1) A scientist runs an experiment, and assigns people to treatment groups and control groups randomly. The scientist estimates the ATE two ways:

   1. The scientist computes
   $$\widehat{ATE}_1 = \sum_{i=1}^{N} \frac{Y_i T_i}{\#Trt} - \sum_{i=1}^{N} \frac{Y_i(1-T_i)}{\#Con}$$
   where T_i are treatment indicators and $\#Trt$ and $\#Con$ denote how many people were assigned to treatment and control respectively. This estimate is unbiased for the ATE.

   2. The scientist assumes that responses were generated by the model
   $$Y_i = \alpha + \beta T_i + \epsilon_i$$
   where ϵ_i are independent and identically distributed with $\mathbb{E}(\epsilon_i) = 0$. The scientist obtains $\widehat{ATE}_2 = \hat{\beta}$ through OLS. Since OLS estimates coefficients unbiasedly, this method obtains an unbiased estimate for the ATE.

   Both methods estimate that the average treatment effect is large and positive.

   True or False: Both methods obtaining large and positive estimates of the ATE gives more evidence that the ATE is positive than if only one of these methods were used.

2) Consider a large medical trial for a new weight loss drug. Before the trial, each patient has their weight, height, and body fat percentage measured. A goodness-of-health score is calculated for each patient based on those characteristics (higher scores are a proxy for worse health). Assume that patients do not have time to manipulate their weight or body fat once selected to participate in the trial. Historically, a histogram of patient goodness-of-health scores closely follows a normal distribution with mean c. It is thought that the effect of the drug varies with the value of this score.

   a) Consider the following mechanism for treatment assignment: Before being assigned to treatment or control, each patient rolls a 6-sided die. For a patient with a score above c, If the die comes up as 1, 2, 3, or 4 and the patient has a score of c or above, that patient takes the weight loss drug, otherwise they receive a placebo. If the die comes up as a 5 or 6 and the patient has a score below c, that patient takes the weight loss drug, otherwise they receive a placebo. Suppose that the die roll is known to the experimenter. Under this setup, what inferences could you make about the effect of the drug on weight loss? Discuss the parameter of interest and the methods used to estimate this parameter. What assumptions are required to estimate this parameter?

   b) Suppose the same set up in part a), except the result of the die roll is unknown to the experimenter. Can the parameter of interest in part a) still be estimated? If so, how? If not, why not?

   c) Suppose that the effect of treatment is thought to be the same for all patients with goodness-of-health scores within the interval $(c-5, c+5)$. Suppose that patients with scores below c are ineligible to receive the weight loss drug. Patients with scores of c or above are given an appointment to receive the the new drug. The drug is administered only once during the trial, and only at this appointment. Some patients fail to arrive at their appointment. Under this setup, discuss at least two types of inference possible for measuring the effect of the drug on weight loss? Discuss the parameters of interest and the methods used to estimate this parameter. What assumptions are required to estimate these parameters? Which estimate will be larger (in absolute value)?

3) Suppose that there is a study with a total of $2n$ subjects. Exactly n of these people are smokers. A height and weight are measured for each subject. Suppose that there are enough people so that the joint distribution of the heights and weights is extremely close to a multivariate normal distribution. The researcher wants to test whether smoking affects 40-yard dash times.

a) A statistician notices some imbalance in the average weight and height between the smokers and the non-smokers. To fix the imbalance, the statistician matches smokers to non-smokers by matching on the Mahalanobis distance with height and weight covariates (with replacement, nearest neighbor). Will the differences in average height and average weight between the smokers and matched non-smokers be as small or smaller as they were before matching?

b) Suppose instead that all subjects in the study are twins. For each set of twins, one twin is a smoker and one twin is a non-smoker, and both twins in each set have the same height and weight. In his analysis, the statistician believes that the smoking sibling in a twin pair is essentially random, though he concedes that some unobserved trait may help explain a twin's propensity for being a smoker.

