4.1 Sampling Distributions

Definition. A *parameter* is a number that describes the population. In statistical practice, the value of a paramenter is not known. A *statistic* is a number that can be computed from the sample of data without use of any unknown pararmeters. In practice, we often use a statistic to estimate an unknown parameter.

Sampling Variability

Definition. The fact that the value of a statistic varies in repeated random sampling is called *sampling variability.*

Note. To see what would happen if we take many samples:

- Take a large number of samples from the same population.
- Calculate the sample proportion $\hat{p}$ for each sample.
- Make a histogram of the values of $\hat{p}$.
- Examine the distribution displayed in the histogram for overall pattern, center and spread, and outliers or other deviations.

Definition. Using random digits from a table or computer software to imitate chance behavior is called *simulation.*

Definition. The *sampling distribution* of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

The Bias of a Statistic

Definition. A statistic used to estimate a parameter is *unbiased* if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The Variablity of a Statistic

Example 4.3. The sampling distribution of $\hat{p}$ for samples of size 100, shown in Figure 4.4(a) (see TM-61) is close to the normal distribution with mean 0.6 and standard deviation 0.05. Recall the 68-95-99.7 rule for normal distributions. It says that 95% of values of $\hat{p}$ fall within two standard deviations of the mean of the distribution. So 95% of all samples give an estimate $\hat{p}$ between

$$\text{mean } \pm (2 \times \text{ standard deviation}) = .6 \pm (2 \times .05) = .6 \pm .1.$$

For samples of size 2500, Figure 4(b) (see TM-61) shows the standard deviation is only 0.01. So 95% of these samples will give an estimate within about 0.02 of the mean, that is, between 0.58 and 0.62. An SRS of size 2500 can be trusted to give sample estimates that are very close to the truth about the entire population.

Definition. The *variablity* of a statistic is described by the spread of its sampling distribution. The spread is determined by the sampling design and the size of the sample. Larger samples give smaller spead. As long as the population is much larger than the sample (say, at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size.

The Language of Probability

Definition. We call a phenomenon *random* if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions. The *probability* of any outcome of a random phenonenon is the proportion of times the outcome would occur in a very long series of repetitions

Example. If we roll a 6 sided die, the probability of getting a 4 (say) is 1/6. Flip a (fair) coin and the probability of "heads" is 1/2.

Note. Some facts about probability:

- Any probability is a number between 0 and 1.
- All possible outcomes together must have probability 1.
- The probability that an event does not occur is 1 minus the probability that an event does occur.
- If two events have no outcomes in common, the probability that one or the other occurs is the sum of their individual probabilities.

Example. Flip a coin twice. The possible outcomes (called the *sample space*) are: HH, HT, TH, TT. The probability of getting at least one H is 3/4. The probability of getting no H is 1/4.