RESEARCH METHODS IN ARTIFICIAL INTELLIGENCE
Assignment 1

Authors: Ryan Dorland, Laith Agabria, Johannes Leppäkorpi, Aleksandra Bobrova
Student numbers: s3219992, s4036328, s3856348, s3660141
Date: April 18, 2025

1 INTRODUCTION

In the digital age, scientific information is more accessible than ever before, yet so is misinformation. News articles that report on scientific studies are the main source that shapes public understanding, especially among non-experts. However, research has shown that even educated readers often struggle to differentiate between accurate and flawed science reporting [1]. This is particularly concerning given the amount of fake news and the increasing pressure on individuals to make informed decisions about science-related topics such as health, technology, and the environment.

Academic programs in scientific and mathematical fields often include statistics courses aimed at teaching data analysis skills and developing statistical literacy. A key assumption is that completing such coursework equips students to become more discerning consumers of scientific news. However, despite this assumption, little empirical research has been conducted to test whether performance in academic statistics courses actually correlates with the ability to identify flawed science reporting in practice.

To address this gap, the present study investigates whether students who have performed better in their statistics courses are better at detecting mistakes in scientific news articles. Participants, graduates with degrees in computer science, read 14 news articles, half of which contained intentional errors in how scientific findings were reported. Participants were asked to judge whether each article was flawed or accurate.

One might expect a positive relationship between statistical proficiency and performance on this task. However, it is unclear whether formal education translates into real-world critical evaluation skills, especially in domains like journalism where statistics may be embedded in complex narratives. To test whether the students with higher average grades in their statistics courses (GradeStats) are better at identifying flawed articles (Performance), we use a univariate linear regression model and test the significance of the slope to assess the association.

The null hypothesis is that there is no association between the average grades in the statistics courses of science students and their performance in identifying flawed science reporting, so the slope of the linear regression function is zero. The alternative hypothesis is that there is an association, meaning that the slope is different from zero. In other words, performance in statistical courses influences the prediction of critical evaluation ability in the context of science journalism.

By investigating this relationship, the study seeks to contribute to the ongoing conversation about how to better prepare students for interpreting scientific claims in everyday media. If a strong association is found, it could support efforts to enhance the depth, rather than just the breadth, of statistical education.

2 DATA SIMULATION AND STATISTICAL TESTING

We simulated a univariate linear regression model with no association between *Performance* and *GradeStats*. Thus, the null hypothesis that the regression slope is zero was generated to be true. The alternative hypothesis states that the slope is not equal to zero, which might happen occasionally due to the randomness in sampling.

Performance ranges from 0 to 14 (integer values) and *GradeStats* ranges from 5.5 to 9.5. The response variable is truncated and discrete, whereas linear regression assumes that its conditional standard deviation is normal. We could not fully account for this fact and simply generated random noise from a truncated normal distribution $\mu = 0$, $\sigma = 2$, bounded between -7 and 7. *GradeStats* was drawn from a truncated normal distribution ($\mu = 6.5$, $\sigma = 2$) to reflect realistic grade distributions.

We set the expected performance to 7, which results in the following model:

$$Performance = round(7 + 0 \times GradeStats + \epsilon) \tag{1}$$

Here, ϵ is the random noise and the slope is set to zero to ensure that no correlation exists. The built-in *round* function was used to make the values of the variable discrete.

To estimate the Type I error rate, we repeated this simulation 1 million times with a sample size of 237 (big enough to account for the noise assumption limitations). For reproducibility, the seed of 42 was used. The histogram of resulting p-values (Figure 1) shows an approximately uniform distribution, which aligns with the expected results of observing no pattern in the data.

An example scatter plot of one simulation is shown in Figure 2.

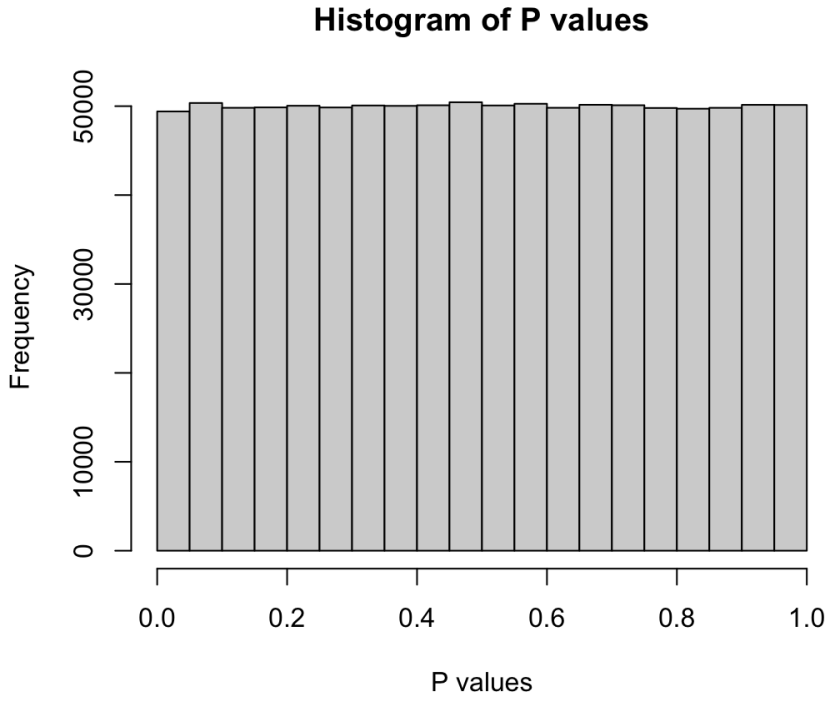

Figure 1: The histogram of P values with a sample size of 237 over 1 million repetitions. The distribution resembles the uniform distribution, where every P value is equally likely. This happens only when the null hypothesis is true, which aligns with the expected results.

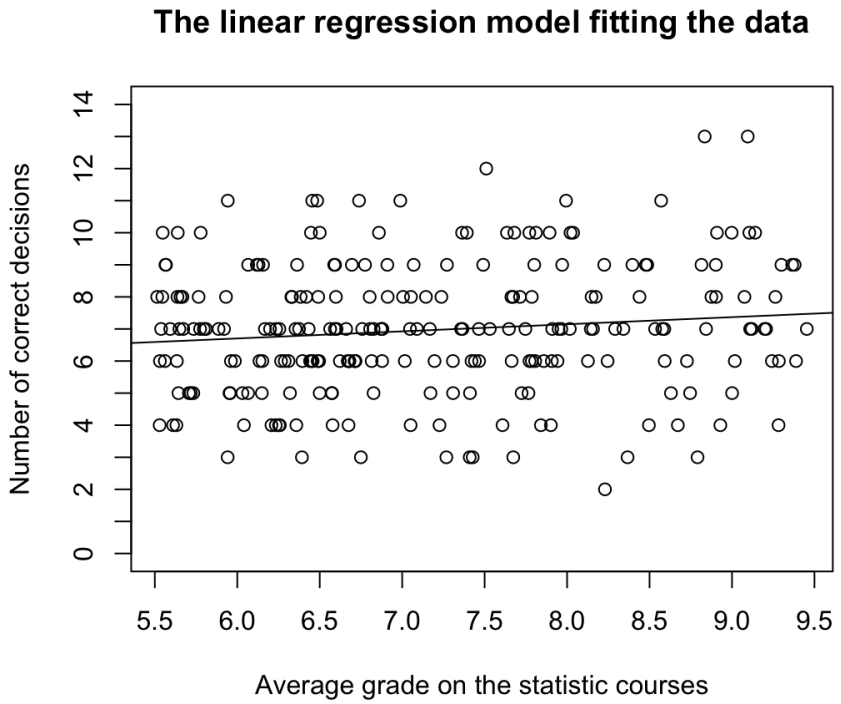

Figure 2: The simulated discrete data of one experiment which fits the linear regression model. The slope is close to zero but positive, due to the random noise in the generation.

