# Written Report – 6.419x Module 1

**Name:** ekanshupadhyay

- **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:***

Procedure:

Participants: Recruit a sample of individuals who are eligible for the vaccine. The sample should be representative of the population you want to generalize the results to.

Randomization: Divide the participants into two groups using random allocation, one group will receive the vaccine and the other will receive a placebo.

Blindness: Ensure that neither the participants nor the experimenters know which group each participant has been assigned to. This can be achieved by using a placebo that is indistinguishable from the real vaccine, and by having a third party allocate the participants to the groups.

Administration of the vaccine or placebo: Administer the vaccine or placebo to the participants following the standard protocols and instructions.

Follow-up assessments: Conduct follow-up assessments at predetermined intervals (e.g. 1, 2, and 6 months after vaccination) to evaluate the incidence of polio in the two groups. The assessments should be performed by individuals who are unaware of which group the participants belong to.

Conclusion: Based on the results of the analyses, draw conclusions about the effectiveness of the vaccine and its ability to prevent polio.

*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 each of the NFIP study, and the Randomized controlled double blind experiment above there are two different data number that shows effectiveness of the vaccine

In NFIP study, Grade 2 students who got vaccine had a size of 225000 and only 25 out of 100000 of them got polio i.e 0.025% people. That means out of 225000 only 56.25 students got polio as compared to Grade 1 and 3 students who didn't got any vaccine. They had 54 per 100000 that is more than double the percentage.

In Randomized controlled double blind experiment, the data was significant. Those who were treated with vaccine were 200000 and only 28 per 100000 got polio. Apart from that those who were on salt injection had 71 per 100000 and who didn't approve for consent had 46 per 100000.

These data goes on to prove the effectiveness of vaccine.

*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, differences in age or other demographic factors between groups in an experiment can certainly influence the results and make the estimates unreliable. In the example above, if children of different ages have different degrees of susceptibility to polio, it would be a confounding variable that could impact the results of the experiment and make it difficult to determine the effect of the intervention being studied.

To prevent such differences from affecting the reliability of the results, it is important to control for these variables through proper experimental design. One way to do this is to use a randomized controlled trial (RCT) and randomize participants into the different groups. This helps to ensure that any differences between the groups are randomly distributed and not systematically biased towards one group or another. Additionally, it may also be useful to use appropriate statistical methods to adjust for any demographic differences between groups, such as using regression analysis to control for age as a confounding variable.

In conclusion, it is important to be aware of any potential confounding variables that may impact the results of an experiment and to use proper experimental design and statistical methods to control for these variables in order to ensure reliable and valid results.

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 fact that the children in the National Foundation for Infectious Diseases (NFIP) study knew whether they received the vaccine or not could potentially bias the results. This is because the children and their caregivers may alter their behavior or report their symptoms differently based on their knowledge of whether they received the vaccine or not. For example, if a child knows they received the vaccine, they may be more likely to report fewer symptoms of the disease, even if they actually have them.

To prevent this type of bias, researchers often use a double-blind study design, where neither the participant nor the person administering the intervention knows whether the participant received the treatment or the control. In this type of design, the information is kept confidential from both the participant and the person administering the treatment, reducing the chances of bias.

By using a double-blind study design, the results are less likely to be influenced by the participants' or the researchers' expectations, beliefs, or behaviors, and are more likely to provide an accurate and reliable representation of the effects of the intervention.

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:***

There are many factors that can contribute to the reduced infection rate besides the vaccine itself, such as changes in behavior . For example , Suppose there's an outbreak of a new strain of flu in a community and a vaccine is introduced to combat the spread of the virus. After the vaccine is rolled out, the number of reported flu cases decreases significantly. However, there are other factors that could have contributed to the decrease in flu cases. For instance, people might have changed their behavior, such as avoiding crowded places or washing their hands more frequently , all of which can help reduce the spread of the flu. To control for these factors and ensure that the reduction in infection is indeed due to the vaccine, we can use a study design called a randomized controlled trial (RCT).

In an RCT, participants are randomly assigned to either receive the vaccine or a placebo. Both groups are then monitored for infection rates. If the infection rate is significantly lower in the group that received the vaccine, it is likely that the vaccine is the cause of the reduction in infections. This study design helps to control for other factors, such as changes in behavior or access to medical treatment that might also be contributing to the reduction in infection.

In addition, the study should have enough participants and be conducted over a long enough period of time to ensure that any potential biases are eliminated. This helps to provide a more accurate estimate of the vaccine's efficacy.

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 group were the group of people who refused to participate in whole experiment and in the experiments didn't received any type of care but control group got Salt vaccine. This can be the reason for their increased polio rate. Salt injection can be a sole reason for increased rate of 71 per 100000 as compared to 46 per 100000 for no consent.

