DATA: A POWERFUL WEAPON THAT CAN BACKFIRE

ANDREOZZI, CAMILLA
Statistics, College of Science and Engineering

ABSTRACT

In recent years, rapid technological advancements have led to an unprecedented surge in information availability. While data and analytics offer powerful tools for decision-making, improper handling can result in misinterpretations and biases, turning potential profits into losses. The shortage of data skills along with overconfidence in the accuracy of numbers, increases the risk of errors as quickly as the innovation they stem from.

This article addresses the problem by examining the impact of sampling bias on a mock toy company, 'Giocattolo'. The firm relies on a Data Provider for retail sales across Europe, yet several flaws (such as missing prices due to privacy reasons, incompatible product categorisations, and sampling bias due to limited coverage) go unnoticed by untrained personnel. A model simulation on raw data showcases how the unrepresentative retail sample, that only included high-revenue stores, results in an overoptimistic total revenue and the consequent six-figure net error. In fact, the naive estimation methods can inflate revenue projections by as much as 30%, resulting in significant misinterpretation. Uncertainty bands are proposed as a tool to convey the uncertainty inherent in biased estimates, offering a more informed basis for decision-making.

In an era where data is abundant, but expertise is scarce, the true advantage lies not in the information itself, but in the ability to interpret it correctly. Learning from a real-life example, this article highlights common oversights in data analysis and the necessity for proper statistical interpretation when research is the ambition.

INTRODUCTION

Sampling, the process of selecting a subset of individuals from a larger population, is the most crucial aspect of every research study. Since studying an entire population is often unfeasible due to time and money constraints, researchers exploit a sample to draw generalisable conclusions. However, if this sample is not representative of the population, the study's conjunctures will be corrupted by biases and therefore lack validity (Simundić, 2013). Research on bias and uncertainty, particularly through the lens of human decision-making and probability assessment, has explored how the mind processes likelihoods and outcomes when faced with incomplete or skewed data (Kahneman et al., 1982). Likewise, statisticians have been exploring the effects of a non-representative sample on the quality of the research frameworks. Such biases, stemming from cognitive predispositions or sampling errors, can deeply affect the interpretation and applicability of findings, underscoring the importance of rigorous statistical strategies to mitigate these influences and enhance reliability.

Nonetheless, when it comes to data, it's easier to overlook the uncertainty of the sampling process it is generated from. Data is the plural of 'datum', a Latin word coming from the verb 'dare' that means 'given': 'something that is certain to be'. In fact, the attractive aspect of data is that (without any other information) it is certain. However, like people, data is heterogeneous and often biased due to the partial resolution of the sample that it stems from (Baeza-Yates & Murgai, 2023). Consequently, the root of the problem lies in carefully doubting the quality of the data collection and investigating the steadiness of its assumptions (Cao et al., 2022). This becomes particularly relevant in a business context where the statistical knowledge required for this scrutiny is neither utilised nor actively sought. This disregard is due to limited resources, lack of expertise, or a misconception that statistical methods are too complex or unnecessary. Previous studies have shown how, while largely advocated by academic literature, the use of proper tools such as probability sampling, a method ensuring every individual or data point has an equal chance of selection, is quite scarce in marketing investigation (Sarstedt et al., 2018). Flawed research omits details about their sampling methods or relies on non-probability sampling without implementing adjustments to account for unequal selection probabilities, incomplete coverage, or sampling variations (Sarstedt et al., 2018). This article focuses on the abilities of a hypothetical toy firm to correctly extract valuable information from digital data on market sales in different countries, bought from a Data Provider, and the implicit and explicit obstacles the company may find.

© by Camilla Andreozzi. This article is licensed under a Creative Commons Attribution 4.0 International License. You should have received a copy of the license along with this work. If not, see https://creativecommons.org/licenses/by/4.0/

METHODOLOGY

First, this section provides a clear explanation of the business scenario and the assumptions behind the model argued in this article, outlining how sampling challenges can arise in a realistic, practical and market-driven environment. The firm in question will be named 'Giocattolo', abstracting from any real-world company, while still evoking a realistic business context. Next, the core assumptions guiding the model are outlined. These assumptions are essential for understanding the conditions generating sampling bias and its impact on the accuracy of revenue estimations. The analysis involves running a model using R (a statistical computing tool and programming language widely used for data analysis) that replicates the firm's estimation process in different plausible scenarios. The simulation reveals that the firm's revenue estimation is frequently inaccurate due to unrepresentative sampling and highlights the scenarios in which estimation errors are most pronounced. Lastly, uncertainty bands, numerical ranges within which the true value of a parameter is expected to fall, are proposed as a statistical tool to address sampling bias. While uncertainty bands do not eliminate sampling bias, they can reveal the extent of imprecision in the estimations and provide a more transparent basis for interpreting results. This methodology aims to assess and illustrate the consequences of sampling bias in market research, especially when it relies on digital data, encouraging readers to focus on quality estimation and to view data analysis as a skill that requires ongoing training and scrutiny.

