Question 1.1.1:
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.

Answer:
The target of the study is running a randomized, controlled and double-blind experiment as mentioned in the lecture. This can be achieved by two different aspects: The choice of the participants and the architecture of the test.

The participants should be randomly chosen from a set of a population, in which we want to analyze for example the effects of an injection. If we want to test the immune answer of all Europeans we should make sure, that all European countries are considered. Different parameters like the sex or age of a person should be considered as well. From that pool we draw a set, which gets the injection of the vaccine (treatment group), and a different set, which gets an injection of a placebo (control group).

The architecture of the test should take into account, that neither the patient nor the experimenter should know, in which of the two groups individuals belong to (double-blindness). In addition to the e.g. death rate due to a disease a parameter of general death rate due to other reasons should be included to make sure that the two groups can be compared to each other.

Question 1.1.2:
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.

Answer:
In case of the first study from 1954 the groups cannot be compared without doubts. On the one hand, the vaccinated group of grade 2 students seem to be comparable with the "control group" of grade 1 and 3 students, who did not get any vaccination. But the differing age of participants might have effects on the polio rate. This could be minimized, if the groups would consist of more comparable individuals. On the other hand participants could behave in a different manner if they knew about their vaccination. This could be reduced by a double-blind experiment.

There are more reliable conclusions possible from the second experiment. The treatment group can be compared with the control group, because the randomization and the double-blind experiment, minimizes the two effects described before.

Nevertheless, the polio rates of participants, who did not consent to the vaccination, and the vaccinated students seem similar in both experiments. But here again, the age structure is not considered (or mentioned), which decreases the reliability of conclusions.

Question 1.1.3a:
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.

Answer:
The statistics cannot eliminate these doubts. If the test result would show separate numbers of the "control group" of grade 1 and 3 students, this effect could be indicated, but not be proven. Different factors could impair the immune system of the individuals in a different manner. Individual factors like age-differing immune answers could be a cause different polio rates. Other factors could be based on the environmental circumstances. For example there is a lesson in the second grade, which grade 1 or 3 students do not attend. Swimming lessons or demanding sports could have different effects on the health of students.

The best way to overcome these issues is the use of randomization when drawing the individuals for the different groups. Important is furthermore, that the age of the students, for which the experiment is made for, is precisely named in the publication.

Question 1.1.3b:
Let us examine how reliable the estimates are for the NFIP study. A train of potentially problematic but quite possible scenarios cross your mind:
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.

Answer:
The double-blindness is an important factor in experiments. Students are likely to behave in a different manner if they know that they are vaccinated, or not vaccinated respectively. They might be more cautious if they know, that they did not get a vaccination, or analog inversely less cautious if they did. This behavior could effect for example the number of friends students visit per week.

The students behavior can only be as comparable as possible if both, the treatment and the control group, think they are vaccinated. This is achieved by using a placebo next the corresponding drug. The experimenters as well should not know who gets a placebo or the drug to minimize the bias from this side of the experiment.

Question 1.1.3c:
Let us examine how reliable the estimates are for the NFIP study. A train of potentially problematic but quite possible scenarios cross your mind:
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.

Answer:
A possible scenario of reducing the chance of getting infected with polio is being in a class full of vaccinated children. These are less likely to get infected and thus less likely to infect classmates. With an increasing rate of immune students the chance of getting an infection reduces.

This effect can be reduced by analyzing individuals infection history separated from other participants. For example only a small amount of students of individual schools could be asked to attend in the experiment. By considering a large amount of schools the necessary number of participants can be achieved.

Question 1.1.4:
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?

Answer:
The no-consent groups of both experiments had comparable infection rates, whereas the comparability should not be overestimated due to differences in the study setup. In both cases the students knew that they were not vaccinated and could have behaved more cautious to prevent infections.

They control group of the second study contrarily might have thought that they were vaccinated and could have behaved less cautious. Students could have been meeting with friends like the vaccinated group, but without the sufficient security of a vaccination. The cause of the differing numbers could based in these different behaviors.

Question 1.1.5:
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?

Answer:
This conclusion is logically correct if only these two numbers would have been published. On the other hand the reduction of polio rate in the treatment group in comparison to the control group is considerable.

If a high stake of parents would behave like this in the next experiments, chances are that there are not enough participants to draw reliable conclusions from adequate sample sizes. This scenario would lead to poor results or even to the situation that the experiments cannot be executed. Thus the vaccination will not be tested enough to be released.

Question 1.3a-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?

Answer:
This approach has a high chance of producing results, which indicate dependencies between some variables and the early childhood education. But the essence of the article "ASA Statement on Statistical Significance and P-Values" states that the exclusive use of the statistical significance as parameter on which policies are evaluated is a poor approach.

Nevertheless, the sampling of large data sets with multiple different variables might lead to assumptions, which are supported by a p-value below the widely used limit of 0.05. But the scientific research should not stop with this discovery. There is still a good chance that the outcome happened because of coincidence. Therefore the low p-value must not be mistaken for "truth".

