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."

  Issues include the lack of a proper time frame ("during the pandemic"), the lack of a proper scale definition (what does 1 vs 5 mean?) and the lack of reliability from only asking about one point of data for the pandemic. An improved version could be "Over the past month, how often have you felt stressed (1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Always)? Stress includes feeling overwhelmed, anxious, or unable to cope."

- "How many hours do you study per week? (open-ended)"

  Issues include unrealistic estimates and distinction between different study activities. An improved version could be " On average, how many hours per week do you spend on each of the following?

  - Attending live online classes: _ hrs
  - Watching recorded lectures: _ hrs
  - Reading/assignments: _ hrs
  - Group study: _ hrs

  "

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.

  The sampling approach is noot representative of the full student population as it may introduce (self-)selection bias: those who respond first may have different characteristics (e.g., more motivated, less stressed) than those who respond later or not at all. However, it is practical and easy to implement.

- All the third-years.

  This sampling approach is biased as it excludes students from other years, which may have different experiences and stress levels (coverage bias). However, it is practical if the university wants to focus on a specific cohort and student lists are available.

- A stratified sample by faculty (Science, Humanities, Engineering, Business), selecting every tenth student from an alphabetical list within each faculty.

  Stratification by faculty improves representation across faculties, especially if sample sizes are proportional to faculty sizes. However, selecting every tenth student from an alphabetical list may introduce bias if certain groups are clustered alphabetically (e.g., common last names). This method remains practical to implement but requires access to reliable enrollment lists and accurate faculty assignment student lists.

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.

(**Q1**) Both provided datasets should primarily be considered secondary data. The Exam Performance dataset was collected as part of routine academic administration for grading and reporting purposes, not specifically for this study. Similarly, the Mental Health dataset was collected for internal follow-up and resource planning within the university's mental health services. Although students were informed their data could be reused for research purposes, the data was not originally gathered to answer the specific questions raised in this analysis.
(**Q2**) The most evident ethical issue with manipulating these datasets is the risk of violating student privacy. The datasets contain two types of sensitive information: mental health service usage and exam performance. Even though a Student ID is used as an identifier, linking the two datasets increases the risk of re-identification. Additionally, while students were informed of potential research use and given an opt-out option for the mental health dataset, this does not automatically guarantee that they fully understood how their data would be combined, analyzed, or interpreted.

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?

(**Q3**) Quantitative data is required as the question involves analyzing the (cor)relation between two numerical variables: Attendance and Exam Score.
(**Q4**) This question can be answered with the provided datasets, actually only the Exam Performance dataset is needed, specifically the fields Attendance and Exam Score.
(**Q5**) Since the lecture was online; students might have logged in but not listen to the content (especially is attendance is mandatory), leading to an overestimation of actual engagement and distort the relationship with exam scores. The analysis might lead to spurious correlations if attendance is not a true reflection of engagement. For example, students might appear to attend but not actually learn, distorting the relationship with exam scores.

ii. Did students who attended more lectures experience less stress during the pandemic?

(**Q3**) The datasets describe attendance using quantitative data and stress can be related to the use of the mental health service, which is represented yes/no categories (qualitative). However, as this is binary data, it can be transformed into numerical values (0/1) to perform mathematical operations.
(**Q4**) This question cannot be answered with the provided datasets. While the use of the mental health services is probably related to stress, it is not equivalent. Thus there is no direct measure of stress in any of the provided datasets.

iii. Did students who used mental health services performed less at the exam than those who did not?

(**Q3**) Quantitative data is required since the question requires comparing exam scores (a numerical measure) between two groups (students who used mental health services vs. those who did not).

(**Q4**) This question can be answered with the provided datasets: the Exam Performance dataset includes Exam Score, and the Mental Health dataset includes Service Usage. These datasets can be combined using Student ID to compare exam scores between the two groups.

(**Q5**) Data Quality Issue: The Mental Health dataset excludes 100 students who opted out of data collection and 150 part-time students. This could introduce bias, as students who opted out might differ systematically from those who did not (e.g., they might have higher or lower stress levels). Note that the question can only be adressed for full-time students and part-time students might be systematically different (e.g., older, working professionals with different stress levels or academic performance).

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?

(**Q1**) We simply weight the old average and add the knowny discovered one $\mu_{\text{adjusted}} = \frac{1100\times12+100\times15}{1200} = 12.25$.

(**Q2**) The adjusted average (12.25) is higher than the sample average (12). This suggests that the sample underestimated the true population average because the in-person exam takers performed better. The misalignment arises because the sample excluded a higher-performing subgroup.

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.

