

Chapter 5

Sampling Distributions

5.1 Sampling Distributions for Proportions

The location, spread and shape of the distribution of the sample proportion can be obtained by the laws of probability. If \hat{p} represents the sample proportion of successes in an SRS of size n drawn from a large population having population proportion p of successes, then the mean and standard deviation of \hat{p} are

$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

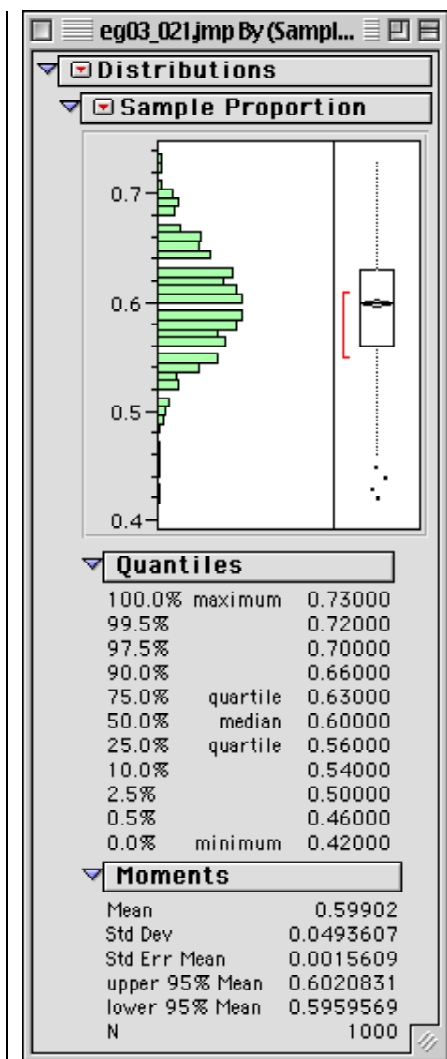
Example 5.1 Are attitudes toward shopping changing?

An SRS of 100 U.S. adults asks whether they agreed or disagreed that “I like buying new clothes, but shopping is often frustrating and time-consuming.” The proportion of the sample that agrees is to be calculated. It is a statistic and varies from sample to sample. The mean and standard deviation of this sample proportion are

$$\mu_{\hat{p}} = p = 0.6$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.6)(0.4)}{100}} = 0.048989$$

In Chapter 3, we conducted a simulation to imitate the process of taking 1000 SRSs of size 100 from a population in which 60% of the individuals agreed with the statement. The resulting distribution of the 1000 sample proportions approximates the sampling distribution of \hat{p} .

We can compare the mean and standard deviation of the simulated distribution with the values that we have just calculated. In Chapter 3, the **Distribution** command gave



The empirical results (mean = 0.59902 and standard deviation = 0.0493607) are very close to the true values. In addition, the shape of the histogram is consistent with the DeMoivre-Laplace theorem—when n is large, the sampling distribution of the sample proportion is approximately normal. As an exercise, you will be asked to construct a normal quantile plot to confirm this.

5.2 The Sampling Distribution of a Sample Mean

The laws of probability can be used to derive the mean and standard deviation of the sampling distribution of \bar{x} of an SRS of size n from a population with mean μ and standard deviation σ . The mean and standard deviation of \bar{x} are

$$\mu_{\bar{x}} = \mu$$

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

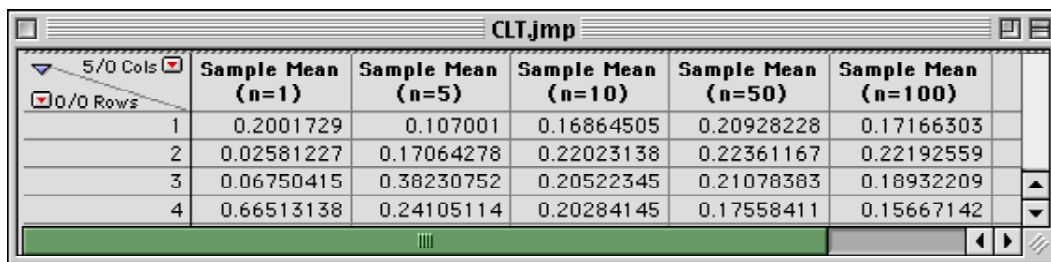
The *Central Limit Theorem* tells us that the distribution of \bar{x} will be approximately normal no matter what the shape of the population distribution as long as the sample size is large enough (and the standard deviation is finite).

To better understand these results and the concept of a sampling distribution, we will simulate the sampling distributions for means of SRSs of different sizes from a very skewed distribution.

Example 5.2 Simulating the sampling distribution of a sample mean

1. Open the JMP data table **CLT.jmp** on the IPS CD-ROM.
2. Select **Rows** ⇒ **Add Rows**.
3. Type **1000** and press **OK**.

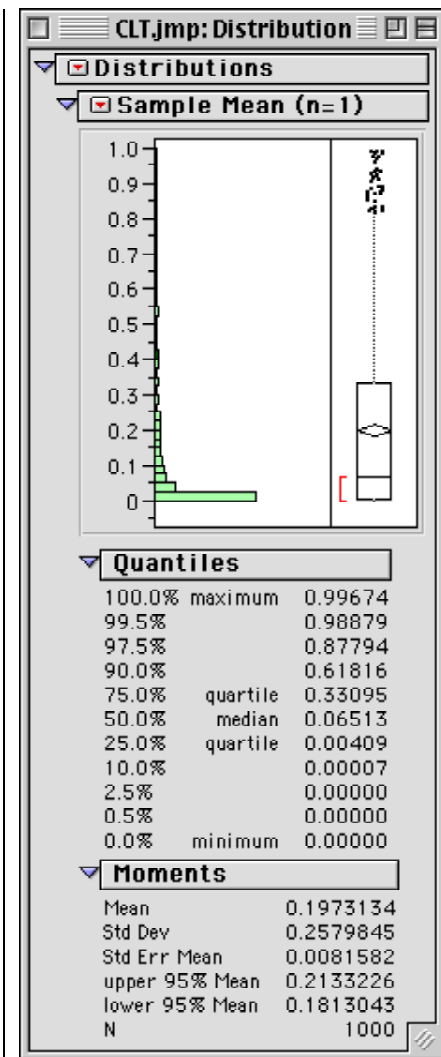
Here is a subset of one possible data table. Since the data values are generated at random, your table will have different values.



