PS 236: Causal Inference
Problem Set 2

UC Berkeley, Fall 2008

Due: Thursday, October 9

Your solutions must be submitted in hard copy to my mailbox in the Political Science main office by 4pm on the due date. No late assignments will be accepted. Clean `R` code should be submitted separate from the solutions requested below.

1 Randomization inference and empty calories

In 2005, Coca-Cola introduced Coca-Cola Zero, a sugar-free version of its acclaimed soft drink. The two beverages' similar tastes were highlighted by a marketing strategy suggesting that Coca-Cola executives were going to sue the Zero division for "taste infringement." My refined palate, however, was surely able to distinguish between the two.[1] I was given ten soda samples and asked to identify those of Coke Zero (Z) and those of the original variety (C). The results are presented below.

Trial	Actual	Prediction	Trial	Actual	Prediction
1	C	C	6	Z	C
2	C	C	7	C	C
3	Z	Z	8	C	C
4	C	Z	9	Z	Z
5	Z	Z	10	Z	Z

a. Suppose that I was told that this was a fixed-margins experiment; *i.e.*, there would be five samples each of Coke and Coke Zero. What is the probability that I have no ability to distinguish between the two varieties (*i.e.*, a "sharp null") and the number of successful identifications that I made or a higher amount would have arisen by chance?

b. Suppose that this was a binomial randomization experiment and each flavor had equal probability of appearing. Assume that I knew that this was the framework applied. What is the probability that I have no ability to distinguish between the two types?

c. How would your answer to part (a) change if I did not know that it was a fixed-margins experiment? When would your answer stay the same and when would it change?

d. How would your answer to part (b) change if I did not know the framework applied in part (b). How would that change your answer to that question?

[1] Actually, I am the target market for this product. Coke Zero uses the same formula as Coca-Cola Light, but young adult men see "light" and "diet" as hallmarks of products for women. Hence, the moniker "Zero."

2 Randomization inference simulation

My friend Mike also claims to be able to identify Coke from its calorie-free counterpart. You give him ten cups to sample but mistakenly tell him that they were generated by a fixed margin process. Instead, you used a binomial randomization procedure. Below are his results.

Trial	Actual	Prediction	Trial	Actual	Prediction
1	Z	Z	6	C	C
2	C	C	7	Z	C
3	Z	Z	8	C	Z
4	C	Z	9	Z	Z
5	C	C	10	C	C

To test the hypothesis that Mike is unable to distinguish between the two varieties, compare his results as to 1,000, ten-unit draws from a binomial distribution (hint: use `rbinom` with `size` equal to 1). In what fraction of those simulations did Mike's guesses meet or exceed the success that he had in the true distribution? What is the probability that he has no ability to discern the difference between the two drinks?

3 An experimental design simulation

You are designing an experiment where an individual is randomly assigned to treatment or control. Individuals arrive in a stream over time and you do not know *ex ante* how many will actually participate. In any case, you want an equal number of treatment and control units.

Perform a simulation in `R` where the number of individuals that participate N is chosen from a uniform distribution from 100 to 1,000. You would like exactly half of these participants assigned to treatment.

a. Use the `rbinom` command to generate N random draws from a binomial distribution with $p = \frac{1}{2}$. This implies that every individual has an equal probability of entering the treatment or control regime. Run 1,000 simulations. What fraction of simulations produce outcomes within 0.1% of an equal division between treatment and control?

b. Instead of using the algorithm above, apply the following procedure. The first individual is assigned to treatment with probability one-half. When each subsequent individual is about to be allocated to treatment or control, first calculate the fraction of patients already assigned to treatment. If this figure is less than $\frac{1}{2}$, assign the new individual to treatment with probability $p > \frac{1}{2}$. If this fraction is above a half, assign him to treatment with probability $1 - p$. If this fraction is precisely one-half, assign him to treatment with probability one-half. Run 1,000 simulations for values of p from 0.5 to 1, in increments of 0.05. What fraction of simulations produced outcomes within 0.1% of an equal division between treatment and control? (This method was proposed in Efron (1971)—see Rosenbaum (2002)).

c. What would you need to assume about the stream of individuals in part (b) for this procedure to be reasonable? What could you do to ensure these assumptions are met?