FIT5125 Assignment 3 Week 9

Working with hypotheses - Artificial Intelligence and Data Privacy

1. Hypothesis

The greater the value of the service received in exchange, the greater data privacy people are willing to give up.

2. Null Hypothesis

People do not care about how much data they give up, as long as they are getting a service in return at all.

3. Relevant Variables

Independent Variables

Time frame of before/after generative AI boom
Usefulness of services provided

Dependent Variables

Volume of sensitive data
Attempts to obfuscate data entered

Confounding Variables

Participant familiarity with AI/IT in general
General vigilance levels of participant
Value of privacy to participant
Prestige bias of participants
Inaccurate self-perception of participants

4. Investigating Study

Due to budget constraints, a survey will be chosen here, as the cheapest study. A questionnaire will be designed to determine usage patterns of generative AI, search engines, and social media, contrasted with their past usage patterns before AI took off. Data privacy is cross-referenced with social media and search engines data collection policies, and contrasted with how users take steps to maintain their data privacy e.g. by anonymizing their search queries in search engines/generative AI, or sharing limited information with social media.

5. Statistical Tests

Multivariate regression analysis - We assume that there are multiple dependent variables to be predicted by at least one independent variable. As we are not sure if it is a linear, logarithmic or exponential relationship, we will run the statistical tests on all 3 scales to see which fits the data the best.

This statistical test is used because we want to know not just if the dependent variables can be predicted, but how strongly correlated they are to the predicting variables.

6. Limitations

People's answers to a survey may not reflect actions in reality. This is doubly true for data privacy, as it is vague and taking steps to protect it is something many people may not even know how to do. It is vulnerable to prestige bias, as people may feel compelled to show that they care about data privacy when they really don't. Sampling bias is also difficult to avoid - this is a topic that is likely to vary widely on the target audience, and the most notable audience to easily miss is the tech illiterate people, as they are unlikely to even participate in any such study. To top it off, "value of services provided in return" is inherently vague and subjective - everyone has a different opinion on what's valuable to them.

Given enough budget for a future investigation, an analysis of people's real behaviours may be undertaken to remove these variables.

7. Alternate Narrative

An alternative narrative could be that a greater volume of user data collected leads to a better quality of service provided. In this scenario, the independent variable would be the volume of data collected, and the dependent variables would be the quality and usefulness of services to the participants.