CASE STUDY

Dabbling with Aggregate and De-Aggregated Data: Role of Mr John Stats

Effulgence

Vol. 19, No. 2

July - December, 2021

Rukmini Devi Institute of Advanced Studies

E-mail : effulgence@rdias.ac.in, Website : www.rdias.ac.in

<http://effulgence.rdias.ac.in/user/default.aspx>

<https://dx.doi.org/10.33601/effulgence.rdias/v19/i2/2021/65-70>

Pardhasaradhi Madasu¹

Abstract

Sometimes as decision-maker many of the managers face paradoxical situations. Few situations exemplify the Yule-Simpson Effect, in which the trend appearing in the aggregate data may be reversed in the data of sub-categories. The paradoxical situation could be observed in many real-life situations say in the case of professional tennis tournaments. Wright (2012) in his research states that in a data set of over 55,000 individual tennis matches, 5% of the matches exhibit the paradox of the Yule-Simpson Effect. The results of the study state that this kind of situations promote gambling-related activities. Carlson (2019) states that Simpson Paradox a classical example of association paradoxes and further mentions that the Yule-Simpson Effect occurs when the marginal association between two categorical variables is qualitatively varied from the partial association between the same two variables after controlling for one or more other variables. This variation in results would break the myth of many that statistical relationships to be unchallengeable. SEP (2021) states that the paradox is not only a surprise for statisticians but also serves as a lens through which the role of probabilities in data analysis, causal inference, and decision-making can be understood. The present case provides an opportunity to understand the paradox from a business point of view. Further, the case also gives the participants the scope of deliberating on the finer points of statistical analysis.

Keywords: *Simpson Paradox, Causal Inference, Trend-Reversal, Statistical Analysis and Decision-Making.*

INTRODUCTION

It was around 8.00 pm on the fourth Friday of November 2020 that Mr John Stats (hereinafter referred to as John) started to rewind the happenings

of the day at the company's corporate office. John is employed as a data scientist at All-Bright Pvt. Ltd – ABPL¹. Thinking on the day's happenings, somewhere John felt that the issue that could be solved using common sense is being prolonged. He

1. Associate Professor (Finance Area), Siva Sivani Institute of Management (SSIM), Hyderabad
mpardhasaradhi@ssim.ac.in

pretty well knows that everyone is not good at number crunching and statistical analysis but he also feels that sometimes statistical lens has to be used for solving issues that come up during business operations. John advocates that all budding managers should understand that the interpretation based on aggregate data and segmented data could differⁱⁱ. John also acknowledges that generally the perspectives of the marketing team and other functional teams would not sync and that is the cause of today's deliberations at John's office. In his previous job, John learnt from his Ex-Boss that numbers and analysis should be the basis for decision making but not the perceptions of individuals. In contrast, John's ex-boss also cautioned John about McNamara's quantitative fallacyⁱⁱⁱ.

John's thoughts are revolving around the claim of Mr Rocky (Head-Marketing) and the negation of the said claim by Ms Dimple (Head-Accounts & Finance). John intuition says that the issue could be solved by de-aggregating the data and checking for the presence of Simpson's Paradox. John is confident that Mr Rocky would share the dis-aggregate data tonight and he could find reasons for dis-agreement by the next day. John agreed to present the segmented data and its analysis on coming Monday such that the issue of Rocky's claim of incentive and negation of the claim by Dimple is put to an end. John wants to elaborate more on how contradicting claims could be proved with the same data as advocated by Simpson's Paradox^{iv}.

About Mr John Stats

John (32-year-old) is a management graduate from one of the very premier B-Schools in Hyderabad. In his masters, John's majors were Accounting and Finance. However, due to the changing job profiles, he started to equip himself with skills from the data sciences domain. He attended a few online data sciences courses offered by Harvard Business School (HBS). Being interested in the data sciences stream John started to read a lot of stuff related to fallacies and biases in data analysis. John has a habit of writing

extensive notes on the topics he liked and felt useful. As of now, John's subject notebook reveals that he is interested in few topics such as Gambler's Fallacy, False Causality, Overgeneralization, McNamara Fallacy, Cobra Effect, Simpson's Paradox and Data Dredging (Mayo, 2019). John started to research these fallacies wherever possible. This exposure has given him a sense of direction and confidence to handle data for decision making.

John joined ABPL two months ago. His primary role is conducting marketing research and providing support to the accounting and finance team with spreadsheet modelling. He is expected to provide technical support in data analysis and visual presentations. This is where his role has come up in resolving the debate between Rocky and Dimple. Both of these senior executives are considered as foundation pillars of ABPL and are very good at their functional roles. They both are respected for their integrity and value system.

