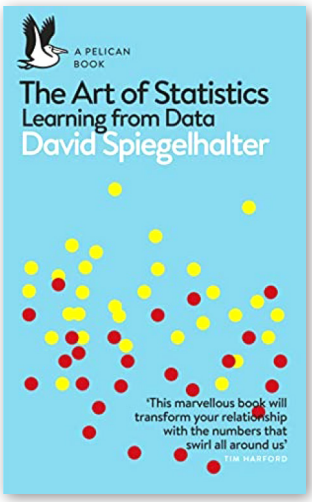
Journal of Multidisciplinary in Social Sciences
Journal homepage : http://jmss.dusit.ac.th

Book Review: The Art of Statistics Learning from Data

Nattapan Tantikul & Wanvitu Soranarak



Book Title: The Art of Statistics Learning from Data
Author: David Spiegelhalter
Publication Date: February 2020 Publisher: Pelican (UK)
ISBN: 9780241258767

"Statistics is the grammar of Science", a well-known quote by Karl Pearson who was a prominent British mathematician and statistician and was famously known for Pearson's r (Pearson correlation coefficient), Pearson distribution and Pearson's chi-squared test. Pearson emphasizes the importance of statistics which is the field of learning from data, and the important tool we use to convert raw data into reliable, relevant, and useful information. Due to the rapidly and tremendously increasing volume and complexity of data and information nowadays, the role of statistics in our lives becomes more crucial.

This book consists of fourteen chapters which provides knowledge on Statistics ranging from PPDAC (Problem, Plan, Data, Analysis and Conclusion) cycle, variables, sample and population distribution, central tendency, variability, correlation, deductive and inductive inference, internal and external validity, causation, statistical model and regression, classification and prediction, probability, relationship between probability and statistics, Poisson distribution, Binomial distribution and Normal distribution, Central Limit Theorem, estimates and intervals, hypothesis testing, statistical significance, Type I and Type II error, Bayes' theorem, likelihood ratio, reproducibility crisis, the problem with P-value to data ethics.

In sum, firstly, the author pointed out the importance of data science and data literacy which is the capability to read, apply, analyze, and communicate with data. Nowadays, in the big data era, it is hard to deny that the data literacy skill is one of the most essential skills, because it can empower everyone to build knowledge, make decisions, and communicate meanings to others. Besides, in tandem with the growing importance of data literacy, the way of teaching modern Statistics is

changing by moving its attention from mathematical theories and statistical techniques to problem-driven approach. PPDAC, a new data analysis cycle which is composed of problem, plan, data, analysis, and conclusion, is applied to teach students how to use data efficiently and morally to understand and solve real-world problems.

Secondly, the author described many types of variables such as binary, categorical, and continuous variables, how to choose appropriate descriptive statistics such as mean, median, mode, percentile, range, standard deviation, Pearson's correlation coefficient, and Spearman's rank correlation, and how to present descriptive results by using graph or infographic effectively. Although this part is relatively easy, it is good to make us aware of the inefficient usage of mean, the most frequently used measure of central tendency, when there are outliers or skewed distributions.

Thirdly, he explained the differences between deductive and inductive inference including any common bias within these approaches, the importance of internal and external validity, and the relationship between statistics and parameter. Fourthly, the distinctions between correlation and causation, and many ways to increase internal validity such as using random selection and random assignment, using control group, and adding extraneous variables were clarified. Fifthly, the author focused on regression analysis, one of the most famous statistical models, by explaining many things from least-square regression line, regression to the mean, response variable, explanatory variable, regression coefficient, multiple linear regression, and logistic regression.

Sixthly, he talked about algorithms that were used for classification and prediction. While classification is the process of finding a good model to predict the categorical class of objects, prediction is the process of finding a good model to predict continuous valued functions. Moreover, he also warned of overfitting problem, and algorithm challenges such as lack of robustness and implicit bias that we should concern. Seventhly, bootstrapping statistics was introduced to provide a way to derive the estimates of standard errors and confidence intervals when the population distribution was unknown. Eighthly, the author briefly described probability theory, how to use probability in Statistics, and how to calculate confidence interval.

Ninthly, hypothesis testing, null hypothesis, alternative hypothesis, p-value, statistical significance, Type I and Type II error, as well as the misuses and misconceptions concerning p-values were discussed. In case of the p-value problems that might be called by p-hacking or data dredging, the author showed the six principles, announced by American Statistical Association (ASA) to provide recommendations for improving the proper use and interpretation of the p-value, and tried to elaborate on each principle by giving us some examples to make it easier to understand. The six principles by American Statistical Association (2016) are presented as follows:

"1) P-values can indicate how incompatible the data are with a specified statistical model.

2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4) Proper inference requires full reporting and transparency.

5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis."

This part makes us realize that incessant criticisms and arguments against p-value from statisticians around the world finally have received great response. While p-value is a valuable and useful tool to give the strength of evidence against the null hypothesis, at the present time, it is recognized as providing very limited information, and overreliance on the p-value to support broader hypotheses is too dangerous. Thus, when using p-value, we should be aware of its limitations and misconduct; for example, stopping collecting data once p-value less than 0.05, using covariates to get p-value less than 0.05, or transforming the data to get p-value less than 0.05. Moreover, if necessary, we had better find some suggestions on any other statistical approach to augment or replace the p-value.

Tenthly, the author introduced Bayesian statistics, Bayes' theorem, Bayesian inference and Bayes factor, which is one of those alternative analyses that can be used to augment or substitute the p-value. For many people who are not statisticians, I think this is the part that most of us might be unfamiliar with. Consequently, the content in this topic might not have enough details for us to clearly understand. To use it properly, additional knowledge is required.

Finally, the author provided some examples of the misuse and abuse of Statistics in plenty of quantitative research in many topics such as reproducibility crisis, choosing sampling methods with limited time and budget constraints, leading questions, too small sample size, ignoring extraneous variables, P-hacking, presenting a post hoc hypothesis (HARKing), and questionable interpretation and communication practice. In addition, he also suggested how to improve the quality of Statistical practice, how to develop communications in Statistics, and how to assess Statistical inferences efficiently by evaluating reliability of research design, source of data and interpretation.

Overall, the Art of Statistics Learning from Data provides the reader the basic statistical principles for how to obtain knowledge from data, introduces the fundamental topics in modern statistics such as data visualization and data-analytics, and gives some suggestions for approaching statistical problems to reduce the likelihood of misuse of statistics, misleading numbers, and misinterpreting data. By explaining with interesting present-day examples and studies, and avoiding using complex statistical formulas, this book is highly recommended for anyone who is endlessly interested in statistics or other related fields and can be perfectly assigned as a supplementary material for undergrads with some basic statistical knowledge. It would provide students a greater understanding of statistics by presenting great examples with clear explanations. In addition, this book is unquestionably useful for skilled professionals as a revision or to brush up on modern statistics. It would not just help them to do statistics efficiently but also help them to interpret statistics appropriately. As academic professionals, this book reminds us that statistics is a powerful tool; therefore, we must present and communicate our statistical findings to the public honestly. Lastly, although there are some technical terms and hard stuff, this book is definitely enjoyable and you will be rewarded for getting to the end.

References

Spiegelhalter, D. J. (2020). *The Art of Statistics Learning from Data.* London, United Kingdom: Pelican, an imprint of Penguin Books.

American Statistical Association. (2016). *American Statistical Association releases statement on statistical significance and P-values.* Retrieved from https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf