Written Report – 6.419x Module 1

Name: (Felipe Mehsen Tufaile)

- **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. Select locations **across the country** where the risk of polio is high. The goal is to avoid any bias that may be related to geolocation.
2. Select children that have received parental consent to be vaccinated and choose the control and treatment group from this population. The goal is to avoid any bias that may come from parental background.
3. **Randomly sample** children from this sub-population in other to form the treatment and control groups. Following the approach in the previous steps while also selecting an enough number of individuals in both treatment and control groups will guarantee that these two groups will be very similar from each other, expect by the fact the treatment group will receive the vaccine and the control wont.
4. Give the vaccine to the children in the treatment group and give a placebo to the children in the control group. This is to avoid any psychological effect that may be generated by the knowledge of (not) having received the treatment (blind experiment).
5. Guarantee that doctors that will examine these children for polio do not know the group each children belong. That is to avoid any bias in the diagnosis procedure (double blinding).

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:

The numbers that show effectiveness of the vaccine are the polio rates per 100,000. As this number are shown with a common denominator, they are comparable even though population sizes may differ.

Therefore, in the NFIP study, the fact thar the polio rate per 100,000 is lower for the Grade 2 (vaccine) group suggests that the vaccine is the driver for this low rate. Similarly in the Randomized Controlled Double-Blind Experiment, the polio rate per 100,000 is lower for the Treatment (vaccine) group, which corroborates to the hypothesis that the vaccine in fact helps to prevent polio.

However, although the difference in the polio rate per 100,000 might suggest that the vaccine is effective, it is still recommended to validate this hypothesis using statistics. In that sense, one approach is to assume a distribution model for the number of cases of polio in each group and check the null hypothesis that all groups (treatment and control) may have come from the same distribution against the alternative hypothesis that these two groups belong to distinct distributions.

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-3). (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. (Maximum 200 words)

Solution:

According to [1], polio is a contagious disease. That means that the chance of contracting the disease will change significantly depending on the amount of contact that children of a certain age might have with each other in the school. Therefore, if we suppose that Grade 1 correspond to children younger than 2 years old (nursery), Grade 2 correspond to children older than 2 years old but younger than 6 years old (kindergarten) and Grade 3 correspond to children older than 6 years old and 10 years old, it is likely that **Grade 2 will have a higher transmission rate of polio due to more human contact (children playing with each other) on daily activities** (Grade 1: children are learning how to interact with each other; in Grade 2: children are allowed to play with each other; Grade 3: children start having their own personal space and to have classes in classroom). In this sense, the experiment will be biased on Grade level.

To avoid the problem described above, one option would be to select only children from Grade 2 (that are more likely to contract polio) to participate in the experiment.

(b-3) (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. (Maximum 200 words)

Solution:

Yes. The knowledge that one has taken the vaccine can bias the result. According to [1], although it may be unlikely that individuals could protect themselves from the disease just by the idea of haven taken a vaccine, that has been proven the case in another research [2] that showed that patients suffering from severe post-operative pain have experienced prompt relieve after receiving a "pain killer" made of completely neutral substances. To avoid this issue, it is common to conduct "blind" experiments, in which individuals that are participating in the experiment do not know if they are in the treatment or control group. To accomplish that, individuals in the control group usually receive a placebo, for example an injection of salt dissolved in water. Following this procedure, individuals have the perspective that anyone could have received the treatment. Therefore, individuals in the treatment and control group are equally likely to experience side effects due to psychological reasons.

(c-3) (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. (Maximum 200 words)

Solution:

In many situations, the "willingness of getting vaccinated" might be a sort of **selection bias**. That is, people that are willing to get vaccinated might also be people that usually care more about hygiene, live a healthier life, go to the doctor regularly, etc. This, in turn, could lead to a lower chance of developing certain diseases. In this sense, it could be the case that the group of people that gets vaccinated have a lower infection rate even though this effect is not due to the vaccine itself, but the healthier habits of the population in this group. To avoid this type of bias, people in both treatment and control group should be selected from the same group of people that are willing to get vaccinated. This procedure corrects for this **confounding variable** and guarantees a **randomized control trial (RCT).**

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? (Maximum 200 words)

Solution:

The reason for that is also due to **selection bias**, but the logic behind it is inverse to the one in the previous response. According to [1], lower-income families are less likely to consent to treatment compared to higher-income families. Therefore, it would be expected to have more lower-income families in the non-consent group. Children in lower-income families are also more likely to live in less hygienic surroundings, hence more likely to contract a mild case of polio early in their childhood. This, in turn, may lead to the production of antibodies in the organism of these children, which will protect them against more severe case in the future (this could be interpreted as if children from lower-income families were being "naturally vaccinated" early in their childhood due to the surrounding where they live).