CEP THE BRAINCHILD OF ROCKY

ABPL is a twelve-year-old company with six stores/branches operating across the state of Telangana and ABPL has a policy of opening one store in each of its six market segments. The major market segments of ABPL are Hyderabad, Warangal, Nizamabad, Khammam, Nalgonda and Karimnagar. ABPL primary business is electronic products and the stores enjoy a good reputation all over the state. Mr. Rocky takes care of the marketing and sales of all these six stores. Ms. Dimple is very much involved in the maintenance of accounts, financial analysis and compliance with GST.

Rocky from day one of ABPL inception started to archive data of customers. He has created a database of customers by using MS - Access (MS Office Suite) and he is very meticulous in updating the database every day. At the beginning of the financial year (FY 2021 - 22) Mr. Rocky has come out with a Customer Engagement Program (CEP) and wanted to introduce the CEP in all the stores of the ABPL. Mr Rocky did not

believe much in the concept of marketing research or the application of statistics for taking decisions. Rocky perceived that in this kind of business hard work matters rather than some kind of statistical analysis and decisions based on these analyses.

Rocky's new CEP is designed around the usage of the Business App of WhatsApp for the promotion of ABPL products. Rocky and his team have designed a plan where through a registered business account; ABPL could send notifications on discounts, exchange offers and new product arrivals to the customers. The feature of Chatbot of WhatsApp business^v is the favourite of Rocky and he is confident of engaging customers directly and can also automate the messages. Further, Rocky feels customer grievances and customer queries on the usage of electronic products could also be done through WhatsApp API. Rocky and his team are convinced that in the matter of customer grievances and queries, WhatsApp is a great alternative to the traditional IVR based telephone conversations.

When the idea of introducing WhatsApp CEP has come up for Board's discussion, the Board suggested not to rush in with the new program with all the customers. One of the Board members was pointing out that customer engagement is needed but the conversion rate is the crux in today's world. Further, the Board member was insisting on deciding on few metrics to measure the effectiveness of customer engagement programs.

To this particular suggestion, Rocky stated that during the WhatsApp message promotion process a promotion code would be given to the customer and they could use this code to purchase products. The number of purchases made using this code is considered as conversion rate and the higher the usage of the code the more effective is the CEP. The Board has advised using the concept of Control Group (CG) and Test Group (TG) for testing the effectiveness of the CEP. One of the Board members has pointed out to Mr Rocky that using test and control groups to understand the effectiveness of any new initiative is a routine practice in marketing research.^{vi}

Initially Mr. Rocky was tentative but on the recommendation from the senior Board members, he agreed to the concept of control and test groups. To conduct the experimental design Rocky has also agreed to categorise the total customers of all six branches into six categories based on the duration of their relationship with ABPL (Table 1). Further, Rocky requested that if the new CEP showed good results his team have to be incentivized and the Board agreed.^{vii} The new program was expected to be run for a period of three months from January 2021 to March 2021 and the results were expected to be analysed in the first week of April 2021. The board felt if the new customer engagement program was 'effective' they would use the same program for the upcoming festive season as well as in the new stores they are proposing to open in the State of Karnataka.

Table 1 : Categorization of Customers

Category	Criteria
A	Customer of ABPL for 10 – 12 Yrs.
B	Customer of ABPL for 8 - 10 Yrs.
C	Customer of ABPL for 6 – 8 Yrs.
D	Customer of ABPL for 4 – 6 Yrs.
E	Customer of ABPL for 2 – 4 Yrs.
F	Customer of ABPL for 0 – 2 Yrs.

Implementation of CEP and Related Issues:

Rocky and his team have divided the total customer database relating to the six stores into two parts in the ratio of 40:60. Test Group (TG) consisted of 40% of the customers and Control Group (CG) consisted of 60% of the customers.^{viii} One of the important observations made in this regard was that only the aggregate level split of 60:40 was made. The database maintained by Rocky showed that the total customers of all the six stores put together are 421,890 out of which 168,790 customers would be exposed to the new WhatsApp based CEP. The program took off from January 1st 2021 and was implemented for three months viz. March 31st 2021. The number of customers who represent both the groups' viz. CG and TG are tabulated (Table 2). At the aggregate level, 30,300 customers from TG have made new purchases and

have used the promotional code sent through WhatsApp. On the other hand, 41,650 customers of CG have made new purchases from ABPL stores. During analysis, Rocky has found that approx. 18.01% of 1,68,790 customers in TG has made a purchase decision based on the WhatsApp campaign and 16.46% of customers in CG has purchased with ABPL without any exposure to the campaign. The new sales made by ABPL with the WhatsApp campaign in TG stores is 1.55% more when compared to the CG. These observations of Rocky made him conclude that the new CEP initiative is successful and should be implemented in near future in existing stores as well as the new stores to be opened in Karnataka. The same results were tabulated and sent across as email to all the Board members along with a copy marked to Ms Dimple for processing of incentives as agreed by the Board.

