Written Report – 6.419x Module 1

Name: ajland

Problem 1.1 The Salk Vaccine Field Trial

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. (Maximum 200 words)

Solution: A randomized controlled double-blind experiment consists of these named components; namely, randomly allocating patients to either a treatment or control group, ensuring that important factors such as age, gender, race, etc. are controlled for, and blinding the administrator and patient from which group they are in. The procedure is loosely as follows: on the first visit, the experimenter informs the participant about the nature of the test, asks for consent, and conducts a screening survey. Upon review, select candidates are invited to return and are randomly and uniformly sorted into the treatment and control groups with a vaccine or saline solution prepared for them, respectively. When they return, the experimenter will administer the predetermined solution, of which they are both ignorant. The final phase of the experiment would consist of regular check-ups wherein the patient is tested for polio and the results are recorded. This is a general outline, but other techniques, such as stratification, which is a control measure, could be discussed.

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. (Maximum 200 words)

Solution: The Grade 2 (vaccine) group in the NFIP study is meant to show the effectiveness of the vaccine compared to the Grade 1 and 3 (no vaccine) group. It is difficult to assess the efficacy of the vaccine due to potential lack of control between age groups, lack of randomness given from assigning participants to a control or treatment group, and lack of blindness of the participants. In the latter study, the Treatment in comparison with the Control group is meant to show the efficacy of the vaccine. In this case, the results do suggest vaccine efficacy against polio due to the apparently random, controlled and blinded nature of the experiment, coupled with the large sample sizes and seemingly different contraction rates.

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. (Maximum 200 words)*

Solution: I believe that such a difference could influence the result. For example, if Grade 1 students are more susceptible to polio than Grade 2 and 3 students, then you might expect the rate for the "Grade 1 and 3" group to be the average of non iid samples. In this case, it would result in a higher average rate, leading one to conclude the vaccine having an even greater effect than it does. One experimental design choice that can remove this effect is to stratify all Grades into their own treatment and control groups [1], which could move one towards a more reliable iid assumption.

- *(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. (Maximum 200 words)*

Solution: I believe not blinding the children could bias the results. For example, the children or parents may decide they do not have to be as careful about contracting polio, thus biasing the results towards a larger contraction rate. Similarly, children who did not receive the vaccine might continue to exercise normal precautions. An experimental design choice that prevents this would be to blind the students against knowing they received or did not receive the vaccine by offering, for example, a saline solution as in the later study as an alternative to vaccination.

- *(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. (Maximum 200 words)*

Solution: It may well be that those who consent to vaccination live a different lifestyle in terms of health compared to those who do not consent. These lifestyle

differences can easily show themselves in the results, and, in this case, even be shown to lower contraction rate. To prevent this, one can create a treatment and control group, where participants in each group have already consented to receiving the vaccine (or saline solution). In this way, the sampling of individuals will more similar and the bias will be eliminated.

4. (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? (Maximum 200 words)

Solution: It could be possible that some in the control group believed it more likely that they received a vaccine, especially compared to a group that knew they did not receive the vaccine, resulting in a change of behaviour.

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? (Maximum 200 words)

Solution: Assuming a clean laboratory environment, this is not a correct conclusion to make as you are averaging the polio rates which potentially come from two different distributions. This is not appropriate since the average number does not have a clear interpretation when compared to the No consent group. Further, this difference could be due to the inherent noise involved with both estimators and the result could have easily been reversed. If a large number of parents act this way in next year's trial, it would reduce the sample size of the treatment and control groups, thus making it harder to distinguish the effect of the vaccine.

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? (Maximum 200 words)

Solution: This approach will likely produce poor policies since it is subject to problems about multiple hypothesis testing. Specifically, in the scenario where all variables are regressed on but all variables actually have no bearing on education outcome, you would expect about $n\alpha$ of them to show significance (when none of them should), where n is the number of variables and $\alpha = 0.05$ is a common setting. Using this naive approach, you could not confidently say which variables are truly causal. Furthermore, unless a careful experiment were carried out, the findings could only be said to correlate with education outcome at best.

(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? Hint: You might need to design some experiment. (We recommend 250 words. Maximum 350 words)

Solution: This modified procedure will not help. Perhaps surprisingly, the Type I error rate does not change as the sample size increases even though the estimator may be consistent. To see this, it may help to give a concrete example. Assume that $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2)$ and you want to test the following hypotheses:

$$H_0 : \bar{X}_n = 0$$

$$H_A : \bar{X}_n \neq 0$$

Thus, the test statistic (**under the null**) will be

$$T = \sqrt{n}\frac{\bar{X}_n}{\sigma} \sim \mathcal{N}(0,1)$$

The key thing to note is that even though the asymptotic variance decreases as $n \to \infty$, we are **still** drawing from an $\mathcal{N}(0,1)$ under the null. Therefore, even though our friend can collect more data, if all the null hypotheses were true, then one can still expect around $n\alpha$ of these to be false positives. This example is not too far from the multiple testing scenario where each parameter is assumed to be

normally distributed and the test statistic is regarded as t-distributed under the null. Conversely, the advantage of a large number of samples will increase the likelihood of rejecting the null under the alternative hypothesis. Considering each of these outcomes, there would be no way to tell if a variable is truly significant in the regression against education outcome. Finally, there is still a necessary distinction between correlative and causal relationships, as mentioned in the previous response.

