**Written Report – 6.419x Module 1**

**Name:** 4sowmya

* **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:**  To run a randomized controlled double-blind experiment to determine the effectiveness of the vaccine, start with a large cohort (or units) of patients. Then, randomly split the cohort into treatment and control group. The treatment group is treated with the vaccine. The control group is treated with the placebo drug. The patients were observed for the outcome. To make this experiment double-blinded, the subject who are participating in the trial and the researcher who will interact with the subject were blinded from the information that, whether the subject under study is in the treatment or control group. The double-blinded study minimizes the risk of various bias such as observer bias or confirmation bias which might influence the result. The double blinded experiment should also follow ethical and practical constraints. In the case of NFIP population, the experimenter offers the vaccine, and the participants has an option to consent or reject the offer. The consent participant where later placed in the treatment or control group in a double-blinded trial. Hence the participant was aware that they are part of a clinical trial. But participants with consent were blinded on where they were placed in treatment or control group.

*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:**

In both the experiments, the rate of Polio is the quality of interest to analyze the effectiveness of the vaccine.

We should also consider the distribution of the population between the group. In case of NFIP study, the one who is not offered the vaccine is of different age group, this can create bias as human body at different age may handle the medication differently. Hence it is optimal to compare the results of the group who were offered the vaccine. The rate of polio in the group who took the vaccine is 25 per 100,000 compared to 44 per 100,000 in no consent group. In the case of randomization control trial, all the participants were offered the vaccine and only the participants who accepted the offer were put in treatment/control group. Here there is no other information about the participants is given (like age, gender…). The treatment and control groups are equally distributed in numbers. It will be optimal to compare the consent group only. Looking at the results, the treatment group had less infection rate than the control group and the difference is 43 (71-28). In both experiments, 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.*

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

**Solution:**

Yes, there could be other factors that influence the result. These factors are called confounding factors. The confounding factors are external variables that can be identified to influence the outcome of the experiment. Age as mentioned in the question is one of the confounding factors. Here are some other possible confounders:

* Kids from educated family background might be able to consent than the kids from uneducated family background.
* kids who are from high income neighborhood may be able to afford maintain highly sanitized environment than kids who are in low-income neighborhoods. Since Polio is an infectious disease, the environment in which the participants are might be an influencing factor.

Stratified randomized controlled trial can be a solution to reduce the effect of these confounding factor. Grouping the kids based on age and the environment might reduce the bias caused by the confounding factors.

*(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, the unblinded approached could cause the bias. The kids who have consented to take the vaccine might become more cautious to health and can start making more healthier choice in the way of life. This will be a confounding factor affecting the outcome. The rate of Polio in the kids who took the vaccine is much smaller than the rate of infection in the unvaccinated kids. This confounding factor can be placed as an argument against the effectiveness of the vaccine. To reduce this kind of bias, the double-blinded experiment approach will be appropriate. Randomly choosing the treatment and control group within the consent group will reduce the cognitive bias.

*(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:**

In the randomized controlled trial double-blinded experiment is the main aspect to eliminate or reduce the biases that could influence the outcomes other than vaccine. Here the participants are selected randomly so that there could be a chance of having good mix of cohort in both treatment and control group. Neither the kids who consented to get the vaccine, nor the researchers administering the vaccine know if a specific participant is getting the actual vaccine and salt injection. Hence any difference in the infection rate will be attributed to the vaccine and not the idea of the 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 description about the Polio vaccine trial does not provide much information about the participants except that they are children in grade 1,2,3 and there was a process of consent before administering the vaccine. We see that the rate of infection in the no-consent group is lower than the no vaccine group in both experiments. There could be cognitive bias that influenced this result. The parent who did not consent on receiving this vaccine are aware that the vaccine might have a choice of reducing the infection and still did not consent. There could be multiple reasons behind it. Some might not be comfortable on a new drug that is not proven yet. Income and educational background might also have influenced the decision to no consent. At the same time, the awareness that the vaccine may be effective could have influenced the parents to take extra care of the kids like providing more cleaner environment. Since Polio is an infectious disease, a cleaner environment will reduce the risk of spreading infection.

*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:**

Parent’s decision is based on the information and understanding they received from their participation the previous year. Parents who had consented for the vaccine do not know if their child is in the placebo or vaccine group. The child in the placebo group is at the same risk as the child with no consent. This can create some misconception and concerns about the vaccine being less effective on the consent group. If a large group of parents act with this misconception in the next year’s trial, the vaccine might not have been successful. There could have been delay in the vaccine approval or in worst case, we might have had a vaccine for this disease. Hence it is important for the clinical trials/experiment to present the process and evaluation results – clearly to the public so that the misconceptions can be reduced.