	Sample Mean (n=1)	Sample Mean (n=5)	Sample Mean (n=10)	Sample Mean (n=50)	Sample Mean (n=100)
1	0.2001729	0.107001	0.16864505	0.20928228	0.17166303
2	0.02581227	0.17064278	0.22023138	0.22361167	0.22192559
3	0.06750415	0.38230752	0.20522345	0.21078383	0.18932209
4	0.66513138	0.24105114	0.20284145	0.17558411	0.15667142

Each row imitates selecting an SRS from a very skewed distribution. There are five columns. The first column contains the mean for each sample of size 1. Hence, these are random values from the population distribution and their distribution will approximate the population distribution. Let's use JMP to look at it.

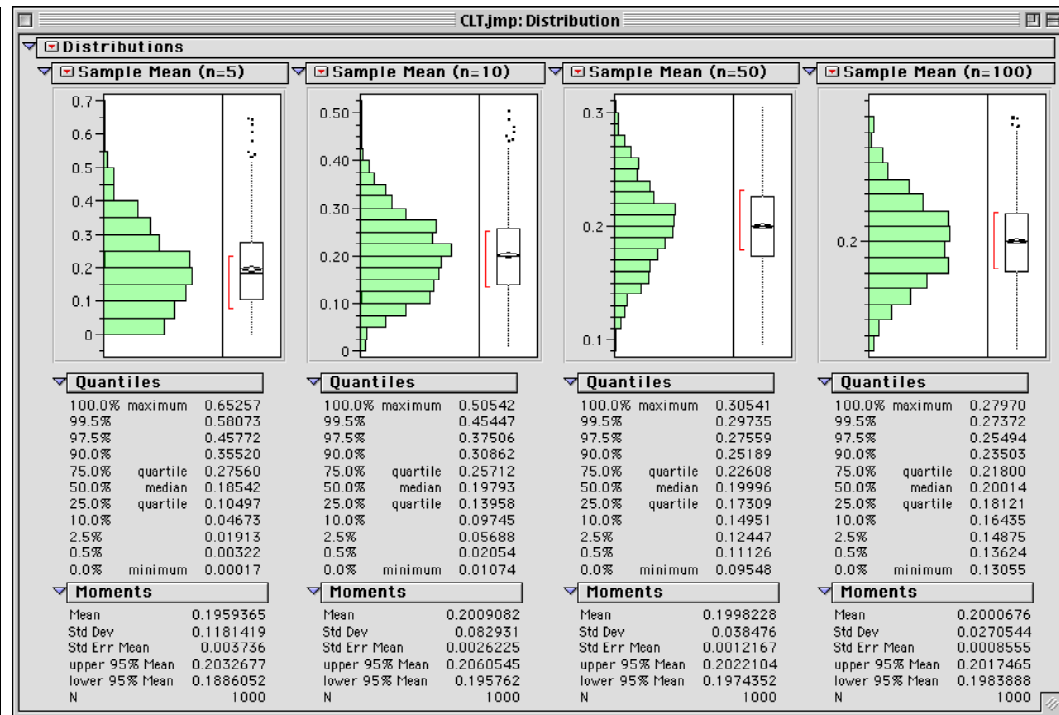
4. Select **Analyze** ⇒ **Distribution**.
5. Select **Sample Mean (n=1)** and press **Y, Columns** and **OK**.



Thus, the mean μ and the standard deviation σ of the population distribution are approximately 0.1973134 and 0.2579845, respectively. (Of course, since the data values are generated at random, your results will differ slightly). You can also see that the distribution is strongly skewed toward larger numbers. Waiting times for service (e.g., a highway tollbooth lane) often follow such a distribution.

Now, let's look at the distributions of the sample mean for SRSs of sizes 5, 10, 50 and 100.

6. Select **Analyze** ⇒ **Distribution**
7. Select **Sample Mean (n=5)**, ..., **Sample Mean (n=100)** and press **Y, Columns** and **OK**.



First, look at the means for each distribution and compare them to the mean of the first column, which approximates the population mean. Notice that they are almost equal.

Second, consider the standard deviations of these distributions, 0.1181419, 0.082931, 0.038476, and 0.0270544, respectively. Note that they become progressively smaller as expected. According to the formula, these standard deviations should each be approximately $1/\sqrt{n}$ times the population standard deviation. For the simulation shown here, the population standard deviation is approximately 0.2579845 and so the standard deviations of these sampling distributions should have been approximately 0.115374, 0.081582, 0.036485, and 0.025798, respectively. They are remarkably close to the corresponding empirical standard deviations. How do the empirical standard deviations of your simulation compare with the theoretical results?

Now look at the shapes of the distributions. Notice that they tend to look more like a normal curve as the sample size increases even though the population distribution is very skewed. That's what the famous *Central Limit Theorem* says!

5.3 Exercises

Are attitudes toward shopping changing? In Example 3.21 and 5.1, you simulated the sampling distribution of the sample proportion of the survey respondents who agreed with the statement. Use JMP to construct a normal quantile plot to confirm that the normal approximation is reasonable.