Statistical Inference — Major Concepts

Make sure that you understand all of the terms in bold.

Populations versus Samples

- **Statistical inference:** using **samples** to understand **populations**.
- **Sample error** is the gap between a sample **statistic** and the popluation **parameter**.
- There are two sources of sample error: **bias** and **random error**.

Bias

- Large samples don't fix bias.
- Only way to avoid sample bias is to take a **simple random sample** (SRS).
- Even with an SRS, you still need to avoid other (non-sample) sources of bias.

Random Error

- Larger sample have less random error (because of the **law of large numbers** and the **central limit theorem**).
- Math (**confidence intervals** and **hypothesis tests**) can quantify random error.

Confidence Intervals

- Use these to **estimate** a population parameter.
- The **confidence level** tells you how confident you are that the interval contains the relevant population parameter.
- A confidence interval formula typically has the form:

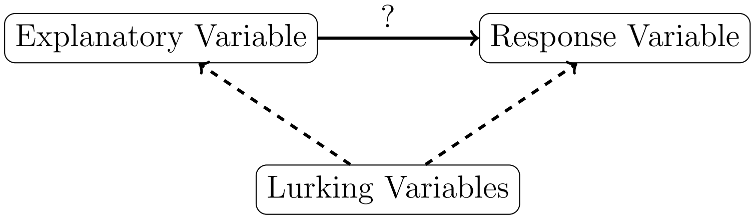
  best guess ± **margin of error**.

Hypothesis Tests

- Can answer a yes/no question.
- A **null hypothesis** must be a specific claim about the population.
- Each of the following are equivalent:
  1. The **test statistic** (such as a **z-value** or a **t-value**) is extreme.
  2. It has a low **p-value** (lower than the **significance level** which is usually 5%).
  3. You should reject the null hypothesis.
  4. The result is **statistically significant**.
  5. The result probably wasn't due to random chance.

Association is not Causation

- The only way to establish causation is with **randomized controlled experiments**. You can **control** all **lurking variables** by **randomly assigning** the individuals to different **treatment groups**.
- **Observational studies** can't rule out all lurking variables.



- A lurking variable that is associated with both the **explanatory** and **response variables** is called a **confounding variable**.