JARI JOKINEN

Problem 1.1: Requires a written report (16 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*)

The goal of the experiment would be to find a link between the treatment variable of offering the possibility of getting vaccinated (i.e. intention to treat) and the outcome variable of not getting polio because of this offered option. The experiment would need two relatively large groups of people - the treatment group and the control group - in order to detect a treatment effect that is small. The experiment would need to be randomized, controlled and made double-blind in order to minimize the problem of possible confounding. After taking into account these three factors it would be possible to identify the causal effect by ensuring that in expectation the treatment and the control group are identically distributed in terms of any relevant feature. The group features would be controlled through stratification which would enable a subgroup analysis inside the groups and thus avoid the possibility of the differences between the groups coming through the differences in the covariates rather than through the treatment itself. This would mean controlling for demographic variables such as age, sex, or living location for example. Double-blindness with the observers and the experiment subjects would help to minimize the possible human behavior that could influence the outcome. The possible causal link should be tested with hypothesis testing by testing the expected mean of the treatment group against the expected mean of the control group and to see if there is a statistically significant difference between them. The null hypothesis would state that there is no difference between the means so that there wouldn't be a difference whether a person was offered a vaccine against polio or not regarding getting infected with polio. The alternative hypothesis would state that the expected mean of the control group is larger than the expected mean of the treatment group, indicating that offering the vaccine has an effect in reducing the risk of getting infected with polio in people. Finally, consent should be made between the participants of the experiment of the possible risks and consequences of the study and the study should be subject to review by an appropriate review board as it would involve human subjects.

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*)

For the NFIP study the treatment group consists of Grade 2 (vaccine) + Grade 2 (no consent) which total 350 000 people and the control group consists of Grade 1 and 3 (no vaccine) of 725 000 people. We can calculate the total number of polio cases in each group so that we get 391.5 polio cases in the control group (54*7.25) and 111.25 polio cases in the treatment group (25*2.25+44*1.25). By comparing just these two numbers we can see even without a test that there is a big difference between the two groups regarding the cases of polio (391.5>111.25) which points out that the vaccine is effective.We can further adjust the amount of the treatment group by multiplying Grade 2 (vaccine) and Grade 2 (no consent) to equal 725 000 and keeping the same proportions (0.64 and 0.36) so that the group consists of 725 000 to match the control group. Now we get 230.89 cases of polio in the treatment group (725000*0.64/1000000)*25 + (725000*0.36/100000)*44) which is still almost two times smaller than the amount of polio cases in the control group (391.5). The difference is so large that we don't need to perform a test to confirm whether there is a difference between the two groups. The study points to the direction that the vaccine is effective.

Making similar assumptions and calculations with the Randomized Controlled Double-Blind Experiment we get the size of the treatment group 550 000 and 200 000 for the control group. The polio cases in the treatment group consist of a total of 217 polio cases (28*2+35*4.6) compared to the total of 142 (71*2) cases in the control group. Now the difference is smaller than before and fewer cases in the control group (142<217) and in order to make a comparison we should perform a test to see if the difference is significant or not. The model follows a Bernoulli distribution where the null hypothesis is that the average rate of polio 0.00039 (217/550000) for the treatment group equals the average rate of polio 0.00071 (142/200000) for the control group (control = treatment). The alternative hypothesis is that these two rates differ (control > treatment). We can conduct a Fisher's exact test that doesn't assume knowledge about the true probability of polio in the control population. This test is based on a hypergeometric distribution where the test statistic is the number of people exposed to polio among the treated individuals (python code: scipy.stats.fisher_exact([[217, 142], [550000, 200000]]). We get a p-value of 0.00000012 which is so low that we can confirm on a alpha=1% significance level that the null hypothesis is rejected and that the vaccine is effective.

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

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)

The children being of different ages and sex, also come from different backgrounds which can be measured with demographic factors such as the living location of the family, wealth of the family, are parents together or divorced, how well the family is educated, and the size of the family for example. All these factors can influence the behavior of the children and the risk of getting contaminated with the polio virus. These factors can be controlled through stratification so that for each group - control and treatment - gets sampled proportionately from each subgroup that has been prevised according to these different demographic categories.