3 QUESTIONABLE RESEARCH PRACTICES

3.1 ROUNDING P-VALUES DOWN

In scientific research, the interpretation of the findings is highly influenced by the way these results were reported. Even small decisions, such as how p-values are presented, can have a significant impact on the conclusion. An example of such questionable research practice is rounding p-values down, a form of p-hacking. Using a significance level $\alpha = 0.05$, a researcher would round down some of the found p-values that are essentially close to this threshold from above. This practice is not inherently questionable; however, it is a questionable research practice if it is purposefully done after conducting tests to find significant results.

We intentionally simulate such researchers' behavior by rounding the observed P-values down if they are close enough to the threshold from above. This is done in the same loop of repeated simulations from Part B, where we fit a linear regression model to each randomly generated dataset and calculate the corresponding P-value of the observed test statistics. However, this time, every P-value is passed to a custom `modified_round` function, which converts values in the open interval from 0.05 to 0.06 to the significant ones by assigning them a value of 0.05. As it can be seen from Figure 3, this practice changes the true distribution of P-values, which was essentially uniform. In comparison with Figure 1, the more significant results became more likely in Figure 3. Due to the artificial shift of P-values from the adjacent bin just above 0.05 into the bin with $p < 0.05$, simple rounding resulted in a higher Type I error rate, compared to the original distribution. Subsequently, this demonstrates how certain reporting choices can influence statistical results, highlighting the importance of transparency in scientific writing.

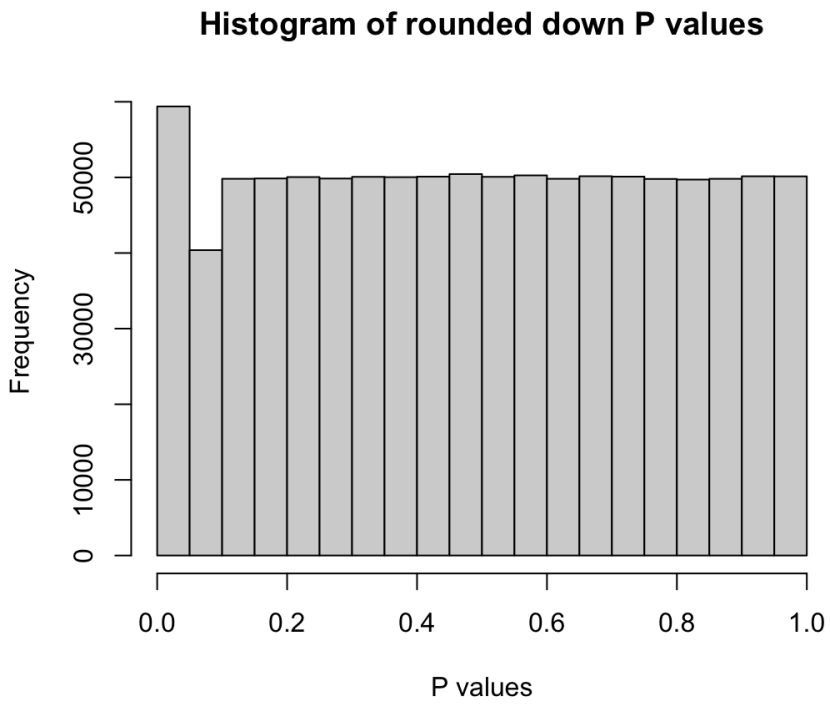

Figure 3: The histogram of rounded down P values with a sample size of 237 over 1 million repetitions. The distribution shows a sharp spike just below 0.05 value, indicating a misleading impression that the results are significant. According to the data, it is plausible to reject the null hypothesis.

3.2 SEQUENTIAL TESTING WITH OPTIONAL STOPPING

In statistical hypothesis testing, the null hypothesis is typically assumed to be true unless the evidence from data suggests otherwise. However, the p-value, which quantifies the strength of evidence against the null hypothesis, can fluctuate due to random chance, even when the null hypothesis is true. Sequential testing refers to the process of analyzing data as they are collected, without a fixed sample size predetermined at the start. In such cases, testing continues until the results meet a predefined significance threshold, such as $\alpha = 0.5$. This adaptive testing can be efficient, but it introduces a major issue: as the number of tests increases by expanding the sample size, the likelihood of observing statistically significant results increases as well, even if the null hypothesis is true. This is known as the optional stopping problem.

