Non-Probability Sampling

Nonprobability sampling is a sampling technique in which some units of the population have *zero* chance of selection or where the probability of selection cannot be accurately determined. Typically, units are selected based on certain non-random criteria, such as quota or convenience. Because selection is non-random, nonprobability sampling does not allow the estimation of sampling errors, and may be subjected to a sampling bias. Therefore, information from a sample cannot be generalized back to the population. Types of non-probability sampling techniques include:

Convenience sampling. Also called accidental or opportunity sampling, this is a technique in which a sample is drawn from that part of the population that is close to hand, readily available, or convenient. For instance, if you stand outside a shopping center and hand out questionnaire surveys to people or interview them as they walk in, the sample of respondents you will obtain will be a convenience sample. This is a non-probability sample because you are systematically excluding all people who shop at other shopping centers. The opinions that you would get from your chosen sample may reflect the unique characteristics of this shopping center such as the nature of its stores (e.g., high end-stores will attract a more affluent demographic), the demographic profile of its patrons, or its location (e.g., a shopping center close to a university will attract primarily university students with unique purchase habits), and therefore may not be representative of the opinions of the shopper population at large. Hence, the scientific generalizability of such observations will be very limited. Other examples of convenience sampling are sampling students registered in a certain class or sampling patients arriving at a certain medical clinic. This type of sampling is most useful for pilot testing, where the goal is instrument testing or measurement validation rather than obtaining generalizable inferences.

Quota sampling. In this technique, the population is segmented into mutually-exclusive subgroups (just as in stratified sampling), and then a non-random set of observations is chosen from each subgroup to meet a predefined quota. In **proportional quota sampling**, the proportion of respondents in each subgroup should match that of the population. For instance, if the American population consists of 70% Caucasians, 15% Hispanic-Americans, and 13% African-Americans, and you wish to understand their voting preferences in an sample of 98 people, you can stand outside a shopping center and ask people their voting preferences. But you will have to stop asking Hispanic-looking people when you have 15 responses from that subgroup (or African-Americans when you have 13 responses) even as you continue sampling other ethnic groups, so that the ethnic composition of your sample matches that of the general American population. **Non-proportional quota sampling** is less restrictive in that you don't have to achieve a proportional representation, but perhaps meet a minimum size in each subgroup. In this case, you may decide to have 50 respondents from each of the three ethnic subgroups (Caucasians, Hispanic-Americans, and African-Americans), and stop when your quota for each subgroup is reached. Neither type of quota sampling will be representative of the American population, since depending on whether your study was conducted in a shopping center in New York or Kansas, your results may be entirely different. The non-proportional technique is even less representative of the population but may be useful in that it allows capturing the opinions of small and underrepresented groups through oversampling.

Expert sampling. This is a technique where respondents are chosen in a non-random manner based on their expertise on the phenomenon being studied. For instance, in order to understand the impacts of a new governmental policy such as the Sarbanes-Oxley Act, you can sample an group of corporate accountants who are familiar with this act. The advantage of this approach is that since experts tend to be more familiar with the subject matter than non-experts, opinions from a sample of experts are more credible than a sample that includes both

experts and non-experts, although the findings are still not generalizable to the overall population at large.

Snowball sampling. In snowball sampling, you start by identifying a few respondents that match the criteria for inclusion in your study, and then ask them to recommend others they know who also meet your selection criteria. For instance, if you wish to survey computer network administrators and you know of only one or two such people, you can start with them and ask them to recommend others who also do network administration. Although this method hardly leads to representative samples, it may sometimes be the only way to reach hard-to-reach populations or when no sampling frame is available.

Statistics of Sampling

In the preceding sections, we introduced terms such as population parameter, sample statistic, and sampling bias. In this section, we will try to understand what these terms mean and how they are related to each other.

When you measure a certain observation from a given unit, such as a person's response to a Likert-scaled item, that observation is called a response (see Figure 8.2). In other words, a **response** is a measurement value provided by a sampled unit. Each respondent will give you different responses to different items in an instrument. Responses from different respondents to the same item or observation can be graphed into a **frequency distribution** based on their frequency of occurrences. For a large number of responses in a sample, this frequency distribution tends to resemble a bell-shaped curve called a **normal distribution**, which can be used to estimate overall characteristics of the entire sample, such as sample mean (average of all observations in a sample) or standard deviation (variability or spread of observations in a sample). These sample estimates are called **sample statistics** (a "statistic" is a value that is estimated from observed data). Populations also have means and standard deviations that could be obtained if we could sample the entire population. However, since the entire population can never be sampled, population characteristics are always unknown, and are called **population parameters** (and not "statistic" because they are not statistically estimated from data). Sample statistics may differ from population parameters if the sample is not perfectly representative of the population; the difference between the two is called **sampling error**. Theoretically, if we could gradually increase the sample size so that the sample approaches closer and closer to the population, then sampling error will decrease and a sample statistic will increasingly approximate the corresponding population parameter.

