MITx - 6.419x

Data Analysis: Statistical Modeling and Computation in Applications: Spring 2021

Homework 1: Written Report by J. Andrew Seidel

Problem 1.1 The Salk Vaccine Field Trial

The first polio epidemic hit the United States in 1916. By the 1950s several vaccines against the disease had been discovered. The one developed by Jonas Salk seemed the most promising in laboratory trials. By 1954, the National Foundation for Infantile Paralysis (NFIP) was ready to try the vaccine in the real world.

They ran a controlled experiment to analyze the effectiveness of the vaccine. The data is shown in the first table below (grade refers to educational stage).

NFIP Study:

	Size	Polio rate per 100,000
Grade 2 (vaccine)	225000	25
Grade 1 & 3 (no vaccine)	725000	54
Grade 2 (no consent)	125000	44

From this table, you interpret that the experiment was run the following way:

Vaccines were offered to Grade 2 students, but some Grade 2 students did not consent and opted out of the offered vaccine.

Vaccines were not offered to Grade 1 and Grade 3 students.

The experiment was later repeated as a randomized controlled double-blind experiment. The data is shown in the second table below.

Randomized Controlled Double-Blind Experiment:

	Size	Polio rate per 100,000
Treatment (vaccine)	200000	28
Control (Salt Injection)	200000	71
No consent	350000	46

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)

- Representative and Consent: From a group that is as representative of the broader population in question (as practically as possible), ask for participant consent for the vaccine study.
- Statistically Large Enough Sample: Given the incidence of polio, ensure this consenting group (the study size) is sufficiently large, such that counts provide a satisfactory degree of precision.
- Stratified Randomization and Control: Of those who consent, use stratified random sampling to split the group evenly into a treatment group and a control group. Stratified samping could be based a number of potential factors to address any potential significant differences in proportions within the consenting group, including gender, race, geographic region, family income, pre-existing health conditions, age, etc. The treatment and control groups should be comparable.
- Double Blind: Administer the vaccine injection to the treatment group and a placebo injection to the control group. Neither those administering the treatment or placebo, nor the individuals consenting to the study should have knowledge which injection they received
- Objective Outcome: Determine objective standards for testing the two groups for the diagnosis of polio and record the outcomes.

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)

- NFIP: The observed control group, consisting of Grade 1 & 3 students not receiving the vaccine, had a polio rate of 54 per 100,000, while the treatment group, consisting Grade 2 students, who consented and received the vaccine, had a polio rate of 25 per 100,000, which is a 53.7% reduction in the rate of polio. Using the CDF function on either a binomial (H0: p=54/100,000, HA: p<54/100000) or poisson (H0: lambda = 54, HA: lambda < 54) distribution model, we find that the probability that the vaccine doesn't reduce polio is very low (p < 9e-6 for both). Similarly, using Fisher Exact and Likelihood Ratio tests, we find the p < 3e-9 to be very small and that the vaccine is effective.

- Randomized Controlled Double Blind: The control group had a polio rate of 71 per 100,000, while the treatment group had a polio rate of 28 per 100,000, which is a 60.6% reduction in polio. Using the CDF function on either a binomial (H0: p=71/100,000, HA: p<71/100000) or poisson (H0: lambda = 71, HA: lambda < 71) distribution model, we find that the probability that the vaccine doesn't reduce polio is very low (p < 6e-9 for both). Similarly, using Fisher Exact and Likelihood Ratio tests, we find the p < 5e-10 to be very small and that the vaccine is effective.

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:

3.1: (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)

- Yes, such differences, including age, or other disproportionate factors including gender, ethnicity, or health history, etc. between Grade 2 from Grade 1 & 3 children could influence results.
- Change the NFIP experiement design as follows to prevent this difference between groups to from making the estimate not reliable: Ask all children in Grades 1, 2 & 3 for consent. Of this group that consents, use stratified random sampling across factors (including age, gender, ethnicity, etc.) to evenly split the consent group into the treatment group and the control group.

