Causal Inference

Outline of solutions for problem set 1.

1. (a)-(c). See R program for the exact and simulation–based calculation of p–values. Let Z_i be the indicator equal to 1 when child i is assigned to vitamin A. The t–statistic t is

$$(\bar{y}_1 - \bar{y}_0)/\sqrt{S_1^2/N + S_0^2/N}, \text{ where}$$

$$\bar{y}_1 = \sum_{i=1}^{2N} Z_i Y_i(1)/N, \quad S_1^2 = \frac{1}{N-1}\sum_{i=1}^{2N} Z_i(Y_i(1) - \bar{y}_1)^2,$$

and similarly for $\bar{y}_0$ and S_0^2. Here, $N = 6$, the t–distribution with unpooled variances has $N - 1$ degrees of freedom, and here gives a two-tailed p-value= $.36$.

(d). In the method in part (b) approximates the probability distribution of the assignment mechanism. Part (c) also uses the assignment mechanism, and, in addition, approximates the values of the potential outcomes at each assignment arm by a distribution.

2. See the program for the exact and simulation–based calculation of p–values. Let $Y_{i1}(0)$ and $Y_{i1}(1)$ be the potential outcomes for the first member of the pair, and $Y_{i2}(0)$ and $Y_{i2}(1)$ be the potential outcomes for the second member of the pair. The t–statistic is :

$$(\bar{y}_1 - \bar{y}_0)/\sqrt{S^2/N}, \text{ where}$$

$$\begin{aligned}\bar{y}_1 - \bar{y}_0 =& \sum_{i=1}^{N}[Z_i(Y_{i1}(1) - Y_{i2}(0)) + (1 - Z_i)(Y_{i2}(1) - Y_{i1}(0))]/N \text{ and} \\ S^2 =& \frac{1}{N-1}\sum_{i=1}^{N}\Big[Z_i(Y_{i1}(1) - Y_{i2}(0))^2 + (1 - Z_i)(Y_{i2}(1) - Y_{i1}(0))^2\Big] \\ &- \frac{N}{N-1}(\bar{y}_1 - \bar{y}_0)^2.\end{aligned}$$

The t–distribution has $N - 1$ degrees of freedom, and gives $p = .36$. Finally, as in question 1, method (b) approximates method (a) through the assignment distribution alone, whereas method (c) also approximates the pairwise differences in outcomes by a distribution.

3. The first reason relates to the quality of matching. A matched pairs design will not be appropriate if the potential outcomes of the two members within each pair are not similar enough. In that case, and because there are quite fewer possible assignments in matched pairs randomization compared to complete randomization, the p-value of an matched pairs analysis will likely be larger.

Secondly, a design D1 is better than design D2 if D1 is more likely than D2 to produce such data that will give precise estimates. On the other hand, once a particular data set is produced, the importance of which design generated the data set is largely irrelevant to precisionof estimation, although it is still relevant to validity. More on this issue will be discussed later on ignorable assignment mechanisms.

4. There are $2N$ units, N of which receive the treatment and N who receive control. The average treatment effect (ACE) is

$$\text{ACE} = \sum_{i=1}^{2N} (Y_i(1) - Y_i(0))/(2N).$$

The estimate is given by $\bar{y}_1 - \bar{y}_0$ as defined in part 1. Taking expectations over the randomization distribution of $\{Z_i\}$ we get

$$\begin{aligned} E(\bar{y}_1 - \bar{y}_0) &= \sum_{i=1}^{2N} E[Z_i Y_i(1)/N - (1 - Z_i) Y_i(0)/N] \\ &= \sum_{i=1}^{2N} E(Z_i) Y_i(1)/N - E[(1 - Z_i)] Y_i(0)/N = \text{ACE}, \end{aligned}$$

because $E(Z_i) = \frac{1}{2}$ for all i. For pairwise randomization, we use the expression given in part 2 for $\bar{y}_1 - \bar{y}_0$, and, by noting that $E(w_i) = \frac{1}{2}$ for all i for this assignment too, we get the required result.

