Investigating the Behavior of the Sample Mean

1. What is a sampling distribution?

The main idea of statistical inference is to take a random sample from a population and then to use the information from the sample to make inferences about particular population characteristics such as the mean (measure of central tendency), the standard deviation (measure of spread) or the proportion of units in the population that have a certain characteristic. Sampling saves money, time, and effort. Additionally, a sample can, in some cases, provide as much or more accuracy than a corresponding study that would attempt to investigate an entire population—careful collection of data from a sample will often provide better information than a less careful study that tries to look at everything.

In this lesson we will study the behavior of the mean of sample values from a different specified populations. Because a sample examines only part of a population, the sample mean will not exactly equal the corresponding mean of the population. Thus, an important consideration for those planning and interpreting sampling results, is the degree to which sample estimates, such as the sample mean, will agree with the corresponding population characteristic.

In practice, only one sample is usually taken (in some cases a small "pilot sample" is used to test the data-gathering mechanisms and to get preliminary information for planning the main sampling scheme). However, for purposes of understanding the degree to which sample means will agree with the corresponding population mean, it is useful to consider what would happen if 10, or 50, or 100 separate sampling studies, of the same type, were conducted. How consistent would the results be across these different studies? If we could see that the results from each of the samples would be nearly the same (and nearly correct!), then we would have confidence in the single sample that will actually be used. On the other hand, seeing that answers from the repeated samples were too variable for the needed accuracy would suggest that a different sampling plan (perhaps with a larger sample size) should be used.

A sampling distribution is used to describe the distribution of outcomes that one would observe from replication of a particular sampling plan.

2. Objectives

The lesson is designed to illustrate aspects of sampling distribution concepts and interpretation. By the end of the lesson students should:

- Know that estimates computed from one sample will be different from estimates that would be computed from another sample.

- Understand that estimates are expected to differ from the population characteristics (parameters) that we are trying to estimate, but that the properties of sampling distributions allow us to quantify, probabilistically, how they will differ.

- Understand that different statistics have different sampling distributions with distribution shape depending on (a) the specific statistic, (b) the sample size, and (c) the parent distribution.

- Understand the relationship between sample size and the distribution of sample estimates.

- Understand that the variability in a sampling distribution can be reduced by increasing the sample size.

- See that in large samples, many sampling distributions can be approximated with a normal distribution.

3. Startup Instructions

On a Vincent workstation

```
% add lisp
% add stat
% sd_module
```

On a PC

- Click on the Lisp-Stat icon in the program manager window

- Click on the ci.lsp icon in the Lisp-Stat window

On a Macintosh

- Start up xlispstat, by clicking on the XLispStat icon

- Pull down the File menu and select Load

- Select the folder Teach

- Select ci.lsp

4. The sd_module interface

Immediately after starting the sd_module, 4 windows appearing on the screen will be mostly blank except for some control tools which we will use to run the sd_module. The window in the upper left contains controls to choose a distribution, choose sample sizes, and run the sampling simulations. Each of the other windows contains slide-bar controls to adjust the number of bins in the histograms of the sample statistics and to adjust the smoothed estimate of the sampling distributions.

Figure 1 shows the `sd_module` windows after a simulation that was used to compute means from samples of size 5, 16, and 50 from an exponential distribution. Before running a simulation, the user must choose the distribution and corresponding parameter(s). The low and high sample sizes can also be adjusted by the user (the defaults are 4 and 50). In Figure 1, the low sample size was changed by the user from 4 to 5 and the intermediate sample size was chosen automatically as $\sqrt{50 \times 5} \approx 16$. The simulation is started by clicking on the "Start Simulation" button. After the simulation is completed, the user can adjust the number of histogram cells and the density smoothing parameter to give visually appealing histograms and estimated densities in each of the windows.
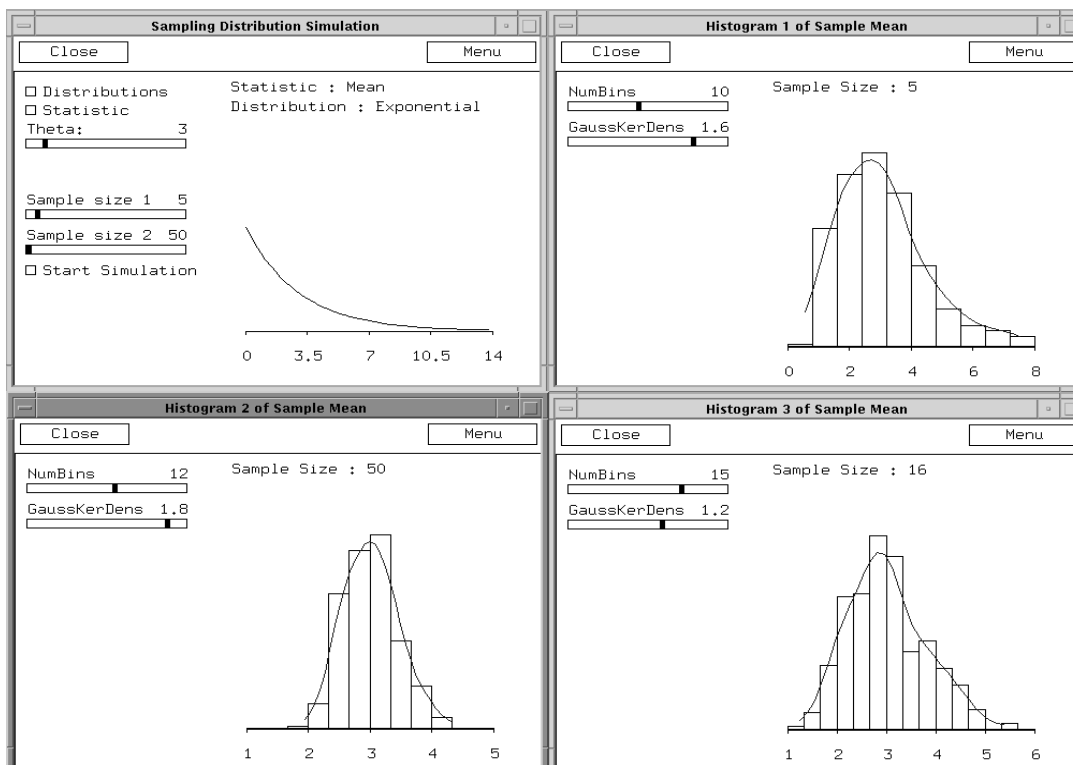