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 reasoning is not correct. We conduct a hypotheses test to test if the risk of polio was lower among those who did not consent than

the risk of polio among those who consented.

| Data Summary | | |
|---|---|---|
| | n | Proportion |
| p1 | 100000 | 0.00049 |
| p2 | 100000 | 0.00046 |

The null and alternative hypotheses are
Ho : P1 = P2 P1 and P2 are the population proportions for
Ha : P1 > P2 polio infected children whose parents consented and
did not consent respectively
Let the level of significance be 5%, hence $\alpha = 0.05$

Using the formulae

$$\text{Pooled Proportion } \hat{p} = \frac{p1*n1 + p2*n2}{n1+n2}$$

$\hat{p} = 0.000475$

Standard Error SE

$$\text{Standard Error (SE)} = \left((\hat{p}) * (1 - \hat{p}) * \left(\frac{1}{n1} + \frac{1}{n2}\right)\right)^{0.5}$$

SE = 0.000097

Test Statistic Z-statistic

$$\text{Z-statistic} = \frac{p1-p2-pdiff}{SE}$$

But pdiff =0 since null hypothesis has p1=p2
Z-statistic = 0.3079

p-value
For z = 0.3079, we find the Right Tailed p-value using Excel function NORM.S.DIST
p-value = 1 - NORM.S.DIST(0.3079, TRUE)
p-value = 0.3791

Decision
0.3791 > 0.05
that is p-value > α
Hence we DO NOT REJECT Ho

Decision
0.3079 < 1.6449
that is Z-statistic < Z-critical
Hence we DO NOT REJECT Ho

Conclusion
There does not exist enough statistical evidence at $\alpha = 0.05$ to show that
the risk of polio was lower among those who did not consent than
the risk of polio among those who consented

If large group of parents will act this way in trial then sample for trials will be very less and effectiveness of vaccine can't be measured.

- **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:***

This approach is unlikely to produce the intended good policies. While it may seem appealing to include as many variables as possible and use regression analysis to determine which variables are statistically significant, this approach has several limitations. For one, it can lead to overfitting, where the model fits the data too closely and does not generalize well to new data. This can result in false positive findings, leading to inaccurate conclusions about the effects of different variables on educational outcomes.

Additionally, this approach does not take into account the complexity of the educational system and the interrelated nature of the variables that influence educational outcomes. It is unlikely that a single variable, or even a small set of variables, will have a substantial impact on educational outcomes. Instead, it is likely that many factors, including family background, school resources, and social and cultural factors, interact in complex ways to influence educational outcomes.

*(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:***

No, simply obtaining more data does not guarantee that the procedure will find the true effects. In fact, having more data can sometimes lead to even more complex relationships and increase the risk of false positive findings. The best way to determine the true effects of different variables on educational outcomes is to conduct well-designed experiments, such as randomized controlled trials (RCTs). RCTs allow researchers to manipulate the independent variable of interest and control for confounding variables, making it possible to determine cause-and-effect relationships. Additionally, it is important to use appropriate statistical methods to control for confounding variables and to assess the generalizability of the results to other populations and settings. For example, multi-level models can be used to account for the

hierarchical nature of the educational system and the influence of both individual and contextual factors on educational outcomes. In conclusion, while having more data can be useful, it is not a guarantee of accurate results. To determine the true effects of different variables on educational outcomes, it is important to use well-designed experiments and appropriate statistical methods.

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

***Solution:***

Correlation does not imply causality. The relationship between chocolate consumption and the number of Nobel prize laureates in a nation could be due to many other factors, such as education, income, and culture. Therefore, the lab cannot conclude that there exists a relationship between chocolate consumption and intelligence based solely on the significant correlation. Further studies, such as experimental or causal inference methods, are needed to establish causality.

*(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 study the relationship between chocolate consumption and intelligence, the lab can conduct a randomized controlled trial (RCT) where participants are randomly assigned to consume chocolate or a control food, and the effect on intelligence is measured. They can also use observational methods, such as collecting data on a large sample of individuals and using regression analysis to control for confounding variables.

*(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:***

A p-value of less than 0.05 suggests that the difference in time to solve the maze puzzle between the two groups is statistically significant. However, this does not necessarily mean that the difference is practically significant, or that it is due to chocolate consumption. It could be due to other factors such as chance, or a difference in the composition of the diets. Further experiments are needed to establish causality.

*(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 not correct. The lab should report all the results, including those that are not statistically significant. Hiding results can lead to publication bias and affect the validity of the study. The lab should also mention any potential confounding variables and the limitations of their study.

