**TOPIC 3**

**DRAWING CONCLUSIONS FROM STUDIES**

**Population** – entire group of interest

**Sampling Frame** – group from which the sample is extracted

**Sample** – Observational Units

**Census** - Sample name if Observational Units are entire population

Population ≥ Sampling Frame ≥ Sample[[1]](#footnote-1)

If one wants to do a study to find relationships between characteristics in a population, they need to

1. Decide on a Sampling Frame
2. Sample the population (from that frame)

When we want to find directional relationships between variables, often variables play one of two roles

**Explanatory Variable** – variable whose effect is being studied (aka *independent* variable)

**Response Variable** – variable whose response to the explanatory variable (aka *dependent* variable)

This relationship is often straightforward when the explanatory variable occurs before the response variable. If characteristics are measured simultaneously, however, the relationship may not be obvious.

**Population Segment –** a portion of a population that has a specific value of a characteristic

**Representative Sample –** a sample withvalues of characteristic similar to the population

**Sampling Bias** – Samples that overrepresent or underrepresent population segments

**Convenience Sample** – a sample that is easy to obtain

**Voluntary Response** – population members decide for themselves to respond or not

**Nonresponse –** in a voluntary sample can even ruin an unbiased sample plan

Things that can go wrong that will affect the results:

**Biased Sample**

A sample can be biased, which means that with respect to characteristics that may affect the variable under study, the sample does not represent the population. (i.e. the proportion of characteristics in the sample is different than that of the population). One must work hard to try to eliminate bias from a sampling process.

Example: Surveys regarding new technology done only by landline might obtain samples from a population that is older or reluctant to adopt new technologies

 **Lurking Variables**

 A variable may affect the response variable, but may not be considered in the study.

  **Confounding Variable[[2]](#footnote-2)**

A variable for which one can not distinguish the effects from that of the explanatory variable. i.e., we don’t know which variable is causing the effect. More concisely, it is a variable which correlates with both the explanatory and response variables.

Mitigate the effects of these problems:

**Control** for the effect of variables that might affect the response variable. Often these variables are held constant.

**Randomization** is very important for mitigating the effects of confounding variables. Anyway, you should randomize. When you are done randomizing, randomize some more. Keep randomizing until you think you might faint. If you don’t randomize, then a bias response can be assumed feasible.

 Examples:

Sometimes you can’t control the sampling process. I this case you can’t draw cause-and-effect conclusions. This is an **Observational Study .**

**Symbols**

 Population Sample

 *Parameter*[[3]](#footnote-3) [[4]](#footnote-4) *Statististics*[[5]](#footnote-5)

Size N n

Proportion π $\hat{p}$

Mean µ $\overbar{x}$

Standard Deviation σ s

1. ≥ means “greater than or equal to” [↑](#footnote-ref-1)
2. To “confound” the situation, these terms (lurking and confounding) are used interchangeably in the real world. You will have to figure out how they are used in your domain, that’s all. [↑](#footnote-ref-2)
3. Population Parameter values are almost never known. One requires a very narrow population. N sometimes is known, but rarely the others. [↑](#footnote-ref-3)
4. Greek letters are pronounced ‘pie’ (pi), ‘myu’ (mu), and sigma, respectively [↑](#footnote-ref-4)
5. pronounced p-hat, and x-bar respectively. [↑](#footnote-ref-5)