Figure 1: *Sampling Distributions Teaching Lesson. The* `sd_module` *windowss after a simulation that was used to compute means from samples of size 5, 16, and 50 from an exponential distribution.*

5. Warm-ups

To gain some familiarity with the `sd_module`, try the following:

- Use the `distribution` button to explore the shapes of the different "parent distributions."

- For each parent distribution, notice the effect that changing the parameter values [done by moving the appropriate slide bar(s)] has on distribution shape.

- Choose a distribution from the list (e.g. a chi-square with 1 degree of freedom) from which to sample. Then click on the "Start Simulation" button and wait for the simulation to run (this could take up to several minutes depending on the kind of computer that you are running). Approximately how many seconds did it take for the simulation to finish (signaled by the appearance of the new set of histograms)? It will be interesting to compare with class mates to see how large are the differences from one computer to another.

- After the simulation is completed, notice the different shapes of the 3 different histograms. Experiment with the number of histogram cells and with the "GaussKerDen" smoothing constant value in the windows to see if you can find a most visually appealing smoothed density estimate in each window.

6. Exercises

For the following exercises, choose the mean as the statistic of interest and choose the low sample size to be 4 (the smallest possible) and the high sample size to be 200 (the largest possible). When asked to "run the simulation" you may want to run the simulation more than once to check the consistency of your conclusions

1. Choose a normal distribution with a mean $\mu = 100$ and $\sigma = 10$ as the parent distribution and run the simulation. Study the histograms in the 3 windows.

    (a) Statistical theory tells us that the means of samples from a normal distribution should follow also a normal distribution. Compare the *shapes* of the histograms in the 3 windows. How do they differ? Explain how your conclusions relate to the statistical theory about sampling from a normal distribution.

    (b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

2. Choose a uniform distribution as the parent distribution and run the simulation.

    (a) For a uniform distribution parent distribution, statistical theory tells us that the means of samples from a normal distribution will *not* follow exactly a normal distribution, but that the normal distribution could provide a good approximation under certain conditions. Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a uniform distribution?

4

(b) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

3. Choose the chi-square distribution.

   (a) Explain how the shape of the density of a chi-square distribution changes as you change the number of degrees of freedom from 1 to 20. Choose the chi-square with 1 degree of freedom. Run the simulation.

   (b) Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a distribution that is chi-square with 1 degree of freedom?

   (c) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

4. Choose the Cauchy distribution. Run the simulation.

   (a) Describe the shape of the parent distribution.

   (b) Compare the *shapes* of the histograms in the 3 windows. How do they differ? What can you conclude about using the normal distribution to approximate the distribution of means from samples from a Cauchy distribution.

   (c) Compare the spread in the 3 different sampling distributions. How does sample size affect spread in these sampling distributions? How does your observation agree with what is predicted by statistical theory?

7. Solutions to Exercises

1. The results from theory for sampling from a normal distribution can be stated simply and concisely: The mean of a sample from a normal distribution with mean μ and standard deviation σ will follow a normal distribution with mean μ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

   (a) Although contaminated with random "noise," the shapes are all similar and approximately normal, as suggested by statistical theory because the sample mean from a normal distribution also follows a normal distribution.

   (b) The spread in the distribution decreases with increasing sample size. This is as expected from theory where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

2. Note that the uniform distribution has a shape that is quite different from that of a normal distribution, but it is a symmetric distribution.

(a) For the two larger sample sizes, the sampling distributions appear (again, except for the random "noise") to be shaped like a normal distribution. The approximation improves as the sample size gets larger.

(b) Again, the spread in the distribution decreases with increasing sample size, and, this is as expected from theory because, even if the underlying distribution is not normal, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

3. (a) The chisquare distribution is highly skewed to the right for small degrees of freedom. As the degrees of freedom increase, the distribution becomes more symmetric. This agrees with theory which says that chisquare is approximately normal for large degrees of freedom.

(b) With only one degree of freedom, the parent distribution is very skewed to the right. The sample means, however, have distributions that are more symmetric. As the sample size increases, the sampling distributions become more symmetric.

(c) As with the other distributions, the spread in the distribution decreases with increasing sample size, and, this is as expected from theory because, even if the underlying distribution is not normal, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

4. The Cauchy distribution is the same as a t distribution with one degree of freedom. The distribution does not have a mean or a variance (i.e., the usual mathematical forms leading to the definition of a distributions mean or variance turn out to be infinite). This will cause some strange behavior in the simulations. In particular, the sample means will often have extremely large deviations from the center of the Cauchy distribution (the distribution is symmetric and does, of course, have a median).

(a) Symmetric with long tails.

(b) The histograms are difficult to interpret. The sampling distribution is so spread out that a histogram, with many equally-spaced cells will have most of its cells (all but 3 or 4) with only one observation in it. Most of the observations from a Cauchy will be near to its median, but some of the sample means will be far, far, away.

(c) See above. The the Cauchy distribution does not meet the conditions for the central limit theorem to hold. Sample means of Cauchy random variables will not, even in very large samples, follow a normal distribution.