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:

No, their conclusion was incorrect. They failed to consider that the group of children that had parental consent and participated in the experiment is different from the group that didn't have parental consent and didn't participate in the experiment. The difference is due to family background. As mentioned in the previous question, children that didn't receive parental consent are less likely to contract more severe cases of polio since they may have been exposed to the virus and produced antibodies in early childhood. The correct analysis would have been to compare the children that received parental consent and took the placebo (polio rate per 100,000 of 71) against the children that received the consent and took the vaccine (polio rate per 100,000 of 28). If more parents think that participating in the experiment leads to a higher risk of contracting polio, they might remove their children from experiment, reducing the number of individuals in the treatment and control group in the following year. Eventually, if the number of children that drop off the experiment is very significative, the experiment becomes impracticable due to insufficient individuals.

- **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:

Although it might seem a good idea to include all variables in the model and then select the ones that are shown statistically significant, this procedure might lead to the selection of **false significant features.** To control the error of finding a false significant variable, i. e. reject the null hypothesis H0 when H0 is true (type I error), we set the significance level alpha, **alpha = P(reject H0 | H0 true)**, to a small value, usually alpha = 0.5. However, even though we choose a small value for alpha, this value is greater than zero.

Therefore, testing multiple hypothesis will eventually lead to the discovery of a false significant feature, since we expect to commit type I error 5% of the time, in average.

(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? (We recommend 250 words. ~ Maximum 350 words).

Solution:

Adding more that data may improve the chances to find true effects up to a point. If the original dataset used to study the relationship between some variables and children's education outcome is small so that it doesn't fully capture the true distribution for each feature, then adding more data allows to better represent the distribution of each feature in the dataset. Therefore, the chance of committing type I error, P(reject H0 | H0 true), and type II error, P(retain H0 | HA true), is reduced. However, if enough data is provided and the true distribution of each feature is already being captured in the dataset, adding more variables will not further increase the chance of finding true relationships. The chance of committing type I error will still be controlled by the alpha value and running excessive hypothesis tests will still result in the discovery of false significant features at a rate of alpha (typically 5%) in average.

On the other hand, if we assume that the distribution of some features may change with time (e.g. daily temperature distribution as we experience global warming) adding more data allows us to capture the change in the distribution of the given feature. However, the same idea is still valid: if we collect enough data to represent the new distribution of the feature in question and the distribution is assumed to remain constant, adding more data will not improve the chances of finding true relationships.

(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:

We would need more information about the design of the test to conclude in chocolate consumption and intelligence are related. Considering the information we have, if researchers collected data on chocolate consumption and Nobel prize laurates in each nation and then performed various tests (one for each nation) it is likely that some of the tests will end up being false significant (type I error), given the alpha level chosen. Depending on the number of tests performed, this will be true even if the alpha is a very small value. Now, what sometimes happen is that researchers will only talk about on their papers about those tests that came out as significant since they want to publish an interesting discovery.

(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:

Considering the researchers are performing multiple test (one for each nation) to study the relationship between chocolate consumption and intelligence, they should adjust the significance level (or equivalently adjust the calculated p-values) to better control for type I error. Two options to implement such adjustment are **Family-wise error rate (FWER)** and **False Discovery Rate (FDR).** In the first one, researchers control the size of FWER, by choosing significance levels of the individual tests to vary with the size of the series of tests (e.g. FWER Holm-Bonferroni Correction: p(i) <= alpha / (m-(i-1))). In the second, researchers

control the expected proportion of false discoveries among all discoveries made, this procedure also leads to choosing different significance levels of the individual tests depending on the size of the series of tests (e.g. FDR Benjamini-Hochberg Correction: p(k) <= (k/m)*alpha). However, FDR has a high power compared to FWER.

(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:

Considering that the two population (treatment group and control group) of mice only differs from each other with regard to the diet that they received, so that a randomized control trials (RCT) are guaranteed, and that the test is based on the difference of the average time each group (treatment and control) takes to finish the puzzle, then it is possible that chocolate consumption leads to improved cognitive power in mice. The design of the experiment is now evaluating whether the distribution of the time to finish the puzzle are the same for the treatment and control groups (H0) or is lower for the treatment group (HA). As the design of this test supposably guarantees randomized control trials and do not perform multiple tests it is more likely that significant results are indeed true.