EMPIRICAL CONTEXT: DATA AND ITS PROVISION IN THE MARKETING WORLD

Giocattolo's revenue is directly tied to the sales of its products, making it essential for the Production and Marketing departments to access accurate market data. In recent decades, technology and digital services have grown exponentially (Cassard & Hamel, 2018), while the number of experts remains relatively small compared to the vast user base. Just as ancient societies once attributed natural phenomena they couldn't explain to the actions of gods (van der Sluijs, 2009), modern business employees attribute exaggerated powers to technologies they don't fully understand, believing these tools will deliver transformative results simply by their presence. It is probably upon this temple of faith in the latest technology that that Data Selling businesses have emerged. Firms such as 'Acxiom', 'Plaid' or 'Quantcast' are the so-called 'data-as-a-service' (DaaS) companies that, by being paid a consistent sum, provide access to data on-demand. Getting hold of the data can be as easy as a transaction but building something meaningful from it is not so straightforward. First, the tables contain high numbers of different types of variables (with substantial increase if historical comparisons are present), making it hard not only to store but also to visualise and manipulate. Secondly, the quality of such data depends on many different factors like the presence of faulty sensors, possible mis-recordings and the reliability of the methods. Lastly, but most importantly, privacy issues emerge. One of the primary concerns for DaaS companies is ensuring that data has been collected with the explicit consent of individuals, especially if it includes personal information. This complication is usually mitigated by anonymising the observations but complicates Giocattolo's efforts to recognise specific customer behaviours. Moreover, while in Europe strict guidelines are provided by the General Data Protection Regulation (GDPR, European Parliament and Council, 2016), legislation may vary from country to country, making Giocattolo's expansionists research and international data challenged by ulterior complexity. Effective data utilisation requires more than technical expertise in running statistical models; it demands professionals with a deep understanding of the business context, the ability to communicate effectively with data scientists, and the skill to translate analytical insights into actionable strategic decisions.

SAMPLING BIAS FROM POOR DATA COLLECTION

The database bought by Giocattolo comprises data on the name of the toy sold, the manufacturer, several classification variables (e.g., puzzle, building blocks), the channel it was sold through (e.g., supermarket, toy specialist), target age, a range of prices it was plausibly sold for and many other data points collected on a 4-week basis. The first issue is that, because of privacy protections, true prices are often omitted from the data to avoid revealing sensitive financial information about small businesses, making it harder to accurately gauge market positioning and pricing trends (Sadler, 2020). This protection extends not only to small-scale manufacturers but also to smaller sales channels, such as local toy shops, which may be excluded altogether. Secondly, the classification of toys, essential for market segmentation and trend analysis, is determined by the Data Provider rather than Giocattolo, leading to discrepancies in categorisation that complicate the alignment between Giocattolo's internal analysis and the external database they rely on. Lastly, the key problem is the country coverage. Not only privacy issues vary from state to state, but the DaaS is only able to obtain information on a non-complete set of shops and sources.

Consequently, the database Giocattolo believes contains sales from all possible shops in Country X, is only from a flawed and skewed sample of them. Once the word sample enters the conversation, statistics and probability follow as bearers of bad news. In an ideal scenario, a dataset should capture every single shop, product, and customer interaction within that scope. However, data is the consequence of a sample (Sudman et al., 1988) and sampling approaches bring statistical risks. This is particularly relevant here, as Giocattolo's Data Provider relies on what's known as a 'convenience' sample (Steelman et al.,

2014, Benfield & Szlemko, 2006). A probabilistic random sample excludes instances on a random basis, this means that no type of information is excluded on a systematic basis (Cochran, 1977). Vice versa, a convenience sample consists of data that is easily accessible to the Provider (MBJ, 2013). The Data Provider might pull data from larger retail chains or specific regions where it has established partnerships and from whom it's easy to get information, while smaller shops or rural locations could be underrepresented. This approach is faster and cheaper, but it introduces substantial bias: the database disproportionately represents larger or urban-based stores, while failing to capture trends in smaller retail environments (Levy & Lemeshow, 1999). Giocattolo's employees are not able to recover this. Instead, they are provided with coverage percentages (e.g., 52% of Italy, 64% of France) and simply adjust their estimations. For example, to compute the total revenue of plastic dinosaurs in Spain, they sum the individual quantities multiplied by the corresponding prices and then divide by the coverage percentage; the statistical approach of 'expanding the mean'. In a bias-exempt framework, this would be perfectly fine. However, this sample is flawed by an over-representation of the large-scale channels. As a result, whatever estimate Giocattolo will produce, will be systematically too large and the data analysis will rest on a shaky foundation.

