Statistics Review: Homework/Analysis

MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications

Ellick Hou - 3/1/21

table of contents

Problem 1.1 The Salk Vaccine Field Trial

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*)

➡ Ans:

(1) Define treatment variable and outcome variable
- treatment variable = "offer vaccine"; outcome variable="Polio rate"

(2) Pre-divide students into groups by:
- Grade 1 / Grade 2 / Grade 3

(3) Sample proportionately from each group.

(4) Inform subjects and their parents of the trial's risks and possible consequences and ask for their consent.

(5) Randomly choose half of the subjects, who consented to participate, to be the control group (taking salt injection) and the other half to be the treatment group (taking vaccine).

(6) Keep the above information in the experimenters.

Problem 1.1.2

(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)

➡ Ans:

(1) In NFIP study
- Polio rate in Grade 2 (vaccine) is lower than Polio rate in Grade 1 and 3 (no vaccine).
- Vaccine is effective. $p - value \approx 2.2 * 10^{-9}$ in Fisher's Exact test.

(2) In Treatment (vaccine) in Randomized controlled double-blind experiment.

π : polio rate per 100,000; H_0 : Null hypothesis;

H_A : Alternative; T : test statistic;

$H_0 : \pi_{treatment} = \pi_{control}$; $H_A : \pi_{treatment} < \pi_{control}$

$\#$ poliot in treatment $= 56$

$\#$ poliot in control $= 142$

$T \sim$ hypergeometric(400000, 200000, 198)

$p - value = P_{H_0}(T \leq 56) \approx 4 * 10^{-10}$

- Polio rate in treatment group is lower than Polio rate in control group
- Vaccine is effective. $p - value \approx 4 * 10^{-10}$ in Fisher's Exact test.

Problem 1.1.3

Let us examine how reliable the estimates are for the NFIP study. A train of potentially problematic but quite possible scenarios cross your mind:

- **(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*)
  - Ans:
    - Yes, maybe the vaccine has an especially better chance to prevent polio in Grade 2 students than in other grades. If so, we will overestimate the overall efficacy of the vaccine. The overestimation will lead to a better chance of rejecting the null hypothesis and result in a latent type-I error rate that can not be discovered in the hypothesis test setting.
    - We can do stratified randomization in RCT, pre-dive subjects into groups by their age, and sample proportionately from each group to prevent this kind of bias.

- **(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*)
  - Ans:
    - Yes, children who take the vaccine might just hang out with children who do too. If the vaccine does prevent polio, the group with higher risk goes even higher due to the infectiousness. And we might overestimate the effect of the vaccine.
    - Double-blind experiment, providing placebo treatment to the control group and not letting any subjects and people who might interact with them know.

- **(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*)
  - Ans:
    - Perhaps the subjects who choose vaccination are more concerned about their own health and have better personal hygiene result in a lower infection rate.
    - We can
      1. Inform subjects and their parents of the trial's risks and possible consequences and ask for their consent.
      2. Offer a large number of treatments.
      3. Randomly choose half of them to be the control group(placebo).
    - Assuming the proportion of the health care level and the personal hygiene level fix in the population. By Law of Large Numbers, we can expect the difference in averages of any relevant feature is small between the control group and the treatment group.

Problem 1.1.4

(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*)

➡ Ans

- In the NFIP study, the lower polio rate might be due to the randomization or the age difference between groups.
- In Randomized Controlled Double-Blind Experiment. Because the refuse rate is much higher (46.7%(no-consent/total) vs. 35.7%(no-consent/total grade 2)). The lower polio rate might be due to the subjects in the no-consent group who don't want to take the risk to be in the control group. Their parents might want to make sure their child can receive treatment.

Problem 1.1.5

(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?(Maximum 200 words)

➡ Ans

- No, if they only make a conclusion in 46/100,000 vs. 49/100,000, the evidence is not sufficient enough to say participating in the experiment has a higher risk.
- However, because join the experiment has a 50% chance to be in the control group and have a higher chance to get polio, if they want to make certain to prevent polio, they probably should not join the experiment and find another way to get vaccine.
- The next year's trial would have a bias in the population of those who do take the risk to be the control group.

Problem 1.3 The ASA Statement on p-Values: Context, Process, and Purpose

This problem requires you to submit a report with written answers to each of the given parts below. Limit your answers for each part to 300 words.

- Problem 1.3.(a)
  - (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?
    - ➡ Ans
      - Yes. According to the ASA Statement on p-Values' principles number 4. Cherry-picking promising findings lead to a spurious bias result. Since the expected value of type-I and type-II errors is highly related to the number of hypothesis tests been done and which correction method was used. Researchers should not only choose those statistically significant results to report. It's important to inform readers of how risky accepting their conclusions is by disclose these details.

  - (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?
    - ➡ Ans
      - Yes, if the additional data is expanding on the sample size, more variables might not help. By the Law of large numbers, the larger sample size results in greater confidence in our estimators. Therefore we can use more strict significant levels to validate our hypotheses resulting in less type-II error rate.