(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:

The approach is not correct. As mentioned in question 1.3(b-2), running multiple tests may result in the discovery of false significant relationships (type I error) since we expect to have a rate of false positive discovery of alpha (typically 5%). Now, if researchers only publish significant results and omit non-significant results, there is a chance they are only reporting false positive discoveries in their paper. However, if another researcher attempts to reproduce the experiment to validate the discovery, this researcher may fail to successfully reproduce it. In general, we should careful if we read about the discovery of a statistically significant relationship when the relationship is very specific like: “orange cars are less likely to have serious damages that are discovered only after purchase” (lecture notes).

(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 are two problems with the first version of the article title: 1. The article doesn't mention the significance level used in the trial; 2. The article could have overestimated the effect of the new drug by choosing the words "curing disease". To avoid misleading statements like the first title, it is important to refer to the significance level alpha, along with a fair description of the significant result. For example, if the experiment showed that patients that received the drug live longer than patients that didn't receive the treatment, stating that the drug has strong effect on curing the disease might overestimate the effect of the drug. The drug could be acting to delay the effects of the disease, but not cure it.

The second title does refer to the significance level used in the trial. However, it might be still overestimating the effects of the drug.

(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:

It depends. Assuming that in the context of a regression a "large effect" means a large coefficient for the regressor "HR spending", finding a large coefficient could be due to the magnitude of the values for "HR Spending" and not necessarily due to a strong relationship between "Revenue" and "HR spending". To overcome this problem, we can standardize each variable, so they all have similar magnitude. If we standardize "HR spending" and obtain a much smaller coefficient, along with a p-value larger than alpha, then the boss conclusion may be right. However, there is still a chance that after applying standardization we find a large coefficient for "HR Spending" (compared to the coefficient of other regressors), and we retain H0 (p-value larger than alpha) when the alternative hypothesis HA is true (type II error). If that is the case, then the boss conclusion is wrong.

(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. A statistically significant discovery is generally considered to be true (at least statistically supported) if the result of the replication is also significant and matches the result of the original

experiment. Considering that the relative frequency of observing this same statistically significant result during replications falls below the significancy level chosen.

(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 ideal. Reporting only statistically significant results can be misleading, especially if there are a lot of tests being carried out. As discussed in question 1.3(b-5), running multiple tests may result in the discovery of false significant relationships (type I error), since we expect to have a rate of false positive discovery of alpha (typically 5%). If a researcher only publishes significant results and omit non-significant results, there is a chance he is only reporting false positive discoveries in his paper. If this is the case, the experiment will not be replicable, invalidating the claimed discovery in the paper. For this reason, it is important to report tests that were not significant, so that other researchers can make a better-informed conclusion as to whether the claimed finding is valid or not.

(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 50 words. Maximum 150 words)

Solution:

True. Even though the model is wrong, there is a chance that this model will generate a result that matches the truth and is statistically significant. From the standpoint of the model, this would be interpreted as a false positive result (type I error) since, for the model, this outcome would have been achieved by chance. However, if we try to replicate the test using the wrong model it is likely that we will not get a statistically significant result again for a null hypothesis that is consistent with truth. In such case, that would confirm that the model is actually wrong.

- **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:

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

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

Solution:

Starting with:

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

We can re-write the expression in other words:

$$PPV = \frac{P(reject\ H_0 |\ H_0\ false, at\ least\ once)}{P(reject\ H_0, at\ least\ once)}$$

Now, considering that for a single test, the probability of retaining H_0 when H_A is true is β, the probability that in at least one of the "n" repetitions we reject H_0 when H_A is true:

$$1 - P(type\ II\ error, for\ all\ \text{"n"}) = 1 - P(retain\ H_0 \mid H_A\ true, for\ all\ \text{"n"}) = 1 - \beta^n$$

Adjusting for the pre-study probability of a relationship being true (H_0 being False), yields:

$$P(reject\ H_0 \mid H_0\ false, at\ least\ once) = (1 - \beta^n) * \left(\frac{R}{R+1}\right)$$

On the other hand, the probability of finding significant at least once is:

$$P(reject\ H_0, at\ least\ once) = P(reject\ H_0 \mid H_0\ false, at\ least\ once) + P(reject\ H_0 \mid H_0\ true, at\ least\ once)$$

Assuming "V" as the number of type I errors, the probability of making at least one false discovery in "n" independent tests is:

$$P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^n$$

Adjusting for the pre-study probability of a relationship being false (H_0 being True), yields:

$$P(reject\ H_0 \mid H_0\ true, at\ least\ once) = [1 - (1 - \alpha)^n] * \left(\frac{1}{R+1}\right)$$