3.2: (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)

- Yes, children knowing whether they were vaccinated (or not vaccinated) can bias the result.
- Studies on the placebo effect have shown a relationship between a patient expectations and their health responses. Administering a vaccine may set positive expectations for the patient, potentially resulting in a positive health response, and not administering the vaccine may set a different expectation and potentially a different health response, perhaps neutral or negative. By making the experiement double blind, where neither the administrator, nor the patients are aware of whether they are receiving the vaccine, and by administering a placebo to a control group, we can help prevent this type of bias by elminating differences in administration that could set patient expectations and influence health responses.

3.3: (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)

- See answer above from prior question regarding double blind design, placebo effect and the effect of setting expectation on health responses. The act of "getting vaccine", whether it is an actual vaccine, or a placebo, still influences expectations and potential health responses versus not being administered either a vaccine or a

placebo. By blindly administering the vaccine to the treatment group and also blindly administering a placebo to the control group, we standardize the administration across both groups and elminate this bias.

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)

- There may be factors that have a different proportionate representation between the control group and the no-consent group, that could influence the likelihood of contracting polio - an infectious disease, as well as the likelihood of consenting to the study.
- Parents with children who are less likely to contract polio may also be less likely to consent to the study, and similarly, parents with children who are more likely to contract polio may be more likely to consent to the study, which could explain in part why the non-consenting group had a lower rate of polio. Factors guiding parents choice to consent or to not consent may include: parents level of education, family income, access to private transportation, access to private healthcare, population densities across groups, as well as any differences in the local manner or language that was used to present the vaccine and solicit consent.
- Some portion of the difference between rates of polio between the consenting group and non-consenting group could also be explained simply by randomness of the study and sampling.

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?

- No, the parents incorrectly concluded that refusing consent to administer the polio vaccine to their child would reduce the likelihood of the child contracting polio. The consequences of a large group of parents refusing consent would be a reduction of children receiving the actual vaccine, and a likely higher rate of polio among those children who did not consent, relative to those children who received the vaccine.

Problem 1.3: Statement by the American Statistical Association about p-values (Wasserstein and Lazar: The ASA's statement on p-values: context, process, and purpose)

(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?

- No, this is not likely to produce good policies:
- As detailed in The ASA's Statement on p-Values: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold. Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "p < 0.05") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making."
- P-values can indicate how incompatible the data are with a specified statistical model, but "p-hacking" large numbers of variables leads to a spurious excess of statistically significant results. Valid scientific conclusions based on p-values and related statistics cannot be drawn without also presenting to the policy maker how many and which analyses were conducted, and how those analyses were selected for reporting.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result, and by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

(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?

- No, additional data will not help find the "true effects."
- As detailed in The ASA's Statement on p-Values: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.... Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough."
- Also, additional data, defined as additional participants in the study, may provide additional precision around estimates and help if the study isn't sufficiently large enough for statistical significance, but will not alter the experiment design nor change the design flaws cited above.
- Additional data, defined as changing the experiement design to include new factors or features collected within the study that may have causal relationships with the dependent variable could help find the "true effects", but numerous poorly chosen factors may also increase the likelihood of spurious statistical significance, which does not help find the "true effects".

(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?

- Yes! There is certainly a proven relationship between chocolate and winning a Nobel... no, no, no I'm completely joking. Sorry - it has been a long week and I figured you the grader (like me) could use a laugh right now. Hope you're enjoying the course too. This is totally a candidate for the 10% "valuable comments" grading portion :) Don't pretend you're not smiling!
- To be clear - the economist should **NOT** conclude that there exists a relationship between national chocolate consumption and national Nobel prize laureates.
- As detailed in The ASA's Statement on p-Values: "By itself, a p-value does not provide a goodmeasure of evidence regarding a model or hypothesis. Researchers should recognize that a p-value without context or other evidence provides limited information.... For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible."
- This relationship is likely the result of confounding variables that influence both the independent variable (national chocolate consumption) and the dependent variable (national count of Nobel Laureates). Examples of confounding variables might be higher national average income, which allows for higher consumption of luxury foods like chocolate and affords higher quality education, which might thereby result in higher likelihood of producing Nobel Laureates.
- We shouldn't conclude chocolate consumption itself increases the likelihood of producing Nobel Laureates, especially when other approaches are both appropriate and feasible.