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)

Knowing that you would get treated or not could influence the behavior of people afterwards and thus bias the results. If people make the choice to get vaccinated it can mean that you are more concerned of health and safety reasons that affect the risk of getting infected with polio virus compared with the people that decline for one. These people could thus already be minimizing the risk of getting exposed to the virus and become even more committed to following a healthy and safe lifestyle after getting vaccinated. On the other hand, declining to get a vaccine could mean that you don't take the risk of the virus so seriously and would not change your behavior in a way that would decrease the risk of getting contamination. To counter this bias you should do an experiment where you divide people into two groups that are unaware whether they were given a vaccine or placebo so that their behavior wouldn't change after the experiment has been concluded.

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)

The people that opt for a vaccine might differ from the control group of people by being generally more mindful and better educated of health and sanitation issues in order to minimize the risk of being exposed to the virus. If the people in the two groups haven't been stratified correctly according to different demographic factors that influence the risk of getting exposed to the virus (i.e. such as financial wealth or education) this could bias the outcome that the vaccine was the reason for the result as there would be hidden confounding variables that were the actual reason or greater bias towards the result.

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)

For some people, the polio virus causes temporary or permanent paralysis, which can be life threatening. However, for most people polio won't show up as symptoms so these people might even not know that they're infected. These people with low symptoms will fight off the infection that causes symptoms such as fever, sore throat, headache, abdominal pain, aching muscles and the general feeling of being sick as in you would do when struck by flu: that is, by resting in bed, by taking painkillers and letting the illness eventually pass away through time. This means physically stronger, healthier and younger people in general will fight better against the symptoms and hence not become severely ill. Furthermore, being mindful about your own health, what you eat and drink as also of sanitation (e.g. washing your hands) decrease the risk of being infected with polio. These factors could be more looked after in a better educated and wealthier family of a child than in a low educated and poor family. Finally, all these factors together if not controlled properly may have contributed for the no-consent group having a lower rate of polio compared to the control group.

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)

Based on the trial numbers it seems both groups were exposed at a similar rate to polio whether their children got vaccinated or not. However, the sample size could be relatively too small to detect a treatment effect that is small. Therefore, you can't make statistically significant conclusions based on these trials and would need to increase the sample size in order to draw conclusions that would show a possible effect of the vaccine. If a large group of parents would refuse their children to participate in a similar experiment in the following year, this would decrease the chance of getting statistically significant results as the sample size might not reach the required large size. Furthermore, these actions could lead to a decrease in the credibility of scientists and doctors that research and test for polio vaccines as there might not be enough people for holding the experiment trials in order to scientifically prove that a certain vaccine helps to minimize the risk of getting contracted with the polio virus and thus getting the vaccine for the general population. This would thus increase the risk of the whole population of getting exposed to the virus and increase the risk of people getting seriously ill or eventually dying.

Problem 1.3: Requires a written report (25 points).
Read the statement by the American Statistical Association about p-values (Wasserstein and Lazar: The ASA's statement on p-values: context, process, and purpose) Limit answer for each part to 300 words.

(a-1) 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?

This way of planning for a study will probably produce statistically significant results when multiple tests are being produced while extracting the insignificant variables and then keeping only the most significant ones that eventually pass the requirement of the set p-value threshold, whether that be 1%, 5% or 10%. However, a policy decision should not be based solely on whether a p-value passes a certain specific threshold and by conducting multiple analyses of the data and reporting only those with certain p-values renders the reported p-values essentially uninterpretable, as this sort of study where you conduct tens or hundreds of different tests is bound to find statistical significance at some point somewhere -

perhaps in a place where there shouldn't be. Rather than falling in this trap, emphasis should be put on a coherent approach which emphasizes principles of good study design and conduct which means having a reasonable theoretical framework where to base the study design and how to conduct it. This framework should act as a benchmark to reason which variables should be included in the study and which perhaps not. This means understanding the context of physical and mental development of a child in early childhood and where the selection process of each variable should be debated based on the arguments for the inclusion of the variable and arguments against it. Moreover, a variety of numerical and graphical summaries of data should complement the main argument. Finally, all results should be interpreted in the context underpinned with logical and quantitative understanding of what the data summaries and conclusions mean.

(a-2) 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?

By obtaining more data, the chance of getting a statistically significant result increases, as it will be easier to conduct a test that produces at least a noticeable minor effect in size. However, this might still not produce the true effects and not estimate the actual size of the effect correctly. In order to make a coherent picture of the situation in order to make policy decisions, when making inferences of the data, the estimation of the size of effects should be stated together with the uncertainty surrounding the estimates.

(b-1) 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?

