Written Report – 6.419x Module 1

Name: (vatsalya243001)

- **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:

1. Experimental Design – We'll measure whether each student gets infected with polio (the outcome variable) and whether a student gets the vaccine (treatment variable). We split the students into two groups, each with 20,000 kids. The control group is randomly chosen from Grades 1 and 3, while the treatment group is randomly chosen from Grade 2, This clear division helps us easily track results.
2. Randomized Control Trial (RCT) - The treatment group receives the vaccine, while the control group gets a salt injection as a placebo. All other conditions, such as diet, sleep, and exercise, stay the same between the groups. This setup allows us to run a two-sample hypothesis test later to see if the vaccine really makes a difference.
3. Stratification - We assign treatments randomly but use stratification to ensure fairness. Students are split based on health status, age, and ethnicity, so our treatment group represents a good mix of these factors. This method helps us avoid sampling biases and keeps our groups similar across important dimensions.
4. Blinded Experiment - We implement a double-blind protocol where neither the students nor the experimenters know who's getting the vaccine and who's getting the placebo. This approach prevents any cognitive biases from influencing the results, ensuring our findings are based purely on the vaccine's effectiveness without external influence.
5. Monitoring and Data Collection - Throughout the study, we will continuously monitor both groups for Polio infections and any adverse reactions. Detailed records will be kept to ensure accurate data analysis at the end of the trial. Regular check-ins and health assessments will be conducted to maintain the integrity and reliability of the data collected.

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 the given experiment, researchers conducted a randomized controlled double-blind study to evaluate the effectiveness of a vaccine. The study involved two groups: the treatment group (which received the vaccine) and the control group (which received a salt injection). The goal was to determine whether the vaccine could reduce the incidence of polio.

Hypothesis Testing: To assess the vaccine's effectiveness, we set up the following hypotheses:

> Null Hypothesis (H_0) - The polio rate for the treatment group is equal to the polio rate for the control group ($\pi control = \pi treatment$).
>
> Alternative Hypothesis (H_1) - The polio rate for the treatment group is greater than the polio rate for the control group ($\pi control > \pi treatment$).

Test Statistic and Model - We calculate the test statistic T, which represents the number of polio cases among the treated individuals. The model used is a hypergeometric distribution. Specifically, we compute the probability under the null hypothesis:

$$P(H_0) = \frac{\binom{20000}{56}\cdot\binom{20000}{142}}{\binom{40000}{198}}$$

P-Value and Conclusion - Using Fisher's Exact test, we find that the p-value is approximately 3.82e-10. Since this p-value is significantly smaller than the common significance level $\alpha = 0.05$, we reject the null hypothesis. In other words, we have strong evidence to conclude that the vaccine is effective in reducing the polio rate.

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.

  Solution:

  Yes, such a difference can influence the test result, as the control and treatment groups need to be identical in all respects except for the treatment (the vaccine) to accurately determine the vaccine's effectiveness.

  For instance, consider if children in the age group of the students in Grade 1+3 (treatment group) are significantly more susceptible to polio than children in the age group for the students in Grade 2 (control group). This scenario would lead to a higher number of polio-affected students in the treatment group and a lower number of polio-affected students in the control group. Consequently, this disparity would elevate the p-value, and in the worst case, we might fail to reject the null hypothesis (and incorrectly conclude that the vaccine is not effective given the data). In this situation, age acts as a confounding variable.

  To mitigate this issue, stratification is a recommended solution. Stratification involves dividing the students into strata (e.g., different age groups) and then performing random stratified sampling to form the control and treatment groups. This method ensures that there is no age bias between the control and treatment groups, making them equivalent in all respects.

Moreover, it's crucial to ensure that other potential confounding variables, such as health conditions, socio-economic status, or prior exposure to the virus, are also evenly distributed between the groups. By using random stratified sampling and carefully considering other confounding factors, we can create more reliable and valid experimental results. This approach ensures that any observed effect can be more confidently attributed to the vaccine itself rather than other differences between the groups.

- *(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.*

Solution:

Yes, knowing their vaccination status can bias the results. For example, vaccinated students might feel a false sense of security and reduce preventive behaviors, potentially spreading polio to unvaccinated students. This could lead to a higher infection rate in the treatment group, misrepresenting the vaccine's effectiveness.