This behavior can be seen in Figure 4, where the p-value fluctuates with an increase in sample size. Ultimately, if we make our sample size large enough, we will observe the desired results and report them. From the figure, it is evident that statistically significant results may occur even with very small sample sizes, such as $N \approx 40$. However, since our initial size was 237, the effect of smaller samples is not the focus of our analysis. Nevertheless, these results are highly unreliable and also highlight how certain research practices can lead to misleading results.

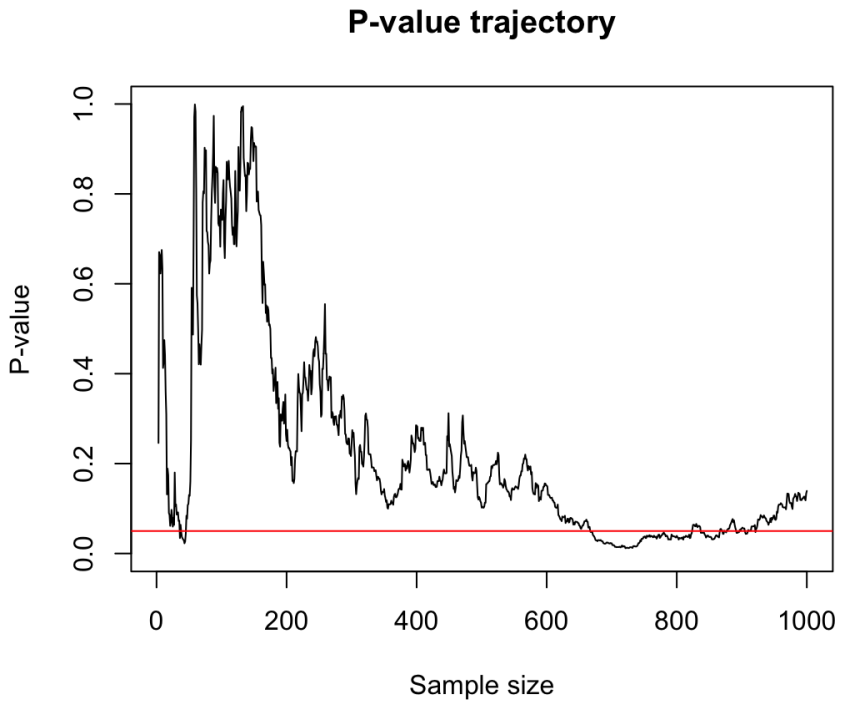

Figure 4: Sequential testing curve, which shows p-values as a function of sample size under optional stopping. The random seed was set to 201. The red curve indicates the .05 level. Although most of the tests do not present significant results, the test around $N = 700$ crosses the $p < .05$ threshold. Therefore, selectively stopping data collection after observing the desired p-values can introduce the Type I error, leading to false-positive results.

The results of Figure 4 were generated by running a single experiment with a huge sample size of `max_sample_size = 1000`. Then, by iterating over sample sizes from 2 to 1000, we progressively took growing slices in each iteration and computed the corresponding p-values. This simulates the process of incrementally expanding the sample size over time, which eventually produces a statistically significant result. The sample of size 700 with the p-value below the threshold was used to illustrate the linear regression curve with its corresponding correlation, as shown in Figure 5. It can be seen that the line has a positive slope, which might incorrectly suggest a relationship between the variables. Even though the effect emerged due to pure random variation, one might intentionally choose these results to report in their paper. Consequently, the usage of optimal stopping increases the risk of false positives in research, leading to misleading conclusions in science.