It is easy to find variables that seem to correlate together by not seeing the complete big picture; that is, not understanding and controlling for the confounding variables that affect this correlation. Spurious correlation is often caused by a third factor that might not be apparent at the time of examination. A simple example would be to not control for the elapsed time, as it is reasonable to predict that both the number of Nobel prize laureates will increase through population growth and because each year new prize winners will be announced. This however does not lead to actual correlation between the variables.

(b-2) 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?

No, you couldn't claim a correlation between sugar and coco consumption and the development of intelligence and brain growth. In order to make such a claim, you would need to exclude all the other outside confounding variables that might have an effect on intelligence and brain growth in order to make conclusions of a precise effect and at best, the conclusions would still include uncertainty. Also, even if there would be correlation between the consumption of sugar and coco regards the development of intelligence and brain growth that wouldn’t imply a causal effect, as the direction could go either way.

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

You could perform a study based for example on a survey conducted between a certain, long enough (e.g. 3-5 year), time interval where you try to control for all the possible confounding variables that relate to effects of chocolate consumption on intelligence or try to run a more controlled study where you observe people directly. In both experiments, you would divide people of enough sample size (e.g. 20-40000) to two groups of people - the treatment and the control group - and include people only of a certain age group (e.g. young children 10-12 olds) in order to better control all the possible confounding variables. The people would be randomly allocated so that the two groups should be similar in all factors that could otherwise influence the outcome, including those factors that we don’t know about. If possible people should not know which group they are in. Then, the treatment group would receive a certain amount of chocolate in order to consume it and the other would receive a placebo of enough amount so that they wouldn’t consume any extra chocolate on their own (i.e. refrain from consuming chocolate altogether). If the people receiving the chocolate would refrain from eating it, they should be counted to which they were allocated anyway. If possible those assessing the final outcomes should not know which group the subjects are in and ‘intelligence‘ should be carefully framed in a way that it would be able to quantify and measure it.

(b-4) 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? (300 words)

The difference in solving the maze could be only some seconds or a rather short time difference to make any decisive conclusions. Moreover, the sample size is quite small to make general conclusions. Finally, there could have been other confounding variables affecting the result such as what the mice had been fed and doing before the experiment, the conditions in the lab (i.e. human behavior towards the mice, living conditions, etc) which could have affected the behavior of the mice. There might be an association with chocolate consumption and feeling satisfied for example and this feeling of satisfaction could help the mouse to find its way out of the maze, but the same feeling of satisfaction could be reached by consuming some other ingredient like sugar.

(b-5) 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?

The P-value is the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis. It thus tests all the assumptions how the data were generated and not just the targeted hypothesis it is supposed to test. All the test results should thus be reported because the intermediate analysis results are part of the assumptions about the conduct of the analysis, and therefore, influence the resulting p-value. If they aren't reported you will get significant results no matter what your features are, because for example if you test 100 features with 100 different tests on a significance level of 5 % you are bound to find at least 5 significant features.

(c) 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"?

The title is misleading as the p-value doesn't show the actual size of the effect and a 95% confidence interval does not mean there is a 95% probability that this particular interval contains the true value. Moreover, the probability of the p-value does not refer to the hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. When interpreting the p-value, it should rather be seen as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model were correct. We cannot thus conclude that a new drug proves over 95% success rate of drug X on curing disease Y.

(d) Your boss wants to decide on the company's spending next year. He thinks letting each committee debate 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? (300 words)

The past doesn't give you enough information to make a prediction on the future that wouldn't also contain some amount of uncertainty. Relying solely on the past performance might miss the current situation where the company is situated in relation to its customers, suppliers and competition. As the company's revenue depends on all these factors, the decision whether to increase spending on the HR sector should be based on a thorough analysis of the company's current value chain that encompasses all the relevant stakeholders and the anticipations of the future. When relying only on the past information, critical information of the present, for example customer needs and market growth, is missed and the spending decision carried out today could only show its true effect in the future when a certain time would have passed. Finally, an increase in HR spending is usually a result of an expansion in the activities of a company, which usually is a result of larger revenues and not vice versa.

(e) Even if a test is shown as significant by replication of the same experiment, we still cannot make a scientific claim.