MODEL SIMULATION IN DIFFERENT MARKET SHARE AND SAMPLE PROPORTION SCENARIOS

The data for this model is entirely artificial to avoid privacy concerns, ensuring that no real company or consumer data is disclosed. The quantities do not match real-world values exactly; instead, the focus is on proportions and error ratios, making magnitudes of revenue or units sold irrelevant to our analysis. This method relies on assumptions; crucial to the computation themselves but also to the understanding of the results. We will fix a certain value as the true value for total revenue in the toy market for a hypothetical 'Country X'. Additionally, the Data Provider's coverage of this country will be set at a specific level of 50%. These values are fictional and fixed as inputs for the simulation. The model itself stands on the computation of what we will call the 'naive estimate.' This is the simple calculation that the firm might use, in which the weighted sum of prices is divided by the given coverage percentage to extrapolate an estimate for 100% market coverage in Country X.

Our main assumption is that the 50% covered by the Data Provided is not a heterogeneous probabilistic sample but rather a homogeneous convenience sample, with a prevalence for large stores over small ones. To quantify the influence of large versus small channels, we introduce the concept of 'market share by channel', which reflects the proportion of total sales attributed to different sales channels. Typically, large channels account for a substantial market share, with their average revenues higher than those of small shops (Investors.com, 2024). Thus, ranging the true market share of large retailers in our simulation (from 60% to 95%), it will affect the extent to which revenue estimates are inflated. Another key parameter that will allow us to capture different market dynamics is the degree of representation of large versus small channels in the Data Provider's sample. This proportion, ranging from 60% to 100%, reflects the extent to which big retailers dominate the sample, and the assumptions stated before. Table 1 summarises the variables used.

Parameter	Description	Range	Assumption
True Total	The total sales value of all toy products in Country X.	Fixed value	Represents the actual market value.
Coverage	The proportion of retailers reached by the DaaS.	Fixed value	The sampling scheme within this coverage is not a probabilistic random sample.
Market share of large channels	Proportion of total market sales captured by large retailers.	50% to 95%	Assumes large retailers control a significant share of the market, with variations to test different scenarios.
Large channel representation in sample	Proportion of the market sample that is made up of large channels.	50% to 100%	Assumes that the sample has a disproportionate representation of large channels compared to smaller ones.
Naive Estimate	Estimate based on the available sample data without adjusting for sampling bias.	Computed via formula	Naive estimate is calculated by dividing the sum of quantities and prices by the coverage.

True Estimate	True estimate of total sales, adjusted for the real market share of large channels.	Computed	True estimate is derived by multiplying the market share by channel with the true total.

Table 1: showing the variables used in the model above

RESULTS

The model simulation reveals that when the sample is disproportionately represented by large retail channels, the naive estimation of total market revenue tends to overestimate the true total. This overestimation is particularly pronounced when the market share of large channels is high (ranging from 60% to 95%), which can be seen in Figure 1, and when the sample closely mirrors the actual market share (up to 100%). In scenarios where large channels make up a significant portion of the sample (around 80-100%), the naive estimate can overestimate the total by as much as 30% or more.

The magnitude of the mistake decreases as the market share of large channels decreases or as the sample includes more smaller channels (see Figure 2). The simulation highlights how such overestimations arise from sampling bias, with the proportion of error being directly linked to the extent to which large retailers dominate the sample. This suggests that, under the model assumptions, when there is uncertainty in the real composition of the sample, misestimation is highly probable.

