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) Data Types and Quality

The university has tasked you with providing insights related relations between the students performances, their stress and their appreciation of the lecture. First, you examine the provided datasets.

(**a**) Identify whether the provided dataset are primary or secondary. Justify your answer.
(**b**) Can you identify an ethical issue with manipulating the provided datasets? Justify your answer.

The next goal is to evaluate whether the specific questions asked by the university can indeed be answered with the provided datasets and to identify potential challenges in the analysis. For each of these questions, address the following:

(**c**) Discuss whether quantitative or qualitative data would be required. Justify your answer.
(**d**) Indicate whether the question can or cannot be answered with the provided datasets. If it can be answered, specify the dataset(s) and the content of the dataset (the fields) that you need to answer the question. If it cannot be answered, explain why. In that case, skip the next question.
(**e**) Discuss one potential data quality issue that could affect the analysis for this question.

This is the list of received questions:

i. Is there a relationship between attendance and exam scores?
ii. Did students who attended more lectures experience less stress during the pandemic?
iii. Did students who used mental health services performed less at the exam than those who did not?
iv. Were part-time students more likely to have needed mental health services than full-time students?

(2) 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,

(3) Scope Misalignment and Adjustments

The Exam Performance Data excludes 100 in-person exam takers. Suppose the average exam score for the 1100 students in the dataset is $\mu = 12$ with a standard deviation of $s = 3$. You later discover that the average score for the 100 in-person exam takers is $\mu_{\text{in-person}} = 15$, with a standard deviation of $s_{\text{in-person}} = 2$.

(**a**) Compute the adjusted population average μ_{adjusted} that accounts for the missing in-person exam takers.

(**b**) How does this adjustment change your understanding of student performance? What does it suggest about the potential misalignment between the sample and the target population?

As we just saw, the computing the adjusted means from the initial one is quite straightforward, but what about the standard deviation? The goal of the next series of qeustion is to prove that the variance of the entire population of 1200 students, $\sigma^2_{\text{population}}$, satisfies the following inequality:

$$\sigma^2_{\text{population}} \geq \frac{1100 \times 9 + 100 \times 4}{1200}.$$

Essentially, the result is an application of the law of total variance, which describes how the variance of a population can be decomposed into the variance within subgroups and the variance between subgroup means. Let us prove it for a discrete, finite population divided into two subgroups. Consider a population of N divided into two subgroups:

- Subgroup 1: n_1 individuals with mean $\bar{x}_1$ and variance s_1^2.
- Subgroup 2: n_2 individuals with mean $\bar{x}_2$ and variance s_2^2.

The overall population mean is μ, and the overall population variance is σ^2.

(**c**) Recall the formula for the overall population variance σ^2 in terms of the individual data points x_i.

(**d**) Split the sum into two parts: one for Subgroup 1 and one for Subgroup 2.

(**e**) For each subgroup, add and subtract the subgroup mean ($\bar{x}_1$ or $\bar{x}_2$) inside the squared term. Rewrite the expression for σ^2 to include terms involving $(\bar{x}_1 - \mu)$ and $(\bar{x}_2 - \mu)$.

(**f**) Expand the squared terms in the expression to separate the contributions from:

- The variance within each subgroup (i.e., $(x_i - \bar{x}_1)^2$ and $(x_i - \bar{x}_2)^2$).
- The squared difference between each subgroup mean and the overall mean (i.e., $(\bar{x}_1 - \mu)^2$ and $(\bar{x}_2 - \mu)^2$).
- The cross terms involving $(x_{1i} - \bar{x}_1)(\bar{x}_1 - \mu)$ and $(x_{2i} - \bar{x}_2)(\bar{x}_2 - \mu)$.

(**g**) Show that the sum of the cross terms over all data points is zero.

(**h**) Deduce that the overall variance σ^2 can be written as:

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - \mu)^2 + n_2(\bar{x}_2 - \mu)^2}{n_1 + n_2}$$

This is the law of total variance for two subgroups. Interpret each term in this expression.

(**i**) Conclude that

$$\sigma^2 \geq \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

Is the inequality tight? When does it become an equality? What do you learn from that?

(**j**) Apply the inequality to the exam score data $\sigma^2_{\text{population}}$.

(4) 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.

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

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.

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.

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