(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?

- No, they should not conclude that there exists a relationship between chocolate consumption and intelligence. Please also see answer above regarding confounding variable related to chocolate consumption and Nobel Laureates.
- As detailed in The ASA's Statement on p-Values: "By itself, a p-value does not provide a goodmeasure of evidence regarding a model or hypothesis."
- Further, the initial hypothesis addresses the relationship between sugar and coco consumption and their effect on development of intelligence and brain growth, yet the experimental design addresses national chocolate consumption and the count of Nobel Laureates, which are variables at a national level that may differ significantly from the hypothesis at the individual level. For example, national chocolate consumption may have no relationship to an individual Nobel Laureate's personal chocolate consumption habits. Similarly, there are many ways to measure development of intelligence and brain growth, which many only be weakly related to relatively sparse

and binary incidence and Nobel Laureates. The absence of being a Nobel Laureate doesn't mean the absence of intelligence and brain growth.

(b-3). (1 point)

In order to study the relation between chocolate consumption and intelligence, what can they do?

- To study the relation between chocolate consumption and intelligence:
- Change the scope of the variables from the national level to the individual level, or some other grouping that reduces the potential for disproportionate chocolate consumption and/or intelligence within the group.
- Identify an objective measure for chocolate consumption (eg grams per day, over a certain period).
- Identify an objective measure for intelligence, (eg: IQ or other cognitive test for intelligence), ideally one with a wide range of possible values (beyond just Nobel Laureate or not).
- In an observational study, record these measures. Preferrably, conduct a random controlled test, by prescribing a measured amount to test group (with consent). Or ideally, if possible, conduct a paired test within the same individuals, whereby they are measured over randomized different periods of time with different chocolate consumption / no chocolate consumption, and test for intelligence after each period.
- Conduct the study

(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?

- Not necessarily: they should not conclude that chocolate consumption leads to improved cognitive power in mice, but they could conclude that it is likely that chocolate consumption changes the time required for mice to solve the maze. Also, they didn't indicate above if the difference is an increase or decrease in time, so we can't assume improvement with lower time.
- As detailed in The ASA's Statement on p-Values: "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone."
- To support the claim of improved cognitive power in mice, they would need to show a relationship between the reduced time required to solve a maze and improved cognitive power.
- As a counter example, reduced time required to solve the maze may be a result of phyical/metabolic changes within the mice group with chocolate added to their diet, such that the mice solve the maze simply by traveling faster, but may still make more mistaken turns or repeated mistaken turns within the maze, which would indicate higher physical ability but lower cognitive power.

(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?

- No, it is not the correct approach to only write about statistically significant results.
- As detailed in The ASA's Statement on p-Values: "Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable." We should avoid "p-hacking".
- Also: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." and "large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise." We shouldn't necessarily discard potentially important findings solely on the basis of their p-value being "unimpressive."

(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"?

- No, neither article title is appropriate.
- Regarding "New trial shows strong effect of drug X on curing disease Y": As detailed in The ASA's Statement on p-Values: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." "Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects." The small p-value does not necessarily represent a "strong effect". Treatment effectiveness, however it is defined, may also be different from curing a disease.
- Regarding "New drug proves over 95% success rate of drug X on curing disease Y": As detailed in The ASA's Statement on p-Values: "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone." The p-value below 0.05 does not represent a 95% success rate.

(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?

- No, "We shouldn't increase our spending on HR" is incorrectly reasoned.
- As detailed in The ASA's Statement on p-Values: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." The study indicates a large effect and the business decision should not be made based only on the p-value.
- Similarly from the ASA: "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." The low p-value for HR spending does not mean there is a small effect, or that HR spending is unimportant.

(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?

