Sampling Terminology

Components of a data set

Element — Entities on which data is collected. Also, "*individual.*"

- *e.g.,* In a population study, each person in the sample is an element

Variable — Characteristic of interest for the elements.

- *e.g.,* the age and height of each person are variables

Observation — The set of measurements collected for a particular element.

Statistic vs Parameter

Parameter — A calculated value for an entire population (μ is a parameter)

Statistic — A calculated value for a sample of the population ($\bar{x}$ is a statistic)

Types of attributes

Nominal (Categorical)

Attribute is a name; no ordering is implied

- Football jersey numbers; player #16 is not necessarily better than player #1
- Person names; "Bill" is not a better name than "Cholmondeley." (Well, alright, it is.)

Ordinal — Ordered, but the distance between ranks is not significant

- Amount of pain a patient feels. (1 < 2 < 3, but the difference in pain between 1 and 2 isn't necessarily the same as between 2 and 3)
- Movie ratings

Interval — Ordered; the distance between ranks is significant

- Temperature (the difference between 10° and 30° is the same as between 40° and 60°)

Ratio — Ordered; the ratio between ranks is significant; attribute has an absolute zero value.

- *i.e.,* It makes sense to say that rank 2 is twice the value of rank 1.
- Height, weight, degrees Kelvin (but not Celsius).

Sampling types

Random — Sample is randomly selected from the population. Every member of the population has an equal chance of being selected.

- *e.g.,* Pick 10 names from a hat containing all the names of a class.

Systematic — Every *n*th member of the population is selected.

Convenience — Sample selected because it's readily available.

- *e.g.,* Select the first 10 students to leave the classroom.

Cluster — Population divided into groups (clusters).
Randomly select a collection of clusters; measure every element in the selected clusters.

Stratified — Population is divided into groups (***strata***) by some characteristic that could influence the variable being measured (*e.g.,* divided into males and females).
A sample is taken from every stratum.

Self-selected (Voluntary)
Sample members volunteer.

Bias

Unbiased — If a large set of measurements is unbiased, then the mean will be close to the true value.

Selection bias — Bias due to unrepresentative samples

- Undercoverage - Some population members not adequately represented in the sample
- Non-response - Individuals don't respond. Those who don't repond differ in meaningful ways from those who do. Also, low response rate, so that the survey response likely doesn't represent the entire population.
- Voluntary Response Bias- Sample members are self-selected volunteers.

Response bias — Bias resulting from problems in the measurement process.

- Leading Questions - Wording of the question invites a particular response
- Social Desirability - People answer in such a way as to cast them in a favorable light.