(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? (Maximum 200 words)

Solution: They should not conclude that there exists a causal relationship between chocolate consumption and intelligence. The downfall of this approach is that it is merely an observational study, which can correlate variables but not confidently determine causal relationships. In this study, there may exist confounders in the data. For example, chocolate consumption and Nobel laureates may correlate with a country's GDP, which may itself be causal link with the number of Nobel laureates. Steps can be taken to try to remove confounders, for example by conditioning on countries with a certain GDP range, but the nature of an observational study is that many variables cannot be critically examined and eliminated.

(b-3). (1 point) In order to study the relation between chocolate consumption and intelligence, what can they do? (Maximum 200 words)

Solution: They could perform a double blinded randomized control trial to study the causal relationship between the two. The particular method of double blinding could be difficult, but the researchers could not disclose what treatment they are studying. Therefore, individuals in the treatment group could receive chocolate and those in the control could receive something else, such as a placebo.

(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 than 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice? (Maximum 200 words)

Solution: Not necessarily. Solving a maze puzzle quickly is not only related to cognitive power, but also to speed. If the tests were performed shortly after feeding, it is possible that a spike in blood sugar for the chocolate eating mice made them move more quickly. Thus, there are other - perhaps more plausible - explanations that could be attributed to the finding. In this study, the researchers are using maze solving time as a proxy for intelligence, but it is important to recognize that it is not a measure of intelligence itself.

(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? (Maximum 150 words)

Solution: This is not a correct approach. According to the ASA statement on p-values, "P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable" [2]. As the statement later says, doing so introduces a "spurious excess" of findings, meaning that selecting based on p-values is misleading.

(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"? (Maximum 150 words)

Solution: Both of these titles are misleading. For the first title, it is important to recognize that statistical significance does not imply a large, or "strong," effect. For

the second, the p-value does not imply the success rate, only whether the observed data is likely under the specified model. It is then possible for drug X to have a very small success rate on curing disease Y and still get small p-values [2].

(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? (Maximum 150 words)

Solution: Again referring to the ASA's statement on p-values, "...large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise" [2]. While 25 data points may seem like it could be enough, it is also important to take into account the variance in HR spending over those years. If it is small, then the data points will be clustered in a small region in the x direction, thus causing the $\hat{\beta}$ to have a larger variance (Recall that $\hat{\beta}_1 \sim \mathcal{N}(\beta, \frac{\sigma}{x^T x})$ in the single parameter case), resulting in a larger p-value. It is also important to note that zero correlation does not imply independence between the two variables (e.g. perhaps the relationship is non-linear). Therefore, the boss's reasoning may not be sound.

(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? (Maximum 150 words)

Solution: False. According to one source, "A scientific claim is a generalization based on a reported statistically significant effect. The reproducibility of that claim is its scientific meaning" [3]. By this definition, even one (properly) significant test could result in a scientific claim.

(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? (We recommend 100 words. Maximum 200 words)

Solution: This is not OK to do. As mentioned before, selectively choosing only significant p-values renders the statistic(s) uninterpretable and is misleading. This is because the Type I error rate will likely allow at least one false discovery through.

If one does not control the error rate somehow, then it is best to report at least the number of tests performed if not the results themselves.

(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? (We recommend 100 words. Maximum 200 words)

Solution: True. Consider the test statistic for a two sample t-test with same sample sizes,

$$T = \sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\hat{\sigma}} \sim t_{2n-1}$$

Now consider the Welch's unequal variances t-test statistic,

$$T = \sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\sqrt{s_1^2 + s_2^2}} \sim t_\nu$$

where $\nu \gtrsim n-1 < 2n-1$. If the model assumes that the population variances are the same, then the test statistic will be the former, which has lighter tails compared to the latter. Therefore, it would be possible to get a significant result under the former but not under the latter while the same hypotheses are being tested.

Problem 1.5

(8). (3 points) Show that the extent of repeated independent testing by different teams can reduce the probability of the research being true. Start by writing the PPV as

$$= \frac{\mathbf{P}(\textit{relation exists, at least one of the n repetitions finds significant})}{\mathbf{P}(\textit{at least one of the n repetitions finds significant})}$$

(Note that this does not include a bias term and you will not need one to answer this question.) (Maximum 200 words)

