Statistics 24/25 Seminar I

Dmytro Tupkalenko

Format of the seminar:

I wanted to make the seminar in the format of article but a little bit less formal, so I made it in the form of blog. Here is the link to the blog itself: Medium blog. All the sources are referenced in the blog. The graphical representation were made by me in R, here is the Github repository with all the working files: Github repository.

My seminar includes the following sections (with short description):

Understanding Simpson's Paradox

(i) Classic Definition

I give the classic definition of Simpson's paradox among with the graphical representation.

(ii) Probabilistic Terms

I give three probabilistic statements that state what should hold for Simpson's paradox to occur.

(iii) Vector Representation

I give a very intuitive and a nice for understanding from my point of view definition in vector-space context.

Real-World Examples

(i) Penguin Body Parameters

This is an example with two numerical variables about the physical parameters of penguins, I found it on this webpage.

(ii) Kidney Stone Treatment

This is an example with two categorical variables about the treatment of Kidney stones, I found it on Wikipedia page.

Claims Regarding Simpson's Paradox

I stated and proved two claims (only one proof is fully shown in the summary) about occurrence of Simpson's Paradox and also connected these two claims to the second example.

- **Claim 1**: If $P(X|Y) = P(X|Y^{\complement})$, the three inequalities cannot hold.
- **Claim 2**: If $P(Y|X) = P(Y|X^{\complement})$, the three inequalities cannot hold.

Proof. Assume that $P(X|Y) = P(X|Y^c)$. Since the three properties should not hold simultaneously we will show that assuming one is true, the other cannot hold together.

Suppose that $P(I|Y) < P(I|Y^c)$, then using the total law of probability for conditional case we can rewrite this as following:
$P(I|X \cap Y) \cdot P(X|Y) + P(I|X^c \cap Y) \cdot P(X^c|Y) < P(I|X \cap Y^c) \cdot P(X|Y^c) + P(I|X^c \cap Y^c) \cdot P(X^c|Y^c)$

Now, note that $P(X|Y) = P(X|Y^c) \iff 1 - P(X|Y) = 1 - P(X|Y^c) \iff P(X^c|Y) = P(X^c|Y^c)$.
Hence, the expression can be further simplified to:
$0 < P(X|Y) \cdot (P(I|X \cap Y^c) - P(I|X \cap Y)) + P(X^c|Y^c) \cdot (P(I|X^c \cap Y^c) - P(I|X^c \cap Y))$.

Then, assuming that $P(I|X \cap Y) > P(I|X \cap Y^c)$ and $P(I|X^c \cap Y) > P(I|X^c \cap Y^c)$; we get a contradiction since $P(X|Y), P(X^c|Y^c) > 0$.
Consequently, the three statements cannot hold all together □

The second proof is very similar, but also uses Bayes' theorem.

Preventing Simpson's Paradox

Here I describe how to avoid Simpson's paradox at the experiment planning stage by using stratified sampling. And also speculate when it is possible to use this strategy.