If a sample is truly representative of the population, then the *estimated* sample statistics should be identical to corresponding *theoretical* population parameters. How do we know if the sample statistics are at least reasonably close to the population parameters? Here, we need to understand the concept of a **sampling distribution**. Imagine that you took three different random samples from a given population, as shown in Figure 8.3, and for each sample, you derived sample statistics such as sample mean and standard deviation. If each random sample was truly representative of the population, then your three sample means from the three random samples will be identical (and equal to the population parameter), and the variability in sample means will be zero. But this is extremely unlikely, given that each random sample will likely constitute a different subset of the population, and hence, their means may be slightly different from each other. However, you can take these three sample means and plot a frequency histogram of sample means. If the number of such samples increases from three to 10 to 100, the frequency histogram becomes a sampling distribution. Hence, a sampling

distribution is a frequency distribution of a *sample statistic* (like sample mean) from a *set of samples*, while the commonly referenced frequency distribution is the distribution of a *response* (observation) from *a single sample*. Just like a frequency distribution, the sampling distribution will also tend to have more sample statistics clustered around the mean (which presumably is an estimate of a population parameter), with fewer values scattered around the mean. With an infinitely large number of samples, this distribution will approach a normal distribution. The variability or spread of a sample statistic in a sampling distribution (i.e., the standard deviation of a sampling statistic) is called its **standard error**. In contrast, the term standard deviation is reserved for variability of an observed response from a single sample.

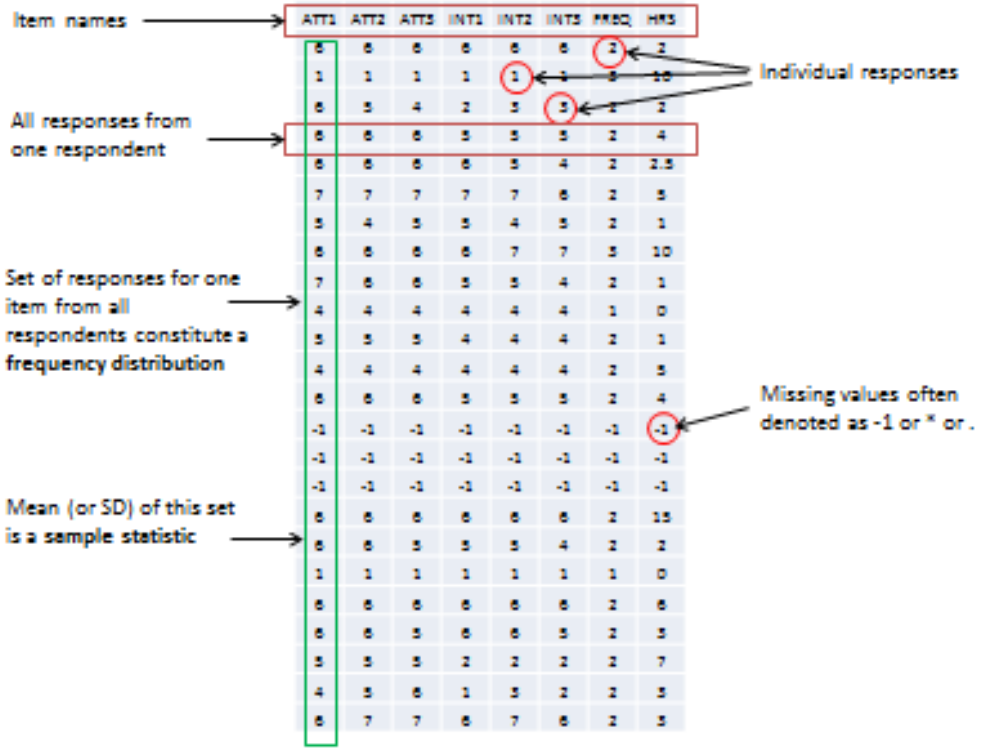

Figure 8.2. Sample Statistic

The mean value of a sample statistic in a sampling distribution is presumed to be an estimate of the unknown population parameter. Based on the spread of this sampling distribution (i.e., based on standard error), it is also possible to estimate confidence intervals for that prediction population parameter. **Confidence interval** is the estimated probability that a population parameter lies within a specific interval of sample statistic values. All normal distributions tend to follow a 68-95-99 percent rule (see Figure 8.4), which says that over 68% of the cases in the distribution lie within one standard deviation of the mean value ($\mu \pm 1\sigma$), over 95% of the cases in the distribution lie within two standard deviations of the mean ($\mu \pm 2\sigma$), and over 99% of the cases in the distribution lie within three standard deviations of the mean value ($\mu \pm 3\sigma$). Since a sampling distribution with an infinite number of samples will approach a normal distribution, the same 68-95-99 rule applies, and it can be said that:

- (Sample statistic $\pm$ one standard error) represents a 68% confidence interval for the population parameter.

- (Sample statistic ± two standard errors) represents a 95% confidence interval for the population parameter.
- (Sample statistic ± three standard errors) represents a 99% confidence interval for the population parameter.

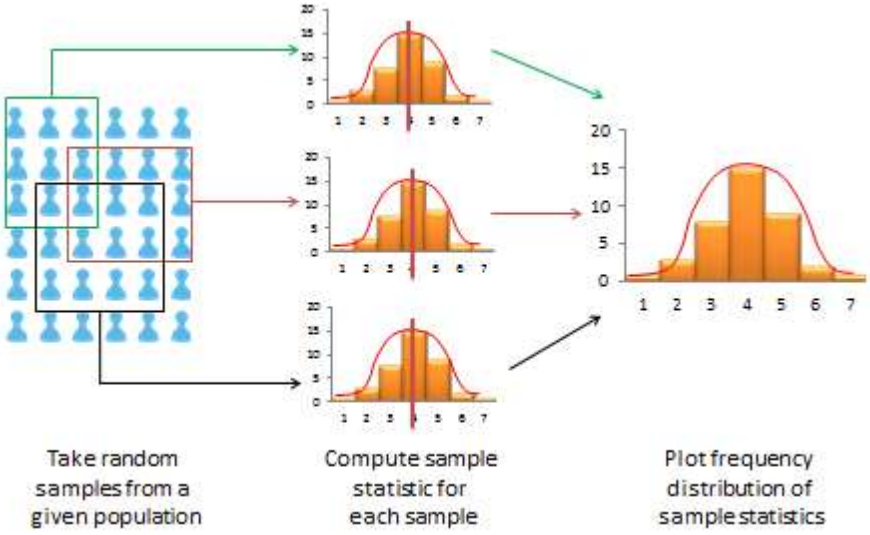

Figure 8.3. The sampling distribution

A sample is "biased" (i.e., not representative of the population) if its sampling distribution cannot be estimated or if the sampling distribution violates the 68-95-99 percent rule. As an aside, note that in most regression analysis where we examine the significance of regression coefficients with $p<0.05$, we are attempting to see if the sampling statistic (regression coefficient) predicts the corresponding population parameter (true effect size) with a 95% confidence interval. Interestingly, the "six sigma" standard attempts to identify manufacturing defects outside the 99% confidence interval or six standard deviations (standard deviation is represented using the Greek letter sigma), representing significance testing at $p<0.01$.

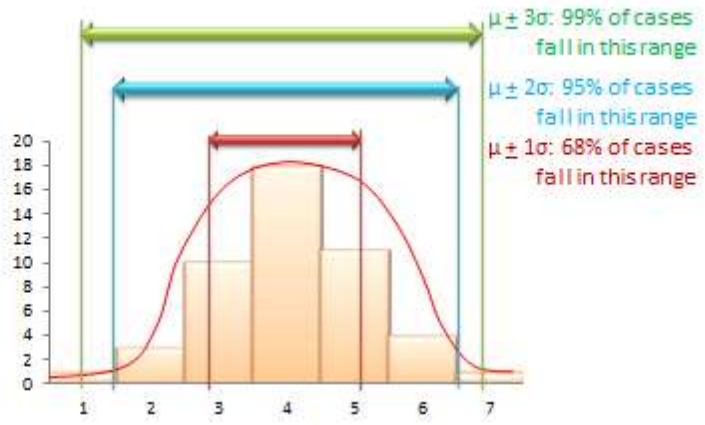

Figure 8.4. The 68-95-99 percent rule for confidence interval