To prevent this bias, a double-blind experimental design is crucial. In a double-blind study, neither participants nor experimenters know who receives the vaccine or placebo. Here's how to implement it:

1. Blinding Participants: Ensure children do not know if they are vaccinated or receiving a placebo to prevent behavioral changes.
2. Blinding Administrators: Ensure healthcare providers do not know who receives the vaccine to avoid biased treatment.
3. Use of Placebos: Utilize placebos resembling the vaccine to maintain blinding.
4. Random Assignment: Randomly assign participants to control or treatment groups to avoid selection 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.*

Solution:

Other confounding variables, such as overall health status (e.g., immunity), healthy diet, good sleep, and proper exercise, can lead to reduced infection rates. If the students in the control and treatment groups differ in these respects, it could falsely appear that the vaccine is the cause of reduced infections. To eliminate such biases, the experimental design should ensure that the control and treatment groups are identical concerning these variables. This can be achieved by providing the same diet, sleep schedule, and exercise routine for all participants and ensuring participants have similar health statuses. If matching health status is not feasible, stratified sampling can be used to evenly distribute these variables across both groups. Additionally, random assignment of students to control and treatment groups can prevent selection bias. By controlling these factors, we can more accurately assess the vaccine's effectiveness, isolating its impact from other health-related influences.

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?

Solution:

It's plausible that the lower rate of polio in the no-consent groups compared to the control group could be attributed to factors beyond chance. This might include differences in exposure to polio-infected individuals or other variables not accounted for in the study design. Further investigation into these potential factors could provide a clearer understanding of this outcome.

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?

Solution:

The conclusion drawn by these parents may not be accurate, as the difference in infection rates between the control and no-consent groups could be due to chance. Additionally, there's a notable decrease in infections among the treatment group compared to the control group. If a large number of parents opt out of the trial in the following year, fewer students will receive treatment, potentially leading to a significant rise in the number of infections among students. This underscores the importance of carefully considering the implications of such decisions on public health outcomes.

- **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:

Relying solely on statistical significance to inform policy decisions in education may lead to oversimplified and potentially ineffective outcomes. While statistical significance is an important aspect of data analysis, it should not be the sole determinant in policymaking.

Instead, policymakers should consider a range of factors including study design, measurement quality, external evidence, and the validity of assumptions underlying data analysis. A well-designed study accounts for factors such as sample size, randomization, control groups, and potential biases, ensuring more robust findings.

Moreover, the accuracy and reliability of measurements used in the study are crucial to avoid introducing noise and bias. External evidence, such as previous research or expert opinions, provides valuable context and insights that help interpret findings. Additionally, policymakers must assess whether assumptions made in statistical analysis hold true in the specific context, as violations can invalidate results. By incorporating these considerations alongside statistical significance, policymakers can develop more informed and effective policies that address the complexities of improving educational outcomes in early childhood.

(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:

While acquiring more data can indeed reduce the likelihood of observing relationships by mere chance, leading to potentially more accurate inferences, it doesn't guarantee the discovery of true effects.

The misconception lies in assuming that statistical significance alone equates to practical significance. As elucidated in the paper's fifth point, p-values merely indicate the probability of observing a particular result given the null hypothesis, without considering the effect's size or practical relevance.

Consequently, as the sample size increases, even the most minuscule effects can yield statistically significant results, simply due to the heightened precision of measurement or the sheer volume of data.

Thus, while larger datasets may render it easier to reject the null hypothesis, this doesn't necessarily signify the presence of meaningful effects. To discern genuine effects from statistical noise, researchers must diligently design experiments, control for potential confounders, and consider alternative hypotheses. This comprehensive approach ensures that statistical significance aligns with practical significance, fostering more reliable and actionable conclusions in research and policymaking.

(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?

Solution:

No, they should not conclude that there exists a relationship between chocolate consumption and intelligence solely based on the statistically significant correlation. Correlation does not imply causation, as outlined in point 3 of the paper. While there may be a statistically significant correlation between sugar/cocoa consumption and intelligence development, this does not establish a causal relationship.

Additionally, the study overlooks potential confounding variables and fails to account for other relevant factors that could influence intelligence and brain growth. Therefore, further research is needed to explore these variables and conduct controlled studies to determine causality accurately.

