Written Report – 6.419x Module 1

Name: Sandipan Dey

- **Problem 1.1**

1. (2 points) How would you run a randomized controlled double-blind experiment to determine the effectiveness of the vaccine? Write down procedures for the experimenter to follow.

Solution:

- **Experimental design**: Define the outcome variable as whether a given student gets infected by Polio and the treatment variable to be whether someone is offered the vaccine. Create two groups of students (each of size 20k) – namely, control and treatment groups, by randomly sampling from the Grade 1+3 and Grade 2 students (population), respectively.
- **Randomized Control Trial**: Setup a RCT, where the vaccine (the treatment) is applied to the treatment group, while the control group are given the salt injection (a placebo), everything else are same between the groups (e.g., receives same diet, sleep, exercise etc.), so that we can setup a 2-sample hypothesis test to determine the effectiveness of the vaccine later.
- **Stratification**: Assign treatments at random using stratification, by splitting patients into groups according to certain demographic features of the students, e.g., ensure that the treatment group is representative across factors such as health status, age, and ethnicity, and is similar along these dimensions as the control group, to avoid biases during sampling.
- **Blinded experiment**: No student will know whether he is part of control or treatment group. The experimenter will also be unaware whether a student to whom an injection is applied belongs to control or treatment group (i.e., whether it's vaccine or salt injection). This will prevent a variety of cognitive biases that could influence the experiment.

2. (3 points) For each of the NFIP study, and the Randomized controlled double blind experiment above, which numbers (or estimates) show the effectiveness of the vaccine? Describe whether the estimates suggest the vaccine is effective.

Solution:

Randomized Controlled Double-Blind Experiment			
	Size	Polio rate per 100,000	# Polio
Treatment (vaccine)	200000	28	**56**
Control (Salt Injection)	200000	71	**142**
Total	400000		**198**

From the experiment results, we can see that the polio rate for treatment group (0.00028) is less than the polio rate for the control group (0.00071). Whether it's statistically significant or not, we need to setup a one-sided hypothesis test with the null hypothesis $H_0: \pi_{control} = \pi_{treatment}$ *against the alternative hypothesis* $H_1: \pi_{control} > \pi_{treatment}$.

- *Test statistic T: number of polio affected among the treated individuals*
- *Model is a hypergeometric distribution:*

$$P_{H0}(T = 56) = \frac{\binom{20000}{56}\binom{20000}{142}}{\binom{40000}{198}}$$

```
fisher_exact([[56, 142],[20000 - 56, 20000 - 142]],'less')
# (0.39266566103399375, 3.817439740833869e-10)
```

- *By Fisher's Exact test, the p-value = 3.817439740833869e-10 < 0.05 at significance level α=0.05, as shown above.*
- *Hence, we can reject the NULL hypothesis and conclude that the vaccine is effective.*

3. Let us examine how reliable the estimates are for the NFIP study. A train of potentially problematic but quite possible scenarios cross your mind:

- *(a) (2 points) Scenario: What if Grade 1 and Grade 3 students are different from Grade 2 students in some ways? For example, what if children of different ages are susceptible to polio in different degrees?*

  Can such a difference influence the result from the NFIP experiment? If so, give an example of how a difference between the groups can influence the result. Describe an experimental design that will prevent this difference between groups from making the estimate not reliable.

  Solution:

  Yes, such a difference can influence the test result, since the control and treatment group needs to be same in all respect except the treatment (the vaccine), in order to conclude whether the vaccine is effective or not.

  For example, let's say the children of in the age group of the students in Grade 1+3 (treatment group) are much more susceptible to polio than children in the age group for the students in Grade 2 (control group). Then it will result in higher numbers of polio-affected students in treatment group and lower number of polio affected students in control group. As a result, it will increase the p-value, and in the worst case we may not be able to reject the null hypothesis (and wrongly conclude that vaccine is not effective given the data). Here age is working as a confounder variable.

  Stratification is a solution to this. We need to first split the students into strata (e.g., ages of different groups) and then perform random stratified sampling to create the control and the treatment groups, so that there is no bias for student age in the control and treatment, they are same in all respect.

