11 Chapter 7 - Sampling Distributions

Chapter 7 and 8 introduce a set of concepts that are used in inferential statistics. The material covered in these chapters will be applied in Chapters 9 and 10. These concepts are important and require that they are covered separately.

1. Previously, we discussed the characteristics of probability distributions. We found that we can obtain a great deal of understanding about the distribution if we know the values of μ and σ (in case of the normal distribution) and p (for the binomial distribution). These are refer to as the **parameters** of interest.

2. There are a number of good reasons for analyzing a sample of a population versus analyzing the complete population. Cost and obtaining full accessibility to the information that describes the population are the two main reasons. If, on the average, a detailed examination of 10% of a production batch provides the same information as an examination of 100% of each batch, it is economically unreasonable to examine a proportion larger than 10%.

3. The questions we will address in this chapter is, "How does one prepare a sample that reflects the characteristics of the population? "

4. Types of samples

   (a) Random samples

      i. In a simple random sample, each element (or each unit of analysis) has the same chance of being selected as others in the population. A sample that is not random is biased. A biased sample has a lower probability of reflecting the characteristics of the population.

      ii. Other sample techniques from the simple random sample. If used correctly, these methods will not bias the sample.

         A. Stratified random sample - the population is divided into subgroups and samples are extracted from these subgroups. Generally, the subgroups are based on demographic differences. In a simple random sample, a researcher may take 5% of the population. Using a stratified random sample approach, a researcher divides the population into subgroups. Different proportion of the populations are extracted from each subgroup.

         B. Cluster sample - the population is divided into groups to reduce the time or cost of the data collection. An example of cluster sampling is to identify potential survey respondents by location and choose samples of persons living in the same locale.

C. 1 in k systematic random sample - again, this is a cost saving method that is used as long as it does not bias the result. Every kth person is selected from the population.

iii. If a researcher uses poor research methods to generate a random sample, these methods can result in a biased sample. Here are some types of methods that can generate biased samples.

A. Distributing surveys to a random sample and accepting a low response rate. The risk of a low response rate is that the response group doesn't reflect the population.

B. Collection techniques reaches only a subset of the full sample. Even with a 100% response rate, the methods will produce a biased sample.

C. Wording/Interviewer bias. The choice of words used in the survey and the choice of interviewers can bias the results.

iv. Every technique of collecting data (e.g., face-to-face interviews, phone interviews, mail, Internet, etc.) has its limitations. The better quality research groups will use more than one techniques to reduce the risk of systematically excluding persons from their sample. Surveys are pre-tested to determine if any survey or interviewer bias exists.

(b) Non-random methods (convenience sample, judgement sample, quota sampling). Data sets created by any of these methods cannot be used for making inferences.

5. Statistics and Sampling Distributions

(a) When a researcher draws or extracts a sample, she calculates a value called a **statistic**. In some cases, the statistic computed is the mean value ($\bar{x}$), in other cases, it is the proportion of sample fitting a given characteristic (p).

(b) For a given statistic, the question to consider is, "If one were to draws samples from this population, what is the probability that one will obtain the same value calculated by this researcher?" A **Sampling distribution** is the probability distribution for possible values of the statistic when random samples of size n are drawn from the population.

(c) The book describes three ways of obtaining a sampling distribution.

i. Derive the distribution mathematically using the laws of probability (Examples 7.3 and tables 7.5 are examples of this method).

ii. Approximate the distribution empirically by drawing a large number of samples. Suppose one wanted to work with a sample of 100 neighborhoods ("census tract is equivalent to a neighborhood") from the Year 2000 US Census. The information of interest is the mean rental value of the two bedroom apartments in the neighborhoods. To obtain a sampling distribution, one would randomly extract a sample of 100 neighborhoods, compute the mean rental value for the sample and repeat this step many times. The histogram of describing the distribution of mean rent values would approximate the sampling distribution for this research.

iii. Use statistical theorems (such as the Central Limit theorem) to derive exact or approximate distributions.

6. The Central Limit Theorem

   (a) If random samples of n observations are drawn from a non-normal population with finite mean μ and standard deviation σ, then, when n is large, the sampling distribution of the same mean $\bar{x}$ is approximately normally distributed, with mean and standard deviation (also known as the standard error of the mean (SE)).

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

   Under certain conditions, the means of random samples drawn from a population tend to approximate a normal distribution.

   (b) Conditions

      i. If the population can be represented by a normal distribution, the sampling distribution of $\bar{x}$ will be normal.

      ii. If the population can be represented by a symmetric distribution, the sampling distribution of $\bar{x}$ becomes normal for small values of n (for samples that are small relative to the population).

      iii. If the population can be represented by a skewed distribution, the sampling distribution of $\bar{x}$ becomes normal for large values of n(for samples that are large relative to the population).

   (c) Key theorem for inferential statistics - if the sample size is large enough (given the shape of the distribution) and the conditions needed for the Central Limit Theorem hold ...

      i. the researcher with the non-normal population can make use of the set of statistical tools only available for the normal distribution.

7. Tools for determining probabilities that a sample mean $\bar{x}$ is of a given value.

   (a) Find the population mean and standard deviation (μ, σ).

   (b) Compute the mean and standard deviation of the sample distribution ($\mu_{\bar{x}}$, $\sigma_{\bar{x}}$).

(c) Convert $\bar{x}$ to a z score using the following function

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

(d) Use the table in the back of the book to calculate the probability.

8. Sampling distribution of the sample proportion

(a) Recall from the previous chapter → Let x be a binomial random variable with n trials and probability p of success. The probability distribution of x approximates the normal with $\mu = np$ and $\sigma = \sqrt{npq}$.

(b) There is a similar outcome in sampling. Let's assume that the sampling distribution has the following characteristics. For a sample, the probability of successes is equal to the number of person with this characteristics (x) over the total number of persons in the sample (n) or

$$\hat{p} = \frac{x}{n}$$

where $\hat{p}$ is the probability of success derived from the sample.

(c) For the sampling distribution

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

where q = 1 - p.
If np > 5 and nq > 5, the sampling distribution can be approximated by the normal distribution.

i. Convert $\hat{p}$ into a z score and calculate the probability.

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$