PS 236: Causal Inference
Problem Set 2

UC Berkeley, Fall 2008

Solutions

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?
Solution: How many different orderings of five Zero cups in ten cups are there?

$$\binom{10}{5} = 252$$

Since this is a fixed-margins experiment, I can only get an even number of cups correct (*e.g.*, if I incorrectly identify a Zero cup, there will be a Coke cup that I also misidentify). Hence, the only way to do as well as I did or better is to get eight or ten cups correct. There is only

[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."

1 way to get all ten correct. How many ways are there to get eight correct? I must choose four of five Zero cups correctly and the same for Coke cups:

$$\binom{5}{4} \cdot \binom{5}{4} = 25$$

There are 26 (1 + 25) ways to do as well as I did or better, hence the probability that my success arose by chance is

$$\frac{26}{252} = 0.103.$$

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?
Solution: How many ways are there to order ten cups of either Coke or Zero?

$$2^{10} = 1024.$$

How many ways are there to get ten, nine, or eight cups correct?

$$\binom{10}{10} + \binom{10}{9} + \binom{10}{8} = 1 + 10 + 45 = 56.$$

Hence, the probability that my success or better came through chance is

$$\frac{56}{1024} = 0.055.$$

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?
Solution: In the case above, your inference would not change. To perform your inference, you would compare my guesses to the possible outcomes under the randomization type. This is precisely what we did in part (a). Your results would have been different, however, if I did not know that it was a fixed-margins experiment and did not guess the number of cups equaling the actual margins used (*i.e.*, I didn't guess five of each variety). In this case, I may have gotten nine cups correct, for example, an impossibility if I had followed the fixed-margin assumption. Again, no matter what I knew, you would compare every potential allocation of cups that you as the experimenter could have made against my fixed slate of guesses.

To put this explanation formally, the significance level for the number of correct guesses (our test statistic here) T is (see the Randomization Inference section slides for more information):

$$\Pr(t(Z,r)) = \sum_{z\in\Omega} \mathbb{I}\{t(z,r) \geq T\}\dot{\Pr}(Z = z).$$

Whether I knew of not, the set Ω and the probability of any z arising do not change because these values are based solely upon the randomization strategy and not my information. If I did know the set-up or I happened to guess the correct margins, then $t(z,r)$ and T would only take even values (*i.e.*, if I get one Coke wrong, then I got one Zero wrong). But, if I don't get the margins correct, these these numbers could take odd values and the set of potential test statistics would 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?
Solution: Your inference would not change and you would follow the same procedure as in part (b). Using the equation given in part (c), you notice that the distribution of potential test statistics does not change based upon my knowledge.

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?
Solution: The probability that he has no ability to discern the difference between the two varieties is approximately 17%. (see separate `R` code).

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?
Solution: Only 3.5% of simulations landed within the window (see separate `R` code).

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)).

Solution:

p	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
% in window	3.5	18.0	33.9	42.8	52.5	57.7	63.9	67.0	70.8	72.9	76.7

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?

Solution: You need to assume that the size of the treatment group and thus the probability of treatment assignment for each unit is independent of its unobserved characteristics; *i.e.*, selection on observables must hold. Since the size of the treatment group is itself random, then this condition is likely to hold. One way that it may not is if you do the following: you observe that the treatment group is too small and you need more participants, so you offer more money to new participants. This lures in units with higher opportunity costs and thus biases the experiment.