- True. As detailed in The ASA's Statement on p-Values: "By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. Researchers should recognize that a p-value without context or other evidence provides limited information." "data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible." * The significance and replicability of an experiment alone provide only limited information related to that specific experiment and do not necessarily indicate an appropriate model or hypothesis to support scientific claims. Additional context and evidence is needed to support such scientific claims.

(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?

- No, it is not okay to report only the statistically significant ones. As detailed in The ASA's Statement on p-Values: "Proper inference requires full reporting and transparency P-values and related analyses should not be reported

selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable." "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided."

- It is important to include all the tests and hypothesis, including those that are statistically insignificant, even it means the report will be long.

(g). (2 points)

If I see a significant p-value, it could be the case that the null hypothesis is consistent with truth, but my statistical model does not match reality.

True or False?

- True, a null hypothesis could be consistent with the truth, however an incorrectly applied statistical model that does not match reality, may incorrectly yield a significant p-value. Often, we do not know the true underlying distribution, and for examply, we may use a model that assumes a lower amount of dispersion that results in smaller p-value than the true p-value and thereby overstates significance of the p-value, when in fact the p-value is not significant, and the null hypothesis is truth and should not be rejected.

Problem 1.5: Why Most Published Research Findings Are False, John P. A. Ioannidis

(No more than ~100 words, include equations if necessary.)

(8). (3 points) Show that the extent of repeated independent testing by different teams can reduce the probability of the research being true.

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

We incorporate multiple studies, n , within Table 3: Research Findings and True Relationships in the Presence of Multiple Studies

Research Finding	True Relationship		
	Yes	**No**	**Total**
Yes	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t003

"Why Most Published Research Findings Are False, Table 3: Research Findings and True Relationships in the Presence of Multiple Studies, John P. A. Ioannidis")

$$PPV = \frac{P(\text{relation exists, at least 1 of n is significant})}{P(\text{at least 1 of n is significant})}$$

$$PPV = \frac{P(\text{relation exists, at least 1 of n is significant})}{P(\text{relation exists, at least 1 of n is significant}) + P(\text{no relation exists, at least 1 of n is significant})}$$

From table 3 above:

- Numerator: $P(\text{True Relationship, True Research Finding}) : c\,R\,(1-\beta^n)/(R+1)$

- Denominator:
  $P(\text{True Relationship,True Research Finding}) + P(\text{False Relationship,True Research Finding})$