Problem 1.3.(b)

- (B-1) **(2 points)** A economist collects data on many nation-wise variables and surprisingly find that if they run a regression between chocolate consumption and number of Nobel prize laureates, the coefficient to be statistically significant. Should he conclude that there exists a relationship between Nobel prize and chocolate consumption?
  - Ans
    - Firstly, what kind of coefficient exactly? I guess correlation coefficient it is.
    - No, the correlation coefficient says nothing about causality. But he can say there exists/not-exists a linear relationship between them.

- (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?
  - Ans
    - No, they can only say something about chocolate consumption and the number of Nobel prize laureates, not the intelligence.

- (b-3) **(1 points)** In order to study the relation between chocolate consumption and intelligence, what can they do?
  - Ans
    - Maybe reduce nation-wise chocolate consumption to personal-wise chocolate consumption and track the subject's own intelligence to avoid ecological correlations and identify the real effect to individuals.

- (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 then 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice?
  - Ans
    - No
      - By principles number 1, the p-values they found only indicate how incompatible the data are with there null hypothesis.
      - By principles number 3, "A conclusion does not immediately become "true" on one side of the divide and "false" on the other."
        - They need more evidence and context to come to the conclusion, maybe by telling us which group is faster first.
        - Another study shows high fat high sugar diet reduces mice's acute running-wheel distance (Heather L. Vellers 2017), which contains more experiment and analysis details. This study might be the indirect evidence of the time difference between the two groups is not due to the cognitive power difference.

- (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?
  - And
    - No
      - By principles number 4, "data dredging, significance chasing, significance questing, selective inference, and p-hacking, leads to a spurious excess of statistically significant results in the published literature"
        - Since they are chasing the significance and doing multiple hypothesis testing without p-value correction, suspicion of the result is come from type-I error is reasonable. If they want to report the relationship of what they found, they should design another experiment to avoid model and dataset bias.

Problem 1.3.(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"?

➡ Ans

- Both of the titles are not appropriate.
  - The first title is not appropriate because the p-value does not measure the size of an effect but how incompatible the data are with the null hypothesis.
  - The second title is not appropriate because the not-affect chance below 5% not equal to over 95% success rate.

Problem 1.3.(d)

(1 points) 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?

➡ Ans

- No, regression only indicates the mathematical relationship, not about the causality.

Problem 1.3.(e)

(1 points) Even if a test is shown as significant by replication of the same experiment, we still cannot make a scientific claim. True or False?

➡ Ans

- True. A good scientific claim should include a good study design, understanding of the phenomenon under study, interpretation of results in context, and proper logical and quantitative understanding of data summarizations. A good claim can reproduce the same result many times, but reproducing the same result many times does not make an experiment become a good claim.

Problem 1.3.(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?

➡ Ans

- It's not OK, because the context and the number of hypothesis he tested all matter to his conclusion is valid or not. He can put those not significant result in the supplemental materials.

Problem 1.3.(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?

➡ Ans

- True. Because we use the p-value to estimate whether a hypothesis is not true in population, but we don't know what the population really is. The statistical model we build on the observations is not guaranteed to match the reality.

Problem 1.5 Why Most Published Research Findings Are False

Include your answer to this part in your written report.(No more than ~100 words, include equations if necessary.)

- 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

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

  ➡ Ans

$$PPV_n = \frac{P(\text{relation exists at least one of the n repetitions finds significant})}{P(\text{at least one of the n repetitions finds significant})}$$
$$= \frac{P(\text{relation exists})P(\text{claim relation n times} \mid \text{relation exists})}{P(\text{relation exists})P(\text{claim relation n times} \mid \text{relation exists}) + P(\text{no relation})P(\text{claim relation n times} \mid \text{no relation})}$$
$$= \frac{P(\text{relation exists})P(\text{claim relation n times} \mid \text{relation exists})}{P(\text{relation exists})P(\text{claim relation n times} \mid \text{relation exists}) + P(\text{no relation})(1 - P(\text{not claim relation after n times} \mid \text{no relation}))}$$
$$= \frac{\frac{R}{R+1} * (1-\beta^n)}{\frac{R}{R+1} * (1-\beta^n) + \frac{1}{R+1} * \left(1-(1-\alpha)^n\right)}$$
$$= \frac{R * (1-\beta^n)}{R * (1-\beta^n) + \left(1-(1-\alpha)^n\right)}$$

$0 \le \alpha, \beta \le 1$

for n=1,2,... $\begin{cases} PPV_{n+1} > PPV_n & \text{if } 1-\beta < \alpha \\ PPV_{n+1} \le PPV_n & \text{else} \end{cases}$

Let $n \to \infty$, $(1-\beta^n) \to 1$, $(1-\alpha)^n \to 0$

$$PPV \overset{n\to\infty}{\to} \frac{R}{R+1} \le \frac{R * (1-\beta^n)}{R * (1-\beta^n) + \left(1-(1-\alpha)^n\right)}$$

- Problem 1.5.(9)

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

  ➡ Ans

  - $1-\beta \le 0.05$

Problem 1.5.(10)

(5 points) Read critically and critique! Remember the gold rule of science, replication? For the third table in the paper, if researchers work on the same hypothesis but 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? What would be the PPV?

➡ Ans

❖

Problem 1.5.(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 likely to be false than true?

➡ Ans

❖ Yes, 1-PPV always > 0

Problem 1.5.(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.

➡ Ans

❖ R

❖ R will get lower and lower, and PPV will decrease and approach to 0.