PharmaSUG 2025 – Paper SI-046

The Current State of Teaching Biostatistics in Academia: Challenges and Software Solutions

Lida Gharibvand, Ph.D., Loma Linda University

ABSTRACT

The traditional approaches to teaching and learning biostatistics have gone through evolutionary stages. This is mainly due to the advancement of computing and modern approaches of data analytics, including but not limited to the methodological impacts of statistical machine learning. However, the growth and impact of biostatistics programs, both at the undergraduate and graduate levels, heavily depend on building strong and relevant foundations. It is from that perspective that it becomes necessary to further evaluate the interaction of the main areas of mathematics, statistics, and computing. In this paper, we discuss the relevance and importance of incorporating modern approaches to data analytics, specifically in conjunction with biostatistics training and research. Chiefly, we aim to discuss the importance of computing and statistical thinking, as the building blocks of ideas associated with tackling projects that have real-life implications. We will focus specifically on the design of experiments and surveys, data collection, data visualization, model interpretation, and data-driven decision-making. We review these concepts from structural, logistical, and budgetary points of view. This paper explores these challenges and discusses the emerging role of software in improving biostatistics education.

INTRODUCTION

Biostatistics education has undergone notable transformation because of the recent advances in data science and computing. However, its effectiveness remains rooted in a strong multidisciplinary foundation in statistics, mathematics, and programming. This paper highlights how integrating modern tools enhances the teaching of core biostatistics concepts like study design, data visualization, statistical modeling, and data-driven decision-making, particularly through cost-conscious and practical methods.

METHODOLOGY

1. Challenges in Teaching Biostatistics

Several challenges exist in the current teaching of biostatistics:

1.1. Outdated Curricula One of the most significant issues is that many academic institutions continue to teach biostatistics using outdated curricula that focus heavily on classical statistical methods, such as hypothesis testing, without incorporating newer, more advanced methodologies including up to date software solutions. With the emergence of fields like bioinformatics, machine learning, and data science, there is a growing need to incorporate computational methods that allow students to handle large datasets and perform more complex analyses (O'Leary et al., 2020). Academic institutions must enhance their biostatistics programs by integrating modern statistical methodologies, appropriate study designs, and practical applications.

1.2. Lack of Emphasis on Practical Skills Many biostatistics courses place a disproportionate emphasis on teaching the theory but neglect the practical skills needed to analyze the vast amount of real-world data. Students often struggle to translate their theoretical knowledge into practice, especially when working with large datasets or real-life case studies (Bickel & Brown, 2018). The lack of practical hands-on experience using statistical software is another barrier to their ability to apply their knowledge effectively.

1.3. Insufficient Training for Non-Statisticians Many students in health sciences, biology, and medicine need a working knowledge of biostatistics but do not have the academic background and the time to master the mathematical foundations. Programs often fail to cater to non-statisticians, offering a one-size-fits-all approach that leaves many students underprepared for the statistical demands of their respective fields (Rodriguez, 2021).

1.4. Shortage of Skilled Instructors The shortage of qualified biostatistics instructors further

exacerbates these challenges. As the demand for biostatistics education grows, so does the need for educators who can teach both traditional and modern methods, as well as mentor students in applying these methods to real-world data to generate useful data-based insights.

2. Importance of Biostatistics in Academia

Biostatistics is integral to understanding complex biological data and is used extensively in epidemiology, clinical trials, genetics, and various other fields. With the rise of precision medicine, machine learning, and big data, the role of biostatistics is expanding significantly. Therefore, academic programs must prepare students not only for theoretical knowledge but also for the practical applications of biostatistics in real-world settings.

According to a study by Begg et al. (2019), there is a growing recognition of the importance of biostatistics in graduate and undergraduate health sciences curricula, yet many institutions lag in updating their teaching methods to reflect modern advancements in the field.

3. Importance of Appropriate Study Design

Choosing an appropriate study design—whether observational (cohort, case-control), experimental (randomized controlled trials), or qualitative (interviews) is vital. Incorrect designs can lead to biased or invalid results. It is worthwhile placing special emphasis on the development of student skills to make the right first step in selecting the correct study design in various scenarios involving specific data analysis objectives.

4. Challenges in Sample Size Calculation

Determining an appropriate sample size is essential for the validity and reliability of research findings. However, the process involves several interrelated challenges that can compromise study outcomes if not properly addressed.

4.1 Uncertain and Inaccurate Assumptions

Sample size estimation relies on input parameters such as effect size, standard deviation, event rates, and dropout rates. In many cases, these parameters are based on limited or outdated information, making them prone to error. Underestimating variability or overestimating effect size can lead to an underpowered study, while overly conservative estimates may result in excessive sample sizes, increased costs and ethical concerns.

4.2 Complex Study Designs and Adjustments

More sophisticated study designs—such as cluster-randomized trials, repeated-measure studies, or those with multiple endpoints—require additional considerations in sample size calculations. Intra-class correlations, design effects, and statistical corrections for multiple comparisons (e.g., Bonferroni adjustment) can significantly inflate the required sample size.