- Denominator: $c\,R\,(1-\beta^n)/(R+1) \quad + \quad c\,(1-(1-\alpha)^n)/(R+1)$

- Denominator: $c(R+1-(1-\alpha)^n-R\,\beta^n)/(R+1)$

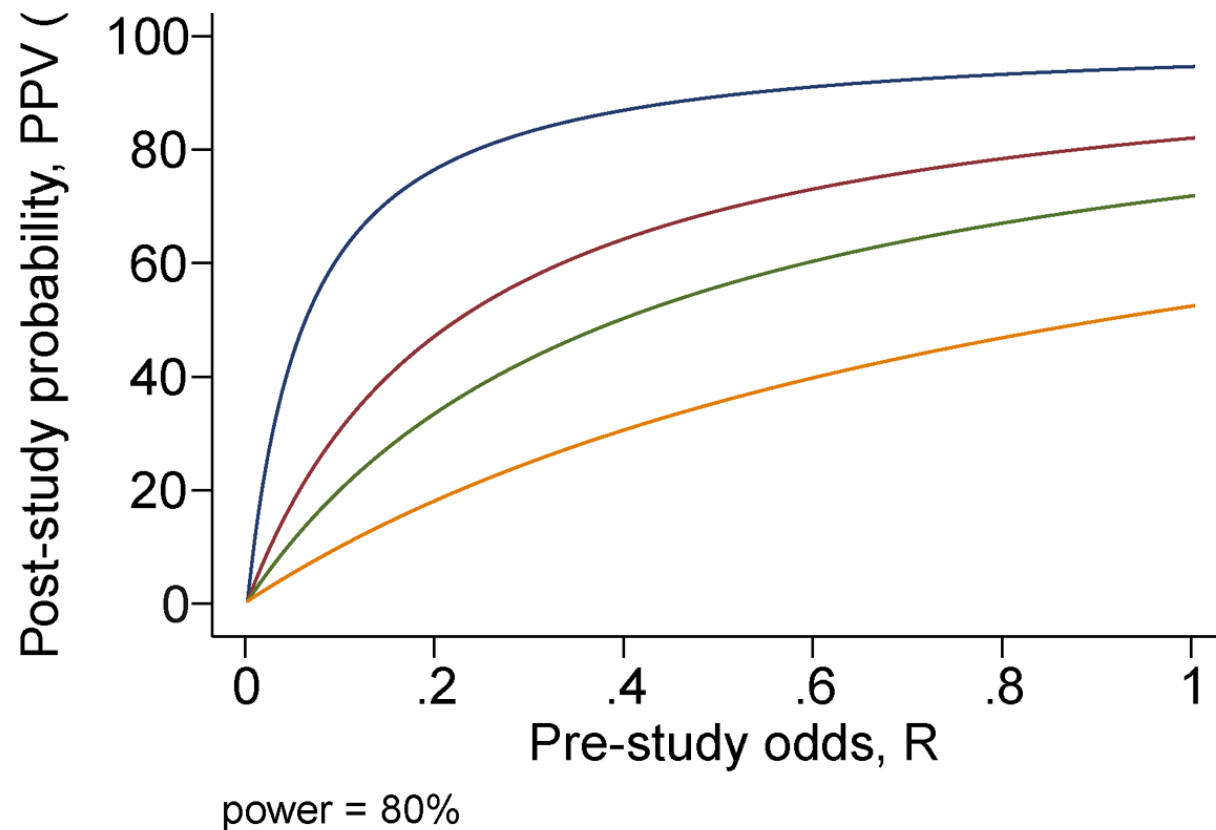$$PPV = \frac{c\,R\,(1-\beta^n)/(R+1)}{c(R+1-(1-\alpha)^n-R\,\beta^n)/(R+1)}$$

- we can cancel the terms c and $(R+1)$ in both the numerator and denominator, leaving:

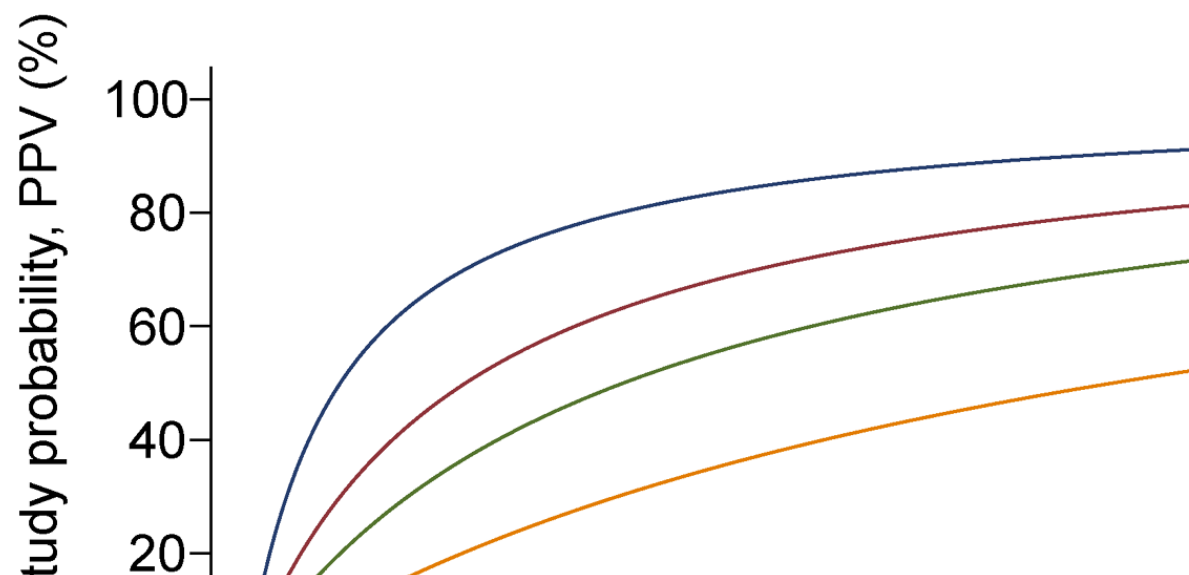$$PPV = \frac{R(1-\beta^n)}{R+1-(1-\alpha)^n-R\,\beta^n}$$