Table 2 : Number of Customers in Each Bin of CG and TG

Table 2(a): Number of Customers in Each Category of Control Group (60%)						
A	B	C	D	E	F	Total
6200	111000	19000	61500	41300	14100	253100
Table 2(b): Number of Customers in Each Category of Test Group (40%)						
A	B	C	D	E	F	Total
18000	89000	16700	16000	19200	9890	168790

After going through the email, Board members have congratulated Rocky and his team with an assurance of incentive amount to be credited at the earliest. However, Dimple, on the other hand, neither congratulated nor agreed to the incentive claim of Rocky. Dimple feels that the data at aggregate may or may not be indicating the true picture and wants to have a look at the granular data. But Rocky feels that Dimple is somehow not happy with the progress of his team. This is where John was called in for a meeting in the Board room.

The Board members felt somewhere ABPL did not fully use the power of data analysis all these years and now having John on Board they want to use his expertise in this regard. One of the Board members has

also reminded the other that this was not the only discrete incident where there were problems in properly understanding the data and results. The Board member also suggested that somewhere statistical significance of the claims made by any of the functional team have to be tested.^{ix} Based on this many opined that to date only absolute figures or percentages were compared and decisions were made. No attempt was made to see whether the differences are statistically significant to base decisions on them.^x For this view, Rocky got upset and says that he and his team's work is being demeaned with some sort of number game and statistical myths. At the end of the meeting, John stated that he would resolve the issue by EOD of Monday provided Rocky shares with the granular data by that day's evening.

John's Task on Monday:

John is not only interested in resolving the present issue but he has a long-term perspective. He feels that standard formats of data collection, data categorization and data analysis has to be established across ABPL. John knows that ABPL is destined to grow in near future and if the data collection, analysis and interpretation are not streamlined there may be chances of misinformed decisions. John's thoughts are revolving around how to sensitize the Board, employees and team leaders of ABPL on the role and importance of statistical analysis and how best he can

train them on using a simple spreadsheet for analysing the data as per the pre-determined formats and statistical tests. John also wants to discuss with the core team the concepts such as level of confidence and level of significance etc. such that the data analysis would be acceptable by all. John feels he got a good amount of work on the next day morning such that he has the presentation ready for Monday. While he was deeply involved in these thoughts, he heard a beep sound on his mobile and the Gmail notification showed that it was an email from Rocky and the subject of the mail was: "Segmented data for further analysis – Regards Rocky" (Table 3).

Table 3 : Number of Purchases Made in Each Bin of CG and TG

Table 2(a): Number of Customers in Each Category of Control Group						
A	B	C	D	E	F	Total
1150	22000	10000	3000	3000	2500	41650
Table 2(b): Number of Customers in Each Category of Test Group						
A	B	C	D	E	F	Total
2500	16100	8500	500	1200	1600	30400

REFERENCES

- Carlson, B. W. (2019). Simpson's Paradox. Encyclopedia Britannica. Retrieved March 12, 2021, from <https://www.britannica.com/topic/Simpsons-paradox>
- Import.io. (2019, October 22). What is Data Aggregation? Examples of Data Aggregation by Industry. Retrieved March 15, 2021, from www.import.io
- Mayo, M. (2019). Five Statistical Traps Data Scientists Should Avoid. Retrieved from KDNuggets: www.kdnuggets.co,
- SEP. (2021, March 24). Simpson's Paradox. Retrieved March 26, 2021, from Stanford Encyclopedia of Philosophy: www.https://plato.stanford.edu/entries/paradox-simpson/#Conc
- Tenny, S., & Abdelgawad, I. (2020, July 10). Statistical Significance. Retrieved March 18, 2021, from NCBI Resources: www.https://www.ncbi.nlm.nih.gov/

books/NBK459346/

- Villegas, A. (2018, September 10). Control Groups: An Effective Way to Measure Campaign Success. Retrieved March 10, 2021, from UpSellIt: blog.upsellit.com
- Wright, B. (2012). Best of N Contexts: Implications of Simpson's Paradox in Tennis. Florida State University Libraries.

-
- i The firm name (company or business) and the characters are fictitious in nature and are created to make the participants more involved in the case. However, the context and the decision situations highlighted in the case are from real-world observations that the author has encountered while providing business consultancy.
 - ii Aggregate data is a form of summary data.

