**Written Report – 6.419x Module 1**

**Name:** (eeshan89)

**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. (Maximum 200 words)*

**Solution:**

Following are the steps for the experimenter to setup a clean double-blind randomized controlled trial:

* **Participants selection & background**: ensure that a sufficiently large sample of participants / children are selected for the experiment, and represent a diverse range of demographics, ages etc.
* **Randomization**: split the participants randomly into two groups of equal size with the treatment group revieing the vaccine, and the control group receiving the placebo (saltwater injection)
* **Double-blind**: neither the participants, nor the parents, nor the experimenters administering injections should know who is in each group to prevent any bias in reporting or care
* **Monitoring**: all participants should be monitored for a specific measurement window, and polio cases for each group should be recorded

*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:**

For both the studies, the “polio rate per 100,000” are the primary metrics that show the number of polio cases in that group but normalized to be per 100,000 (since the group sizes are unequal).

For the NFIP study, when comparing the three numbers, grade 2 (vaccine) group had the lowest polio rate at 25 per 100,000. And for the RCT, when comparing the three numbers, treatment (vaccine) group had the lowest polio rate at 28 per 100,000. In both studies, we see that the vaccine treated group had the lowest polio rate.

However, the NFIP study had some bias because it was not randomized and the grade 2 (vaccine) group doesn’t have a comparable control group that had consent but no vaccine. So, the NFIP results are less trustworthy compared to the RCT study.

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

**Solution**:

Yes, the differences in ages for the children across the three grades can affect the result of the NFIP study. This is shown in the following ways:

* Grade 2 (vaccine) group consists of children who provided consent AND received the vaccine, but there is no equivalent control group of Grade 2 children who provided consent AND DID NOT receive the vaccine.
* Grade 2 students that did not provide consent may be systematically different from Grade 2 (vaccine) group.
* Grade 1 and Grade 3 (no vaccine) groups are not comparable to the Grade 2 (vaccine) group due to the difference in ages.

The best way to setup an experiment to account for the age differences would be to randomize children within the same grade where the ages are very similar. The randomization should be at the level of “offering vaccine” to ensure that we have similar rates of “no consent” with the grade. Now the only difference between the treatment and control group would be the intervention of the vaccine treatment.

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

**Solution:**

Yes, due to the study not being double-blind, the results were biased. Children (or parents) in the treatment group may change their lifestyle / behavior due to the fact they are getting the vaccine, which may independently influence the outcomes. The same may happen in the control group as well. Vaccinated children may feel protected and engage in riskier activities, whereas the control children may start taking more precautions.

The aspect of experimental design that would prevent this bias would be to introduce a “**placebo control**” (or the salt injection), so that neither the participants nor the administrators of the experiment do not know which injection is being treated to which participant.

* *(****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. (We recommend 50 words. Maximum 200 words)*

**Solution**:

The participants of the NFIP study self-selected into getting the vaccine, which means that the reduced infection rates could be correlated with other factors that are characteristics of those families, such as higher household income, access to healthcare, better hygiene practice etc. which could independently reduce infection rates as well.

The aspect of experimentation design that would prevent this bias would be to completely **randomize** the participants across all background & demographic factors, so that all individuals have an equal opportunity of ‘getting a vaccine’.

*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?* (*We recommend 50 words. Maximum 200 words*)

**Solution**:

The ‘no consent’ groups may be systematically different from the consent groups due to **self-selection**. This group had opted out of the experiment prior to the randomization (in the second study), so they were not comparable to the other groups. Their lower infection rates compared to control group could be due to the characteristics of the group itself – such as higher household income, healthier children, better access to healthcare etc.

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

**Solution**:

The parents’ assessment was incorrect. As explained in the previous question, the lower polio rates observed in the ‘no consent’ group was likely explained due to other confounding factors and self-selection. Furthermore, the difference itself was very small, and the ‘no consent’ group clearly had higher infection rates compared to the ‘vaccinated’ group which only had a 28 per 100000 infection rate.

If a large group of parents opt-out of the trial, then the results of next year’s trial will become biased since the sample of consenters may not be representative of the general population anymore. It will reduce the generalizability of the trial results. This might also lead to long-term negative consequences where there is less trust in vaccination, which will lead to an even faster growth of infectious diseases like polio.

**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?* *(We recommend 50 words. Maximum 200 words)*

**Solution:**

No. This approach would be cherry-picking or ‘p-hacking’ only selective outcomes that show statistical significance, which goes against the principles of proper scientific process of providing the full context & results of the study. Regressing on a large number of variables also increases chances of false positives if not corrected for multiple hypothesis testing. Finally, significance alone does not tell us anything about the magnitude of the effect, which is critical for policy making.

