Written Report of Module 1 - Statistics Review

Tien Minh Dam {damminhtienchl@gmail.com}

June 2024

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). From this table, you interpret that the experiment was run the following way: (1) Vaccines were offered to Grade 2 students, but some Grade 2 students did not consent and opted out of the offered vaccine. (2) 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. The "No consent" group here means they are people who refused to participate in the whole experiment. In this problem, you will compare these two studies.

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. (100 words. Maximum 200)*

   Answer. To determine the effectiveness of the polio vaccine through a randomized controlled double-blind experiment, a large, diverse sample of individuals should be randomly selected and assigned into two groups: the treatment group, which receives the polio vaccine, and the control group, which receives a placebo such as a salt injection. Double-blinding should be implemented so that neither the participants nor the experimenters know which treatment is being administered to prevent bias. Both groups should receive their respective treatments under identical conditions. Participants should be monitored over a defined period, with the incidence of polio recorded for both groups. Statistical analysis would then compare the polio rates between the two groups to assess the vaccine's effectiveness. Ethical considerations include obtaining informed consent from all participants and ensuring their right to withdraw at any time. We hope this method ensures the reliability and validity of the results by minimizing bias and providing a clear measure of the vaccine's efficacy.

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

Randomized Controlled Double-Blind Experiment		
	Size	**Polio rate per 100,000**
Treatment (vaccine)	200000	28
Control (Salt Injection)	200000	71
No consent	350000	46

□

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. (100 words. Maximum 200)*

   Answer. In the NFIP study, the effectiveness of the vaccine is shown by the polio rates per 100,000: Grade 2 students who received the vaccine had a rate of 25, compared to 54 in Grade 1 and 3 students who did not receive the vaccine, and 44 in Grade 2 students who did not consent to the vaccine. This suggests the vaccine reduced the polio rate among those who received it.

   In the randomized controlled double-blind experiment, the polio rate was 28 per 100,000 in the treatment group (vaccine) and 71 per 100,000 in the control group (salt injection). This also indicates a significant reduction in the polio rate among those who received the vaccine, suggesting it is effective.

   Both studies show that the vaccine considerably lowers the incidence of polio, demonstrating its effectiveness. □

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

   (a) *(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. (100 words. Maximum 200)

   Answer. Yes, differences between Grade 1, Grade 2, and Grade 3 students can influence the results of the NFIP experiment. For example, if younger children (Grade 1) have weaker immune systems compared to older children (Grade 3), they might be more susceptible to polio, leading to higher infection rates regardless of the vaccine. This could falsely suggest that the vaccine is less effective.

   To prevent these differences from influencing the results, a more reliable experimental design would be to use a randomized controlled trial within a single grade. Randomly assign students within Grade 2 to either the treatment group (vaccine) or the control group (no vaccine). This approach ensures that any age-related susceptibility to polio is evenly distributed between the groups, making the comparison between vaccinated and unvaccinated students within the same grade more reliable. This design controls for age-related variables and provides a more accurate estimate of the vaccine's effectiveness. □

(b) (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. (100 words. Maximum 200)

Answer. Yes, the lack of blinding in the NFIP study could bias the results. Knowing who received the vaccine and who did not could influence both the behavior and the reporting of polio cases among participants. For example:

- Children who received the vaccine might engage in riskier behavior, assuming they are protected, potentially increasing their exposure to the virus. Conversely, unvaccinated children might be more cautious, reducing their exposure.
- Parents and healthcare providers might be more vigilant in monitoring and reporting symptoms of polio in unvaccinated children, leading to higher reported rates in the unvaccinated group.

To prevent such biases, the experiment should be conducted as a double-blind study:

1. Double-blind setup, neither the participants nor the experimenters know who is receiving the vaccine or the placebo. This eliminates the possibility of behavior changes based on knowledge of vaccination status.
2. Use a placebo (such as a salt injection) that appears identical to the vaccine. This ensures that all participants and administrators treat the conditions similarly.
3. Implement uniform procedures for monitoring and reporting polio cases across all groups to ensure that any differences in polio rates are due to the vaccine's effectiveness rather than differential reporting or behavior.

□

(c) (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. (50 words. Maximum 200)

Answer. Our examples:

- Children and parents who consent to vaccination might be more health-conscious, practicing better hygiene and taking more precautions, which reduces their risk of infection.
- Participants might experience a psychological benefit from believing they are protected, which can sometimes manifest in physical health improvements.

