# Assignment 2

Submit your answers on April 3. We will discuss in class on that date, and grades will be returned at the end of term.

## 1 Simulations - Study Design

You have been asked to design a randomized controlled trial of a new treatment. You are told that the expected distribution of outcomes in the placebo arm is $N(0, 50)$.

1. Using a standard formula, calculate the statistical power of this study to detect a difference in means of 1.0, assuming sample sizes of 100, 500, and 1000. Assume a two-sided null hypothesis and $\alpha = 0.05$. (Hint: this is really easy)

2. Estimate the same statistical power using simulation. Simulate the study, with the characteristics above, and compute the power. Justify the number of simulation runs you use (see the attached paper by Morris et al). How do your results compare to those in Q1?

3. Re-run your simulation, but with $\sigma^2 = 100$ in the treatment arm only. How does this change your calculations?

## 2 Simulations - Confounding

1. Read the attached paper by Setoguchi et al.

2. On page 547, the paper outlines a simulation structure. Following their data generation process, using only the n=2000 setting for scenario A (main effects only, all associations linear), generate 1000 simulated datasets.

3. Estimate the crude odds ratio for the outcome $Y$ as a function of exposure $A$ using logistic regression. Present yoru results in an appropriate display (follow Morris for suggestions)

4. Estimate the adjusted odds ratio, adjusting for $W_1, \ldots, W_{10}$.Repeat excluding $W_5 - W_7$. Present your results in an appropriate way. What do you conclude about excluding $W_5 - W_7$?

# 3 Survival Data - Analyses

Let's consider data from patients with advanced lung cancer from the North Central Cancer Treatment Group (Loprinzi et al., 1994). In this study, prognostic variables measured by patient-completed questionnaires were evaluated. We will focus on a binary indicator of ability for self-care: create a binary indicator of having a physician-rated Karnofsky performance score of greater than 80 as the exposure. The data set includes several factors for which the analysis can to be adjusted (why adjust?). The data set can be found from the R survival package, called lung.

1. Fit exponential and Weibull survival models to these lung data (using, e.g., the eha package), and compare the results.

2. Fit a Cox proportional hazards model to these data.

3. Bonus question: Compute at least one type of residuals (Schoenfeld, martingale, Cox- Snell, etc.) for a Weibull (and/or Cox) model. Provide code and a sample residual plot where (i) the model fits the data well, and (ii) the model does not provide a good fit to the data.