Therefore,

$$PPV = \frac{(1 - \beta^n) * \left(\frac{R}{R+1}\right)}{(1 - \beta^n) * \left(\frac{R}{R+1}\right) + [1 - (1 - \alpha)^n] * \left(\frac{1}{R+1}\right)}$$

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

(9). (2 points) 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:

PPV - Bias standpoint (easy to demonstrate through calculations):

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

$$\frac{\partial(PPV)}{\partial u} = \frac{R(\alpha + \beta - 1)}{[(1-u)(\alpha - \beta R) + R + u]^2}$$

Hence,

$$\frac{\partial(PPV)}{\partial u} \geq 0 \rightarrow \alpha \geq 1 - \beta$$

Therefore, PPV tends to decrease with increase "u", unless $\alpha \geq 1 - \beta$. Assuming $\alpha = 0.05$, PPV tends to decrease if $\beta > 0.95$.

PPV - Increasing teams testing standpoint: Similarly**,** PPV tends to decrease with increasing "n", unless $1 - \beta < \alpha$.

In other words, when the power of a study is very low, we might expect the PPV to not decrease with increasing bias or increasing number of teams testing the same hypothesis. Since the Power of a test is related to sample size, all other things being kept constant, this scenario may be expected in areas of research where only small sample sizes are available for studying (e. g. lack of information, new field of study).

(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:

Starting with:

$$PPV = \frac{P(relation\ exists, all\ n\ repetitions\ finds\ significant)}{P(all\ n\ repetitions\ finds\ significant)}$$

We can re-write the expression in other words:

$$PPV = \frac{P(reject\ H_0 |\ H_0\ false, for\ all\ "n")}{P(reject\ H_0\ |\ H_0\ true, at\ least\ once)}$$

The probability that we reject H_0 when H_A is true for all "n", adjusted for the pre-study probability of a relationship being true (H_0 being False) is:

$$P(reject\ H_0 |\ H_0\ false, for\ all\ "n") = \ (1-\beta)^n * \left(\frac{R}{R+1}\right)$$

Assuming "V" as the number of type I errors, the probability of making at least one false discovery in "n" independent tests is:

$$P(V\ \geq 1) = \ 1 - P(V = 0) = 1 - \ (1-\alpha)^n$$

Adjusting for the pre-study probability of a relationship being false (H_0 being True), yields:

$$P(reject\ H_0\ |\ H_0\ true, at\ least\ once) = [1 - \ (1-\alpha)^n] * \left(\frac{1}{R+1}\right)$$

On the other hand, the probability of finding significant for all "n" is:

$$P(reject\ H_0, for\ all\ "n") = \ P(reject\ H_0\ |\ H_0\ false, for\ all\ "n") + P(reject\ H_0\ |\ H_0\ true, at\ least\ once)$$

Therefore,

$$PPV = \frac{(1-\beta)^n * \left(\frac{R}{R+1}\right)}{(1-\beta)^n * \left(\frac{R}{R+1}\right) + [1 - \ (1-\alpha)^n] * \left(\frac{1}{R+1}\right)} = \frac{(1-\beta)^n R}{(1-\beta)^n R + [1 - \ (1-\alpha)^n]}$$

(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 a scenario that there is no bias and teams are not racing for the same test, the PPV is:

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

Considering this equation, it is possible to note that a research finding is more likely to be true than false only if $(1-\beta)R > \ \alpha$. That is, if the power of the test, adjusted by the pre-study ratio of the number of "true relationships" to "no relationships" is greater than the significance level. This would be the case for studies in fields with great number of studies for example. However, publications will still be more likely to be false than true when the field of research is considerably new and there is a small number of studies that has been published in that field so far.

(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, alpha or beta? 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:

Since p-value is defined as the probability under the assumption of the null hypothesis (no effect or no difference of obtaining test results at least as extreme as the result observed), decision based on p-values will affect alpha, the control parameter for type I error. Considering the scenario of no bias and teams are not racing for the same test, the adapted PPV would be:

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

Therefore, a research finding would be more likely to be true than false only if $(1-\beta)R > p$. We see that the PPV will than be greater as the calculated p-value becomes smaller.

Reference

[1] D. Freedman, R. Pisani and R. Purves, “Statistics,” Fourth Edition, Norton & Company, 2007.

[2] H. K., Beecher, “Measurement of Subjective Responses,” Oxford University Press, 1959, pp.66-67.