To eliminate biases not due to the vaccine itself, the experiment should include:

1. Randomization: Randomly assign participants to either the treatment (vaccine) group or the control (placebo) group.
2. Double-Blind Procedure: Neither the participants nor the experimenters know who receives the vaccine or the placebo.
3. Placebo Control: Use a placebo that mimics the vaccine but has no therapeutic effect.

□

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? (50 words. Maximum 200)*

   Answer. The lower rate of polio in the no-consent groups compared to the control groups could be due to self-selection bias, where individuals who opted out might be more health-conscious and practice better hygiene, reducing their exposure to polio. Additionally, socioeconomic factors could play a role, as those who refuse consent might come from backgrounds with better access to healthcare, inherently lowering their risk. Behavioral differences, such as more cautious health practices, could also contribute. To address this, ensuring randomization within the study design can help evenly distribute such factors across groups, isolating the vaccine's effect. □

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? (100 words. Maximum 200)*

   Answer. The parents' conclusion that participation in the trial exposed their children to a higher risk of polio, based on a slightly higher rate of 49 per 100,000 compared to 46 per 100,000 in non-participants, is not necessarily correct. The difference is minimal and likely due to random variation rather than an actual increase in risk. Additionally, the trial aims to provide a long-term benefit by demonstrating the vaccine's effectiveness, which can significantly reduce polio incidence over time.

   If a large group of parents refuses participation in the next year's trial, it could introduce bias, making the study population unrepresentative of the general population and skewing results. This would reduce the statistical power of the trial, making it harder to detect significant differences between the vaccinated and control groups, potentially delaying public health advancements. Effective communication about the trial's purpose and addressing parental concerns are essential to maintain participation and ensure the study's validity and public health benefits. □

Problem 1.3 Regarding the statement by the ASA about p-values

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) (Note that the statement itself appears after the editorial.)

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? (50 words. Maximum 200)*

Answer. The colleague's approach of including as many variables as possible in a regression analysis and selecting those that are statistically significant to inform policy is not likely to produce good policies. The ASA statement on p-values highlights several key issues with this approach: Firstly, including numerous variables increases the likelihood of finding some statistically significant results by chance alone, leading to false positives. Secondly, a p-value does not measure the probability that the studied hypothesis is true or the probability that the data were produced by random chance alone. Over-reliance on p-values can lead to erroneous conclusions. Moreover, scientific conclusions should not be based solely on whether a p-value crosses a threshold like 0.05. Other factors such as study design, data quality, and external evidence must be considered to ensure robust and reliable conclusions.

A better approach involves focusing on the context and quality of the data, using methods that emphasize estimation over testing, and considering multiple sources of evidence to guide policy decisions. This holistic approach can help prevent the pitfalls of relying solely on p-values and produce more effective educational policies. □

(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? Hint: You might need to design some experiment. (250 words. Maximum 350)

Answer. Gathering more data can indeed improve the accuracy and reliability of statistical inferences, reducing the likelihood of finding relationships that occur purely by chance. However, simply increasing the data size does not automatically ensure that the procedure proposed by the friend will find the true effects. Several factors must be considered to ensure the robustness of the findings.

There are some potential pitfalls:

- More data can enhance statistical power, but only if the data is of high quality. Poor-quality data can introduce noise and bias, leading to incorrect conclusions.
- Adding more variables can lead to overfitting, where the model fits the noise in the data rather than the true underlying patterns. This can result in poor generalization to new data.
- As mentioned earlier, including numerous variables increases the risk of type I errors (false positives). With a larger dataset, the chance of detecting spurious correlations can increase if not properly controlled.

We suggest to do some experiments to test procedure:

1. Simulated data generation
   (a) Create a simulated dataset with known variables and predefined relationships (true effects).
   (b) Include a mix of significant and non-significant variables.
2. Incremental data addition
   (a) Start with a small sample size and apply your friend's procedure to identify significant variables.
   (b) Gradually increase the sample size, applying the same procedure at each step, and track the identified significant variables.

3. Evaluation metrics - use several metrics to see the differents
4. Statistical controls
   (a) Apply corrections for multiple comparisons, such as the Bonferroni correction or False Discovery Rate (FDR) control, to adjust p-values and reduce type I errors.
   (b) Use cross-validation to assess model performance and prevent overfitting.