Solution:

$$
\begin{aligned}
PPV &= \frac{\mathbf{P}(\mathrm{A},\,\mathrm{I})}{\mathbf{P}(\mathrm{I})} \\
&= \frac{\mathbf{P}(\mathrm{A},\,\mathrm{I})}{\mathbf{P}(\mathrm{A},\,\mathrm{I}) + \mathbf{P}(\mathrm{B},\,\mathrm{I})} \\
&= \frac{\mathbf{P}(\mathrm{A},\,\mathrm{I})}{\mathbf{P}(\mathrm{A})\mathbf{P}(\mathrm{I}|\mathrm{A}) + \mathbf{P}(\mathrm{B})\mathbf{P}(\mathrm{I}|\mathrm{B})} \\
&= \frac{\mathbf{P}(\mathrm{A})\mathbf{P}(\mathrm{I}|\mathrm{A})}{\mathbf{P}(\mathrm{A})\mathbf{P}(\mathrm{I}|\mathrm{A}) + \mathbf{P}(\mathrm{B})\mathbf{P}(\mathrm{I}|\mathrm{B})} \\
&= \frac{R/(R+1)(1-\beta^n)}{R/(R+1)(1-\beta^n) + 1/(R+1)(1-(1-\alpha)^n)} \\
&= \frac{R(1-\beta^n)}{R(1-\beta^n) + 1 - (1-\alpha)^n}
\end{aligned}
\tag{1}
$$

Where A is the event "relation exists", B is the event "no relation", and I is the event "at least one of the n repetitions finds significant."

(9). (2 points) What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming $\alpha = 0.05$.) (Hint: Please treat the two issues separately.) (Maximum 100 words)

Solution: According Ioannidis, this would happen if $1-\beta < 0.05$ for both cases [4]. Under these under-powered conditions, the PPV will slightly increase with bias. If $1-\beta = 0.05$, there would also be no decrease in PPV as it would remain constant.

(10). (5 points) Read critically and critique! Remember the golden rule of science, replication? For the third table in the paper, if researchers work on the same hypothesis but only one team finds significance, the other teams are likely to think the results is not robust, since it is not replicable. In light of this, how would you model the situation when multiple teams work on the same hypothesis and the scientific community requires unanimous replication? What would be the PPV? (You do not need to include a bias term for this question.) (Maximum 100 words)

Solution: The stated situation could be modeled as,

$$\begin{aligned} PPV &= \frac{\mathbf{P}(\text{relation exists, all of the n repetitions finds significant})}{\mathbf{P}(\text{all of the n repetitions finds significant})} \\ &= \frac{\mathbf{P}(\text{A, N})}{\mathbf{P}(\text{A, N}) + \mathbf{P}(\text{B, N})} \\ &= \frac{\mathbf{P}(\text{A, N})}{\mathbf{P}(\text{A})\mathbf{P}(\text{N|A}) + \mathbf{P}(\text{B})\mathbf{P}(\text{N|B})} \\ &= \frac{\mathbf{P}(\text{A})\mathbf{P}(\text{N|A})}{\mathbf{P}(\text{A})\mathbf{P}(\text{N|A}) + \mathbf{P}(\text{B})\mathbf{P}(\text{N|B})} \\ &= \frac{R/(R+1)(1-\beta)^n}{R/(R+1)(1-\beta)^n + 1/(R+1)(\alpha)^n)} \\ &= \frac{R(1-\beta)^n}{R(1-\beta)^n + \alpha^n} \end{aligned} \tag{2}$$

Where A and B remain the same as before and N is the event "all of the n repetitions finds significant." As long as $1-\beta < \alpha$, the PPV in this framework monotonically increases to 1.0 as $n \to \infty$.

(11). (3 points) Suppose there is no bias and no teams are racing for the same test, so there is no misconduct and poor practices. Will publications still be more likely to be false than true? (Maximum 200 words)

Solution: This can still depend on a number of things, such as the topics discussed in the Corollaries. Here, Ioannidis cites low power as a cause for false research findings, which can come from both low sample sizes and small effect sizes. He also cites a greater number and lesser selection of hypothesis tests as a cause for less likely to be true research findings; this relates to R, which defines the pre-test probability of a relationship existing. In general, a PPV greater than 50% is difficult to achieve [4]. Therefore, in many fields, research findings are more likely to be false than true.

(12). (2 points) In light of this paper, let's theoretically model the problem of concern in Problem 1.3! Suppose people base the decision to making scientific claim on p-values, which parameter does this influence? R, α, or β? Describe the effect on the PPV if scientists probe random relations and just look at p-value as a certificate for making scientific conclusion. (Maximum 200 words)

Solution: This decision would not effect α or β as these are determined prior to the test. This decision would effect R since the number of true null hypotheses would inflate, thereby decreasing R significantly. Looking at the equation for the PPV, a smaller R will cause the PPV to decrease. If R becomes vanishingly small, then so will the PPV. Thus, using p-values as the sole criteria for making scientific conclusions would result in probable false findings.

References

[1] Wikipedia contributors, "Stratification (clinical trials) — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Stratification_(clinical_trials)&oldid=1045786378, 2021, [Online; accessed 29-May-2024].

[2] R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values: Context, process, and purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016. [Online]. Available: https://doi.org/10.1080/00031305.2016.1154108

[3] J. P. de Ruiter, "The meaning of a claim is its reproducibility," *Behavioral and Brain Sciences*, vol. 41, p. e125, 2018.

[4] J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, p. null, 08 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.0020124