STOR 390: HOMEWORK 8

Riley Harper

April 20, 2024

Abstract: This homework explores various aspects of data privacy and fairness in statistical analysis, integrating topics such as differential privacy and the ethical responsibilities surrounding data interpretation. By exploring the Randomized Response technique, we demonstrate its effectiveness in maintaining ε-differential privacy and protecting individual responses in sensitive surveys, where $\varepsilon = \ln(3)$. Additionally, we engage with regression models to control for confounding variables such as age and smoking status, illustrating how these factors can influence the outcomes in studies related to lung capacity.

Question 1

Prove that Randomized Response Differential Privacy is $\varepsilon-$differentially private. In particular, show that $\varepsilon = ln(3)$.

The Randomized Response technique is a method used to protect the privacy of individuals in surveys, particularly for sensitive questions. The method involves the following steps,

1. The respondent flips a fair coin.
2. If the coin lands heads, they flip another coin,
    - If the second coin lands heads, they answer "Yes".
    - If the second coin lands tails, they answer "No".
3. If the coin lands tails, they answer truthfully.

Let f denote the true answer, where $f = 1$ for "Yes" and $f = 0$ for "No". The probability of responding "Yes" ($Y = 1$) is calculated as follows,

$$P[Y = 1 \mid f = 1] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$$

$$P[Y = 1 \mid f = 0] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

The definition of ε-differential privacy is given by,

$$\varepsilon := \ln\left(\frac{P[Y = y \mid f = 1]}{P[Y = y \mid f = 0]}\right)$$

If we let $y = 1$,

$$\varepsilon = \ln\left(\frac{P[Y = 1 \mid f = 1]}{P[Y = 1 \mid f = 0]}\right) = \ln\left(\frac{\frac{3}{4}}{\frac{1}{4}}\right) = \ln(3)$$

Therefore, the Randomized Response technique is ε-differentially private with $\varepsilon = \ln(3)$.

Question 2

Recall the example of *regressing out* a confounding variable we used in class. Here, we established that not only is 'age' related to lung capacity, but so too is 'smoke'. Interpret the model coefficients produced below, and in particular explain why including both 'age' and 'smoke' helps to control for confounding. (*Hint: This should involve interpreting your coefficients* ***marginally***.)

Correlation: 0.756459

Model Summary:

```
Call:
lm(formula = fev ~ age + smoke, data = fev)

Residuals:
Min 1Q Median 3Q Max
-1.6653 -0.3564 -0.0508 0.3494 2.0894

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 0.367373 0.081436 4.511 7.65e-06
***
age 0.230605 0.008184 28.176 < 2e-16 ***
smoke -0.208995 0.080745 -2.588 0.00986 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5651 on 651
degrees of freedom
Multiple R-squared: 0.5766, Adjusted
R-squared: 0.5753
F-statistic: 443.3 on 2 and 651 DF, p-value:
< 2.2e-16
```

Including both 'age' and 'smoke' is essential for controlling the confounding effects these variables may have on lung capacity (FEV)-forced expiratory volume. Age is a primary determinant of lung capacity. Generally, lung capacity may increase in younger ages, peak, and then decline as part of the natural aging process. By including 'age' in our regression, we adjust for these physiological changes with the aim of isolating the impact of other factors like smoking. Also, smoking status has been linked to lung health; smoking can significantly reduce lung capacity. Including this variable helps to quantify its impact on lung capacity while controlling for age differences among subjects. Including both 'age' and 'smoke' in the model allows us to mitigate the confounding influence one might have on the other. For example, if older individuals in the study are more likely to be smokers, not adjusting for age could lead to overestimating the impact of smoking on lung capacity.

The coefficients from our regression model give us the marginal effects of each variable,

- The coefficient of 'age' (0.230605) tells us lung capacity is expected to change by 0.230605 units with each additional unit change (in years) of age, holding smoking status constant.
- The coefficient of 'smoke' (-0.208995) tells us us the lung capacity is expected to change by -0.208995 units with each additional unit change in smoking, holding age constant.

Question 3

Describe one experimental design technique that could be used to mitigate *known* confounders or batch effects (as opposed to latent batch effects).