*(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:***
The first title "New trial shows strong effect of drug X on curing disease Y" is not appropriate, as it overstates the results and the causal relationship between the drug and the cure. The second title "New drug proves over 95% success rate of drug X on curing disease Y" is also not appropriate, as the p-value does not give the success rate, and it assumes causality based on a single study. A more appropriate title would be "New randomized controlled trial finds significant treatment effect of drug X on disease Y."

*(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 not correct. A lack of statistical significance does not necessarily mean that there is no relationship between the variables, but rather that there is not enough evidence to support it. Other methods, such as effect size estimation and confidence intervals, should also be considered when making decisions based on the results of a regression analysis.

*(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. Even if a test is shown as significant by replication of the same experiment, it is important to consider other factors such as generalizability, confounding variables, and sample size. Replication of a study provides evidence of the validity of the results, but it does not prove that the results are true in all cases.

*(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:***
No, this is not okay. Reporting only the statistically significant results can lead to publication bias and affect the validity of the study. All tests and hypotheses, including those that are not statistically significant, should be reported in a paper. This allows for transparency and a complete understanding of the study design and results.

*(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. A significant p-value means that the observed data are unlikely under the assumption that the null hypothesis is true. However, the null hypothesis may not reflect the true relationship between the variables, and the statistical model may not accurately reflect the reality of the data. Therefore, it is important to interpret p-values in the context of the study and to consider other factors such as effect size, confidence intervals, and causal inference methods.

- **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*

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

***Solution:***
PPV, or Positive Predictive Value, is a measure of the accuracy of a research result. It represents the probability that a positive result from a test is actually true, given that the test result is positive. In the context of repeated independent testing by different teams, we can express PPV as follows:

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

Where n represents the number of independent tests conducted. The denominator represents the probability that at least one of the tests produces a significant result, regardless of whether the relation actually exists. The numerator represents the probability that the relation actually exists, given that at least one of the tests produces a significant result.

As the number of independent tests (n) increases, the denominator also increases, making it more likely that at least one of the tests produces a significant result. However, if the relation does not actually exist, the probability of obtaining a false positive result remains constant. Hence, as the number of independent tests increases, the probability that the research finding is true decreases.

*(9).* ***(2 points)***

*What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming a = 0.05.)*

***Solution:***
Bias or increasing teams testing the same hypothesis can lead to an increased number of false positive results and decrease the positive predictive value (PPV) of a hypothesis test. The PPV measures the proportion of true positive results among all positive results, and a decrease in PPV can indicate that the hypothesis test is becoming less reliable. In order to maintain a high PPV, it is important to minimize bias and reduce the number of teams testing the same hypothesis. Additionally, reducing the significance level $\alpha$ from 0.05 to a lower value can help to reduce the number of false positive results and increase the PPV of the hypothesis test.

*(10).* ***(5 points)***

*Include your answer to this part in your written report. (We recommend  words 50. Maximum  words. 100 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:***
The idea of replication in science is based on the premise that empirical regularities or universal laws can be replicated and verified. There are many ways to do this, but the scientific method is considered adequate. The discovered patterns may not be permanent laws of human behavior and need to be proven through statistical verification via replication. When multiple teams work on the same hypothesis and only one team finds significance, the results may not be considered robust and replicable. When a team makes a statement, it must be justified under all conditions and situations in order to be considered replicable. If other teams work under different conditions, the hypothesis may not be found to be correct and may be subject to objections.

*(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:***
The replication of a hypothesis is based on the premise that empirical regularities or universal laws can be replicated and verified through the scientific method. When researchers work on the same hypothesis and only one team finds significance, other teams may question the robustness and replicability of the results. In order for a hypothesis to be considered replicable, it must be justified in all conditions and situations. In a situation where no bias exists and no teams are racing to test the hypothesis, the team working on it is likely to have objective findings with more thorough verification and research. By avoiding misconduct and poor practices, the hypothesis is more likely to be conducted in the correct manner and published with greater accuracy. Thus, replicability is greatly influenced by the adherence to scientific methods and the avoidance of misconduct and poor practices.

*(12).* ***(2 points)***

*Include your answer to this part in your written report.(We recommend 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? 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.*

***Solution:***
The problem of concern in hypothesis testing is the determination of statistical significance through the use of p values. In these tests, researchers are trying to assess the impact of a certain factor, represented by α and β, which are dependent on p-value. The independent variable R does not have an influence on p and therefore does not impact the relationship between αand β . The decision to make a scientific claim is based on the p values, which reflect the significance of the results.

# 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] B. Gustavii, How to write and illustrate a scientific paper, Cambridge University Press, 2017.

[3] Wikipedia, “Principal component analysis,” Accessed: Sep. 2021. [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis