AP STATS

by Dr. Bob Stephenson

1 Sampling Distributions

The sampling distribution of a statistic is one of the most difficult concepts encountered by students in an introductory statistics course. It is also one of the most fundamental. Without an understanding of sampling distributions, statistical inference becomes a mysterious mix of formulas and steps learned by rote. Sampling distributions are hard to understand because there is so much going on. Students must grapple with not just one idea, but several.

This article presents a series of activities, some in class and some out of class, that allow students to experience and explore the sampling distribution of the sample mean. The in class activity is presented using a hypothetical class of 25 students. However it can easily be adapted to both larger and smaller class sizes. The out of class activity involves the use of the internet and a Java applet developed by David Lane as part of the Rice Virtual Lab in Statistics at Rice University (**www.ruf.rice.edu/∼lane/rvls.html**).

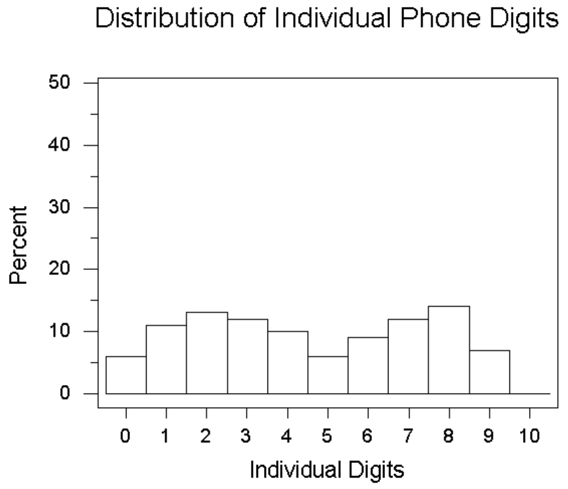
2 In Class Activity

The understanding of sampling distributions begins with a clear distinction between a population and a sample. The population consists of all items of interest, where a sample is simply a few of those items selected from the population. One problem with in class activities is having materials conveniently available. For this activity the students provide the materials by way of their telephone numbers. Specifically, the last four digits of their telephone numbers. Below is a hypothetical class of 25 students and the last four digits of their telephone numbers.

Amanda	2078	Doug	1849	Kathryn	5293	Nathan	4198	Sarah B.	1213
Amy	1176	Jamie	0885	Kelly	8652	Nathaniel	2647	Sarah S.	1325
Angela	7877	Jeffrey	7314	Kimberly	9281	Patrick	2456	Shelley	4839
Audra	0679	Jennifer	5748	Mark	2338	Peter	3678	Theresa	4227
Brianne	3083	Jessie	1492	Miki	3806	Robert	2741	Virginia	3066

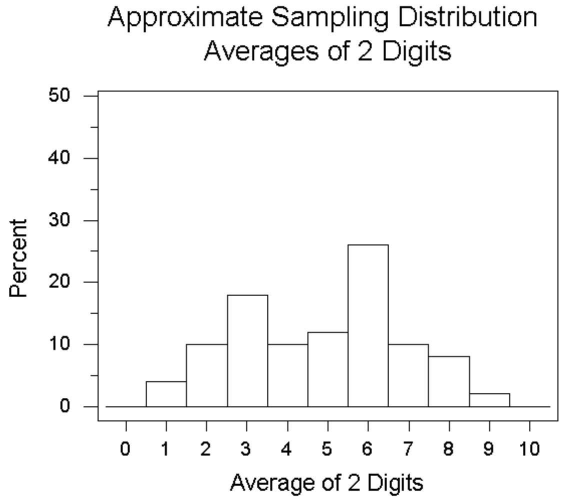
The collection of 100 phone digits will be our population of interest. If you have a large class (100 or more, as I do) each student contributes only the last digit of her/his phone number to the population. For a smaller class you can go to a phone book and get additional numbers to create a larger population.

The first thing to consider is how these digits are distributed. Have students come to the board and tally their digits, create a histogram and comment on the histogram's center, spread and shape.



The distribution of the individual digits is centered around 4.5 with values ranging from 0 to 9. The shape is somewhat bimodal with relatively more 2's and 8's and relatively few 0's, 5's and 9's. This shape is not necessarily typical of the distribution of telephone digits. An important point to make is that this distribution does not change. Any time you look at the entire population you will get the same distribution of values.