After doing these experiments, we can conclude about the results: With increasing data, the friend should observe a higher true positive rate and lower false positive and false negative rates if the procedure is effective. □

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? (100 words. Maximum 200)

Answer. The neuroscience lab should not conclude that there exists a causal relationship between chocolate consumption and intelligence based solely on the statistically significant correlation between chocolate consumption and the number of Nobel Prize laureates. Correlation does not imply causation, and this observed relationship could be influenced by various confounding factors, such as socioeconomic status, education systems, and healthcare quality, which may also contribute to both higher chocolate consumption and a greater number of Nobel laureates.

Additionally, the significant correlation might be a spurious result, especially given the complexities and multifaceted nature of intelligence and brain development. The lab should consider further investigation using controlled experiments or longitudinal studies to explore potential causal mechanisms and rule out confounding variables. Without robust evidence of causation, drawing conclusions from correlation alone can lead to misleading and potentially erroneous scientific findings. □

(3) (1 point) In order to study the relation between chocolate consumption and intelligence, what can they do? (100 words. Maximum 200)

Answer. To study the relationship between chocolate consumption and intelligence, the neuroscience lab can undertake the following steps:

1. Controlled experiments - conduct randomized controlled trials (RCTs) where participants are randomly assigned to consume chocolate or a placebo over a set period. Measure changes in intelligence and brain growth using standardized tests and neuroimaging techniques.
2. Longitudinal studies - track chocolate consumption and intelligence metrics over time in a large cohort to observe long-term effects and potential causal relationships.
3. Control for confounding variables - ensure that factors such as diet, socioeconomic status, education, and health are controlled or matched between groups to isolate the effect of chocolate consumption.
4. Use animal models - conduct experiments on animals to understand the biological mechanisms through which chocolate might influence brain development and intelligence.

5. Meta-analysis - combine data from multiple studies to increase statistical power and validate findings across different populations and methodologies.

□

(4) (3 points) The lab runs a randomized experiment on 100 mice, add chocolate in half of the mice's diet and add in another food of the equivalent calories in another half's diet. They find that the difference between the two groups time in solving a maze puzzle has p-value lower than 0.05. Should they conclude that chocolate consumption leads to improved cognitive power in mice? (100 words. Maximum 200)

Answer. While the lab's finding that the difference in maze-solving times between the two groups of mice has a p-value lower than 0.05 suggests a statistically significant effect, they should be cautious in concluding that chocolate consumption directly leads to improved cognitive power. Here are several considerations:

1. Replication - the experiment should be replicated to ensure the results are consistent and not due to random chance.
2. Effect size - assess the magnitude of the difference to determine if it is practically significant, not just statistically significant.
3. Confounding factors - ensure that no other variables, such as differences in overall health, activity levels, or other environmental factors, contributed to the observed effect.
4. Biological mechanism - investigate the underlying biological mechanisms to understand how chocolate might influence cognitive power.

□

(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? (50 words. Maximum 150)

Answer. The lab's approach of reporting only the statistically significant results while ignoring the non-significant relationship between chocolate consumption and IQ is not entirely correct. This practice, known as "cherry-picking" or "p-hacking," can lead to biased and misleading conclusions. Proper scientific reporting should include all relevant findings, regardless of their p-values, and provide a transparent account of the analysis conducted. This includes discussing why certain variables were not significant and ensuring that the reported significant results are not due to chance or multiple testing errors. Providing a comprehensive and honest account of the study's findings ensures the integrity and reliability of the research. □

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"? (50 words. Maximum 150)

Answer. The title "New trial shows strong effect of drug X on curing disease Y" is misleading because a p-value smaller than 0.05 only indicates statistical significance, not the magnitude of the effect. A more accurate title would be "New trial shows statistically significant effect of drug X on curing disease Y." The title "New drug proves over 95% success rate of drug X on curing disease Y" is inappropriate and misleading, as a p-value does not translate to a success rate and scientific studies provide evidence, not proof. A more suitable title would reflect the statistical significance without overstating the findings, such as "Study finds drug X has a statistically significant impact on curing disease Y." □

