $-42.7 + 1.7{*}L_1 + \text{randNorm}(0,4.7,7)$. This will generate body fat percentages that vary from the predicted value by an amount that is normally distributed with standard deviation 4.7, and should result in a table that looks something like Figure 15. Of course, since we are randomly generating the errors, your values in L_2 will probably be different than mine.



L1	L2	L3 2
30	2.5669	------
32	13.685	
34	18.856	
36	23.594	
38	27.046	
40	26.749	
42	31.314	

L2(1)=2.566860528...

Next, students should calculate the regression line and create a scatterplot of the values.

Have students compare their plots and their regression equations. Remind them that they all "sampled" from the same population, yet their graphs and regression equations are different. There are two parameters of interest in this simulation: the slope of the regression line and the *y*-intercept of the regression line. Here (as in AP Statistics) we will consider only the slope. Unfortunately the steps above are a bit cumbersome for selecting many samples, which is what we need to do to generate an effective simulated sampling distribution. A short calculator program will allow us to repeatedly select at random one person of each waist size from this population as well as calculate the slope of the sample regression equation. (As a reminder, we note that different calculators will have a slightly different syntax in their programming steps.)

Pedagogically, it makes sense to first talk to students about the process we are going to do and then convert the verbal explanation into calculator language. We recommend writing the basic steps on the blackboard as you talk through the process with students, then writing the calculator commands that will execute each step. Our basic process is as follows:

1. Select (say) one person of each waist size in the domain.
2. Calculate the regression equation.
3. Record its slope.
4. Repeat many times.
5. Plot the slopes.

Steps 1 and 2 are pretty straightforward. Leave some space above the calculator commands for step 1 because we'll need to add some things later.

Step 1: Select one person of each waist size.

```
:{30,32,34,36,38,40,42}→L1


:−42.7 + 1.7* L1 + randNorm(0,4.7,7)→L2
```

Step 2: Calculate the regression equation.

```
:LinReg(a + bx)
```

Steps 3 and 4 require a loop to be set up. A "For" loop would be effective here, and it is worth taking a few minutes to explain how this works to students. A "For" loop has a placeholder variable, a beginning index, and an ending index. The command for the loop is something like

```
:For(I,1,50)
:
:
:End
```

When the line with the “For” command is reached, the value of the beginning index (1 in this case) is stored into the variable I. Then the calculator executes the commands that follow until it reaches the “End” command. At this point it jumps back to the “For” line and stores the value 2 in the variable I. This continues until the value 50 is used, at which point the calculator will continue past the “End” command. The value of *I* can also be used within the loop, a property we will exploit as we tell the calculator where to store the value of the slope each time.

L_3`(I)` refers to the “ith” element in List 3. Therefore, we can use the command

```
:b-> L3(I)
```

to store the current slope into the corresponding place in List 3 as we proceed through the loop. So far, the program looks like this:

```
:{30,32,34,36,38,40,42}→L1
:For(I,1,50)
:−42.7 + 1.7* L1 + randNorm(0,4.7,7)→L2
:LinReg(a + bx)
:b→L3(I)
:End
```

Note that we only need to store the *x*-values into L_1 once because they will not change from sample to sample. (The store command is therefore not inside the loop.) Now we need to plot the distribution of L_3. To accomplish this, we need to add two lines to the end of the program:

```
:Plot1(Histogram, L3, 1)
:ZoomStat
```

To avoid complications from previous use of the calculator, we also must add some commands to the beginning to turn off all other plots and equations, and we can add an optional command to ask the user to enter the number of samples to take. The completed program would look like this:

```
:PlotsOff
:FnOff
:ClrList L1,L2,L3
:Disp“HOW MANY”
:Disp“SAMPLES?”
:Prompt S
:{30,32,34,36,38,40,42}-> L1
```

```
:For(I,1,S)
:−42.7 + 1.7* L1 + randNorm(0,4.7,7)→L2
:LinReg(a + bx)
:b→L3(I)
:End
:Plot1(Histogram,L3,1)
:ZoomStat
```

The program may take a few moments to run, but the reward at the end is terrific—a simulated sampling distribution of the slopes! Next, you should calculate the mean and standard deviation of the slopes. (My calculator shows a mean of 1.77 and standard deviation, 0.43.) To reinforce the meaning of this histogram it is probably a good idea to ask some questions such as: “What is represented by the bar furthest to the right?” You want students to recognize that, again, the sampling distribution of the slope is again a game of “What If?” What if this was our sample? Or that? What would be the slope of the regression line? Each sample resulted in a different slope. The bar farthest to the right in the histogram represents the steepest slopes. You might suggest a possible sample that would give a slope that is steeper than normal. Such a sample might have lower than predicted values for smaller waists and higher than predicted values for larger waists. Then follow up with a fill-in-the-blank question like, “Most of the samples resulted in regression line slopes between __?__ and __?__.”

The standard deviation calculated above is the estimated standard error of the slopes, which is an important component of the inference procedures. You might want to review regression output and show where this standard error of the slope appears on the printout.