* **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:**

The process is not correct. The ideal step is to define the null hypothesis first. Then select the variables that are already proven in the literature for the related outcome or include variables that are considered as the cause of the exposure, outcome, or both. Then analyze the correlation and variability of the variable. Independent variables have higher variability. Once the exploratory analysis of the problem is completed, then we go into regress the outcome on the set. Based on the significance of the outcome, from the statistical analysis, the highly significant variables will be considered for the intended use,

*(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:**

Yes, the procedure he proposes can be effective. The law of large number in probability and statistics states that as a sample size increases, their mean gets closer to the average of the whole population. This is because as the sample size increases, the sample becomes more representative of the whole intended population. This will reduce any bias caused by confounding factors. The inference will be much accurate then smaller samples. This can also support subgroup analysis as the cohort is larger.

*(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, just by looking at the chocolate consumption and Nobel prize rate in different countries does not provide enough information on the causal effect. A correlation calculation between two variables does not implies causation. The information in the question states that “they collected data on chocolate consumption and number of Nobel prize laureates in each nation”. This statement is very vague and can cause caveat. It does not say that the data is chocolate consumption by the Nobel laureates in each nation. Nobel laureate and chocolate consumer can be two different population. An experimental design or interventional study would be ideal to make any conclusion on the causal effect

*(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:**

In order to study the relation between chocolate consumption and intelligence, an experimental design or interventional study is needed. One of the following two options can be done:

1. The researcher can set up a study on the chocolate consumption rate among the novel laureates. Collect data on the number of Nobel prize laureates in each nation and get the rate of chocolate consumption among them and then compare to the chocolate consumption rate of regular population. This requires data for a period of time and not at one point of time.
2. Set up an experiment by having the chocolate consumers to take IQ test and estimate their intelligence. The cohort can be randomly split into two groups. The treatment group can be offered chocolate on the regular basis for snack for a period of time and have them take the IQ test at the end, while for the control cohort were not given chocolate. Then compare their test scores to assess the effect of chocolate consumption. The cohort selection should be very critical in this case. There should be steps to reduce bias created by confounding factor.

In both the tests, we should also take steps to eliminate confounding factors such as age, gender, education, other comorbidities etc. The research can also be focused on specific type of chocolate like dark chocolate or white chocolate.

*(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:**

Not yet.

We should also analyze the experiments to see if there are any confounding factors involved. Is there a test on solving a maze puzzle done prior to the experiment is started to get a base value of the participating mice. This will help to understand if there is a change in the intelligence level because of the diet. If the mice with chocolate diet have already been solving the maze puzzle before the start of the diet, then this conclusion is wrong. Also, this problem should also be verified scientifically based on the ingredients in the chocolate and their effects in the health. Not all comparison and conclusion can be confirmed statistically. If the experimenter took those base values and then obtained the p-value <0.05 in the result, then we could come to the conclusion .

*(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:**

Yes, this approach is correct. Since the data is not clear of bias and confounders, it will be good idea to just publish the possible explanations. But it will be good if they could take little more time to make efforts to remove this bias and then publish with a more confirmed conclusion.

*(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:**

I would go with “New drug proves over 95% success rate of drug X on curing disease Y”. This statement state that the drug is very effective, but not a 100% cure while the other title might create misconceptions that the drug is 100% effective. Any publication about the research should provide appropriate information and should reduce caveat on the understand of the causal effect.

*(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, this decision might not be right. Not all results can be statistically significance. The p value can only tell if the result is likely due to chance or not. The significant difference or relationship depends on the size of the sample. It is important to consider that the significant result might not be so significant sometimes. The size of the effect should be considered when discussing the statistical significance.

*(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:**

True. [2] A successful replication does not guarantee that the original scientific results of a study were correct. Also, a single failed replication does not confirm that the study is wrong. Sometime, the replication might follow the same factors missed in the original scientific study which can lead to the same results. It is always important to understand the experiment before making any conclusion of the result.

*(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:**

It is not ok to include only significant results in the paper. This is because, it can cause inference bias. The best practice in writing a statistical analysis paper is to report all the tests and hypothesis done to arrive at the conclusion. This will provide enough evidence on the why the author accepts or rejects the hypothesis. This will be helpful the readers to have clear understanding on the experimental design, process, and analysis of the result. This will also increase the chances of efficient reproducibility of the hypothesis testing.

*(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. P-value only tells us that if the result is due to a chance or not. It is not the only metric to evaluate the study. There size of the effect the variables have in the output is assessed using the R2 value, co-efficient of determinant. There is also a level of domain knowledge needed to assess the significance of the results. The statistical model does not match reality might be caused by some bias or confounders in the dataset used for modeling. The experimental design should follow steps to reduce them so that the result will be closer to reality.