The overall population variance is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

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

Splitting the sum into two subgroups:

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{n_1}(x_{1i} - \mu)^2 + \sum_{i=1}^{n_2}(x_{2i} - \mu)^2\right)$$

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

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{n_1}((x_{1i} - \bar{x}_1) + (\bar{x}_1 - \mu))^2 + \sum_{i=1}^{n_2}((x_{2i} - \bar{x}_2) + (\bar{x}_2 - \mu))^2\right)$$

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

$$\sigma^2 = \frac{1}{N}\Big(\sum_{i=1}^{n_1}\left[(x_{1i} - \bar{x}_1)^2 + (\bar{x}_1 - \mu)^2 + 2(x_{1i} - \bar{x}_1)(\bar{x}_1 - \mu)\right] + \sum_{i=1}^{n_2}\left[(x_{2i} - \bar{x}_2)^2 + (\bar{x}_2 - \mu)^2 + 2(x_{2i} - \bar{x}_2)(\bar{x}_2 - \mu)\right]\Big)$$

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

$$\begin{aligned}
\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)(\bar{x}_1-\mu) &= (\bar{x}_1-\mu)\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1) \\
&= (\bar{x}_1-\mu)\times(\sum_{i=1}^{n_1}x_{1i}-\sum_{j=1}^{n_1}\bar{x}_1) \\
&= (\bar{x}_1-\mu)\times(\sum_{i=1}^{n_1}x_{1i}-\sum_{j=1}^{n_1}\frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i}) \\
&= (\bar{x}_1-\mu)\times(\sum_{i=1}^{n_1}x_{1i}-\sum_{i=1}^{n_1}x_{1i}) \\
&= 0
\end{aligned}$$

The same holds for the other term.

(**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.

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n(\bar{x}_1-\mu)^2 + \sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2 + m(\bar{x}_2-\mu)^2\right) \\
&= \frac{n_1}{N}s_1^2 + \frac{n_2}{N}s_2^2 + \frac{n_1}{N}(\bar{x}_1-\mu)^2 + \frac{n_2}{N}(\bar{x}_2-\mu)^2 \\
&= \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1-\mu)^2 + n_2(\bar{x}_2-\mu)^2}{n_1+n_2}
\end{aligned}$$

Here, $\frac{n}{N}s_1^2$ and $\frac{m}{N}s_2^2$ represent the within-subgroup variances, while the terms $\frac{n}{N}(\bar{x}_1-\mu)^2$ and $\frac{m}{N}(\bar{x}_2-\mu)^2$ represent the between-subgroup variances.

(**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?

By construction, the other terms are non negative. The inequality becomes an equality if and only if the subgroup means are equal to the overall mean, i.e., $\bar{x}_1 = \bar{x}_2 = \mu$. In this case, the subgroup means do not contribute to the overall variance. This inequality tells us that including the excluded group cannot decrease the overall variance below the weighted average of the subgroup variances. In other words, the excluded group introduces additional variability unless its mean is identical to the overall mean.

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

$$\sigma^2_{\text{population}} \geq \frac{1100\times 9 + 100\times 4}{1200} \approx 8.56$$

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.

The table may lead to several classical modification anomalies:

- Insertion anomaly: It is not possible to register a new student in the system unless the student has already used a mental health service. Similarly, a new **ServiceType** or **Provider** cannot be recorded unless at least one student has already used that service.
- Update anomaly: If a student changes faculty (e.g., from Science to Engineering), the **Faculty** value must be updated in every row corresponding to that student. If one row is missed, the database becomes inconsistent. Likewise, if the provider for a service type changes (e.g., *Counseling* is no longer provided by Dr. Smith), this update must be applied to all past rows using that service.
- Deletion anomaly: If the last row corresponding to a given student is deleted, all information about that student (name and faculty) is lost, even if the student is still enrolled. Similarly, deleting the last usage of a given **ServiceType** removes all information about its associated **Provider**.

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

The functional dependencies implied by the description are:

- **StudentID** → **StudentName**, **Faculty**
- **ServiceType** → **Provider**
- **StudentID**, **ServiceType**, **Date** → all attributes, since a student can receive a service only once per day.

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

A superkey is any attribute set that uniquely identifies a row. A candidate key is a minimal superkey. The only candidate key here is {**StudentID**, **ServiceType**, **Date**}, meaning that any supersert of this combination is a superkey such as:

- {StudentID, ServiceType, Date}
- {StudentID, ServiceType, Date, StudentName}
- {StudentID, ServiceType, Date, Faculty}
- {StudentID, ServiceType, Date, Provider}
- etc.

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

The table is not in 2NF because non-key attributes depend on only part of the composite key (**StudentID**, **ServiceType**, **Date**):

- **StudentName** and **Faculty** depends only on **StudentID**
- **Faculty** depends only on **StudentID**

A possible 2NF decomposition is:

- Student(**StudentID**, **StudentName**, **Faculty**)
- Service(**ServiceType**, **Provider**)
- ServiceUsage(**StudentID**, **ServiceType**, **Date**)

(**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.

With provider specialties are added directly to the Service table, the design is **not in 3NF** because a provider may have multiple specialties, introducing a transitive dependency: **ServiceType** $\rightarrow$ **Provider** $\rightarrow$ **Specialty**. A possible 3NF decomposition is:

- ProviderSpecialty(**Provider**, **Specialty**)
- Service(**ServiceType**, **Provider**)

(**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?

To support long-term analysis and students changing faculties, the schema should include a FacultyHistory(**HistoryID**, **StudentID**, **Faculty**, **StartDate**, **EndDate**) table to track faculty changes over time. Note that the Faculty attribute should be removed from the Student table. To handle the stress trend, a **StressLevel** attribute can be added to the ServiceUsage table to log stress levels over time.

(**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.

For reporting, a denormalized table could be: ServiceCountByFaculty(**Faculty**, **Year**, **Month**, **ServiceCount**). This design improves query performance and simplifies reporting dashboards but introduces data redundancy and requires careful maintenance to keep the aggregated data consistent with the underlying transactional tables.