- *(b) (2 points) Polio is an infectious disease. The NFIP study was not done blind; that is, the children know whether they get the vaccine or not. Could this bias the results? If so, Give an example of how it could bias the results. Describe an aspect of an experimental design that prevent this kind of bias.*

  Solution:

  When a student comes to know that he is vaccinated, he may be relaxed psychologically and he may stop being isolated from other fellow students and spread the infectious disease unconsciously to others who

are not infected, thereby increasing the number of infections in the treatment group and messes up the measurement of the effect of the vaccine.

We can use (double) blind experiments, e.g., to prevent a students know whether they are vaccinated or not.

- *(c) (2 points) Even if the act of "getting vaccine" does lead to reduced infection, it does not necessarily mean that it is the vaccine itself that leads to this result. Give an example of how this could be the case. Describe an aspect of experimental design that would eliminate biases not due to the vaccine itself.*

  Solution:

  There may be other confounder variables, such as health status (e.g., immunity), healthy diet, good sleep, proper work-out exercises etc. that can be reason of reduced infection, when the students in the control and treatment group are not same w.r.t. them.

  We must ensure that the students in the control and treatment group are same w.r.t. all these variables, i.e., they get the same diet, sleep, work-out etc. and they have similar health status (if not, use stratified sampling to avoid bias).

  (2 points) In both experiments, neither control groups nor the no-consent groups got the vaccine. Yet the no-consent groups had a lower rate of polio compared to the control group. Why could that be?

  Solution:

  This can simply be by chance, or may be attributed to the contagiousness of polio - where the students who opted out might have been exposed to less polio-infected students, or some other confounder.

5. *(3 points) In the randomized controlled trial, the children whose parents refused to participate in the trial got polio at the rate of 46 per 100000, while the children whose parents consented to participate got polio at a slighter higher rate of 49 per 100000 (treatment and control groups taken together). On the basis of these numbers, in the following year, some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were their conclusion correct? What would be the consequence if a large group of parents act this way in the next year's trial?*

   Solution:

   Their conclusion is not correct since the numbers of infected in control group vs. those in the no-consent group might be simply because of chance and also there is a significant reduction in number of infected students in treatment group, when compared to the control group.

   If the parents act in this way, less number of students will receive treatment, and it will increase the number of infected students significantly.

- **Problem 1.3**

(a-1). (2 points) Your colleague on education studies really cares about what can improve the education outcome in early childhood. He thinks the ideal planning should be to include as much variables as possible and regress children's educational outcome on the set. Then we select the variables that are shown to be statistically significant and inform the policy makers. Is this approach likely to produce the intended good policies?

Solution:

This is because scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold (by point 3 from the paper). The widespread use of "statistical significance" as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process. The p-values alone may not ensure that a decision is correct or incorrect, he should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.

(a-2). (3 points) Your friend hears your point, and think it makes sense. He also hears about that with more data, relations are less likely to be observed just by chance, and inference becomes more accurate. He asks, if he gets more and more data, will the procedure he proposes find the true effects?

Solution:

No, this is as per point 5 in the paper. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough. So if he gets more and more data, he will likely get smaller and smaller p-values, so that he can reject his null hypothesis easily. But this may result in even because of a tiny effect. He needs to carefully design the RCT, get rid of any possible confounder and consider other possible hypothesis.

(b-1). (2 points) A economist collects data on many nation-wise variables and surprisingly find that if they run a regression between chocolate consumption and number of Nobel prize laureates, the coefficient to be statistically significant. Should he conclude that there exists a relationship between Nobel prize and chocolate consumption?

Solution:

No, scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold (again by point 3 from the paper). This result may be because of some other confounder variable that is not controlled in the RCT properly.

(b-2). (2 points) A neuroscience lab is interested in how consumption of sugar and coco may effect development of intelligence and brain growth. They collect data on chocolate consumption and number of Nobel prize laureates in each nation, and finds the correlation to be statistically significant. Should they conclude that there exists a relationship between chocolate consumption and intelligence?

Solution:

No. Correlation does not imply causation. Where there may have been a high (statistically significant) correlation of sugar / coco consumption with the development of human intelligence, there are several other variables (confounders) and many more hypothesis tests that are completely ignored to be included in the study (neither they are controlled), again this is by point 3 in the paper.