An effective experimental design technique to control for known confounders or batch effects is **stratified random sampling**. This method can be defined by dividing a population being observed or surveyed into smaller groups, known as strata, that are homogeneous with respect to one or more characteristics which are known to be confounders.

Implementation of Stratified Random Sampling (SRS):

- **Strata Formation:** The first step is to identify the known confounder(s) which might impact the study outcome. The entire population is then divided into subgroups or strata based on these confounders. Each stratum should be homogeneous regarding the confounder (a few examples of confounders could be age, sex, socioeconomic status).
- **Random Sampling:** From each stratum, participants are randomly selected to participate in the study. This ensures that each stratum is represented proportionally or equally, depending on the study design.
- **Control Over Confounding:** By stratifying the population by known confounders, the impact of these confounders is controlled across the treatment groups, reducing their effect on the outcome variable.

A few advantages to utilizing SRS include increased precision, improved validity, and efficiency. SRS improves the precision of the study by ensuring that the sample is more representative of the overall population, particularly regarding the confounders that are stratified. SRS also enhances the validity of the results by ensuring that the confounding variables are evenly distributed across all study groups, leading to more reliable and reproducible findings. Finally SRS requires a smaller sample size to achieve the same level of precision compared to simple random sampling, making it more efficient in terms of resources and time. SRS is particularly useful in observational studies and surveys where specific subgroups may have different responses or outcomes. By ensuring these subgroups are adequately represented and controlled for in the study, researchers can make more precise inferences about the effects of the variables being studied.

Question 4

I have described our responsibility to proper interpretation as an *obligation* or *duty*. How might a Kantian Deontologist defend such a claim?

From the perspective of Kantian Deontology, ethical behavior is grounded in adherence to universal maxims and the fulfillment of duties, irrespective of the consequences. According to Immanuel Kant, actions are morally right when they are performed out of duty and conform to universal law. This approach can be directly applied to the concept of the duty of proper interpretation. According to Kant's principle of universalizability, an act is morally acceptable if it can be universalized, meaning it can consistently be willed as a law that everyone should follow. In terms of proper interpretation, this means that if the act of interpreting information accurately and responsibly could be universalized, it becomes not only a preferable practice but a moral duty. This stems from the desire for a society where truth and understanding are upheld as common values. Kant's Categorical Imperative demands that we act only according to maxims that could be adopted universally. This imperative implies that misinterpreting or distorting information would be contrary to this principle because it would not be desirable for such practices to become universal law. Thus, a Kantian Deontologist would argue that we have a moral duty to interpret correctly to respect and uphold the rationality and autonomy of all individuals affected by the interpretation. Further, Kant's notion of treating humanity, whether in oneself or in another, always as an end and never merely as a means to an end, reinforces the duty to interpret accurately. This ensures that individuals are not misled or harmed through misinterpretation, thus respecting their autonomy and dignity.

Therefore, from a Kantian Deontological perspective, proper interpretation is not merely a professional responsibility but a moral obligation. This statement aligns with Kant's emphasis on actions that can be universally willed and that inherently respect the autonomy and dignity of others involved. Misinterpretation, by contrast, would undermine trust, dignity, and rational decision-making, thereby violating Kantian moral laws.

Question 5

(Free Points) What was the most informative/interesting unit in this course? The least?

The most interesting unit for me was focused on statistical metrics of fairness, including Statistical Parity (SP) and Disparate Impact (DI). This was particularly engaging due to its relevance in real-world applications and my previous lack of knowledge on the subject. While I found the deep dive into COMPAS also enjoyable and fascinating, the lecture dedicated class-wide discussion felt somewhat repetitive, with the same points being reiterated without adding substantial new insights; I also have concerns about the grading of opinions in such discussions, particularly considering the inherent biases stemming from the instructor's privileged background as a white heterosexual male. This aspect of the course could potentially influence the objectivity and fairness of grading, especially in discussions on sensitive topics like fairness and bias.

The least interesting unit for me was on differential privacy. Despite the clear importance of protecting privacy in data analysis, the presentation and explanation in class were somewhat underwhelming. I felt that more detailed examples or engaging visuals, similar to the exemplary bottle analogy used for SVMs, could have made this topic more accessible and intriguing for all students. Such enhancements to the lecture would've likely helped clarify the concepts covered in E-DP and enhance understanding, making the learning experience more effective and enjoyable.