Problem Set 2[1]
Math Camp, Summer 2023, UCSB
Instructor: Seonmin (Will) Heo (sheo@ucsb.edu)
Due: 11:59 PM on Tuesday, Sep 5, 2023

This problem set will help you review the key concepts from the course so far. You are free to use any software but make sure to type the answers and include the code and the results.

1. Prove that
   (a) $S \in \mathcal{B}$ and $\emptyset \in \mathcal{B}$ in a sample space S with a σ-algebra $\mathcal{B}$ on S.
   (Hint: start with that $\mathcal{B}$ should be nonempty.)
   (b) σ-algebra $\mathcal{B}$ on S is closed under countable intersections.

2. Prove Theorem 2.3.

3. Consider testing for the presence of a disease. The test is very accurate in the sense that if a patient has the disease, the test always comes back positive, i.e., $\mathbb{P}(\text{positive}|\text{disease}) = 1$. Sometimes the test is inaccurate, however, in the sense that the test gives a false positive (a positive value when a person doesn't have the disease) with probability 0.005, i.e., $\mathbb{P}(\text{positive}|\text{no disease}) = 0.005$. If the probability of having the disease is 0.001, i.e., $\mathbb{P}(\text{disease}) = 0.001$, what is the probability a patient has the disease, given they have a positive test?

4. A variable X is lognormally distributed if $Y = \ln(X)$ is normally distributed with μ and σ^2, i.e. $f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$. Let $x = g(y) = e^y$ and $y = g^{-1}(x) = \ln(x)$.
   (a) Derive $f_X(x)$.
   (b) Derive $\mathbb{E}[X^t]$ using $M_Y(t)$. What are $\mathbb{E}[X]$ and $V(X)$?

5. Assume that the n units are volunteers to receive the treatment. Given any $i \in \{1, \cdots, n\}$, $D_i = 1$ if treated, and $D_i = 0$ otherwise. Let $(D_1, \cdots, D_n)$ be a vector stacking the treatment indicators of all units. Treatments have capacity constraints and only $n_1 (< n)$ units can be treated: $\sum_{i=1}^{n} D_i = n_1$.
   (a) What is the number of possible values $(D_1, \cdots, D_n)$ can take?

   We say that treatment is randomly assigned if $(D_1, \cdots, D_n)$ are random variables, and if for any vector of n numbers $(d_1, \cdots, d_n) \in \{0, 1\} \times \cdots \times \{0, 1\}$ such that $\sum_{i=1}^{n} d_i = n_1$,

$$P(D_1 = d_1, \cdots, D_n = d_n) = \frac{1}{\binom{n}{n_1}}$$

   That is, random assignment generates uniform treatment probabilities across units.

   (b) If full randomization is satisfied, then for every $i \in \{1, \cdots, n\}$, what is $P(D_i = 1)$?
   (c) If full randomization is satisfied, then for every $i \neq j$, what is $P(D_i = 1 \ \wedge \ D_j = 1)$? Is it true that unit i getting treated is independent from unit j getting treated?

[1] This problem set is created by Woongchan Jeon and modified by Seonmin (Will) Heo.

6. Download the `lbw.dta` dataset from Stata.

    (a) Run a linear regression of low birthweight (`bwt`) on smoking (`smoke`), controlling for age (`age`) weight at the last menstrual period (`lwt`), and history of hypertension (`ht`). Report the coefficients, standard errors, and confidence intervals.

    (b) Estimate the bootstrapped standard error for this OLS. Use the following procedure:

        i. Get a bootstrapped sample, which is a random sample *with replacement*.

        ii. Run the OLS based using this bootstrapped sample b_1. Let's call the coefficient on smoking $\widehat{\beta}_{b_1}$.

        iii. Repeat the process (i)-(ii) $B = 1000$ times.

        iv. Calculate the bootstrapped variance using the following formula:

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}_b - \bar{\widehat{\theta}}_b\right)^2.$$

        v. Construct confidence intervals by obtaining the 97.5th and the 2.5th percentiles of the B bootstrapped coefficients $(\widehat{\beta}_{b_1}, \ldots, \widehat{\beta}_{b_B})$. Are the bootstrapped confidence intervals similar to the those obtained in (a)?