Homework 01

Goncharov Anton, group 204
Data Science and Business Analytics
Higher School of Economics

December 27, 2023

Question 1

Consider a test for testing the null hypothesis that a proportion is equal to some constant $p_0 = \frac{1}{2}$:

$$H_0 : p = \frac{1}{2}$$

(a) Calculate minimum sample size needed to determine confidence interval with relative precision of 8%, i.e. $p \pm 0.08 \cdot p$.
(b) Derive the formula for MDE as a function of power $(1 - \beta)$, type I error (α), number of observations (N). Calculate MDE in case the researcher has a sample of 500 observations, desired power is 60% and significance level is 5%.
(c) Based on the derivations from (b), find what will be the power of a test in case true value of p is known: $p = \frac{1}{4}$. Sample size is 500 observations and significance level is 5%.

a). Minimum sample size:

$$n = \frac{Z^2}{E^2} \cdot p(1 - p),$$

where n is the sample size, Z is the Z-score corresponding to the desired confidence level, p is the estimated proportion, E is the margin of error, which is the relative precision multiplied by p.
Substituting the values we get:

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot 0.5}{(0.08 \cdot 0.5)^2} = \frac{0.9604}{0.0016} = 600.25$$

Therefore the minimum sample size must be 601 to ensure a CI with 8% relative precision.

b). The minimum sample size can also be expressed through the significance level (α), the power $(1 - \beta)$ and the Minimum Detectable Effect (MDE):

$$n \geq \frac{\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{MDE^2} => n \geq \frac{p(1-p)(z_{1-\alpha/2} + z_{1-\beta})^2}{MDE^2}$$

A few rearrangements, and we get:

$$MDE^2 = \frac{p(1-p)(z_{1-\alpha/2} + z_{1-\beta})^2}{n} = \frac{1/4(1.96+0.25)^2}{500} = \frac{0.25 \cdot 4.88}{500} = 0.00244$$

So, $MDE = \sqrt{0.00244} \approx 0.05 = 5\%$

c).

$$n = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{MDE^2} => (z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{n \cdot MDE^2}{\sigma^2} =>$$

$$=> z_{1-\alpha/2} + z_{1-\beta} = \sqrt{\frac{n \cdot MDE^2}{\sigma^2}} => z_{1-\beta} = \sqrt{\frac{n \cdot MDE^2}{\sigma^2}} - z_{1-\alpha/2}$$

$$=> 1-\beta = P(Z \geq \sqrt{\frac{n \cdot MDE^2}{\sigma^2}} - z_{1-\alpha/2}) => 1-\beta = P(Z \geq \sqrt{\frac{500 \cdot MDE^2}{1/4 \cdot 3/4}} - 1.96)$$

$$=> 1-\beta = 1 - \Phi(\sqrt{\frac{500 \cdot MDE^2}{0.1875}} - 1.96)$$

Question 2

Case Study:
Cryptocurrencies, such as Bitcoin and Ethereum, are digital currencies which employ encryption techniques to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank. Volatile price movements in the past year have resulted in financial regulators leading calls for cryptocurrencies to be reined in for fear of contagion risks if they enter the financial mainstream.
Recently, some retail banks have banned customers from trading in cryptocurrencies using credit cards (due to the bank's liability in the event of default), but still permit transactions using debit cards (as a customer's own funds are used), although some banks are considering also banning the use of debit cards to trade as well to safeguard against customers realising large losses.
One bank is considering imposing a cap on debit card transactions rather than an outright ban (for fear of losing customers to competitors). The bank has invited you to devise an appropriate sampling scheme to research attitudes of its customers regarding the introduction of a cap, including the level of any cap.
Consider the following sampling techniques:
i. Simple random sampling.
ii. Quota sampling.
iii. Stratified sampling.
iv. Cluster sampling.
Explain how each of these sampling methods could be applied to this case study. Describe the merits and limitations of each of these methods in light of the given case.

1. Simple Random Sampling:
The sample is drawn at random, every customer in the bank has an equal probability of being selected for the survey.
+ Merits:
Simple and easy to implement.
Every customer has an equal chance of being selected, reducing bias.
- Limitations:
May not ensure representation of various customer segments.
Could lead to a skewed sample if not executed properly (usually the case).

