Very Brief Introduction to Statistical Sampling and Confidence Intervals

Source: *For All Practical Purposes*, COMAP, seventh edition, chapter 7.

Suppose you have a large population (e.g., adults in the U.S.) and you wish to estimate a particular parameter (such as the proportion of people who find shopping frustrating). Suppose you select a truly random sample of 2500 inhabitants, and 66% agree that they find shopping frustrating. What can we infer about the proportion of adults in the U.S. who find shopping frustrating?

Quoting now from *FAPP*:

Simple Random Sample. A *simple random sample (SRS)* of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

Parameters and Statistics. A *parameter* is a number that describes the *population.* A parameter is a fixed number, but in practice we do not know its value. A *statistic* is a number that describes a *sample.* The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

Sampling Distribution. 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.

Sampling Distribution of a Sample Population. Choose an SRS of size n from a large population that contains population proportion p of successes. Let $\hat{p}$ be the sample proportion of successes,

$$\hat{p} = \frac{\text{count of successes in the sample}}{n}.$$

Then:

- **Shape:** For large sample sizes, the sampling distribution of $\hat{p}$ is *approximately normal.*
- **Center:** The *mean* of the sampling distribution is p.

- **Spread:** The *standard deviation* of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Confidence Interval. A *95% confidence interval* is an interval obtained from the sample data by a method that in 95% of all samples will produce an interval containing the true population parameter.

Choose an SRS of size n from a large population that contains an unknown proportion p of successes. A *95% confidence interval for* p is

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Here $\hat{p}$ is the proportion of successes in the sample and $2\sqrt{\hat{p}(1-\hat{p})/n}$ is the *margin of error*.

This recipe is only approximately correct but is quite accurate when the sample size n is large.

Applying the above to the shopping example we obtain a confidence interval of 0.6 ± 0.0098.