Analysis 01 for MIT 6.419x 1T2021 Data Analysis: Modeling and Applications

Evan Rushton

March 8, 2021

1 PROBLEM 1.1 The Salk Vaccine Field Trial

3a) (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.

A difference between grade 1+3 and grade 2 students can influence the result of the NFIP experiment. If those groups have different polio rates for reasons other than the vaccine, then a rejected null hypothesis of their rates being equal won't necessarily be due to the vaccine, but to other contributing factors related to their ages. An example of such a contributing factor could be that habits of 2nd grade students involve much more germ sharing than that of 1st or 3rd graders. To prevent this difference between groups from hurting the reliaility of the estimate, one should offer the vaccine across all three grade bands such that we have vaccinated students from all of the grades. In this case, since we don't, one way to offset this effect would be to stratify the grade 2 vaccine group into a younger half and an older half and stratify grade 1 and 3 by grade. Then compare the younger grade 2 vaccine rate with the grade 1 rate and the older grade 2 vaccine rate with the grade 3 rate. (179 words)

3b)(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.

Yes, the NFIP study not being blind can bias the result. Students who know they have the vaccine may think they are safe from disease and behave with less caution. Similarly, students who know they don't have the vaccine may think they are less safe and act more carefully. To prevent this kind of bias we could do the study blind or double-blind to limit bias. In order to do that we would need a control group within the grade 2 students who agreed to receive the vaccine who would receive a placebo injection. (94 words)

3c) (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.

In the NFIP study, we could compare the vaccine group to the no consent group. However, differences in polio rates might be due to behavior and attitude differences between those two groups. What makes someone more or less likely to consent to a vaccine study may correlate with other factors that make them more or less likely to contract a virus. Selection bias like this can be avoided if participants in the study are randomly sampled from the population and randomly assigned to groups. (84 words)

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?

The no-consent group could have lower rates of polio than the control groups in both experiements due to random chance. If it is significant, it may be due to differences between the groups. People who don't consent to participate in the study may have lower polio rates than the general population because of their attitudes, dispositions, behaviors, and it could be that being offered to participate in the study influenced their behavior to take more precaution against contracting the disease. (80 words)

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?

The conclusion that there is a higher risk of polio if parents consent to participate is probably not correct. Firstly, those rates are very close in value and the difference could be due to random chance. Additionally, there is nothing inherent to the study design that would put participants at higher polio risk.

If a large group of parents refuse to participate for fear of higher risk of polio, then the sample size will be reduced for treatment and control groups. This will reduce the significance and power of the studyand increase the incidence of type-I and type-II errors. (99 words)

2 PROBLEM 1.3 ASA's Statement on p-value

(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?

Including as many variables as possible and selecting the statistically significant ones to inform policy may lead to disappointing results. The problem with this approach is that by testing many features, let's say 100 features with 100 hypothesis tests, and choosing a significance level of 0.05, then there is an expected value of 5 false significant results out of the 100 features tested. One can protect themself from false significant results when testing multiple hypotheses with the family-wise error rate and the false discovery rate.

The family-wise error rate is the probability of at least one false significant result. For example, the Bonferroni correction is to divide our significance level alpha by the number of hypothesis tests m. In this case the p-value is adjusted by a scalar, the number of tests m. In Holm-Bonferroni, the correction is (m-i+1). (141 words)

(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?

Technically, if enough data is collected such that the number of observarions, $n >> d$, the number of relations, then it is true that inference becomes more accurate. However, it is also true that when we pick a significance level α, that we are picking a number of false significant results we expect to show up. Therefore, if we are performing a large number of tests, say 100 for example, it remains true that we expect 5 of them to be falsely significant at $\alpha = 5\%$ (87 words)

(b-1) (2 points) An economist collects data on many nation-wide variables and surprisingly finds 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?

No. The fact that the economist collected data on many variables suggests that mutliple tests were performed. Therefore, the p-value should be adjusted to account for multiple testing. After correction, if the effect is still found significant, then to truly find a relationship between these variables we need to be able to replicate the results. (55 words) (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?

No. There are a few things wrong with this study design. The first is using the number of Nobel prize laureates in each nation as a reliable metric for intelligence. There are many factors that go into a nation producing Nobel prize laureates other than intelligence. It might be possible to correct for some of the major factors like a nation's GDP and education spending, but it still feels like a very distal and downstream measure for intelligence. The second issue with this study is that they seem to be reporting consumption as a raw total and number of laureates as a raw total per nation as well. This would correlate with population since nations with more people will have higher numbers of each of those variables. Not adjusting for population by considering the per capita ratio is an oversight. Finally, grouping the results by nation introduces many cultural and national factors that are different across each group that are likely to influence both chocolate consumption and the number of Nobel laureates. (172 words)

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

Choose a more proximal measure for intelligence like average IQ scores or an intelligence test. Rather than gathering data by nation, it would be best to conduct a randomized controlled double blind study on the effects of treatment with chocolate versus control on an intelligence test task. (47 words)

(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?

Assuming that they conducted a one-sided test, that with-chocolate maze puzzle time is lower than without-chocolate maze puzzle time as the alternative hypothesis to the null hypothesis that their maze puzzle time is equal, then this result suggests chocolate consumption may lead to improved cognitive power in mice. To make such a conclusion, we need to check the effect size and power of this study. This also assumes that maze time is normally distributed and our sample is large enough and representative of the total population. (86 words)

(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?