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? (50 words. Maximum 150)*

Answer. His reasoning is not entirely correct. A non-significant result in the regression analysis means that there is not enough evidence to conclusively determine the effect of HR spending on revenue, but it does not necessarily mean there is no effect. The large effect size suggests potential importance, but the lack of statistical significance might be due to insufficient data, high variability, or other confounding factors. It would be prudent to consider additional analyses, such as examining other variables, increasing the sample size, or exploring different models, before making a final decision on HR spending. Relying solely on statistical significance can lead to incomplete or misguided conclusions. □

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? (50 words. Maximum 150)*

Answer. False. While replication of the same experiment and obtaining significant results increases confidence in the findings, it is a crucial step toward making a scientific claim. Replication helps to confirm the reliability and validity of the results, reducing the likelihood that they are due to random chance or specific experimental conditions. However, making a robust scientific claim also requires considering other factors, such as the consistency of results across different studies, theoretical plausibility, and potential confounding variables. Therefore, replication is necessary but not solely sufficient for making a definitive scientific claim. □

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? (100 words. Maximum 200)*

Answer. It is not okay for the lab mate to report only the statistically significant results. Selectively reporting only significant findings, known as "p-hacking" or "selective reporting," leads to biased and misleading conclusions. This practice inflates the likelihood of false positives and undermines the integrity of scientific research. To ensure transparency and reliability, all tests and hypotheses conducted should be reported, regardless of their significance. This allows for a comprehensive evaluation of the research and helps others understand the full context and

potential limitations of the study. Full disclosure of all results is essential for maintaining the credibility of scientific findings. □

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? (100 words. Maximum 200)

Answer. True. A significant p-value indicates that the data are unlikely to occur under the null hypothesis, but it does not confirm that the null hypothesis is false. It is possible that the null hypothesis is true, but the statistical model used does not accurately reflect the real-world scenario. This could be due to model misspecification, incorrect assumptions, or other limitations in the model. Therefore, a significant p-value should be interpreted with caution, and other evidence, such as the validity of the model and the assumptions behind it, should be considered when drawing conclusions. □

Problem 1.5: Regarding the paper on why most published research findings are false

Read the paper by Ioannidis on why most published research findings are false (PLoS Medicine, 2005). By answering the questions below you will summarize the paper in your own words. We wish to show that that bias in a scientific field may lead to even smaller probabilities of the research findings being true. We define bias as claiming a result is true regardless of the statistical results. Let u be the proportion of probed relations that are presented as existing because of bias (e.g., systematic bias in how research is conducted in the field may lead to a certain fraction of the conducted studies to find relations regardless of the ground truth).

Therefore, we can express PPV in terms of α, β, R, u as follows:

$$\text{PPV} = \frac{\mathbf{P}(\text{relation exists, and claim relation})}{\mathbf{P}(\text{claim relation})}$$
$$= \frac{uR + (1-u)R(1-\beta)}{u(1+R) + (1-u)R(1-\beta) + (1-u)\alpha}$$

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

Start by writing the PPV as

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

(Note that this does not include a bias term and you will not need one to answer this question.) (50 words. Maximum 100)

Answer. To show that repeated independent testing by different teams can reduce the probability of the research being true, we start with the PPV definition. Let's denote:

- α as the Type I error rate (false positive rate).
- β as the Type II error rate (false negative rate).
- R as the ratio of true relationships to no relationships.

- n as the number of independent repetitions.

We will calculating the prob. as follow:

1. Prob. of finding significance at least once when relation exists:

   $\mathbf{P}$(relation exists, at least one significant) = $1-\mathbf{P}$(relation exists, none significant) = $1-\beta^n$

2. Prob. of finding significance at least once when no relation exists:

$$\mathbf{P}(\text{no relation, at least one significant}) = 1-(1-\alpha)^n$$

3. Total prob. of at least one significant result:

$$\mathbf{P}(\text{at least one significant}) = \mathbf{P}(\text{relation exists}) \cdot (1-\beta^n) + \mathbf{P}(\text{no relation}) \cdot (1-(1-\alpha)^n)$$
$$= \frac{R}{1+R} \cdot (1-\beta^n) + \frac{1}{1+R} \cdot (1-(1-\alpha)^n)$$

PPV Calculation

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

Simplifying:

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

As n increases, $(1-\beta^n)$ approaches 1 and $(1-(1-\alpha)^n)$ also approaches 1, making the denominator larger. This reduces the PPV, showing that repeated independent testing reduces the probability that research findings are true. □

9. *(2 points) What would make bias or increasing teams testing the same hypothesis not decrease PPV? (Assuming $\alpha = 0.05$.) (Hint: Please treat the two issues separately.) (50 words. Maximum 100)*

Answer. Bias would not decrease the PPV if the study has perfect statistical power ($\beta = 0$), as true relationships would always be detected. Mathematically, this can be shown as:

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

If $\beta = 0$:

$$\text{PPV} = \frac{uR + (1-u)R}{u(1+R) + (1-u)R + (1-u)\alpha} = \frac{R}{1+R}$$

Thus, PPV remains unaffected by bias when $\beta = 0$.

