Exploiting Simulation Features of R in Teaching Biostatistics

Zdeněk Valenta

European Center for Medical Informatics, Statistics and Epidemiology, Institute of Computer Science AS CR, Prague, Czech Republic

Conveying or illustrating some fundamental statistical principles may become a challenge when teaching students who do not possess a sufficient theoretical background. More often than not they do rely on a hands-on experience in understanding different theoretical concepts.

One of such concepts in statistics in general is, for example, that of confidence intervals. For example, the idea that the true mean of a normal distribution, a fixed and unknown constant μ, will rest within some interval with random endpoints with a pre-specified probability, say 95%, may to medical students at large represent an equally remote concept such as that of average relative frequency of covering the true population mean μ by such intervals constructed during the process of repeated random sampling from the target population.

Our experience is that explaining similar concepts to medical or other interdisciplinary students using R seems quite rewarding. Students can each perform their own simulations, discuss and compare the corresponding results in the class and observe that though not being identical the results are all consistent with the probabilistic statements they wondered about. It suddenly becomes a relevant learning experience to many of them.

In our classes we endeavoured to illustrate the concept of confidence intervals using a simulated birth weight data set that might actually represent the birth weights of children born in the Czech Republic. An example of the real birth weight data for children born in Boston City Hospital, which served as an inspiration to our simulated data sets, is discussed with regard to introducing the concept of confidence intervals e.g. in Rosner[1].

A sample R code is used to illustrate the meaning of confidence intervals. We wrote an R function counting the relative frequency of occurrences in which the hypothetical true population mean is actually being "covered" by the random confidence intervals during the process of repeated sampling. Examples are shown based on 1000 simulations and a sample size of 100 where the sample birth weights are drawn from a normal population with the mean of 3.2 kg and standard deviation 0.6 kg.

The examples document versatility of R environment which serves very well not only for the research purposes of biostatisticians, but also in teaching biostatistics to medical students and medical professionals, where the hands-on experience may be essential for grasping more difficult statistical concepts.

References

[1] Rosner B.: Fundamentals of Biostatistics, 4th Edition, Duxbury Press 1994, Wadsworth Publishing Company, 10 Davis Dr., Belmont, CA 94002, U.S.A., ISBN 0-531-20840-8