In the article mentioned above this analysis based on the p-value is only one part, on which decisions or scientific researches should be based on. A robust experimentation setup is the first step, which includes a sufficient high number of data points and good quality of measurements. After finding significances in the data, the results should be understood and analyzed in one scope. If the results follow a logical and reproducible manner, there is no reason, why policies should not consider new insights.

Question 1.3a-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?

Answer:
Indeed, with an increasing number of data points, the reproducibility increases. But just like Question 1.3a-1 indicated, a low p-value must not be mistaken for the "truth".

Besides the high number of data points a good measurement quality and solid experimentation setup in general is fundamental. But with more data in terms of number of features the chance of coincidental significances increases as well.

The p-value, which shows how good the data fits to a "specified hypothetical explanation", might lead to important insights. But it should not be mistaken with a true relationship between a cause and it's consequences. If the significance can be supported in a reproducible manner and if it follows logical rules, it can be an very important indicator for an significant behavior of the data with respect to specific hypotheses.

Question 1.3b-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?

Answer:
If the economist would solely rely on the p-value to support his thesis, he would probably find enough "evidence" to make him believe that there is a relationship between chocolate consumption an the number of won Nobel prizes. But by doing this the economist would underestimate possible other explanations and might draw poor conclusions.

In this example one cause of high values in both features (chocolate consumption and number of won Nobel prizes) might be the income of a state. If it is high, more money can be invested in education and research. Equally there can be more money spent on luxury products like candies. This connection would not be drawn by the economist solely by expressing the significance in terms of a p-value. Therefore the correlation should not be mistaken with causality.

This is a good example of a study, in which the p-value is weighted to highly up to the degree, where it is the only explanation of a relationship. But just as in the article "ASA Statement on Statistical Significance and P-Values" mentioned, there are more factors to consider, when a result is to be presented.

Question 1.3b-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?

Answer:
The neuroscience lab wants to state that there is an relationship between the consumption of chocolate and intelligence by analyzing a states consumption and the number of Nobel price laureates.

But despite the correlation between these two variables this approach lacks of some important aspects. Firstly, the researchers use the number of won Nobel prices as a variable for intelligence. Even if mostly smart scientists are winning the award, there is no reason to assume, that even smarter people exist, who do not win a Nobel price. Besides intelligence, environmental factors like education and research funds lead to superior scientific results. Therefore the number of Nobel laureates should not be used as an equivalent for intelligence.

Even if we take the stated connection (number of Nobel prices and intelligence) as given, it was not analyzed if the laureates are consuming more chocolate than average. And this connection would be drawn by the first approach. Therefore the experimental setup is very poorly chosen. The authors want to refute the null-hypothesis that there is no relationship between chocolate consume and intelligence by using data, which is logically not connected to the thesis.

Summarizing, the authors should not publish this insight without further investigations.

Question 1.3b-3:
In order to study the relation between chocolate consumption and intelligence, what can they do?

Answer:
To analyze the connection between chocolate consumption and intelligence, the architecture of the experiment should firstly focus on the data measurement, which is demonstrable valid for this purpose. The consumption of chocolate can either be defined in a (double-blind) test or figured out in a survey.

If the researchers want to show that acute chocolate consumption is beneficial for the intelligence, the double-blind test should be used. The participant gets in one time episode chocolate and in the next one an adequate replacement. After each episode a test should be done to measure the intelligence as objectively as possible (e.g. IQ-test).

If a connection between long-term chocolate consumption and the resulting intelligence of individuals should be analyzed, a survey with a corresponding intelligence test should be executed. But here an comparable control group is necessary. Further variables should be reported as well to show, that different individuals only can be separated with respect to the consumed chocolate.

Either way, the number of data points should be high enough to get robust and reproducible outcomes.

Question 1.3b-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?

Answer:
The general approach of this study seems reasonable. The test is set up, such that the mice's behavior after the consumption of chocolate can be analyzed. And in contrast to earlier questions the connections between the variables seems more comprehensible.

But the number of tested individuals is quite small. Therefore chances are high that a significant result occurred by coincidence. The p-value is an easy to use tool for checking if the null-hypothesis should be rejected. Drawing the conclusion only based on the p-value should be avoided, as mentioned in the article "ASA Statement on Statistical Significance and P-Values".

The insight won in this study could be a good starting point for further investigations, but should not be presented as the ultimate truth.

Question 1.3b-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?

Answer:
Of course the lab could write about different effects on the IQ, if the study shows that the first assumption was not correct. But there are some dangers if a study with multiple features is scanned for correlations. A significance level of 0.05 results in 5 expected significant events out of 100.

A possible way to reduce the number of wrong rejections of the null hypothesis is for example the usage of the Family-Wise Error Rate (FWER) or the False Discovery Rate (FDR). With help of one of these tools we can reduce the risk of rejecting the null hypothesis too soon.

The FWER expresses the probability of making at least one false discovery, whereas the FDR estimates the fraction of making false significant discoveries in comparison with the significant discoveries. One of these values should be mentioned to give the reader any information about this aspect.

Without any of these tests, the result should at best be used for further investigations. Here again, the p-value needs to be explained in a context to be of any value.

Question 1.3c:
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"?