* **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*

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

**Solution:**

[3] The probability that a research finding is indeed true depends on the prior probability of the result being true (before doing the study), the statistical power of the study and the level of significance. Let R be the ratio of the number of true hypothesis and no hypothesis. The pre-study probability is calculated using the formula:

Pre-study probability =

The probability of finding the true relationship reflects the power 1- type-II error (False negatives). The probability of the claim when none truly exist reflects the type-1 error (False Positives).

The PPV is positive predictive value that the post study probability is true after a research finding is claimed to be true based on the achieved formal statistical significance.

PPV =

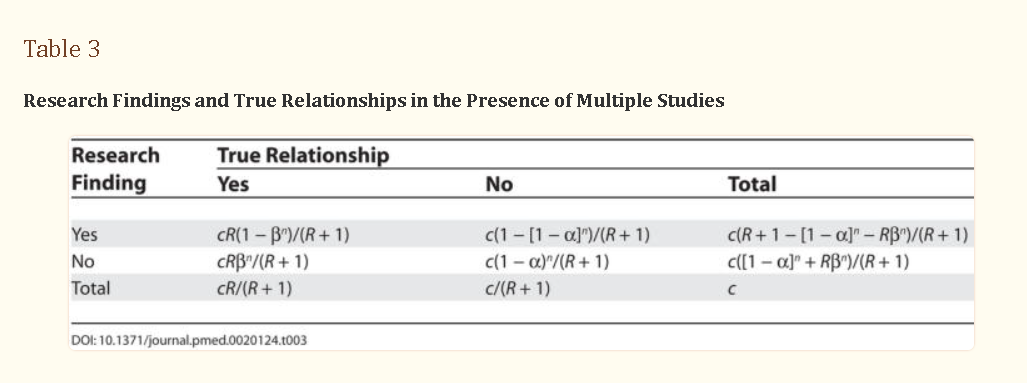
Where,

β = type II error

α = type I error

R = Ratio between true hypothesis and no hypothesis

A finding is most likely to be true if α. When several independent teams perform the study to address the same question, at least one study to claim the statistical significance is easy. The following table from [3] shows the true relationships and research finding in the presence of multiple studies.



*9) (2 points) What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming α=0.05) (Hint: Please treat the two issues separately.)*

**Solution:**

Increased number of teams testing will reduce the PPV unless, 1 − β ≤ α, i.e., 1 − β ≤ 0.05 for most situation.

*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.)*

**Solution:**

There is a concern about both reproducibility and robustness to different analytic approaches. There are research [6] that suggests that archival findings may be less robust than hoped when the same set of observations is used, but different analytic strategy is employed.

It is impossible to know 100% what the truth is in any research question [5]. However, there are several approaches to reduce the post-study probability.

* Large scale evidences should be targeted for the research questions where the pre-test probability is high.
* When multiple teams are addressing the same research question, then the totality of the evidence matters. Diminishing the bias through enhancing the research standards helps.
* Finally, instead of chasing the statistical significance, we should understand the range of R values (pre-test odds).

The PPV = *R*(1 − β*n*)/(*R* + 1 − [1 − α]*n* − *R*β*n*)

Where, n = number of independent studies

β = type II error

α = type I error

*(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?*

**Solution:**

A publication likely to be true or false depends on the dataset used, methodologies and report of the results published. A research finding is true only when the relationship between variables, probability of the cause, statistical significance and the size of the effects are all taken into considerations. If all these steps are followed, then there could be chances that the publication can 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?  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:**

Let’s consider the randomized experiment on 100 mice where chocolate is added to the diet for half of the mice diet and in other half, food of equivalent calories is added to the diet. They find that the difference between the two groups time in solving a maze puzzle has p-value lower than 0.05. It is reasonable to consider that chocolate consumption influences the intelligence. Let’s consider that the pre-test odd is 0.25 and 10 out of 50 mice is hoped to solve the problem.

R = 10/50 = 0.2

The pre-test probability for the study should also be 0.2.

PPV = (1 - β)*R*/(*R* - βR + α)

= (1-5)0.2/ (0.2 – 5\*0.2+5) = 0.137

The research finding is most likely to be true if (1 - β)R>0.05. In our case, (1 - β)R is -0.8 which is less than +0.05. Hence the finding is most likely to be false

**Reference**

[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] Replicability: <https://www.ncbi.nlm.nih.gov/books/NBK547524/>

[3] Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7713993/>

[4] https://en.wikipedia.org/wiki/Positive\_and\_negative\_predictive\_values

[5] Why Most Published Research Findings Are False: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/>

[6] https://www.sciencedirect.com/science/article/pii/S0749597821000200#b0155