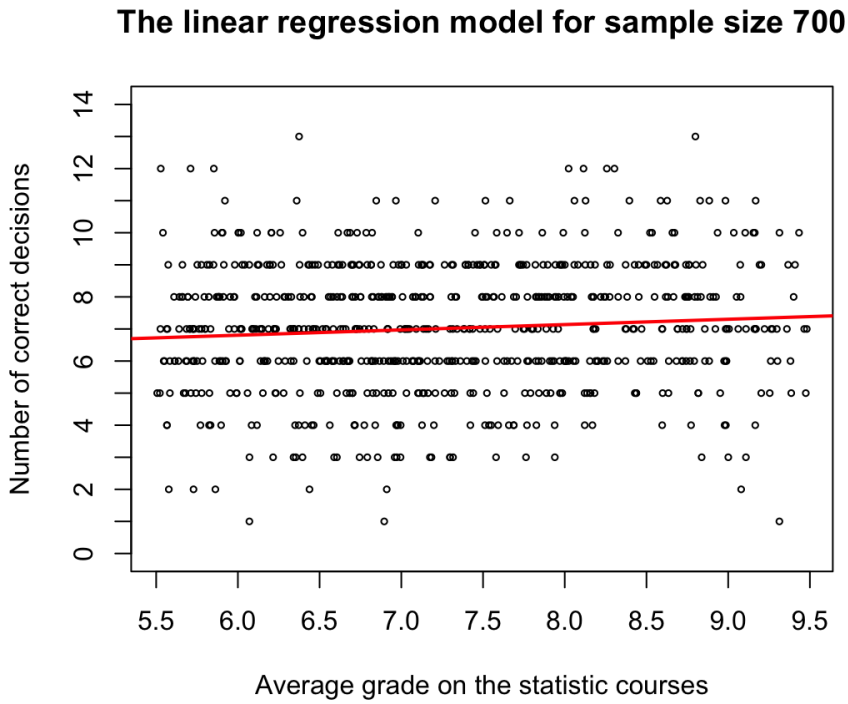

Figure 5: Linear regression curve based on the sample size 700 that produced a statistically significant result ($p < 0.05$). The graph illustrates the positive correlation observed when selectively reporting results during sequential testing.

3.3 CONCLUSION

Both rounding p-values and sequential testing with optional stopping are common yet potentially problematic practices in scientific research. While these methods may seem efficient and without malicious intent, they introduce significant risks of false positives and misleading conclusions, particularly when misused.

Rounding p-values down, as demonstrated in Figure 3, can artificially inflate the appearance of statistical significance, leading to the selective reporting of results that may not truly support rejecting the null hypothesis.

Sequential testing, as illustrated in Figures 4 and 5, can also lead to results that appear statistically significant simply due to the increasing sample size over time. This approach increases the likelihood of finding a p-value below the threshold $\alpha = 0.5$, even if the null hypothesis remains true. When data collection is stopped prematurely based on the p-value threshold, the risk of Type I errors is heightened, which can distort the scientific record.

Both questionable research practices may lead to the publication of findings that are not truly reflective of the data. Researchers must be aware of this issue and adopt strict protocols to ensure that the scientific record remains reliable and accurate.

4 CONTRIBUTIONS

All group mates reviewed all of each other's work before submitting.

- **Ryan**: Wrote most of introduction (part A); worked on R-code for sequential testing (part C); wrote part of sequential testing section and conclusion (part C).
- **Johannes**: Implemented some of the p-value rounding mechanisms and wrote text for it (part C); partially built code for data simulation and statistical testing (part B).
- **Aleksandra**: Wrote the hypothesis in the introduction (part A); worked on R-code for data simulation together with its description (part B); worked on R-code for sequential testing (part C); wrote part of sequential testing section (part C).
- **Laith**: Worked on improving code quality and assisted with the report writing for the p-value rounding (part C).

REFERENCES

[1] Michael Michal, Yinan Zhong, and Priti Shah. "When and why do people act on flawed science? Effects of anecdotes and prior beliefs on judgments of scientific research". In: *Cognitive Research: Principles and Implications* 6.1 (2021), p. 28. DOI: 10.1186/s41235-021-00293-2. URL: https://doi.org/10.1186/s41235-021-00293-2.