

1.5 Collecting Sample Data

*** If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.

*** Randomness plays a crucial role in determining which data to collect.

Two common sources to get data

- **Observational Study:** we observe and measure specific characteristics without attempting to *modify* the subjects being studied
- **Experiment:** we apply some *treatment* and then observe its effects on the subjects

Key elements in design of experiment:

Control effects of variables through:

- Blinding: Subject does not know whether he or she is receiving the treatment or a placebo
- Double Blind: • Blocks • Completely randomized experimental design
- Rigorously controlled design • Matched Pairs Design: using twins

Methods of Sampling

- **Random sample:** Members of the population are selected in such a way that each *individual member* in the population has an equal chance of being selected.
- **Probability Sample:** Each subject has a known (but not necessarily the same) chance of being selected.
- **Systematic sampling:** Select some starting point and then select every *k*th (such as every 50th) element in the population
- **Convenience sampling:** Use results that are readily available or very easy to get
- **Stratified sampling:** Subdivide the population into subgroups that share the same characteristics (such as gender or age bracket), then draw a simple random sample from each subgroup
- **Cluster sampling:** Divide the population into sections (or clusters); randomly select some of those clusters, and then choose *all* members from the selected clusters
- **Multistage Sample Design:** Using a combination of basic sampling methods

Definitions

- **Confounding:** occurs in an experiment when you are not able to distinguish among the effects of different factors
- **Replication:** used when an experiment is repeated on a sample of subjects that is large enough so that we can see the true nature of any effects (instead of being misled by erratic behavior of samples that are too small)
- **Sampling Error:** the difference between a sample result and the true population result; such an error results from chance sample fluctuation (not human or mechanical error)
- **Non sampling Error:** sample data that is incorrectly collected, recorded or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)