No. This is not the correct approach to scientific inquiry. First of all, when publishing their results they need to report all of the tests they performed, not just those that resulted in significant p-values. Secondly, since they performed so many tests they need to correct the p-values for the effects of multiple testing. Either the family-wise error rate or the false discovery rate should be used. Finally, if the lab was solely interested in IQ versus chocolate consumption to begin with, it doesn't make sense for them to test 100 features when they only had one variable of interest. It makes more sense to test 100 variables when you are forming hypotheses around what to study in more detail or look for general trends. (125 words)

(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"?

No. The first title doesn't make sense claiming a "strong effect" when p-value is a measure of significance, not effect size. The second title isn't appropriate because the claim "95% success rate" is not what the p-value smaller than 0.05 means either. The result shows that seeing the measured effect of the drugor something more extreme is only 5% likely. Therefore, it is 95% likely that the drug has a measureable effect on treating whatever symptom of the disease that was measured. From the information in the question, there is no indication that this drug "cures disease Y", it is much more likely that the drug has a measureable effect on one or more symptoms of disease Y. (118 words)

(d) (1 point) Your boss wants to decide on company's spending next year. He thinks letting each committee debate and propose the budget is too subjective a process and the company should learn from its past and let the facts 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?

We don't have enough informatoin yet to decide on spending. We may not have a large enough sample to say it is significant. We should do a power analysis and calculate the sample size we need to show a statistically significant result. We should also look at spending in each of the sectors and base any of our spending decisions on a more complete analyis of all of the data. (70 words)

(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?

False. We can make a scientific claim if our results are reporoduceable. However, if the design of our experiment is flawed, we will see thee same flawed results, and it may be the case that our scientific claim will be shown to be false in a revised/improved study design. (49 words)

(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?

No. This isn't ok. If he does this he will be introducing bias for what might be false significant results. He should include the number of tests along with correction values for family-wise error rate and false discovery rate. This will be a more accurate representation of the results, rather than the skewed representation of the results of only reporting on the significant results. (64 words)

(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?

This is true. We make many assumptions when we specify our models, and it might be that we have made incorrect assumptions in our model that lead to significant p-values when in fact the null hypothesis is true. (38 words)

3 PROBLEM 1.5 Loannidis Paper and PPV

(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

$$\begin{aligned}\text{PPV} &= \frac{\text{P(relation exists, at least one of the n repetitions finds significant)}}{\text{P(at least one of the n repetitions finds significant)}}\\ &= \frac{\frac{cR(1-\beta^n)}{R+1}}{\frac{c(R+1-[1-\alpha]^n-R\beta^n)}{R+1}} = \frac{R(1-\beta^n)}{(R+1-[1-\alpha]^n-R\beta^n)}\end{aligned}$$

PPV tends to decrease as n increases unless $1-\beta < \alpha$. Note that when $1-\beta = \alpha$

the PPV becomes $\frac{R}{R+1}$

$$\begin{aligned}\text{PPV} &= \frac{R(1-(1-\alpha)^n)}{R+1-(1-\alpha)^n-R(1-\alpha)^n} = \frac{R(1-(1-\alpha)^n)}{R(1-(1-\alpha)^n+1(1-(1-\alpha)^n)} \\ &= \frac{R(1-(1-\alpha)^n)}{(R+1)(1-(1-\alpha)^n)} = \frac{R}{R+1}\end{aligned}$$

(9) (2 points) What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming $\alpha = 0.05$.)

As stated earlier, if $1-\beta < \alpha$ then increasing teams testing will not decrease PPV. Also, if $1-\beta \leq \alpha$ then increasing bias will not decrease PPV.

(10) (5 points) Read critically and critique! Remember the gold 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?

If the community requires unanimous replication, then the P(all n repetitions find significant) = $(1-\beta)^n$ rather than $1-\beta^n$.

$$\begin{aligned}\text{PPV} &= \frac{\text{P(relation exists, all n repetitions find significant)}}{\text{P(all n repetition find significant)}} \\ &= \frac{\frac{cR(1-\beta)^n}{R+1}}{\frac{c(1-(1-\alpha^n)+R(1-\beta)^n)}{R+1}} = \frac{R(1-\beta)^n}{(\alpha^n+R(1-\beta^n)}\end{aligned}$$

(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 likely to be false than true?

This depends on the power of the study, the ratio of true relationships to no relationships in the field, and the significance level. From Discussion (3) we have,

$$\text{PPV} = \frac{R(1-\beta)}{R(1-\beta)+\alpha}$$

Evaluating this for varying levels of R, β, and α gives us a sense of the positive predictive value as shown in Table 1. For a low power of 0.2 at a significance level of 0.05, results will start to be more likely true at a ratio of true relationships:no relationships above 0.25. For a higher power of 0.7 at a significance level of 0.05, results will be more likely true even at a ratio of true relationships:no relationships of 0.1. You can play with these values in this desmos graph, https://www.desmos.com/calculator/2egvfu7ljx

Table 1: Positive Predictive Value

PPV	R	$1-\beta$	α
0.286	0.1	0.2	0.05
0.5	0.25	0.2	0.05
0.615	0.4	0.2	0.05
0.688	0.55	0.2	0.05
0.736	0.7	0.2	0.05
0.773	0.85	0.2	0.05
0.583	0.1	0.7	0.05
0.778	0.25	0.7	0.05
0.848	0.4	0.7	0.05
0.885	0.55	0.7	0.05
0.907	0.7	0.7	0.05
0.922	0.85	0.7	0.05

(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.

Since we set the significance level α, we are able to find false significant results by running many tests for a large number of relations. This bias influences α and can be modeled by the parameter u - the proportion of analyses that would not have been findings but end up reported as such - as shown in the derivation in Discussion (7). u depends on how many tests are conducted and ignored in this case and results in a lower PPV as shown in figure 1 of the paper.