For each set of twins s, let $(1, s)$ denote the twin that smokes, and let $(2, s)$ denote the non-smoking twin. Let T_{is} denote random smoking indicators; $T_{is} = 1$ if the ith unit in the sth twin pair smokes, $i = 1, 2$. For this study, for each pair s, the smoking indicators are observed to be $T_{1s} = 1$ and $T_{2s} = 0$. The statistician models the probability that a subject smokes in the following way:

$$\log\left(\frac{P(T_{is}=1)}{1-P(T_{is}=1)}\right) = \alpha + \kappa_1 h_{is} + \kappa_2 w_{is} + \gamma u_{is} \tag{1}$$

where h_{is} and w_{is} are the height and weight of twin (i, s), and u_{is} is the value of an unobserved covariate for that twin. The statistician also assumes that any subject cannot influence any other subject to smoke or not smoke (smoking is independent across all subjects).

Show that, under this model, the probability that subject $(1, s)$ is a smoker is:

$$P(T_{1s} = 1 | T_{1s} + T_{2s} = 1) = \frac{e^{\gamma u_{1s}}}{e^{\gamma u_{1s}} + e^{\gamma u_{2s}}} \tag{2}$$

Hint: Use $P(A|B) = P(A \cap B)/P(B)$, and find an expression for

$$\frac{P(T_{1s} = 1 \cap T_{2s} = 0)}{P(T_{1s} = 0 \cap T_{2s} = 1)} = \left(\frac{P(T_{1s} = 1)}{1 - P(T_{1s} = 1)}\right)\left(\frac{P(T_{2s} = 1)}{1 - P(T_{2s} = 0)}\right)$$

c) Suppose that $0 \leq u_{is} \leq 1$ and that $\gamma > 0$. Find an upper and lower bound (sharper than just 1 and 0) for the probability $P(T_{1s} = 1|T_{1s} + T_{2s} = 1)$. Denote these bounds by p_s^+ and p_s^- respectively. Do the same for $P(T_{1s} = 0|T_{1s} + T_{2s} = 1)$. Comment, in one sentence, on how these bounds change if $\gamma < 0$.

d) Let Y_{is} denote the 40-yard dash time of subject (i, s) in milliseconds. Let Z_s denote an indicator variable for the smoker having the faster 40-yard dash time: $Z_s = 1$ if and only if twin $(1, s)$ had a faster 40-yard dash time than twin $(2, s)$. Let d_s denote the rank of $|Y_{1s} - Y_{2s}|$; higher ranks denote larger absolute values. Assume there are no ties between Y_{1s} and Y_{2s} within any twin pair s, and that $|Y_{1s} - Y_{2s}| \neq |Y_{1t} - Y_{2t}|$ for all distinct twin pairs s, t.

The Wilcoxon signed rank statistic is:

$$W = \sum_{s=1}^{n} d_s Z_s.$$

Let Z_s^+ and Z^- be independent and identically distributed bernoulli random variables (or indicator variables) with $P(Z_s^+ = 1) = p_s^+$ and $P(Z_s^- = 1) = p_s^-$. Consider the following statistics:

$$\begin{aligned} W^+ &= \sum_{s=1}^{n} d_s Z_s^+ \\ W^- &= \sum_{s=1}^{n} d_s Z_s^- \end{aligned}$$

Show that, under the null hypothesis that smoking does not effect 40-yard dash times, the following property holds:

$$\mathbb{E}(W^-) \leq \mathbb{E}(W|T_{1s} + T_{2s} = 1) \leq \mathbb{E}(W^+)$$

e) In fact, it can be shown that under this null hypothesis, for any a:

$$P(T^- \geq a) \leq P(T \geq a|Z_{1s} + Z_{2s} = 1) \leq P(T^+ \geq a) \tag{3}$$

Discuss, in about 3 -5 sentences or so, how property (3) can be exploited to test the exact null of no treatment effect.

Bonus: Prove property (3).