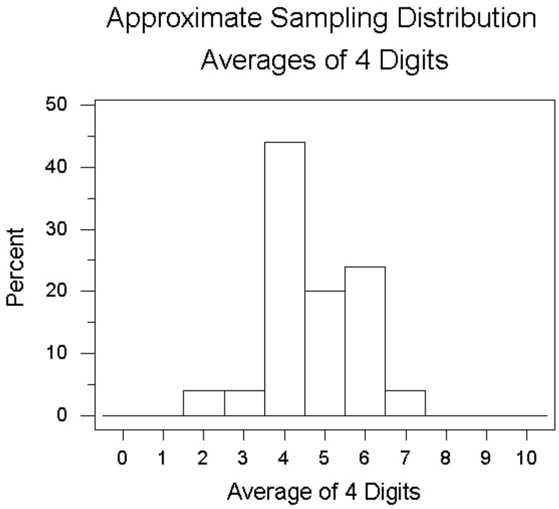
What happens when we sample from our population of 100 digits? Consider samples of size 2. It is convenient to have each student take the first 2 digits of her/his 4 digit number as a sample of size 2. She/he can also use the last two digits as a sample of size 2. By finding the average value for each of these samples, one can have 50 realizations of the average of samples of size 2. These are only 50 of the thousands of possible samples of size 2. Once each student has her/his averages of size 2, the results can be tallied and a histogram of these averages can be constructed and described. Below is the histogram for our class of 25.



The center of this simulated sampling distribution is around 4.5 (similar to that of the original population distribution) but its spread is somewhat less, sample means range from

1 to 9. The shape is tending to mound more towards the center. The percentage of values away from the center is reduced.

Now consider samples of size 4. It is convenient to have each student take his/her 4 numbers as a sample of size 4. By finding the average value for each of these samples, one can have 25 realizations of the average of samples of size 4. Again, the students can go to the board, tally their averages and construct a histogram of those averages. Below is the histogram for our class of 25.



The center of this simulated sampling distribution is around 4.5 (similar to that of the original population distribution) but its spread is much less, sample means range from 2 to 7. The shape is much more clustered (a single mound) near the center.

3 Discussion

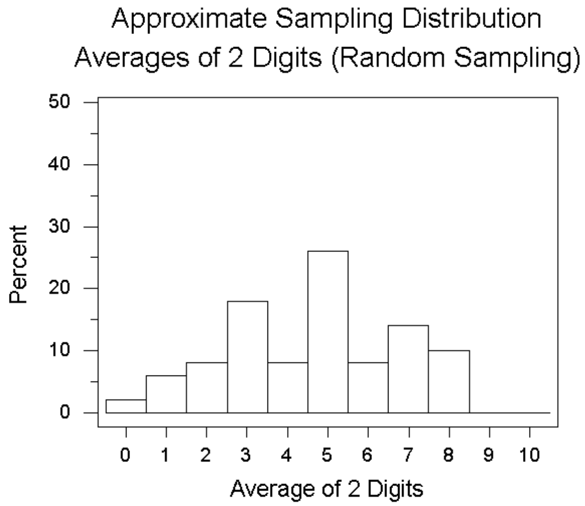
With a class of 100, this activity takes about 30 minutes to complete. Rather than have students come to the board, I have a show of hands for the number of values in each interval class.

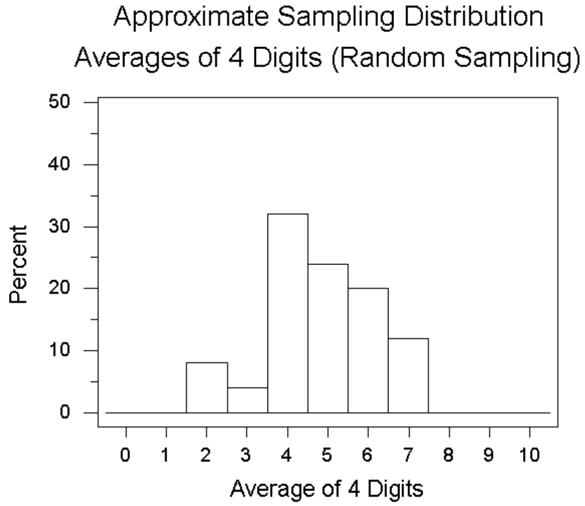
This activity helps to highlight the relationship between the center and spread of the population of values and the center and spread of the sampling distribution of the sample average. It also begins to hint at the Central Limit Theorem. The bimodal shape seen in the population distribution is not apparent in the sampling distribution of the sample average values as the sample size increases.