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

Solution:

To thoroughly investigate the chocolate consumption-intelligence link, the neuroscience lab must rigorously control for diverse treatment variables in RCTs, employing techniques like stratified sampling.

They should explore multiple hypotheses and report significance for each test, enhancing understanding. Linear regression analysis, incorporating all relevant predictors including chocolate consumption, offers deeper insights into the relationship. This comprehensive approach ensures robust conclusions regarding chocolate's influence on intelligence, contributing to scientific understanding.

(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?

Solution:

No, the lab should not conclude that chocolate consumption leads to improved cognitive power in mice based solely on a p-value below 0.05. They must account for potential confounding variables, such as genetic factors, and ensure proper control in the randomized controlled trial (RCT).

This entails ensuring that the control and treatment groups are comparable regarding all relevant variables except chocolate consumption. Additionally, they should explore alternative hypotheses with other treatment variables through RCTs and report the significance for all tests to provide a comprehensive understanding of the observed effects.

(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?

Solution:

This approach is not advisable. As outlined in point 4 of the paper, maintaining the integrity of scientific inference demands complete transparency in reporting findings.

Cherry-picking statistically significant results while neglecting others constitutes a form of p-hacking, a practice that undermines the reliability and validity of research outcomes.

By selectively emphasizing certain findings, researchers risk distorting the scientific process and perpetuating biases in the literature. It's essential to adhere to rigorous standards of transparency and disclosure, disclosing all hypotheses explored, data collection decisions, statistical analyses conducted, and computed p-values.

This comprehensive approach ensures accountability and fosters trust in the research community, safeguarding the integrity of scientific inquiry.

(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"?

Solution:

No, the titles proposed by the journalist are not appropriate. Scientific conclusions and policy decisions should not rely solely on whether a p-value crosses a specific threshold, as highlighted in point 3 of the paper.

The first title, "New trial shows strong effect of drug X on curing disease Y," overstates the findings by implying a strength of effect not necessarily supported by the data.

The second title, "New drug proves over 95% success rate of drug X on curing disease Y," is even more misleading.

It extrapolates the results from a sample to the entire population without sufficient evidence, and inaccurately implies a level of certainty and generalizability that the trial's results do not warrant. Instead, any conclusions drawn from the study should be presented with appropriate caveats, highlighting the trial's specific context, the limitations of the findings, and the necessity for further research to corroborate the results and explore their broader applicability

(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?

Solution:

No, his reasoning is not right. If a coefficient's t-statistic is not statistically significant, it means we lack sufficient evidence to conclude that there is a non-zero effect of HR spending on revenue.

However, this does not imply that HR spending has no effect; it merely suggests that, based on the available data, the evidence is not strong enough to confirm a relationship. Dismissing a potentially large effect solely due to lack of statistical significance can overlook important practical insights, especially if the sample size is small or the study is underpowered. Therefore, the decision should consider other factors, such as strategic importance and potential long-term benefits, rather than relying solely on the statistical significance of the regression result.

(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?

Solution:

“False”

Because if a test is shown to be significant through replication of the same experiment, it strengthens the evidence for a scientific claim. Replication is a fundamental principle in scientific research, as it helps verify the reliability and validity of findings.

Consistent results across multiple studies enhance confidence in the observed effect and support the robustness of the scientific claim. However, while replication increases credibility, it is still important to consider the quality of the studies, potential confounding factors, and the broader context of the research.

(f). (2 point) 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?

Solution:

No, this isn't okay. Only reporting the statistically significant results and ignoring the others is known as "p-hacking" or selective reporting.

This can bias the results and mislead the scientific community by making significant findings seem more common than they are. Transparency is key in science, so it's important to report all tests and hypotheses, not just the significant ones. This ensures a more accurate interpretation of the data, prevents false positives, and provides a complete view of the research. Proper reporting means sharing all data collection decisions, analyses, and p-values to ensure the findings are trustworthy.

(g). (2 point) 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?

Solution:

“True”

If I see a significant p-value, it could be the case that the null hypothesis is consistent with the truth, but my statistical model does not match reality. This is exactly what happened with the effect of chocolate consumption on the Nobel prize study. As per point 1 in the paper, p-values only indicate how incompatible the data are with a specified statistical model.

