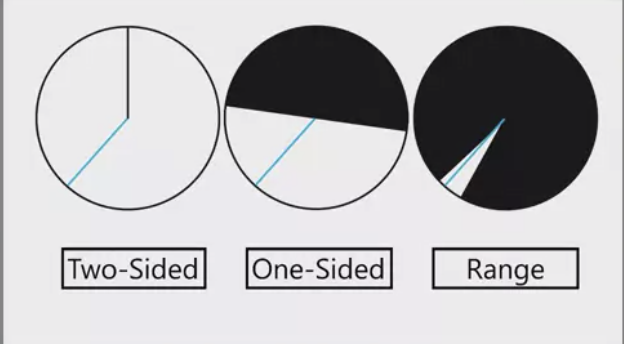
**Week 1**

1. Improving your statistical questions
   1. **Type 3 error**: “The error or probability, of having solved the wrong problem”
      1. Mitroff & Featherington, 1974
   2. Types of modeling
      1. Descriptive
         1. Statistical models to summarize data
         2. Focus on measurement and descriptions, rather than constructs or theory
      2. Explanatory
         1. Statistical models to test causal explanations derived from theories
         2. Theory ⬄ Statistical hypothesis ⬄ Data ⬄
         3. Sample size justifications based on statistical power
         4. Well controlled experiments
         5. Interpreting each significant causal factor
      3. Predictive modeling
         1. Statistical models to predict future or otherwise unknown observations
         2. Train-test splits
         3. Large hold-out datasets
         4. Messy datasets due to real world observations
         5. Difficulty in interpreting models, significant factors can be ignored if it improves prediction!
      4. Exploratory research
         1. What’s going on here?
         2. Example: what happens if you ask people in the subway to give up their seat (Milgram & Sabini, 1978)
            1. Implicit norms in society
         3. Tentative model building
            1. Often through data visualization
   3. Acknowledging state of research
      1. Not always theory-driven experimentation, messier than this scientific ideal
      2. Cyclical approach to theory ⬄ experiment
   4. Statistician fallacy:
      1. Statisticians telling researchers what you want to know
      2. Often happens with likelihoods (i.e., statisticians telling you what you really want to know is the likelihood of the null)
         1. When it is entirely plausible researchers are only concerned with the direction, rather than the effect size, of an effect
      3. Solution: justify the question you’re asking!
2. Do you really want to test a hypothesis?
   1. Hypothesis testing
      1. Metaphor of dart games
         1. Methodological procedure to decide who is better than the rest
         2. A single dart can determine who wins or loses
            1. Might be trivial difference, like the p-value threshold, but we have to have some arbitrary standard
      2. Divide all possible states of the world into predicted and not predicted
         1. “All swans are white” predicted
         2. Often have probabilistic environments
            1. Specify rejection rules which may render statistical interpreted evidence ‘inconsistent’ with probabilistic theory (Lakatos, 1978)
      3. Methodological falsificationism (Neyman-Pearson hypothesis test)
         1. Goal is to not be too often wrong (Neyman & Pearson, 1933)
         2. Don’t focus too much on the particular case (single study) but rather the method for results (multiple studies / theory)
      4. P-value of 0.05
         1. Fisher – convenient convention dictates we will always see flukes in single studies
         2. Not an isolated record, but reliable method of procedure (replicability)
   2. Why would you want hypothesis testing?
      1. Decision how to act
         1. Act is to accept or reject the null (H0)
         2. A “region of doubt may be obtained by a further subdivision of region of acceptance” (Neyman, 1933)
      2. Are your hypotheses “good hypotheses”
         1. **Randomization** is an important argument for the plausibility of the null
            1. Randomization makes null somewhat likely
         2. **Crud** is a strong argument against the null
            1. Crus is systematic noise (as in the case of correlational research)
            2. Makes null not as important
         3. Null-hypothesis does not have to be 0 (nil)
      3. Can you control the error rates?
         1. Can’t make decisions when error rates are huge
      4. Verisimilitude
         1. Truth likeness of theories
         2. Approach to scientific process needs to be transparent (prereg!)
   3. Hypothesis tests may not be appropriate in some cases
3. Risky predictions
   1. Not all predictions are equally exciting
      1. 
         1. Circle represents all possible outcomes
         2. White area represents values predicted under alternative
         3. Black area represents not predicted (null hypothesis)
         4. Blue line is the observed
   2. Severity requirement (Mayo, 2018)
      1. Weak
         1. One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false
         2. **HARKing** – hypothesizing after results are known
      2. Strong
         1. Passing the strong test provides good evidence for the claim
   3. Classic NHST (two-sided)
      1. Not a specific prediction
   4. Classic NHST (one-sided)
      1. More specific (directional) prediction
   5. Minimal effect test (one-sided)
      1. Minimal effect of interest which is not just 0
      2. Can set a threshold of a positive effect size (e.g., 0.2) and say everything under would be a null result!
   6. Equivalence test (range)
      1. Predict a range around 0 that we are predicting (H1) no meaningful difference between two groups
         1. Say the observed data is equivalent to 0, but doesn’t have to be centered around 0
      2. Null hypothesis is now extremely large and small effect sizes
      3. Theoretically relevant effects to be significant, but not theoretically irrelevant effects
         1. Falsified based on clear criteria
      4. Many criticisms of p-values disappear when p-values are calculated for range predictions
         1. No longer have practically insignificant but statistically significant findings
         2. Not every effect will become statistically significant with large samples
            1. Putting a range around 0, rather than testing against 0, you make your findings more practically meaningful