Answer:
This is a dangerous example for the publication of scientific results. At first point it could be mentioned that the result was not a cure for a disease rather than a "treatment effect" of a new drug, which might be a difference. The second headline mistakes the p-value for the percentage of cases, in which the drug did not cure a disease. In both cases the journalist writes about study results which might be untrue.

Nevertheless, a too strict view on the p-value is sometimes the trigger of discussions. But in this case, a p-value of 0.05 is too high for studies relating to drug effectiveness. A better and in this context widely used choice should be a significance level of 0.01. The development and production is a very expensive procedure where decisions based on a weak foundation may lead to serious consequences. The safety of patients is another at least equally important point and reason to formulate stricter limits.

Question 1.3d:
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?

Answer:
Given that the analysis is based on annual balances, we have 25 data points on which we run our regression analysis. Therefore one cannot rely on trends from the statistical point of view. Here is where the knowledge of the specific area gets important.

If the data gives information about correlations in the data set, it is the person with domain knowledge, who has to draw the conclusion. In this example the data scientist may say that there is an correlation between variables, even if it is not statistically significant. But the person in charge has to decide whether this connection is meaningful or not.

One conclusion could be that the boss rejects investments in HR because of the possibility, that this correlation is due to randomness. But in this example he should be informed that it could still be a possibility to increase the revenue.

Question 1.3e:
Even if a test is shown as significant by replication of the same experiment, we still cannot make a scientific claim.

True or False?

Answer:
Given the scenario that there was an significant experiment which should be confirmed in a second study several considerations should be made to report a scientific claim. Firstly the experimentation setup must be robustly executed. If a too small sample size or poor variables are measured, there is no improvement of the claim just by repeating the experiment. For example, if we state that a states caffeine consumption leads to higher income by measuring both, this could result in the same outcome. But here are questionable assumptions made, which are rooted in the experimentation setup. This scenario might actually lead to significance, but it should not be published.

Secondly, the general rule should be considered, that the statistical significance is not the only criterion on which scientific claims should be made. Even it is an important objective factor and over all domains accepted, there should always be an awareness, that a type I error is being made or more in general there is a statistical significance despite the conclusion was drawn incorrectly.

To cut a long story short: There is no simply "True or False" to this question. The answer depends on the circumstances and is not True in general. But testing significant experiments can be an important tool to confirm or refute hypotheses in a scientific way.

Question 1.3f:
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?

Answer:
The person who is following this approach should better explain what he did and why he did it in that way. Without the information about the whole experiment the results are judged in a different manner. As the article "ASA Statement on Statistical Significance and P-Values" states, "p-values and related analyses should not be reported selectively".

If he wants to publish his results he should follow the instructions of the above mentioned article: "Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting."

Only by following this rules and analyzing the chance of mentioning an significant event due to a high number of variables by giving the FWER (Family-Wise Error Rate) or the FDR (False Discovery Rate) he can publish his claims.

Question 1.3g:
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?

Answer:
This question can be answered with True.

Of course the null-hypothesis can be True, even if the p-value indicates that it should be rejected. This is almost the definition of the p-value: Probability that the test outcome is at least as extreme as given in the study.

Another point is mentioned in the article "ASA Statement on Statistical Significance and P-Values". If the p-value indicates correctly, that the null-hypothesis should be rejected there is still a chance, that the hypothesis itself is not correct. For example if a states caffeine consumption correlates with it's income, there might be a correlation but one can hardly explain the causality. Therefore the reality is different of what the experiment would indicate.

Question 1.5.8:
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

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

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

Answer:

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

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

with α as significance level and β as type II error rate.

To show that the PPV decreases with increasing n, we have to check whether the term

$$\frac{\partial(PPV)}{\partial n} < 0, \forall n \in \{1, 2, ...\}$$

holds.

This is indeed true for $1 - \beta \geq \alpha$ according to the article "Why Most Published Research Findings Are False".

Question 1.5.9:
What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming $\alpha = 0.05$*)*

Answer:
The PPV can be increased by increasing the R value. This can be done e.g. by limiting the number of variables, which should be analyzed to the most likely parameters.

A further approach is to increase the power $(1 - \beta)$ of a study. But this way is not possible due to the problem definition that α is hold constant and both parameters interact with each other.

Question 1.5.10:
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.)

Answer:

Question 1.5.11:
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?

Answer:
Both, bias and number of competing teams, have a negative effect on the PPV. Even if both are missing, there might still be a chance that a relationship is published, which does not exist. These two sources of uncertainty are merely minimized, others might still exist.

This risk increases with a high R value. This could root for example in a highly multidimensional data set, in which pairwise correlations are investigated.

Question 1.5.12:
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.

Answer:

The parameter α is chosen study-specific beforehand. β describes the power or the chance of a type II error and is dependent from the data and α. But the p-value influences the parameter R because the type I error is accounted in it.

The PPV decreases as the number of random relations increases. If the p-value is the only criterion to judge whether a relationship is significant, a rate of about α is found (without corrections). Many of those are likely to be non-correctly evaluated. Therefore the PPV tends to be low.