(b-3). (1 point) In order to study the relation between chocolate consumption and intelligence, what can they do?

Solution:

They need to consider many different (all possible) treatment variables (e.g., environmental and genetic factors) that are likely to have effects on intelligence and control them properly in the RCT (e.g., ensure that the treatment and the control group are same w.r.t. them, e.g., using stratified sampling etc.) and repeat the experiment.
Another possibility is to come up with many other more relevant hypothesis and conduct RCTs for them, report the significance for all the tests.
We could also perform a linear regression with all the treatment variables as predictors and the outcome variable as some sort of intelligence score, and report the statistical significance for each of the other predictors are along with that of chocolate consumption.

(b-4). (3 points) The lab runs a randomized experiment on 100 mice, add chocolate in half of the mice's diet and add in another food of the equivalent calories in another half's diet. They find that the difference between the two groups time in solving a maze puzzle has p-value lower then 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice?

Solution:
No, they should consider other confounding variables (e.g., genetic factors etc.) and consider controlling all of them in the RCT (requiring the control and treatment group to be same w.r.t. all of them, these two groups are only different about chocolate consumption in diet). Also, they should come up with many other more relevant hypothesis (with other treatment variables) and conduct RCTs for all of them, report the significance for all the tests.

(b-5). (3 points) The lab collects individual level data on 50000 humans on about 100 features including IQ and chocolate consumption. They find that the relation between chocolate consumption and IQ has a p-value higher than 0.05. However, they find that there are some other variables in the data set that has p-value lower than 0.05, namely, their father's income and number of siblings. So they decide to not write about chocolate consumption, but rather, report these statistically significant results in their paper, and provide possible explanations.

Is this approach correct?

Solution:

No. As per point 4 in the paper, proper inference requires full reporting and transparency. This is a classic example of p-hacking. Cherrypicking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously

avoided. They should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.

(c). (3 points) A lab just finishes a randomized controlled trial on 10000 participants for a new drug, and find a treatment effect with p-value smaller than 0.05. After a journalist interviewed the lab, he wrote a news article titled "New trial shows strong effect of drug X on curing disease Y." Is this title appropriate? What about "New drug proves over 95% success rate of drug X on curing disease Y"?

Solution:

No. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold (again by point 3 from the paper).

The second title is more inappropriate, since we have only seen the sample and we are claiming about the entire population.

(d). (1 point) Your boss wants to decide on company's spending next year. He thinks letting each committee debates and propose the budget is too subjective a process and the company should learn from its past and let the fact talk. He gives you the data on expenditure in different sectors and the company's revenue for the past 25 years. You run a regression of the revenue on the spending on HR sector, and find a large effect, but the effect is not statistically significant. Your boss saw the result and says "Oh, then we shouldn't increase our spending on HR then".

Is his reasoning right?

Solution:

Yes, his reasoning is right. If a coefficient's t-statistic is not significant, we don't need to interpret it at all. We can't be sure that the value of the corresponding parameter in the underlying regression model isn't really zero. That's why even though the predictor variable "spending on HR sector" was found to have a large coefficient in linear regression, no statistically significant linear dependence of the mean of revenue (the response variable) on the predictor was detected.

(e). (1 point) Even if a test is shown as significant by replication of the same experiment, we still cannot make a scientific claim.

True or False?

Solution:

False, if we can replicate the result of the same experiment many times on multiple samples, we may make a scientific claim. Precise inductive inference is impossible and replication is the only way to be sure.

(f). (2 points) Your lab mate is writing up his paper. He says if he reports all the tests and hypothesis he has done, the results will be too long, so he wants to report only the statistical significant ones.

Is this OK? If not, why?

Solution:

No. Again, as per point 4 in the paper, proper inference requires full reporting and transparency. This is again a classic example of p-hacking. Cherrypicking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. They should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.

(g). (2 points) If I see a significant p-values, it could be the case that the null hypothesis is consistent with truth, but my statistical model does not match reality.

True or False?

Solution:

True, this is exactly what happened for the effect of chocolate consumption on the Nobel prize study. As per point 1 in the paper, p-values only indicates how incompatible the data are with a specified statistical model.