5. **Extra Credit.** Under the assumption of complete randomization and under the hypothesis that $Y_i(1) = Y_i(0) + k_0$, the distribution of the statistic $\bar{y}_1 - \bar{y}_0$ is the same as in part 1 except that the number k_0 is added in each of the 924 values. Of course, the true k_0 is unknown. Nevertheless, we can test the hypothesis that $Y_i(1) = Y_i(0) + k$, for some fixed k, in a similar manner as in part 1. Suppose we collect all values k that are not "rejected" at the level 0.05. Then this set of values is a 95% confidence interval for the true value k_0. To find the set of such k's, we need only to look at the .025 and .975 percentile of 924 values of the statistic $\bar{y}_1 - \bar{y}_0$ under the hypothesis $k = 0$ (part 1), which are -3.75 and 3.75 respectively, and add to them the value of the observed statistic, that is, -2.018 (think why). We get that the resulting confidence interval is [-5.77, 1.73].

6. **Extra Credit.** By definition, under the assumption of additivity, the potential outcomes under treatment would be just a shifted version of the potential outcomes under control, $Y_i(1) = Y_i(0) + k$. Therefore, if the spread of the sample values under treatment is different enough from the spread of the sample values under control, that would suggest the treatment effect is not additive. Depending on which measure of "spread" we think will be different under nonadditivity, there are many possible ways to proceed. For example, one could use the ratio of sample variances, $R = S_1^2/S_0^2$, where S_1^2 and S_0^2 are as given in part 1. Then, as in the previous question, we can

test for the hypothesis $Y_i(1) = Y_i(0) + k$, for a fixed k, by comparing the observed value of R, say R^{obs}, to the randomization distribution of R. If all values k are "rejected", or those that are "accepted" are deemed biologically implausible by the researchers, the assumption of additivity would be questionable. [Under additivity, the statistic R also has approximately an F distribution with $(N-1, N-1)$ degrees of freedom, where the approximation is in the same sense as in part 1(d).]

Programs

For the computations it is convenient to create a $2^{12} \times 12$ matrix m0 that contains all possible randomizations of 12 people, even the unbalanced ones. See program. 1. Then, for part 1, we extract m1 as the subset of m0 that corresponds to balanced assignments. We then use this matrix to calculate the exact and simulated p-values.

```
# the data
v0=c(8.62,0.06,1.48,1.72,8.93,2.19,9.57,7.32,2.65,7.53,7.30,7.62)

# m0 will eventually be the matrix of all possible assignments.

m0=matrix(c(1,-1),nrow=2,ncol=1))
#  building block for m0

for(i in 1:11){
l=nrow(m0)
#  keeps track of no. of rows of m0

#  The following two commands build m0. Think why.

m0=rbind(m0,m0)
m0=cbind(c(rep(1,l),rep(-1,l)),m0)}

m1=m0[(apply(m0,1,sum)==0),]
# m1 picks the balanced randomizations

v1=sort(m1%*%v0)/6
# v1 carries all possible outcomes. We get that 125 out of 924 values
# are less than or equal to the observed statistic -2.018, so p=2*(125/924)
# = 0.271

v2=sort(sample(v1,size=1000,replace=T))
# v2 carries 1000 draws from v1 that are then sorted.
# In our simulation we got 124 out of 1000 values smaller than
# or equal to the observed -2.018, so p=2*0.124=0.248
```

2. Under pairwise randomization, there are only 64 rows of m1 that are relevant. We extract these, storing them in m2, and then proceed as in the completely randomized experiment.

```
ind=c(1:64)
# ind will carry the row numbers of m1
# that correspond to a balanced randomization

count=1
for(i in 1:924){
temp=0
for(j in 1:6){
temp=(m1[i,(2*j-1)]+m1[i,(2*j)])^2+temp}
# temp will be zero only for the rows of m1 that correspond
# to pairwise randomizations.

if(temp==0){ind[count]=i
count=count+1}}
m2=m1[ind,]
v3=sort(m2%*%v0)/6
# v3 carries all possible outcomes. We get that 12 out of 64 values
# are less than or equal to the observed statistic -2.018, so p=2*(12/64)
# = 0.375. For the simulation method, we let

v4=sort(sample(v3,size=1000,replace=T))

# In our simulation we got 188 out of 1000 values smaller than
# or equal to the observed -2.018, so p=2*0.188=0.376
```