Many students are confused about the number of samples versus the sample size. We have taken 50 samples of size 2 and 25 samples of size 4 to construct the latter two histograms. It is the size of the sample that is important, not the number of samples. By using a percent scale for the vertical axis on the histograms, the number of samples is de-emphasized.

For larger classes, I have students form groups of 2 and 4. Each member of the group contributes the last digit of his/her telephone number. Most students will form groups by proximity (who is closest to them).

Either with the four digits from each student or groups formed by proximity, it appears that we have violated an important rule for the construction of sampling distributions. We have used convenience samples instead of random samples. Does this make a difference? Have the students come up with a method for randomly sampling from the population. This can be done by having each student write their digits on individual slips of paper. The slips can then be put into a bag, mixed and sampled from. Create multiple random samples of size 2 and multiple random samples of size 4 and construct histograms of the averages. The results you get will be different from the histograms below but should show the same narrowing of spread and clustering (mounding) in the middle.





What the students will find is that in this case the convenience samples work as well as random samples since the assignment of the last 4 digits of a phone number is essentially a random assignment. With most practical sampling situations, convenience samples are prone to bias and random samples should be used.

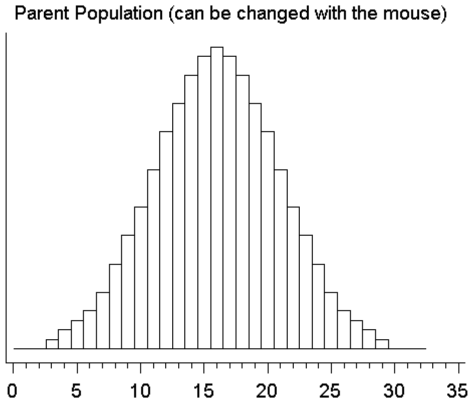
4 Exploring Sampling Distributions on the World Wide Web

Once students have done the in class activity it is time to turn them loose (with a little guidance) to explore sampling distributions using a program developed by David Lane at Rice University. This program allows students to quickly draw many, many samples and to easily change the characteristics of the population and the size of samples. Students will need access to the World Wide Web via a browser that supports frames and Java.

Students should go to the URL:

www.ruf.rice.edu/∼lane/stat_sim/sampling_dist/index.html

Instructions for the use of the sampling distribution Java applet appear on the right side of the page. Click on the Begin button on the left side of the page. This will bring up a Java applet Window with four sets of axes. The top set displays a mounded, symmetric distribution of the parent population. To the left are the population parameters, a mean of 16 and a standard deviation of 5. To the right is a pull down menu where you may select a few other shapes. Make sure that a normal distribution is selected for the first activity. The parent population should look like the histogram below.



1. Click on the Animated Sample button located at the right of the Sample Data axes. The results of a simple random sample of size 5 taken from the parent population will appear on the Sample Data graph. The mean of this sample of size 5 will appear in blue on the Distribution of Means graph.

    - In what ways does the sample look like the parent population?
    - What could you do to increase the likelihood that the sample would look more like the population?
    - Looking at the summary statistics to the left of the Sample Data graph, is the mean of the sample near the population mean?

- If the Animated Sample button is clicked again will the new sample be the same as the current sample? Briefly explain your answer.
- Click on the Animated Sample button to confirm your answer.

Let's focus on the Distribution of Means graph. This graph should contain two blocks representing the means of the two random samples of size 5 that have been selected. The mean of these two sample means is given to the left of the graph. Press the Animated Sample button several times. Now use the 5 samples and 1000 samples buttons to build up the sampling distribution. Each time one of these buttons is pressed, more samples of size 5 are selected from the population. It is important to remember that the sample size (5) does not change. Each sample contains 5 observations from the parent population. We have to simulate enough samples before the Distribution of Means becomes apparent. Reset the applet by clicking on **Clear lower 3**. Use the 10,000 samples button to simulate the sampling distribution of the sample mean for samples of size 5. In Table 1, sketch the 'Distribution of Means' and give the mean and standard deviation (sd).

