Statistics 641, Spring 2018
Take Home Final Exam
Due 4pm, May 9, 2018

1. Suppose that we have 8 subjects in each of two groups and observe the following responses (same as HW 5, problem 2):

   Control ($j = 0$): 0.2, 0.8, 1.9, 2.2, 2.6, 3.9, 8.2, 21.8
   Experimental ($j = 1$): 2.8, 5.1, 7.1, 7.7, 12.3, 18.8, 27.1 39.7

   Denote the the i^{th} observation in group j by Y_{ij} and let μ_j is the *median* for group j and $\theta = \mu_1 - \mu_0$.

   (a) Under the random allocation rule, conduct the randomization test of $H_0\colon \theta = 0$.

   (b) You can test the hypothesis $H_0\colon \theta = \theta_1$ as follows. Transform the observations to $\widetilde{Y}_{i1} = Y_{i1} - \theta_1$ and $\widetilde{Y}_{i0} = Y_{i0}$ with medians $\tilde{\mu}_j$ and $\tilde{\theta} = \tilde{\mu}_1 - \tilde{\mu}_0 = \theta - \theta_1$. Then $H_0\colon \tilde{\theta} = 0$ is equivalent to $H_0\colon \theta = \theta_1$. You can apply the randomization test to the $\widetilde{Y}_{ij}$.

      i. Test the hypothesis $H_0\colon \theta = -3$. Is $\theta = -3$ in a 95% confidence interval for θ?
      ii. Test the hypothesis $H_0\colon \theta = -4$. Is $\theta = -4$ in a 95% confidence interval for θ?
      iii. Test the hypothesis $H_0\colon \theta = -5$. Is $\theta = -5$ in a 95% confidence interval for θ?

      Find the lower limit of a 95% confidence interval for θ.

2. Suppose we conduct a randomized trial and observe

   - binary treatment `z` (0,1)
   - continuous baseline covariate `w`
   - continuous response `y`

   We perform the following analyses (some output deleted):

```
> summary(lm(y~z,data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4295     0.2242   10.84   <2e-16
z             0.7545     0.3170    2.38   0.0199

> summary(lm(w~z,data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9332     0.1678  29.402   <2e-16
z             0.1574     0.2373   0.663    0.509
```

   Finally, we fit the model $y = \alpha + \beta z + \gamma w + \varepsilon$ below. Using the previous analyses and the partial information in the output below, find $\hat{\beta}$ from this model (the `******` in the first column). Is `w` a confounder?

```
> summary(lm(y~z+w,data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3330     0.7620   0.437  0.66341
z             ******     ******   *****  *******
w             0.4250     0.1482   2.867  0.00542
```

3. Suppose that we have two treatments with a 1:1 permuted block randomization with blocks of size 6 (i.e., within each block of 6 we randomly allocate 3 to each treatment). We enroll 12 subjects and in the two blocks we observe the following summary tables:

Group	D	A	Total
1	3	0	3
2	0	3	3
	3	3	6

Group	D	A	Total
1	2	1	3
2	0	3	3
	2	4	6

   (a) Calculate the size of the reference set (all possible allocations of treatments to subjects).

   (b) If x_j is the number of deaths in group 1 for block j, find the sample space for $U(0) = \sum_j x_j - E[x_j]$ and corresponding sampling probabilities under the randomization distribution. (*Hint: x_j has a hypergeometric distribution*).

   (c) Calculate the one-sided randomization p-value for the observed data.

4. Read LaRosa, J.C. *et al.*, (2007) "Safety and efficacy of Atorvastatin-induced very low-density lipoprotein cholesterol levels in Patients with coronary heart disease (a post hoc analysis of the treating to new targets [TNT] study)." *The American Journal of Cardiology*, **100**(5):74752.

   - Interpret Figure 1 in light of Table 1 in this paper and the Figures 2 and 3 of Article 1 (LaRosa et al., 2005) that you read previously. Specifically, what can we say regarding the causal effect of achieved LDL on cardiovascular events. Is achieved LDL a confounder for the causal effect of treatment on the primary outcome?
   - In the discussion (page 751, left column, 3rd full paragraph), the authors note: "This suggests that the observed benefit on major cardiovascular events in TNT is more closely related to achieved on-treatment LDL cholesterol levels than drug dose per se." (Note that by "covaried for" the authors mean "adjusted for.")

