Exercise Sheet: Exam Preparation

Data, Data Storage, Data Collection

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 are asked to study the impact of these exceptional circumstances on student performance.

Part 1 - Suggestions From The University

The university plans to conduct a survey to better understand how the pandemic and the shift to remote learning have affected students. Your role is not to judge the opinions expressed in the questions, but to assess the quality of the data that would be collected and to suggest improvements where appropriate.

(**Q1**) The current survey draft that you receive from the university includes these questions:

- "Have you felt stressed during the pandemic? Answer on a scale from 1 to 5."
- "How many hours do you study per week? (open-ended)"

For each question: identify two distinct issues related to data quality, reliability, bias, or interpretability; then, propose an improved version of the question that addresses these issues.

(**Q2**) The university considers the following sampling approaches:

- The first 100 students to answer.
- All the third-years.
- A stratified sample by faculty (Science, Humanities, Engineering, Business), selecting every tenth student from an alphabetical list within each faculty.

Evaluate each approach in terms of representativeness of the student population (bias) and practicality of the implementation.

Part 2 - Data About The Ethics Lecture

Before launching a large-scale survey among the students, the university's management asks for preliminary insights based on already available data. The goal is to quickly assess whether

there are visible patterns or warning signals related to student performance, attendance, and well-being during this exceptional academic year.

As a result, you are provided with two existing datasets and the university emphasizes that these results will be used only as initial indicators.

1. Exam Performance:
   - Student ID (unique identifier)
   - Faculty (Science, Humanities, Engineering, Business)
   - Exam Score (score at the exam for the Ethics class: 0–20)
   - Attendance (percentage of classes attended)
   - Exam Mode (on-site or remote)

   This dataset comes from the university's academic administration system, used for grading and reporting purposes.
2. Mental Health:
   - Student ID (unique identifier)
   - Service Usage ("Yes" if the student used mental health services, "No" if they did not, and "Unknown" if they did not give explicit consent to appear in the dataset)

   This dataset is extracted from the university's mental health services records. Students were informed that their data could be used for research purposes, with an opt-out option available. It was collected for internal follow-up and resource planning.

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 the 150 part-time students who are not eligible.

First, you examine the provided datasets.

(**Q1**) Identify whether the provided dataset are primary or secondary. Justify your answer.
(**Q2**) 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. 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?

For each of these questions, address the following:

(**Q3**) Discuss whether quantitative or qualitative data would be required. Justify your answer.
(**Q4**) 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.
(**Q5**) Discuss one potential data quality issue that could affect the analysis for this question.

Part 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 $s_{\text{in-person}} = 2$.

(**Q1**) Compute the adjusted population average μ_{adjusted} that accounts for the missing in-person exam takers.

(**Q2**) 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.

(**Q3**) Recall the formula for the overall population variance σ^2 in terms of the individual data points x_i.

(**Q4**) Split the sum into two parts: one for Subgroup 1 and one for Subgroup 2.

(**Q5**) 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 $(\bar{x}_2 - \mu)$.

(**Q6**) 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 $(x_{2i} - \bar{x}_2)(\bar{x}_2 - \mu)$.

(**Q7**) Show that the sum of the cross terms over all data points is zero.

(**Q8**) 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.

(**Q9**) 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?

(**Q10**) Apply the inequality to the exam score data $\sigma^2_{\text{population}}$.

Part 4 - Long Term Implementation Of Mental Health Services

Based on your analysis, the university decided to provide free access to the mental health services to all students but wants to monitor the effect that it has on the students. Your are tasked with the role of setting up the OLTP system that will be used to add additional information over time.

You are provided with an initial table and the information that each student has a unique **StudentID** and belongs to a unique **Faculty**. Additionally, each student can get access to a **ServiceType** only once a day and there is a unique **Provider** for each **ServiceType**.

StudentID	StudentName	Faculty	ServiceType	Date	Provider
S1001	Alice	Science	Counseling	2023-11-01	Dr. Smith
S1001	Alice	Science	Workshop	2023-11-15	Dr. Jones
S1002	Bob	Engineering	Counseling	2023-11-02	Dr. Smith

(**Q1**) Propose examples of modification (insertion, deletion, update) anomalies that could arise in this table.

(**Q2**) List all functional dependencies.

(**Q3**) List all superkeys of this table and identify the candidate key(s).

(**Q4**) Is the table in 2NF? If not, transform it into 2NF.

(**Q5**) Suppose we add provider specialties (e.g., "Anxiety", "Depression") to the Services table. Is the resulting table in 3NF? If not, normalize it to 3NF.

(**Q6**) The university board wants to analyze trends over 5 years, what would you add to your schema to support: comparisons of stress trends across faculties?

(**Q7**) For reporting, The university board wants a table showing the number of services by faculty over time. Propose a denormalized design for this report and explain one trade-off.