Increasing teams testing the same hypothesis: increasing the number of teams testing the same hypothesis would not decrease PPV if the probability of Type I errors (α) is reduced proportionally to the number of tests performed. Mathematically, if α is very small, even with multiple tests, the false positive rate remains low.

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

If α approaches 0, the term $(1-(1-\alpha)^n)$ also approaches 0, thus not significantly increasing the denominator, keeping PPV stable. □

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.) (50 words. Maximum 100)*

Answer. To model the situation where multiple teams work on the same hypothesis and the scientific community requires unanimous replication for acceptance, we need to modify our PPV calculation. The requirement is that all teams must find significance for the result to be considered true. We calculate prob. as follow:

1. Prob. of finding significance by one team when relation exists:

$$\mathbf{P}(\text{relation exists, one team finds significant}) = 1 - \beta$$

2. Prob. of finding significance by one team when no relation exists:

$$\mathbf{P}(\text{no relation, one team finds significant}) = \alpha$$

3. Prob. that all n teams find significance when relation exists:

$$\mathbf{P}(\text{relation exists, all } n \text{ teams find significant}) = (1 - \beta)^n$$

4. Prob. that all n teams find significance when no relation exists:

$$\mathbf{P}(\text{no relation, all } n \text{ teams find significant}) = \alpha^n$$

5. Total prob. that all n teams find significance:

$$\mathbf{P}(\text{all } n \text{ teams find significant}) = \mathbf{P}(\text{relation exists}) \cdot (1 - \beta)^n + \mathbf{P}(\text{no relation}) \cdot \alpha^n$$

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

PPV Calculation

$$\text{PPV} = \frac{\mathbf{P}(\text{relation exists, all } n \text{ teams find significant})}{\mathbf{P}(\text{all } n \text{ teams find significant})}$$

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

Simplifying:

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

□

11. *(3 points) Suppose there is no bias and no teams are racing for the same test, so there is no misconduct and poor practices. Will publications still be more likely to be false than true? (100 words. Maximum 200)*

Answer. Even with no bias and no teams racing for the same test, publications can still be more likely to be false than true due to inherent statistical limitations and the nature of scientific research. We consider key factors:

1. Low Pre-study Prob. (R): If the prior prob. R of a hypothesis being true is low, the Positive Predictive Value (PPV) will also be low. For many exploratory studies, R is often small.
2. Type I and Type II Errors (α and β): Even in well-designed studies, the presence of Type I error (α) and Type II error (β) affects the reliability of findings.

PPV Calculation:

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

Even without bias and misconduct, the statistical structure of research, combined with low pre-study probabilities and the possibility of error, means that many published findings can still be 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. (100 words. Maximum 200)*

Answer. In Problem 1.3, the colleague's approach of including as many variables as possible and making scientific claims based on p-values influences the Type I error rate, α. The p-value threshold (commonly set at 0.05) determines the probability of falsely rejecting the null hypothesis, i.e., claiming a relationship exists when it does not.

If scientists probe random relations and use p-values as a certificate for making scientific conclusions, the effect on the PPV can be modeled as follows:

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

- Type I error rate (α) - by relying on p-values, the decision-making process is directly tied to α. If α is set to 0.05, this means there's a 5% chance of falsely claiming a relationship exists. Probing many random relations increases the likelihood of encountering false positives, thus inflating α.
- Exploratory analysis - in an exploratory analysis with many variables, the effective α increases due to multiple comparisons. This reduces the PPV because the denominator of the PPV equation increases, leading to more false positives.

When decisions are based solely on p-values, the increase in Type I error rate (α) reduces the PPV, leading to a higher proportion of false positive findings. This emphasizes the importance of considering the context, adjusting for multiple comparisons, and using additional statistical measures to ensure robust and reliable scientific conclusions. □