Exercise Sheet: Foundations of Data Collection

Data, Data Storage, Data Collection

Exercise 1 - Student Performance and Mental Health During a Pandemic

In the spring of 2020, as the pandemic forced universities worldwide to shift to remote learning, the introductory course on Ethics at your university – which is required for all students regardless of their field of study – happened entirely online. This course, traditionally popular among students, became even more relevant as ethical dilemmas surrounding public health, privacy, and social responsibility took center stage in public discourse.

The course was designed to accommodate 1200 students from four faculties: Computer Science, Mathematics, Engineering, and Business. Lectures were delivered via video conferencing, and attendance was recorded.Due to the easing of lockdown restrictions but the continued presence of the virus, the final exam was conducted in a hybrid format: students could choose to take it either on-site (with strict sanitary protocols) or remotely.

Recognizing the heightened stress and anxiety caused by the pandemic, the university offered free mental health services to all full-time students. Part-time students, who come back to studies while already having a job, were not eligible for these services due to limited resources.

As a data analyst, you have been tasked with analyzing the impact of these unusual circumstances on student performance. You are provided with the following datasets:

1. Exam Performance:
   - Student ID (unique identifier)
   - Faculty (Science, Humanities, Engineering, Business)
   - Exam Score (0–20)
   - Attendance (percentage of classes attended)
   - Exam Mode (on-site or remote)
2. Mental Health:
   - Student ID (unique identifier)
   - Service Usage (”Yes” if the student used mental health services, ”No” otherwise)

However, the datasets are incomplete. The Exam Performance dataset excludes 100 students who took the exam in person due to a data entry error. The Mental Health dataset excludes 310 of students who opted out of data collection for privacy reasons, as well as the 150 part-time students who are not eligible.

(1) Confidence Intervals for Mental Health Service Usage

You want to estimate the proportion of students who used mental health services during the pandemic. In the Mental Health dataset, the is the record of 210 students having used the services.

(a) Provide a lower bound and a upper bound on the number of studens who might have used the mental health services. Derive a lower bound and an upper bound of the proportion of students who might have used the mental health services.

(b) Compute the sample proportion $\hat{p}$ of students who used mental health services.

(c) The standard error (SE) of the sample proportion $\hat{p}$ measures the variability or uncertainty in the estimate of the true proportion p. It is calculated as:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where n is the sample size. Based on the standard error, the 95% confidence interval is given by

$$\hat{p} \pm 1.96 \times SE$$

Compute the SE for $\hat{p}$. Derive from it the 95% confidence interval for the true proportion p of students who used mental health services.

(d) Interpret the confidence interval: What does it suggest about the prevalence of mental health service usage among students during the pandemic? What decision can we make out of this simple analysis?

(e) Discuss one potential limitation of this confidence interval estimate, considering the data quality issues identified earlier. (Note: this corresponds to content that should be documented when communicating results).

includes records for $1200 - 310 - 150 = 740$ students,

(a) The Mental Health dataset states that 210 students used the mental health services but 310 students opted out of the collection procedure. So the exact number is between 210 and $210{+}310 = 520$. The mental health services were available only to full-time students $1200 - 150 = 1050$, thus the true proportion is between $\frac{210}{1050} = 0.2$ and $\frac{520}{1050} \approx 0.495$.

(b) The sample size if $1050 - 310 = 740$, so the sample proportion is $\hat{p} = \frac{210}{740} \approx 0.284$

(c) We have $\hat{p} = \frac{210}{740}$ and $n = 740$, so $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{210 \times * 530}{740^3}} \approx 0.017$. Then, the 95% confidence interval is $0.284 \pm 1.96 \times 0.017 \approx [0.251, 0.317]$.

(d) We are 95% confident that the true proportion of students who used mental health services during the pandemic is between 25.1% and 31.7%. This suggests that roughly 1 in 4 to 1 in 3 students accessed these services, reflecting the important stress and anxiety during this period. Without additional information (e.g., the proportion before the pandemic), it remains difficult to use this information in a decision-making process.

(e) One limitation of this estimate is the exclusion of 460 students (310 who opted out and 150 part-time students). If these excluded students had systematically different usage patterns (e.g., part-time students might have faced unique stressors but were ineligible for services, or students who opted out might have been more or less likely to use services), the confidence interval could be biased. For example, if students who opted out were more likely to need services, the true proportion might be higher than estimated.

(2) Follow-up on your analysis

The university was highly satisfied with your findings and is now considering implementing changes based on your analysis. Specifically, they are exploring initiatives to expand mental health services and improve attendance tracking.

The university is also considering using attendance data to identify at-risk students for academic support. However, they are aware of potential data quality issues in remote attendance records.

(a) Beyond logging in, what additional metrics or data sources could the university collect to better measure student engagement? Provide at least two examples and explain their advantages.

1. The university could track interactions such as responses to polls, discussion board contributions, or quiz completion rates. These metrics provide a better indication of engagement than mere login records.
2. The university could use learning management system (LMS) data to measure the time students spend on reading materials, watching lectures, or completing assignments. This can help distinguish between passive and active engagement, providing a more nuanced understanding of how students are interacting with the course content.

(b) How could the university validate the accuracy of these new metrics?

One standard solution would be do to a pilot study. This would mean implementing the new metrics in a small number of courses, or some a small number of students and compare them with traditional attendance records. Another validation procedure could be to study the correlations between the new engagement metrics and final exam scores or course satisfaction surveys. This Essentially means redoing the analysis and see whether it provides more accurate results.

The university wants to communicate your findings to faculty and students to justify their decisions. However, they are concerned about misinterpretation or misuse of the data.

(c) What are the key limitations of your analysis that should be clearly communicated to avoid misinterpretation? Draft a short, non-technical summary (2-3 sentences) of your findings that the university could share with students and faculty. Emphasize transparency about limitations related to data quality.

Our analysis suggests that students with lower attendance tended to have lower exam scores, and that mental health service usage was higher among students with low attendance. However, these findings are based on incomplete data, particularly excluding some in-person exam takers and part-time students. We recommend expanding mental health support and improving how we measure engagement, while ensuring student privacy and informed consent are prioritized.

The university is considering integrating your findings into a long-term monitoring system for student well-being.

(e) What additional data should the university collect in the future to improve the robustness of their analyses? Consider well-being metrics and contextualization (demographic) indicators.

Well-being metrics could include regular surveys on stress levels and workload to track mental health trends over time. Demographic data could include students living situations, employment status, or caring responsibilities..

(f) How could the university ensure ethical data collection and usage in this long-term system?

The university needs to obtain informed consent, clearly informing the students about what data will be collected, how it will be used, and their right to opt out. Consent should be obtained separately based on the sensitivity of the data. All data have to be anonymized and stored securely to protect student privacy. Finally, the university should provide transparency, regularly communicating with students and faculty staff about what data is being collected, how it is being used, and what insights have been gained. This also means offering opportunities for feedback and questions.