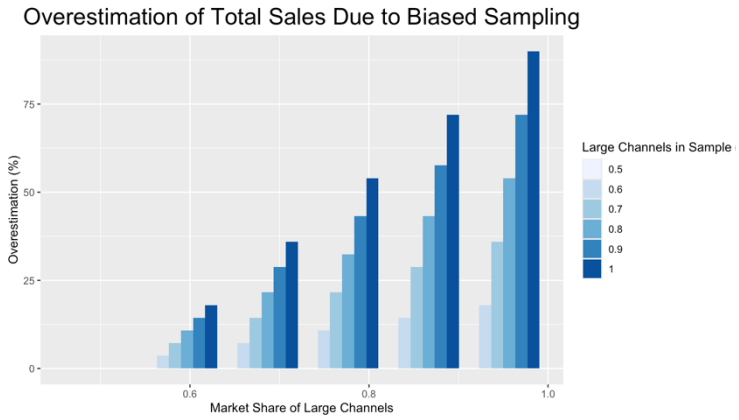

Figure 1: the Bar Plot shows the percentage of overestimation in the naive estimation applied by the firm, for different values of Market Share of large channels (x-axis) and for different representations of large channels in the sample (colour).

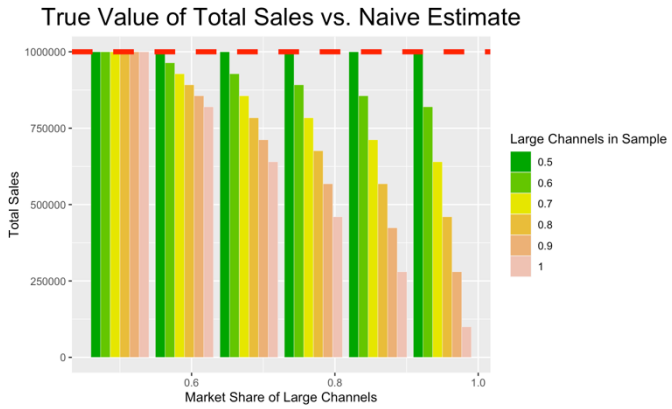

Figure 2: the red horizontal dotted line represents the naive estimation, the vertical bars the true values for the specific scenario combining plausible Market Share and Proportion of Large Channels in samples.

DISCUSSION

The overestimation issue simulated in this study exemplifies the dangers of uncritical reliance on data in digital business environments. Country comparison, essential for international marketing research, becomes unreliable when biased estimates distort the decision-making process. If the naive approach suggested that Country X's market is significantly larger than Country Y's, Giocattolo could overinvest in marketing and inventory for Country X, while underinvesting in Country Y where actual growth potential might be greater. This misallocation of resources could ultimately harm the company's profitability and become particularly problematic in industries like toy manufacturing, where trends are seasonal, and market demands are dynamic.

Statistical laws such as the Central Limit Theorem (Kwak & Kim, 2017) suggest that as the number of observations increases, biases tend to diminish, especially if sampling is random (Cochran, 1977). However, for the non-probabilistic sampling in this case, these laws do not apply. In such contexts, uncertainty bands and confidence intervals offer a valuable tool to quantify variability and provide a range within which the true parameter is likely to fall. While the computation of uncertainty bands is not straightforward, Figure 3 illustrates the formula for constructing a confidence interval for a population mean, where the sample mean ($\bar{x}$) is adjusted by adding and subtracting the margin of error. The margin of error accounts for the variability in the sample (represented by the standard deviation 's') and the size of the sample (n), scaled by the z-score for the confidence level. This range estimates where the true population mean is likely to lie. While uncertainty bands do not correct sampling bias, they highlight the unreliability of estimates, enabling firms to recognise uncertainty and apply better-informed decision-making. Further studies could consider the underlying distribution to use for this uncertainty bands and how to practically implement them in a marketing research framework.

$$\overline{x} \pm z\left(\frac{s}{\sqrt{n}}\right)$$

Figure 3: the formula for the confidence interval of a population mean

CONCLUSION

In this study, a structured methodology was employed to undermine this certainty and explore the implications of sampling bias in market estimation processes. While the availability of data has escalated with technological advancements, interpreting and applying such data effectively remains a significant challenge, especially in marketing. Giocattolo obtains vast datasets of variables such as toy classifications, sales channels, and plausible price ranges. However, extracting insights is challenging due to data complexity, quality issues, and privacy regulations. The most critical issue lies in sampling bias. Data Providers often rely on convenience samples, prioritising data from large retailers while underrepresenting smaller shops. Giocattolo calculates market revenue by expanding totals based on a given coverage percentage. While this technique works under unbiased sampling, it fails when the sample over-represents large-scale channels thereby enlarging revenue estimates. The results of the model simulation demonstrate how this overrepresentation in samples can systematically distort revenue projections. After fixing parameters like total market revenue and sample coverage, it is shown that variable factors like the true market share of large retailers and the proportion of their representation in the sample determine the extent of bias. Confidence interval, although unable to correct biases, provide awareness of the unreliability of estimates based on incomplete or skewed data. Limitations for this study lay in the assumptions behind it. While uncommon, scenarios in which privacy does not interfere with sampling and market share is not disproportionate between large and small channels are plausible. In these cases, the bias could be mitigated, if not removed.