     Comment on whether this claim is justified by the data.

5. Read Bristow *et al.*, (2004) Cardiac-Resynchronization Therapy with or without an Implantable Defibrillator in Advanced Chronic Heart Failure New England Journal of Medicine **350**:2140–2150.

   (a) Consider Table 1. How would you interpret the second sentence of the starred ($*$) footnote?

   (b) Consider Figure 1. Contrast the interpretation of panels A and B to that of panels C and D.

   (c) In light of the results shown in Figure 1(B) (all-cause mortality), how would you interpret the results shown in Table 2?

6. Suppose we're conducting a randomized trial with a survival outcome, hazard ratio e^{μ}, and we want 90% power to detect a hazard ratio of $e^{\mu} = 0.8$ for the experimental group versus control at one-sided $\alpha = 0.025$ using the unweighted log-rank test. We expect that the control group follows an exponential distribution with rate $\lambda = 0.1$/person-years and anticipate uniform enrollment over the first two years and total study length of 4 years.

   We also want to impose an interim monitoring procedure using a one-side α-spending function, $f(t) = 0.025t^2$.

   (a) Find the total number of events required to achieve the desired power, and the total expected sample size for a fixed size trial (no interim monitoring). Plot the expected number of events as a function of time from the enrollment of the first subject to the expected end of the trial (4 years). *(Hints: Use the formula for the probability that a subject experiences an event during the study. If $F \in [0, 4]$ is the time from study start, for $F < 2$, a fraction $F/2$ of subjects will have been enrolled, and $R = F$. For $F \geq 2$, all subjects will have been enrolled and $R = 2$.)*

   (b) We anticipate 5 equally spaced analysis at information fractions $0.2, 0.4, 0.6, 0.8, 1$. Estimate the time after study start at which these fractions of events are expected to accrue.

   (c) Find the critical values, b_k, for Z_k for rejecting $H_0 \colon \mu \geq 0$ if analyses are conducted at the information fractions expected above, using the α-spending function $f(t)$ and plot the resulting boundary.

   (d) Calculate power accounting for the interim monitoring when $\mu = \log 0.8$.

7. Suppose that we conduct the trial designed in problem 5 and observe data at each of five analysis times. Data from each analysis have been compiled into the data file `dataFinal.csv` which contains the following variables:

   - `z`: Treatment variable (0,1)
   - `time1`: Follow-up time at analysis 1
   - `dead1`: Death indicator at analysis 1 (0=alive, 1=dead)
   - `time2`: Follow-up time at analysis 2
   - `dead2`: Death indicator at analysis 2 (0=alive, 1=dead)
   - `time3`: Follow-up time at analysis 3
   - `dead3`: Death indicator at analysis 3 (0=alive, 1=dead)
   - `time4`: Follow-up time at analysis 4
   - `dead4`: Death indicator at analysis 4 (0=alive, 1=dead)
   - `time5`: Follow-up time at the final analysis
   - `dead5`: Death indicator at the final analysis (0=alive, 1=dead)

   For subjects who were still to be enrolled at the time of each interim analysis, the follow-up time and death indicators are both zero. The primary analysis is the comparison of treatment groups using the log-rank test.

   (a) For each interim analysis and the final analysis, compute the log-rank Z-statistic. Use the convention that Z is positive if the difference favors group $z = 1$.

(b) Calculate the information fraction at each of the interim analyses based on the expected full information from problem 5. (Note that in practice, the actual full information will not be known until the completion of the trial). Would the stopping boundaries be crossed at any interim analysis?

(c) Suppose that after the fourth analysis we want to be able so stop for futility if the observed treatment difference is too small and the probability of success at the end of the trial is low. Calculate *conditional* power given the result at the fourth stage and assuming that the hypothesized effect size is true using expected full information.

(d) Compute the information fractions at the interim analyses based on *observed* information at trial completion. Compute the critical value at the final analysis given the actual information observed at trial completion.

(e) What is the conclusion from this trial?