Chapter 1 – What is Statistics?

Read Chapters 1 and 2 in the Text

Harris Poll example (pp.1-2) – The n = 1256 selected adults (*subjects* or *units*) constitute the *sample*, and all (N ≈ 200 million) U.S. adults is the corresponding *population* to which we wish to make an *inference*. Here, we wish to estimate π, the *population proportion* of "yes" responders in the population, and we can use the *sample statistic* p = 57% to estimate π, although we prefer to give a *confidence interval* instead of a *point estimate*. Note that in this example, the variable of interest is *dichotomous* (yes-no). Provided that the sample is a *random sample* (i.e., without *bias*), p = 57% is probably a good estimator of π.

In Statistics,

- **the proper statistical analysis is driven by the nature of the measured (*response* or *dependent*) variable(s)**
- **we use inference to make intelligent statements about population parameters based on sample statistics; in contrast, Mathematics and Sampling Theory use *deduction***
- **the population is often modeled using a *Box Model***

Simulation – For example, suppose that you know that in a population of individuals, π = 0.60 are Female, and you take a random sample (with replacement) from the population of n = 25 people from this population. If you observe y = 14 Females in your sample, then p = 14/25 = 0.56 for this sample. The Box Model in this case is a box with a 60% "1" tickets (corresponding to Females) and 40% "0" tickets (corresponding to Males); from this box, we take n = 25 draws with replacement. The number of tickets in the box is irrelevant (since we sample with replacement) – just that 60% are "1" and 40% are "0" tickets. Notice that the box has nothing to do with our sample (where we saw 56% "1" tickets or Females), and everything to do with the population.

Since we know that $\pi = 0.60$, in this instance the random sampling error (or *random error*) is $p - \pi = 0.56 - 0.60 = -0.04$ (–4%). On p.4, BF continue this process three more times, repeatedly taking samples of size n = 25, and getting additional errors of 4%, –8%, and 20%.

Error Analysis - These random errors can be combined into the *root mean square* (RMS) as on p.5:

$$\text{RMS} = \sqrt{\frac{(-4\%)^2 + (+4\%)^2 + (-8\%)^2 + (+20\%)^2}{4}} = 11.14\%$$

As mentioned on p.5, the RMS is preferred to the MA (mean of the absolute values) due to reasons of differentiability (smoothness). This is the RMS of just four random samples – the RMS of all possible random samples (of the same size, n = 25) is called the standard error (SE). Later, we'll show that $SE_p = \sqrt{\frac{\pi(1-\pi)}{n}}$; SE_p gives us a measure of the expected variability in the p's.

SE's for uncorrelated measurements A and B are combined using the Pythagorean approach on pp.6-7, so that the SE of (A + B) is

$$\sqrt{SE_A^{\,2} + SE_B^{\,2}}$$

Hypothesis Testing and Decision Making – In Statistics, for example when we want to see if a new anti-hypertension drug is better than the old one, we assume no difference between the two (this is called the *Null Hypothesis*), and we then assess how rare or unusual is the data (reality) in the face of this assumption. If the data is assessed to be rare or unlikely, than we can conclude that there is a difference between the two drugs (the *Alternative Hypothesis*). Thus, Statistics uses a 'proof by contradiction' approach and 'reductio ad adsurdum' reasoning.