In this evolving digital landscape, data is perceived as a goldmine for decision-making, yet the overlooking of common complications like sampling bias turns this valuable resource into a potential liability. This highlights the paradox of accessibility in the digital world: while technology enables acuity, it also amplifies the risk of misuse by individuals lacking statistical training. The broader takeaway is the vital need for organisations to inquisitively look at the numbers they are provided with and investigate possible flaws in the data collection. This argument was easily illustrated by building a simple model applicable to different scenarios and showcasing the mistakes miscalculation can make. For Giocattolo and others, recognising the uncertainty means demolishing the idea of data as an infallible tool, acknowledging the skewed estimation and its damage, and then acting upon it.

REFERENCES

Baeza-Yates, R., Murgai, L. (2024). Bias and the Web. In: Werthner, H., *et al. Introduction to Digital Humanism*. Springer, Cham. Benfield, J. & Szlemko, W. (2006) 'Internet-based data collection: Promises and realities', *Journal of Research Practice*, **2**(2), pp.1-15. Bureau of Labor Statistics (2024) Number of employees in the United States' information sector from 2010 to 2023, by quarter (in 1,000s). *Statista*. [Online] [Accessed: 7 November 2024]. Available at: https://www.statista.com/statistics/

BMJ (2013) 'Convenience sampling', *BMJ*, 347.

Cao, R., Koning, R. & Nanda, R. (2024) 'Sampling Bias in Entrepreneurial Experiments', *Management Science*, **70**(10), pp.7283–7307.

Cassard, A. & Hamel, J. (2018) 'Exponential growth of technology and the impact on economic jobs and teachings: Change by assimilation', *Journal of Applied Business and Economics*, 20(2).

Cochran, W.G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Deloitte (2023) Leading retailers worldwide in 2021, by retail revenue (in billion U.S. dollars). *Statista*. [Online] [Accessed: 15 November 2024]. Available at: https://www.statista.com/

European Parliament and Council (2016) 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)', *Official Journal of the European Union*, L119, pp.1–88.

Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Kwak, S.G. & Kim, J.H. (2017) 'Central limit theorem: the cornerstone of modern statistics', *Korean Journal of Anesthesiology*, 70(2), pp. 144–156. Levy, P.S. & Lemeshow, S. (1999) *Sampling of populations: Methods and applications*, 3rd edn. New York: Wiley.

Lumley, T. (2004) 'Analysis of complex survey samples', *Journal of Statistical Software*, **9**(1), pp.1-19.

Lynch, P. (1982) 'Sampling strategies in research: considerations for external validity', *Journal of Research Methods*, **14**(2), pp.101-115.

Investors.com, 2024. *Walmart stock, Costco roar as they create Amazon-like businesses*. [Online] [Accessed: 15 November 2024] Available at: https://www.investors.com/news/walmart-stock-ecommerce-online-retail/?utm.com

Reynolds, N.L., Simintiras, A.C. & Diamantopoulos, A. (2003) 'Theoretical justification of sampling choices in international marketing research: Key issues and guidelines for researchers', *Journal of International Business Studies*, **34**(1), pp.80-89.

Sadler, C. (2020) *Protecting privacy in data releases: A primer on disclosure limitation.* New America. [Online] [Accessed: 15 November 2024] Available at: https://www.newamerica.org/oti/reports/primer-disclosure-limitation/

Sarstedt, M., Bengart, P., Shaltoni, A.M. & Lehmann, S. (2018) 'The use of sampling methods in advertising research: a gap between theory and practice', *International Journal of Advertising*, **37**(4), pp. 650-663.

Simundić, AM. (2013) Bias in research. *Biochem Med* (Zagreb), **23**(1): pp 12-5.

Steelman, Z.R., Hammer, B.I. & Limayem, M. (2014) 'Data collection in the digital age: Innovative alternatives to student samples', *MIS Quarterly*, **38**(2), pp. 355-378. Suchman, E. A. (1962). An Analysis of “Bias” in Survey Research. *The Public Opinion Quarterly*, **26**(1), pp.102–111.

Sudman, S., Sirken, M.G. & Cowan, C.D. (1988) 'Sampling rare and elusive populations', *Science*, **240**(4855), pp. 991-995.

van der Sluijs, M.A. (2009) 'Myth and geology', *Myth & Symbol*, **5**(2), pp. 58-74. doi: 10.1080/10223820902723254