- Numerically from the equation above, and graphically from the charts below from the original paper, we can see that PPV decreases as n increase when the power is greater than alpha: $1-\beta > \alpha$
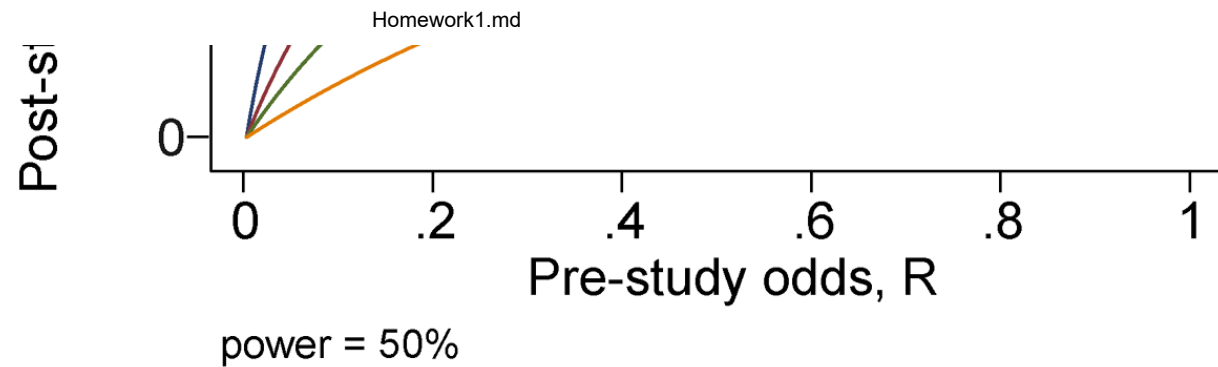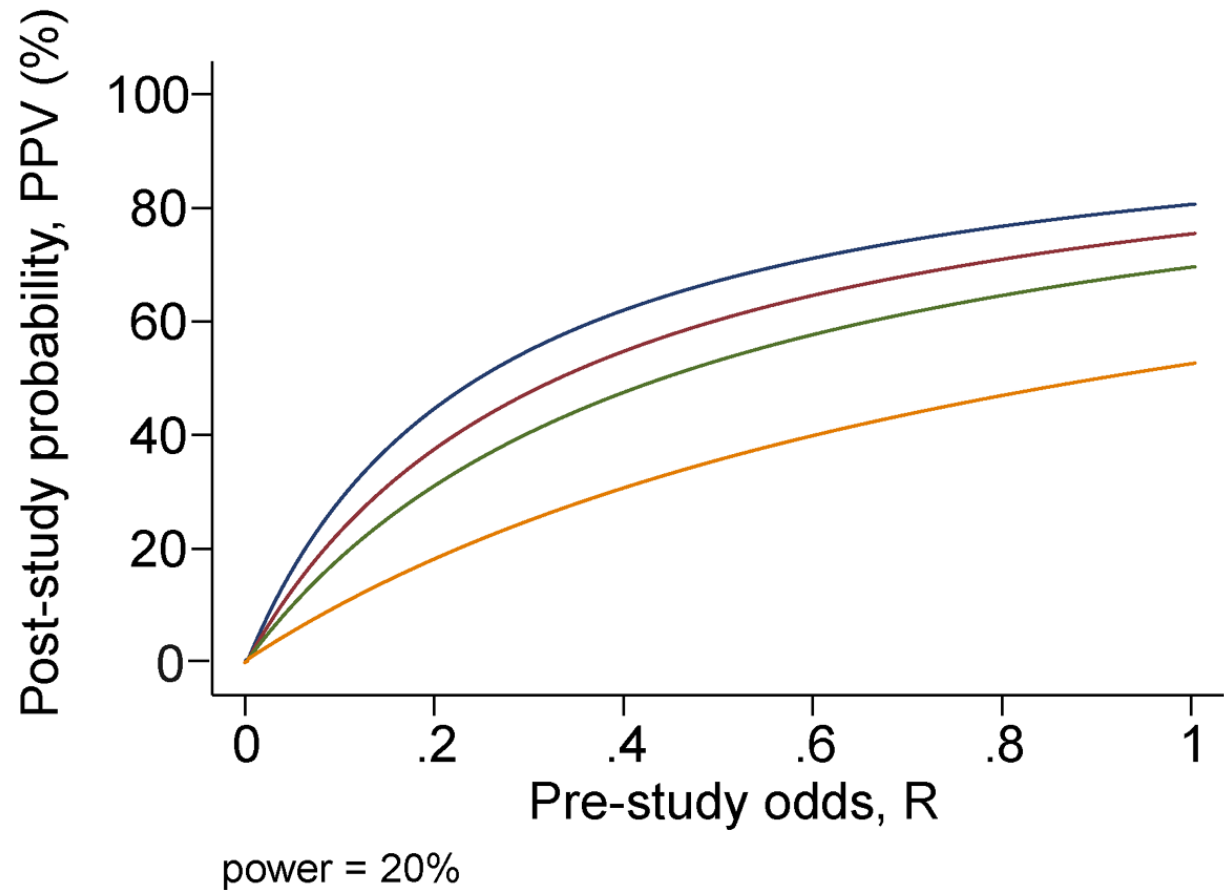
A