*(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:**

It’s not necessarily true that collecting more data will help uncover “true” relationships. It may help reduce the random error and improve precision of the estimates but will not solve the problem of multiple testing. If many variables are reported, some could be significant by chance, especially in large datasets.

To truly identify the casual relationships, the critical factor is to not only increase the amount of data but also improve the design of the data collection. This is where experiments can help uncover some of those relationships. Researchers could randomly assign students or classrooms into treatment and control groups, apply interventions (such as more reading hours, new teaching methods, new technology etc.) and then measure outcomes to find true causal relationships.

The experiments could also be designed as longitudinal studies, where the outcomes for the same children are followed over time, which allows us to discover any long-term or lag-effects of the initial interventions. Adjusting for multiple hypotheses testing using Bonferroni correction (family wise error rate) or Benjamini-Hochberg correction (false discovery rate) should also be done if using many variables.

Overall, more data is helpful but should be accompanied by pre-research to identify causal approaches using experimental studies, adjustments should be made for multiple hypotheses and then the full reporting of the analyses should be made along with all details and not just cherry-picking significant outcomes.

*(b-2). (2 points) A neuroscience lab is interested in how consumption of sugar and coco may affect 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? (We recommend 100 words. Maximum 200 words)*

**Solution**:

No, they should not conclude that chocolate consumption causes higher intelligence. This is a classic case of correlation being mistaken for causation. There are many confounding factors that could explain the association such as higher income countries may consume more chocolate but also invest heavily in education and research infrastructure. On an individual level, chocolate consumption could also be correlated with higher income status, which in turn is correlated with greater educational opportunities. To find a causal relationship, we need to design an experiment with longitudinal measurement (over long periods of time) and control for confounding variables.

*(b-3). (1 point) In order to study the relation between chocolate consumption and intelligence, what can they do? (We recommend 100 words. Maximum 200 words)*

**Solution**:

To properly find a causal relationship between chocolate consumption and intelligence, they will need to design a randomized controlled trial, where participants are randomly assigned to consume different levels of chocolate and include a control / holdout group (that does not consume chocolate). Then they need to measure “intelligence” in some form of standardized measurable outcome (such as test scores). Outside of the experiment, they might need to run longitudinal studies where they follow cohorts of individuals with different levels of chocolate consumption over several years (and control for confounding variables such as background & demographics).

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

**Solution**:

No, a p value < 0.05 shows that the data is incompatible with the null hypothesis but doesn’t necessarily prove that chocolate improves cognitive power. The biggest potential flaw in this study is the control diet, which is stated as “equivalent calories”. Even if the calories are equal, this food could have a negative effect on the control group, hence making the effect of the chocolate look more inflated. In this study, we need to make sure that the control diet consists of a placebo food that is equal calories but has no additional effect of its own.

Furthermore, 100 mice is a very small sample size so the study should be replicated at a larger scale with a proper placebo diet. And besides a p-value, we also need to report the magnitude of the effect along with confidence intervals to understand the uncertainty in the estimates.

*(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? (We recommend 50 words. Maximum 150 words)*

**Solution**:

No. This approach is incorrect and is a form of cherry-picking or p-hacking only those variables with a p-value < 0.05. With a large number of tests, some variables could show up as false positives purely due to chance even if no true relationship exists. Reporting only these variables will be misleading.

The correct approach would be to pre-define test hypotheses around which relationships we are trying to test for and apply multiple testing corrections as Bonferroni or Benjamini-Hochberg to account for the large number of comparisons. Besides the p-value, we also need to provide the magnitude of the effect and confidence intervals to judge the importance of the output.

*(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"? (We recommend 50 words. Maximum 150 words)*

**Solution**:

* p-value does not provide an information about the effect size, so saying that the new drug had a “strong effect” would be incorrect.
* The second title is also incorrect since the p-value does not measure the success rate. That metric comes from the outcome of the trial itself, which could be calculated as # of cured patients / # of treated patients. And this needs to be accompanied by a confidence interval to understand the level of uncertainty.

*(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? (We recommend 50 words. Maximum 150 words)*

**Solution**:

No, his reasoning is incorrect. A not significant result does not prove that HR spending has no effect, it just means that with the current data we cannot disprove that the effect might be due to chance. Since the observed effect was large, there might be a meaningful relationship but there is uncertainty in the estimate. There might a lot of variance in the HR spending so it might be worth doing some exploratory analysis to understand if there are any outliers that might need to be accounted for. Additionally, for important decisions such as budgeting, we should not rely on a single data point and supplement it with additional data including qualitative data (such as surveys or results from panel discussions).