Table 1: Distribution of Means, n=5

Mean	
St. Dev	
Sketch	

- Does the mean of the Distribution of Means differ from the center of the parent population by a lot?
- How does the spread of the Distribution of Means differ from the spread of the parent population?
- Does the shape of the Distribution of Means differ from the shape of the parent population distribution?
  Concentrate on the shape. Where does the distribution mound? Is the distribution symmetric?

2. Repeat activity 1 only this time use the N=5 pull down menu next to the Distribution of Means graph to change the sample size to N=25. Record the description of the

Distribution of Means in Table 2. What do you notice that is different when using random samples of size 25 instead of 5?

Table 2: Distribution of Means, n=25

Mean	
St. Dev	
Sketch	

3. Change the population shape to skewed. Report the mean and standard deviation (sd) and shape (it may help to sketch the distribution) of the parent population in Table 3. Select a sample size of 2 for the Distribution of Means on the third set of axes. Select the Mean and a sample size of 5 for the Distribution of Means on the fourth set of axes. Take 10,000 samples. Report the means and standard deviations and describe and/or sketch the shapes for the Distributions of Means in Table 3. Now select a sample size of 10 for the Distribution of Means on the third set of axes. Select the Mean and a sample size of 25 for the Distribution of Means on the fourth set of axes. Take 10,000 samples. Report the means and standard deviations and describe and/or sketch the shapes for the Distributions of Means in Table 3.

Table 3, Skewed distribution

Sample size	Parent	2	5	10	25
Mean					
Standard Deviation					
Shape					

Compare the Distribution of Means for each sample size to the parent population distribution. How does the mean compare? How does the standard deviation compare? How does the shape compare? Summarize your findings about the relationship between the distribution of the parent population, sample size and the sampling distribution of the sample mean in one or two sentences. How does your summary compare to the statement of the Central Limit Theorem?

5 Discussion

The purpose of the first activity is to have students understand the idea of random sampling from a parent population. The Animated Sample is quite good at visually displaying the selection of 5 values from the parent population. These 5 values produce one realization of the sample mean. Another Animated Sample provides a different set of 5 values and a different realization of the sample mean. By building up the Distribution of the Mean slowly at first, students can see how different samples produce different sample means (the basis of sampling variability). Being able to generate multiple samples (10,000 samples button) greatly speeds the process of building the Distribution of the Mean.

Students can sometimes confuse the number of samples with the sample size. It is important to differentiate these two ideas. The Java applet uses Reps to indicate the number of times samples are taken and N to indicate the size of each sample. By sketching the Distribution of Means once for each sample size, and comparing the distributions for several sample sizes, the confusion with the number of samples generated should be lessened.

The first two computer activities deal with a parent population that is symmetric and mounded in the middle (a.k.a. normal). The Central Limit Theorem does not enter in here because the Distribution of Means is normal, no matter what the size of the sample, when sampling from a normal parent population. The Central Limit Theorem only comes into play when we are sampling from a parent population that does not have a normal shape. The third activity does address the Central Limit Theorem and how large a sample one needs before the Distribution of Means is approximately normal shaped. The activity is nice in that the parent population is always displayed and its shape never changes throughout the activity. It is the shape of the Distribution of Means that changes with the sample size.

There are several other web sites that you can go to in order to explore sampling distributions. Some, but certainly not all, are:

- **statweb.calpoly.edu/chance/applets/applets.html**

  There are several applets at this site that deal with sampling distributions.

  - **Sampling Senators** This applet allows you to draw samples of U.S. Senators and look at characteristics such as: gender, party, and years in office.
  - **Sampling Pennies** The population of interest consists of pennies and the year they were minted. Samples are taken from the population and distribution of the sample mean year is displayed.

  - **Reeses Pieces** This applet has a clever animation that looks at the sampling distribution of a binomial proportion.

- **www.stat.sc.edu/~west/javahtml/CLT.html**

  This applet simulates the rolling of 1 to 5 dice to demonstrate the Central Limit Theorem.

- **www.gen.unm.edu/faculty_staff/delmas/stat_tools/sampling_distributions/samp_dist_tools.htm**

  This website contains tools for exploring sampling distributions. There is computer software, that you can down load, for construction of the sampling distribution of the sample mean. There are also pre-tests and post-tests and instructions for using the software.

Acknowledgement

The author would like to thank Beth Chance for her helpful comments that greatly improved the quality of this column.