(%)

power = 80%

B

power = 50%

power = 20%

——— n=1   ——— n=5   ——— n=10   ——— n=50

"Why Most Published Research Findings Are False, Table 3: Research Findings and True Relationships in the Presence of Multiple Studies, John P. A. Ioannidis")

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

- PPV tends to decrease with increasing number of independent studies, and with increasing bias unless $1 - \beta < \alpha$.
- Assuming type I error significance at 0.05 ($\alpha = 0.05$), if type type II error (β) was greater than 0.95, PPV would not decrease with increasing bias or an increasing number of teams testing the same hypothesis.

(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.)

- See question 8 above for additional introductory context
- We can incorporate the multiple research teams as the n term and model the Positive Predictive Value as follows:
$$PPV = \frac{R(1 - \beta^n)}{R + 1 - (1 - \alpha)^n - R\,\beta^n}$$
- This model generates a lower PPV with a value of $n > 1$ than if only a single team of researchers were addressing the topic (where $n = 1$), subject to the condition that the power of the tests are greater than the type 1 error. See answer 9 for additional details on this assumption.

(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?

- As cited in the paper: "A research finding is more likely true than false if $(1 - \beta)R > \alpha$. Since usually the vast majority of investigators depend on α = 0.05, this means that a research finding is more likely true than false if $(1 - \beta)R > 0.05$.
- If we have a reasonably powerful test, we can still have a relatively low ratio of True Relationships to No Relationships, and still have their product above $\alpha = 0.05$, such that the research finding is more likely to be true than false.

(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.

- The decision to make scientific claims based on p-values directly influences type I errors, the α parameter, as well as type II errors, the β parameter. Both $\alpha and \beta$ are also factors within the calculation for R, and thus also influence the ratio of true relationships to no relationship.

- If scientists probe random relations and just look at p-value, they may be successful in find significant p-values, but the PPV will decline. From corollary 3 in the paper: "The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true."

- Thank you graders! I know this was a lot of work. I appreciate any and all leniency for any questions that weren't clearly answered. Good luck with the rest of the course.