Applets

No discussion of simulating sampling with technology would be complete without mentioning the Internet! Statistics applets are available that offer some nice advantages over purchased statistics software. First, applets are free. For districts with tight budgets, this is an important consideration. Second, they are accessible to anyone who has Internet access. This means many students can access them from home or the library.

There are also a couple disadvantages. Applets sometimes disappear or move unexpectedly so becoming too dependent on them can be risky. They are also relatively inflexible. Applets are designed to demonstrate one concept and are generally not adaptable to different settings. It is often impossible to change settings such as histogram bin widths and sample sizes, or you may have only limited options for the settings.

As of this writing, putting “‘Applet’ [&] ‘Simulation’ [&] ‘Sampling Distribution’” into my browser netted over 850 hits, most with .edu extensions. That’s a lot of places to begin looking for quality applets to utilize in teaching statistics!

References

Ruggles, Richard, and Henry Brodie. 1947. "An Empirical Approach to Economic Intelligence in World War II." *American Statistical Association* 42(237): 72–91.

Teague, Dan. 2003. "The Taxi Problem." http://courses.ncssm.edu/math/Talks/index.htm.

Watkins, Ann E., Richard L. Scheaffer, and George W. Cobb. 2004. *Statistics in Action*:*Understanding a World of Data*. Key Curriculum Press.

Contributors

Chief Editor

Chris Olsen has taught statistics at George Washington High School in Cedar Rapids for over 25 years and AP Statistics since 1996; next year he will be starting AP Statistics at a sister school, Thomas Jefferson High School in Cedar Rapids. He is a frequent contributor to the AP Statistics Electronic Discussion Group and has reviewed materials for the *Mathematics Teacher,* AP Central, *American Statistician,* and the *Journal of the American Statistical Association*. He currently writes a column for *Stats* magazine and is coauthor of *Introduction to Statistics and Data Analysis.* He is a past member of the AP Statistics Development Committee and now serves as the College Board Content Advisor for AP Statistics. He has served as Table Leader and Question Leader at the AP Statistics Reading and has presented numerous workshops in AP Statistics in the United States and internationally.

Authors

Corey Andreasen has taught mathematics at the junior high school, high school, and university levels in Wisconsin over the past 12 years, and he currently teaches at Sheboygan (Wisconsin) North High School. He serves as a director (representing grades 9–12) of the Wisconsin Mathematics Council and is associate editor of the journal *Wisconsin Teacher of Mathematics* and is a contributing member of the author team for *Statistics in Action: Understanding a World of Data*, 2nd Edition. Andreasen holds bachelor's and master's degrees from the University of Minnesota at Minneapolis, and is certified by the National Board for Professional Teaching Standards.

Floyd Bullard is on the faculty at the North Carolina School of Science and Mathematics in Durham, North Carolina. He was a 2001 recipient of the Tandy Award for Excellence in Teaching Mathematics (now called the RadioShack National Teacher Award). Floyd has been an AP Statistics Exam Reader for four years, and his particular interests in statistics include simulations and Bayesian methods. Floyd is currently on a leave of absence to study statistics in a Ph.D. program at Duke University, after which he plans return to teaching.

Roxy Peck has been a professor of statistics at Cal Poly since 1979, serving for six years as chair of the Statistics Department, and is currently in her ninth year as associate dean of the College of Science and Mathematics. She was made a fellow of the American Statistical Association in 1998, and in 2003 she received the American Statistical Association's Founders Award in recognition of her contributions to K-12 undergraduate statistics education. She has authored two leading textbooks in introductory statistics and is a past chair of the ASA/NCTM Joint Committee on Curriculum in Statistics and Probability for grades K-12. She served from 1999 to 2003 as the Chief Reader for the AP Statistics Exam.

Advisors

Rob Gould
University of California, Los Angeles
Department of Statistics
Los Angeles, CA

Rob Gould has been managing the undergraduate program in the Statistics Department at UCLA since 1998. He has taught introductory statistics at UCLA since 1994, and has been a Reader for the AP Statistics Exam for the last five years. He has also been active in the American Statistical Association, serving six years on the Advisory Committee for Teacher Enhancement (including one year as chair) and is currently serving a three-year term as a Member at Large of the Section on Statistics Education board. He is a co-creator and current director of the INSPIRE program, an National Science Foundation-funded distance learning course for beginning AP Statistics teachers.

Robin Levine-Wissing
Glenbrook North High School
Instructional Supervisor of Mathematics
Northbrook, Illinois

Robin Levine-Wissing is currently the Instructional Supervisor of Mathematics at Glenbrook North High School in Northbrook, Illinois, where she also teaches one course in Advanced Placement Statistics. Robin has taught for 29 years in five states. She has been an AP Reader and Table Leader since 2000 and has conducted AP Statistics and Pre-AP workshops and Summer Institutes all across the United States and Hawaii. She has been a national T^3 (Teachers Teaching with Technology) instructor since 1993 and provided training for teachers on graphing calculators and computer software. Robin received the Presidential Award for Excellence in Mathematics in 1993, the Tandy Technology (Radio Shack) Award in 1998, and Clark County (Nevada) Teacher of the Year in 1996. Robin is the coauthor of the new 3rd edition of REA's *AP Statistics Test Preparation Book*.

F07STSF177