1. Data (variables / data / regression / standard errors)
   1. Types of variables (cardinal, ordinal, categorical)
   2. Descriptive statistics (how to find mean, median, and standard deviation)
   3. Data wrangling skills in R – eg. how to generate variables/summarize data based on some criteria (PSet 1)
   4. Covariance and correlation coefficient (formula, differences between the two)
   5. For method field people, remember to memorize the formula of how to calculate the slope and intercepts.
2. Probability (Distributions / expected values)
   1. Joint distribution
   2. Marginal distribution
   3. Conditional distribution 🡪 conditional expected values
   4. How to know whether two random variables are independent from each other?
   5. Properties of expected values and variance
      1. If X ~ N(3, 4^2), Y = X+3, then what is the distribution of Y?
      2. If Y = 3X, then what is the distribution of Y?
3. Sampling distribution
   1. Central Limit Theorem
      1. If X ~ Bernoulli(p = 0.6), what are the mean and standard deviation under this distribution and what is the distribution of X\_bar?
      2. If X ~ Binomial(n = 100, p = 0.6), what are the mean and standard deviation under this distribution and what is the distribution of X\_bar?
      3. If X~ Uniform(min = 3, max = 9), what are the mean and standard deviation under this distribution and what is the distribution of X\_bar?
      4. If X ~ Normal(mean = 2, sd = 3), what are the mean and standard deviation under this distribution and what is the distribution of X\_bar?
   2. X\_bar is an unbiased estimator of the true parameter, why? Because E(X\_bar) = X
   3. What are Z statistics? What is the distribution of Z-statistics?
   4. How to use p-functions in R to find cumulative probability? (pbinom, punif, pnorm)
4. Hypothesis testing
   1. Set up null and alternative hypotheses.
   2. What is the distribution under the null hypothesis?
   3. What is the definition of p-value? (thinking about drawing many repeated samples) If we were two draw many repeated samples, the p-value is the proportion of samples that have a difference in average strong leader rating between Republican and Democratic candidates as or more extreme than 0.23.
   4. How to find p-value in one-tailed test? In two-tailed test?
   5. What does a confidence interval tell you?
5. Causal inference
   1. What is the fundamental problem of causal inference? we can only observe one of the two potential outcomes, depending on whether the subject is treated or not?
   2. How to calculate unit-level treatment effect E(Y1i-Y0i)?
   3. How to calculate average treatment effect using actually observed data (E(Yi|Ti=1)-E(Yi|Ti=0))?
   4. What could cause the difference between unit-level treatment effect and average treatment effect?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Obs | Ti | Y0i | Y1i | Y1i-Y0i | Yi |
| 1 | 1 | 12 | 10 | -2 | 10 |
| 2 | 0 | 7 | 3 | -4 | 7 |
| 3 | 0 | 4 | 4 | 0 | 4 |
| 4 | 1 | 12 | 7 | -5 | 7 |
| 5 | 1 | 5 | 4 | -1 | 4 |

Y=illness severity

E(Y1i-Y0i) = -2.4 🡪true effect: going to hospitals does help patients get better

E(Yi|Ti=1)-E(Yi|Ti=0) = 21/3-11/2 = 1.5 🡪 going to hospitals make patients get worse

People who are sick are more likely to go to the hospital 🡪 there is a correlation between the treatment and the potential outcomes