2. Quota Sampling:
We first establish quotas with which we select customers based on convenience or judgment, ensuring representation from each category. In this case, categories might include different age groups, income levels, or transaction histories.
+ Merits:
Ensures representation from various customer segments.
Relatively simple to implement.
- Limitations:
May introduce bias if the quotas are not well-defined.
The randomness of selection is compromised.

3. Stratified Sampling:
First split the population into equal sub-populations, called strata. Then, we select customers from each stratum by a random procedure, usually SRS (Simple Random Sample). This could involve selecting a certain number of customers from each income bracket.
+ Merits:
Ensures representation from various customer segments.
Allows for more precise analysis within each stratum.
- Limitations:
Requires accurate information about the characteristics of the population.
Can be more complex to implement than simple random sampling.

4. Cluster Sampling:
First split the population into several clusters based on some criteria. Then, we pick a random cluster based on a probability sampling technique. For instance, the bank's branches could be considered as clusters, and a subset of branches could be chosen for the survey.
+ Merits:
Cost-effective, especially if clusters are geographically concentrated.
Allows for more localized studies.
- Limitations:
May lead to a lack of representation if clusters are not diverse.
Potential for intra-cluster similarity, leading to biased results.

Considerations for the Given Case:
Customer Preferences: Simple random sampling might be appropriate if the bank wants to ensure each customer has an equal chance of providing input.
Diverse Customer Base: Stratified or quota sampling could be useful if the bank is concerned about representing the opinions of various customer segments.

Cost and Efficiency: Cluster sampling might be beneficial if the bank wants to minimize costs and effort by focusing on specific branches or regions.
Accuracy: If the bank is primarily concerned about accurate representation, stratified or quota sampling might be more suitable.

Question 4

This assignment asks you to replicate some of the findings from Sheridan and Ball [Sheridan, Ball, 2005].
Source data file: Inflation_Targeting.xlsx. Please note that there are two separate sheets (for developed countries and for developing countries).
(a) Compare average inflation rates in target countries before and after inflation targeting. Based on this result, can we conclude about the impact of inflation targeting on the inflation rate?
(b) Now apply the difference-in-differences method by estimating the parameters of equation (1) from the paper by Sheridan and Ball [Sheridan, Ball, 2005, pp. 249-276]. Estimate for three samples:
- complete selection of countries;
- the developed countries;
- developing countries.
Interpret the results, explain what can be said about the impact of the transition to inflation targeting on the inflation rate in the long term, based on the obtained parameter estimates?

a). After conducting a t-test on the average of inflation rates before and after targeting, we get these results:
Average Inflation Before: 10.9
Average Inflation After: 5.093103448275863
T-Statistic: 2.7615825347309197
P-Value: 0.007764623568338804 **
The P-Value is lesser than 0.01, showing significance at 99% confidence. Thus, the null hypothesis that there is no significant difference in average inflation rates before and after inflation targeting is rejected. However, it's important to note that statistical significance does not imply the size or practical significance of the effect.

b). The model proposed in the paper is:

$$X_{post} - X_{pre} = a_0 + a_1 D + e$$

The difference-in-difference estimate can be obtained from regressing this formula:

$$y = a_0 + a_1 g + a_2 t + a_3 (t \cdot g) + \epsilon,$$

where g is 0 for the control (non-targeting countries) group and 1 for the treatment (targeting countries) group;

t is 0 for before and 1 is after;
tg is the interaction between t and g.
i). Complete selection of countries:
DID estimate: -5.57
P-value: 0.018 * (significant at 95% confidence)

ii). The developed countries:
DID estimate: -0.977
P-value: 0.492 (insignificant)

iii). The developing countries:
DID estimate: -7.456
P-value: 0.021 * (significant at 95% confidence)

So, the average of inflation rates in the developed countries after being targeted has seen insignificant changes throughout the years, while the average decrease of inflation rates in the developing countries has seen significant change of 7.456%. The overall average of inflation rates across all countries has also been significantly reduced by 5.57%.

Results show that the impact of the transition to inflation targeting on the inflation rate varies between developed and developing countries. In developed countries, its suggested that inflation targeting might not have had a notable effect on average inflation rates. However, in developing countries, there is evidence of a significant reduction in inflation rates following the adoption of inflation targeting.

The overall average reduction in inflation rates across all countries supports the idea that, on average, inflation targeting has had a statistically significant impact on lowering inflation rates globally.