- **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

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

Solution:

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

$$\frac{\frac{cR(1-\beta^{n})}{R+1}}{\frac{c(R+1-[1-\alpha]^{n}-R\beta^{n})}{R+1}} = \frac{R(1-\beta^{n})}{(R+1-[1-\alpha]^{n}-R\beta^{n})}$$

PPV tends to decrease as n increases unless $1 - \beta < \alpha$. This highlights the critical interplay between false positive and false negative rates in determining the credibility of research findings. Notably, when

$1 - \beta = \alpha$, the PPV simplifies to $\frac{R+1}{R(1-(1-\alpha)^n)}$.

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

Solution:

To prevent a decrease in PPV despite bias or increased team testing of the same hypothesis (assuming $\alpha = 0.05$), two separate conditions are key. Firstly, if $1 - \beta < \alpha$, additional teams testing the hypothesis will not diminish PPV. Secondly, if $1 - \beta \leq \alpha$, increasing bias will also not reduce PPV. These conditions underscore the importance of maintaining a balance between Type I and Type II error rates and accurately assessing the effects of bias and team testing on research credibility.

(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?

Solution:

To address the scenario where multiple teams work on the same hypothesis and the scientific community demands unanimous replication, a modified approach is needed to calculate the Positive Predictive Value (PPV). Typically, PPV is determined based on the probability of at least one repetition finding significance. However, if unanimous replication is required, the PPV should consider the probability of all repetitions finding significance, denoted as

$$P(relation\ exists,\ all\ n\ repetitions\ find\ significant) = (1 - \beta)^n,\ instead\ of\ 1 - \beta^n.$$

Therefore, the revised PPV formula becomes:

$$PPV = \frac{P(relation\ exists,\ all\ \ n\ repetitions\ finds\ significant)}{P(all\ n\ repetitions\ finds\ significant)} = \frac{cR(1-\beta)^n}{c(1-(1-\alpha)^n+R(1-\beta)^n)} = \frac{R(1-\beta)^n}{(a^n+R(1-\beta)^n)}$$

The solution suggests modifying the PPV calculation to account for unanimous replication, where all repetitions must find significance. This adjustment acknowledges the higher standard demanded by the scientific community when unanimity is required. However, it simplifies the calculation by assuming that unanimity implies perfect replication without considering the potential for variations in experimental conditions, methods, or interpretations among different teams. In reality, achieving unanimous replication across multiple independent studies is challenging due to inherent differences in experimental setups, sample populations, and environmental factors.

(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:

In the absence of bias and competition among research teams, the likelihood of publications being false rather than true hinges on several factors. Studying the power of the study, the ratio of true relationships to no relationships in the field, and the significance level provides valuable insights. By evaluating the Positive Predictive Value (PPV) using the formula:

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

We can gauge the reliability of research findings across different scenarios. For instance, with a low power of 0.2 and a significance level of 0.05, results tend to be more likely true when the ratio of true relationships to no relationships exceeds 0.25.

Conversely, at a higher power of 0.7 and the same significance level, results are more likely true even with a lower ratio of true relationships to no relationships, specifically at 0.1. This analysis underscores the importance of considering multiple factors in assessing the credibility of research findings and highlights the nuanced interplay between study power, the prevalence of true relationships, and the significance level.

(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.

Solution:

In this paper, the reliance on p-values as the sole basis for making scientific claims predominantly affects the parameter α\alphaα, representing the significance level or the acceptable false positive rate. When researchers solely consider p-values, they essentially set a threshold (usually α = 0.05), below which they deem results statistically significant.

However, this approach can lead to a phenomenon known as Type I errors, where false positive results occur, particularly when multiple tests are conducted without proper corrections for multiple comparisons. This bias is encapsulated by the parameter u, which reflects the proportion of analyses erroneously reported as significant despite not meeting true significance criteria. As more tests are performed and disregarded, u escalates, contributing to a decline in the Positive Predictive Value (PPV) as depicted in figure 1 of the paper.

Therefore, the overreliance on p-values without considering the broader context of scientific inquiry can distort research outcomes, leading to inflated false positive rates and ultimately compromising the integrity of scientific findings.

Reference

[1] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

[2] Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124