*(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? (We recommend 50 words. Maximum 150 words)*

**Solution**:

False. If a test result can be replicated, then the result is more credible. But a scientific claim is not just about “significance”, it is also about magnitude and consistency of the effect, the confidence intervals, potential confounders and other assumptions around the data. Replication of significance is necessary but not sufficient in this case.

*(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 statistically significant ones. Is this OK? If not, why? (We recommend 100 words. Maximum 200 words)*

**Solution**:

No, this would be a violation of the core principles of scientific research. Reporting only significant results is cherry-picking or p-hacking, which could lead to misleading results and decisions. With many tests, some results could be false positives due to chance if not corrected for multiple hypothesis testing. If only the “significant” results are published, then the study outcome will be biased.

The ASA statement emphasizes that the p-value should be part of the full context of the study design, data, assumptions and analyses. We need to be transparent and report all the tests performed, along with the magnitude of effect and confidence intervals.

*(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. Significant p-values don’t guarantee that we can reject the null hypothesis. It just shows that the current data is incompatible with specified model. So, this means that if the specified model poorly reflects reality, then the p-value will be misleading. This means we need to understand the model specifications (distribution assumptions, variable selection, nonlinear relationships etc.). It’s possible that we wrongly reject a true null because the model was mis-specified.

**Problem 1.5**

*(8). (3 points) (Include your answer to this question in your written report. We recommend 50 words. Maximum 100 words. Include equations if necessary.) 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*

*PPV= 𝐏(relation exists, at least one of the n repetitions finds significant) / 𝐏(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.)*

**Solution**:

As shown in the paper, it is usually the case that for the same field of research, several research teams are often probing the same or similar questions. Usually, one team gets unilateral attention if there is a claim made.

For n independent studies of equal power, the PPV can be shown to be the following:

As n increases, the denominator grows faster, more false positives accumulate so the PPV decreases. Therefore, the repeated independent testing by increasing number of teams actually reduces the probability that the claimed finding is true.

*(9). (2 points) Include your answer to this part in your written report. (We recommend 50 words. Maximum 100 words. Include equations if necessary.)*

*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.)*

**Solution**:

* If bias (u = 0), then the PPV is not artificially lowered. Only with no bias (full transparency and no selective reporting), we have a PPV that can be as high as possible.
* From the PPV formula for n independent research teams, the false-positive term increases as n increases, so the only way to minimize PPV would be to have 100% power or =0.

*(10). (5 points)*

*Include your answer to this part in your written report. (We recommend 50 words. Maximum 100 words. Include equations if necessary.)*

*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.)*

**Solution**:

In the paper, the PPV formula shows the probability that at least one of the n teams find significance, and we need to find the probability that all n teams find significance.

With pre-study odds R, power and type-1 error , we chance the framework to define:

* If the relation is true and all teams detect it, then
* If the relation is false and all are false positives then

So, then the PPV formula becomes:

*(11). (3 points)*

*Include your answer to this part in your written report. (We recommend 100 words. Maximum 200 words. Include equations if necessary.)*

*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?*

**Solution**:

Even if n = 0 (independent teams researching same test), or u = 0 (no bias), the publications are still more likely to be false than true. This is because the PPV depends on pre-study odds (R) that is characteristic of that field, the power and type-1 error .

If R is small (as in the gene polymorphism example), and power is low (which is also true in most fields), then the denominator of the PPV is dependent on . In these cases, the PPV will still be pretty low even in ideal conditions.

*(12). (2 points)*

*Include your answer to this part in your written report.(We recommend $\sim 100$ words. Maximum $200$ words. Include equations if necessary.)*

*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? or ? Describe the effect on the PPV if scientists probe random relations and just look at p-value as a certificate for making scientific conclusion.*

**Solution**:

Based on the discussion in problem 1.3, if most claims are based solely on p-values then the parameter most affected is the pre-study odds R. If scientists probe random relations, then R will be very small since very few tested relations are true. So, in the PPV formula, if R is close to zero, then PPV is also close to zero. This means that even though we find many statistically significant results, the proportion that reflects true relationships is very low.

It's also good to mention that since a lot of random relations are tested, then the false positives dominate, inflating the , further reducing the PPV.

**References**

[1] 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.

[2] Why Most Published Research Findings Are False, Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLOS Medicine 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>