True or False? (300 words)

This will depend on how both of the studies were conducted, how data was collected, measured and analyzed, and how the analysis results were selected for presentation. The interpretation of the analysis part should be grounded on that all statistical methods and interpretations are premised on the model assumptions such that the model provides a valid representation of the variation we would expect to see across data sets. If these assumptions in both experiments reflect the circumstances surrounding the study and phenomena occurring within it, and the tests were conducted in a sound manner on a sample size that gives a reasonable notion of a claim for a size of an effect and the uncertainty surrounding the estimates, then we could make a scientific claim, as also one purpose for making a scientific claim is to show how the experiment could be replicated in different circumstances. Thus, replicating the same experiment in different circumstances and obtaining statistically significant results could act as a basis for a scientific claim, as making robust conclusions requires multiple studies.

(f) Your lab mate is writing up his paper. He says if he reports all the tests and hypotheses 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? (300 words)

The results should be transparent and complete with all the hypotheses testing that lead eventually to the conclusions of the paper. Without including all the other tests, the conclusions would not state the estimated size of the effects and the uncertainty surrounding the estimates correctly. Furthermore, depending on the size of the tests, the conclusions could be eventually false, and not replicable in other circumstances. In order for validation, the experiments should be able to be carried with a different set of data based on a similar interpretation of model assumptions. Therefore, it would not be alright to report only the statistically significant tests and hypotheses in the paper.

(g) 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? (300 words)

If the statistical model incorporates unrealistic or at best unjustified assumptions this could be the case that when we test to reject the null hypothesis and obtain statistically significant p-values our model might not actually match reality and we would thus be making a type I error by rejecting the null hypothesis. This could happen for example when the assumptions of the model are merely based on mathematical concepts of truth that do not match actual reality. For example it is difficult to make certain assumptions on human behavior and interactions because there are so many variables that need to be controlled and often a long sequence of actions depends on the successful completion of a certain behavior or interaction. Therefore, a simple model on paper might not reflect the actual reality in the world.

Problem 1.5: Requires a written report (15 points)

Read the paper by Ioannidis on why most published research findings are false (PLoS Medicine, 2005) and summarize the paper in your own words. What is the most important lesson you learned from reading this paper? How does Ioannidis get to the conclusion that a research finding is more likely true than false if

$$(1-\beta)R>\alpha$$

(at the beginning of page 697)? What does this mean?

(8) Show that the extent of repeated independent testing by different teams can reduce the probability of the research being true. (no more than 100 words)

As the term n which indicated the number of independent studies is included in the equation as a numerator and denominator, you can see that the second term of the denominator (-R*beta^n) cancel out with the numerator (-R*beta^n). Afterwards, only one term (1-alpha)^n is left in the denominator and increasing n thus will decrease PPV when n reaches towards infinity.

(9) What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming $\alpha=0.05$)

If the bias would be zero, the only terms in the PPV equation would be R/(R+alpha) and if the number of independent studies n would be zero, the PPV equation would be R/(R+1). If alpha would be 1 the two equations would be the same and thus for finding to be true, the prior probability of it being true would only matter.

(10) 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 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?

As a general critique towards the paper, all the referred papers seem to point to medical research studies which usually involve a large number of researchers, a large budget, capital-intensive equipment to perform the study and a long period of several years to perform the research. As the final result is uncertain, it seems natural that these research results end up having significant results as their final conclusions, because if there wouldn't be any results to show, there wouldn't probably be any funding afterwards. As there isn't any other real way to prove that your research matters than hypothesis tests through the evaluation of the p-value, then this is again the natural conclusion of a medical research study. Thus, the problem seems to be more a structural problem in the sector itself, and related to the way funding is given and the capital-intensive investment one needs to make - in humans and equipment - in order to perform this sort of research. As the results of research in general are uncertain, funding should be given in different ways: for research that is expected to get solutions for the market that will generate added revenues and research where the results are not so certain especially in the short term and perhaps difficult to measure, but whose results could bring fruit in the long term. This would include tolerance for non-significant results, and thus "failures", should also be accepted as results.

(11) 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?

The equations of PPV would then transform into (R/R+alpha) and (R/R+1), where then only the prior probability would matter whether publications would likely be true or not. Thus, continuing for example previously researched topics where the ratio of the number of "true relationships" to "no relationships" among those tested in the field is high would increase the likelihood of the publications being true.

(12) 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.

It would influence the beta variable which is noted to be the bias in medical research studies. The PPV would decrease depending on the proportion of probed analyses u and the extent of the bias beta that entails manipulation in the analysis or reporting of findings, typically as a form of selective and distorted reporting. The more concern researchers would base on getting the right p-values leads to influencing the parameter beta that decreases the PPV, that is the probability that the research finding is correct.