
UDC 372.857

J. Wei, bachelor's degree student, (*Kyiv College at Qilu University of Technology, Kyiv*)

I. Hretskyi, PhD in Biology, (*KNUTD, Kyiv*)

THE IMPORTANCE OF LEARNING BIOSTATISTICS WITH R FOR BIOLOGISTS

Biostatistics plays a crucial role in modern biology, providing the tools needed to analyze and interpret complex biological data. For biologists, understanding statistical methods is essential for designing experiments, analyzing results, and drawing valid conclusions. Whether studying population genetics, ecology, or molecular biology, biostatistics allows researchers to make sense of variability in biological systems and test hypotheses with scientific rigor. Biostatistics ensures that the conclusions drawn from biological research are reliable and reproducible, which is vital for advancing scientific knowledge. Furthermore, as the field of biology increasingly relies on big data and computational methods, biostatistics provides the foundation for bioinformatics and systems biology. Biologists equipped with statistical knowledge can better collaborate across disciplines, making them more competitive in the workforce.

One of the challenges in studying biostatistics is its practical application. Despite the importance of theory, biologists often find it difficult to effectively apply statistical methods to real biological data. Biological data tend to be complex, with many variables and sources of variability, which can make it difficult to choose the right statistical tests or to interpret results correctly. Moreover, working with large data sets, such as genomic or ecological studies, adds another layer of complexity.

One of the possible options to improve the study of biostatistics by biologists is the use of modern software products, which can simplify multivariate analysis when working with large data sets. In today's data-driven world, the ability to apply biostatistics using modern software has become essential for biologists. Advanced statistical software like R, Python, SPSS, and SAS significantly enhance the ability to process, analyze, and visualize complex biological data. These tools not only simplify the application of statistical methods but also allow biologists to handle large datasets with greater efficiency and accuracy.

The integration of biostatistics and programming has become increasingly important in biological research. Among the various programming tools available, R has emerged as a leading language for

statistical analysis in biology. Learning biostatistics using R equips biologists with the skills to handle large datasets, perform advanced statistical analyses, and visualize data effectively.

R and the Bioconductor are powerful open-source project and indispensable tools for both statisticians and biologists, given the escalating need for interdisciplinary research in bioinformatics as the life sciences continue to advance. Before Bioconductor, existing open-source bioinformatics initiatives faced challenges with transparency, reproducibility, and the development efficiency of their computing environments tailored for biostatistics. Bioconductor offers more than 2,000 software packages for handling complex biological datasets, including those generated by high-throughput technologies like RNA-sequencing, microarrays, and proteomics. Bioconductor stands out by effectively addressing a broad spectrum of biostatistical requirements—spanning data acquisition, management, transformation, and modeling—thanks to its integration with R's robust capabilities. Its robust infrastructure not only supports scalable, command-driven operations but also encourages community engagement and interaction. The widespread adoption of Bioconductor and R is poised to efficiently satisfy the escalating demand for bioinformatics methodologies. This is especially crucial for managing the complex, high-throughput molecular data produced by cutting-edge assay technologies.

It is important to acknowledge that R and Bioconductor have certain limitations that users should consider. One of the primary challenges is the steep learning curve. R, being a programming language, requires users to have a solid understanding of coding. Another limitation is performance. R is an interpreted language, which means it can be slower than compiled languages like C++ or Java, especially when handling very large datasets. Additionally, while Bioconductor offers a vast range of specialized packages, managing package dependencies and ensuring compatibility between them can be challenging. Bioconductor relies on frequent updates, and packages can sometimes break or become incompatible with newer versions of R. This creates potential issues for reproducibility, as different users may not be able to run the same analysis on different versions of R or Bioconductor without encountering errors.

As the biological sciences continue to rely on computational approaches, proficiency in R not only strengthens a biologist's statistical skills but also opens doors to interdisciplinary collaboration in fields such as bioinformatics, genomics, and epidemiology. Learning biostatistics with R thus ensures that biologists are well-prepared to tackle modern challenges in research and data analysis.