- Business firms often collect data from online customers and visitors to their website. In this context, the aggregate data could be in the form of customer demographics and behaviour metrics, such as average age and number of transactions undertaken. Generally, aggregate data is used to present before the management teams (import.io, 2019).
- iii Quantitative fallacy deals with McNamara's (US Secretary of Defense during Vietnam War) obsession for metrics.
 - iv Simpson's Paradox demonstrates that trends that appear in a dataset when used as 'groups' could be reversed when the data are aggregated.
 - v WhatsApp Chatbot (also referred to as WhatsApp Bot) works as an automated chat system connect to the mobile number with WhatsApp. It is similar to Facebook Messenger Chatbot but works on WhatsApp. Chatbot on WhatsApp helps the business firms send contextual messages and makes the customer communications more secure due to end-to-end encryption.
 - vi In marketing research 'Control Group' is a term used to refer subjects/sample that receives ZERO exposure to the campaign or the experiment on hand. The most common control is a 50/50 split. In this kind of split half of the population is exposed to the new campaign or new initiative and the other half (control group) does not. There are other variants of control group formation methods such as 80/20 or 40/60 and the selection of split for control group depends upon the goal of the campaign. Control groups could be used to different research variables such as time of the day, source of traffic and date range (Villegas, 2018).
 - vii According to Mr Rocky, if the number or percentage of conversions in the test group is more when compared to the control group then the CEP could be considered a success and the team is eligible for incentives.
 - viii One of the most reliable and regularly used methods to assess the campaign's effectiveness is to split the audience/customers into two different groups and later to compare the results. The goal of these kinds of studies is to observe whether the campaign has any impact on the test group or not?
 - ix The importance of statistical significance arises due to the mere fact that the results have to be quantified and decide whether those results are due to chance or due to some factors/variables of interests. In case it is found that the results are statistically significant then the researcher can confidently say that the factors of interest have some impact on the phenomenon being researched.
 - x Generally in any marketing research study, the researcher starts with a hypothesis and processes the data collected to prove the hypothesis he has framed at the beginning of the study. In the present case, the claim of Rocky could be verified by starting with a hypothesis that "The CEP program is effective and would lead to an increase in the sales". To further work on this statement an inverse statement in the form of a Null Hypothesis could also be framed. The researcher always tries to disprove the null hypothesis and when trying to disprove the null hypothesis the researcher will not be 100% sure of the outcome. In that case, the researcher would settle for a lower confidence level for which they want to be sure of the outcomes. If the confidence level is 95% then it means that the researcher is willing to be wrong for 5% of the time about the results of the study. After performing different statistical tests the researcher would compare the p-value based on which the statistical significance of the results is decided. The p-value could be expressed as the probability that the null hypothesis is true based on the existing data collected and processed by the researcher. The study results are said to be statistically significant if the p-value of the data analysis is lower than the alpha (level of significance) (Tenny & Abdelgawad, 2020).

Effulgence: Bi-Annual Management Journal of Management

Indexed in:
Scilit
J-Gate
Crossref
Indian Citation Index
Index Copernicus International

Guidelines for Authors

Perspective

Effulgence is devoted to the publication of research papers/Book Reviews/Case Studies which advance knowledge through research and cases in functional areas of marketing, finance, human resource development, operations management, general management and also macro issues of globalization, regional trade and investment. The selection of papers for publication will be based on their relevance, clarity, and originality, the extent to which they advance knowledge, understanding, application and their likely contribution towards inspiring further development and research.

Types of Contribution

Research Paper/Case Study

The Journal attempts to maintain a balance between purely research-oriented papers and those derived mainly from the experiences of practitioners involved in the different areas of management. In case of research-based articles, papers that present the results of original research analysis are accepted for publication in the Journal. The managerial/policy implications of the study should be highlighted towards the end of the paper.

Book Review/Scan

This section covers abstract of books on management especially in the areas of skill building, managerial effectiveness, and macro-economic issues in development. The summary should not normally exceed 2000 words.

Abstract

The manuscript should be accompanied by an abstract (not more than 150 words) of the article and introduction of authors (name's, designation/s, affiliation/s) on separate sheets.

Text

The Research Paper/Case Study/Book Review should not normally exceed 6000 words. The articles should be in clear and concise English and should be typed on A4 paper in double-line spacing with wide margins. Papers submitted for consideration in Effulgence should be accompanied by a declaration by the authors that they have not been published or submitted for publication elsewhere. Editorial decision regarding articles accepted for processing in Effulgence will be communicated as per schedule in call for papers. We follow blind-refereed review.

Copyright

The manuscript should not appear in another publication.

This will ensure copyright protection for both the authors and Rukmini Devi Institute of Advanced Studies. Note that it is the author's responsibility to obtain permission to reprint long quotations or use tables, figures or graphs previously published with copyright restrictions.

General Outlines of Papers

Your paper should contain the following items in the order shown below:

1. Title, author's affiliation and full address.
2. Abstract (not more than 150 words) outlining the purpose, scope and conclusions of the paper. The abstract should explain why the readers should consider these results important.
3. The text, suitably divided under headings and sub-headings to elucidate the text and to emphasize points.
4. Acknowledgements, if any.
5. References.
6. Tables: Tables (each on a separate sheet) should be numbered consecutively in Roman numerals and should be referred to in the text as Table I, Table II etc. Tables should not duplicate results in graphs.