4.3 Participant Dropout, Non-Compliance, and Resource Constraints

Attrition and non-adherence are common in real-world studies and must be factored into sample size planning. However, estimating dropout rates is often imprecise. Furthermore, ethical and practical limitations, including limited budgets, time constraints, and the burden on participants—can restrict the feasible sample size, potentially compromising study power and generalizability.

4.4 Software Limitations and Regulatory Demands

While various tools are available to assist with sample size estimation, they require proper statistical knowledge to be used effectively. Misuse or misinterpretation of software output can lead to incorrect planning and data analysis. Additionally, regulatory agencies may impose specific power and sample size requirements that may not always align with the institution's resources or recruitment capacity.

5. Dealing with Missing Data

Missing data can substantially affect the analysis and interpretation of large epidemiological and clinical datasets. Ignoring incomplete cases may lead to bias and loss of statistical power (Little & Rubin, 2002). Researchers usually address missing data by only including complete cases in the analysis, those individuals who have no missing data in any of the variables required for that analysis. However, results of such analyses can be biased. Furthermore, the cumulative effect of missing data in several variables often leads to exclusion of a substantial proportion of the original sample, which in turn causes a substantial loss of precision and power. Here are the classification of missing variable:

5.1 Missing completely at random (MCAR)

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing, apart from the obvious loss of information. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck.

5.2 Missing at random (MAR)

If the probability of being missed is the same only within groups defined by the observed data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and we assume MCAR with the type of surface, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability to be included depends on some known property. MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption.

5.3 Missing not at random (MNAR)

Finally, data are said to be missing not at random if the value of the unobserved variable itself predicts missingness. A classic example of this is income. Individuals with very high incomes are more likely to decline to answer questions about their income than individuals with more moderate incomes.

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Patterns of data loss are typically described as either ignorable or non-ignorable (Kuligowski& Gharibvand, 2020).

6. Data Collection and Visualization

Proper data collection is critical for generating reliable data-driven insights. Statistical thinking enables the creation of data collection protocols that ensure accuracy and reduce errors, such as bias or confounding. Data quality is a key consideration here, and computing helps streamline this process by automating data entry, minimizing human error, and structuring large datasets for ease of analysis.

The ability to visualize data effectively is a key outcome of strong statistical thinking and computing skills. Data visualization helps translate complex datasets into understandable formats that facilitate insight generation and communication. Statistical thinking guides the selection of appropriate visual representations—bar

charts, histograms, scatterplots, or more advanced techniques such as heatmaps and network graphs—based on the underlying data structure.

7. Role of Software Tools in Biostatistics Education

7.1 Open-Source Software (R, Python)

Advantages

- Free and Accessible: Ideal for budget-conscious institutions.
- Customizable and Flexible: Extensive libraries for various statistical needs (Wickham, 2016; Chambers, 2008).
- Industry Trend: Increasingly used in data science applications (Van Rossum & Drake, 2009).

Disadvantages

- Steep Learning Curve: Coding skills are essential, potentially overwhelming beginners.
- Lack of Standardization: Community-developed packages may lack consistency.

7.2 Proprietary Software: SAS

Advantages

- Industry Standard: Widely used in pharma, FDA-regulated trials, and clinical research (SAS Institute, 2021).
- User-Friendly: GUI, detailed documentation, and availability of training resources ease the learning process.
- Regulatory Compliance: Trusted in healthcare due to significant reliability and validation protocols.

Disadvantages

- High Cost: Licensing fees are a major barrier (SAS Institute, 2024).
- Limited Flexibility: Less adaptable to newer data science techniques.

CONCLUSION

A hybrid approach that exposes students to a variety of tools is key to developing comprehensive skills and meeting the growing analytical demands in academia and industry. Educators should design curricula that seamlessly integrate theoretical rigor with applied practice, preparing students to confront real-world challenges with advanced statistical thinking and modern techniques. To further enhance training, academic programs must be modernized to include machine learning, study design, and data science, along with hands-on software training. Investing in faculty development on emerging data analysis methods and fostering industry collaborations will also be critical. Additionally, dedicated training on handling missing data and sample size calculation is essential for equipping students with the capabilities required for rigorous data analysis and effective study planning.

REFERENCES

O'Leary, N., Ryan, A. M., & O'Neill, A. (2020). Teaching and learning biostatistics: Current trends and future directions. *Biometrical Journal, 62*(5), 1212–1224.

Bickel, P. J., & Brown, L. D. (2018). The future of statistics in the era of big data. *Proceedings of the National Academy of Sciences, 115*(14), 4657–4660.

Rodríguez, R. N. (2021). Expanding statistical literacy and training for the biomedical workforce. *Statistics in Medicine, 40*(40), 6690–6702.

Begg, M. D., Vaughan, R. D., & Arevalo, M. (2019). Biostatistics as a discipline in a changing academic environment. *Statistics in Biopharmaceutical Research, 11*(2), 174–181.

Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. Wiley.

Kuligowski, A. T., & Gharibvand, L. (2020). Dealing with Missing Data in Epidemiological and Clinical Research.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

Chambers, J. M. (2008). Software for Data Analysis: Programming with R. Springer.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace.

SAS Institute Inc. (2021). *SAS® 9.4 SQL Procedure User's Guide*. Cary, NC.

SAS Institute Inc. (2024). *SAS/STAT User's Guide*. Cary, NC.