A Model of Teaching Statistical Computing

Ken W. Li

***Abstract*—The underlying philosophy of statistical software primarily aims to empower data analysts to concentrate on statistical thinking and leave the computational burden to computers. Teaching how to program statistical software, however, is over-emphasized by some lecturers and students may not develop the skills needed to be competent at justifying and/or interpreting statistical results. For this reason, this article aims to address this issue by developing a model of teaching statistical software to strengthen secondary and tertiary students' capacity to understand statistical processes and conduct statistical investigations. The model consists of six major steps: examining data characteristics, selecting statistical tools, understanding the strengths and weaknesses of statistical software, checking the accuracy of statistical output, interpreting statistical output and presenting statistical results.**

***Index Terms*—Semantic error, statistical context, statistical logic, statistical investigation.**

I. Introduction

Statistical software offers tools for organizing data, visualizing data and analyzing data [1]. Some of the software developers are devoted to designing and implementing computer programs to offer a window-based environment, handy statistical tools and user-friendly features, such as pull-down manuals, an on-line Help menu, hyperlinks to statistical glossary, fancy computer output and eye-catching graphical displays without taking into account whether such tools can serve statistical purposes [2]. For instance, a cone chart is more eye-appealing but produces visual illusion in such a way that direct comparisons between the lengths of vertical bars can be distracted by its base area according to the hierarchy of perceptual accuracy [3], [4]. On the other hand, a bar chart is more vivid and clearly shows varying frequencies of data as presented in different heights of the bars. In addition, it would be better to arrange the heights of the bars either in ascending or descending order of data magnitude so that their heights can be compared more easily.

Ideally, the aim of using statistical software is to empower data analysts to concentrate on statistical thinking and graphing and leaving the computational burden to the computers, but students make mistakes in statistical calculations and have lax criteria for accepting a solution as plausible. For instance, many students accept negative sums-of-squares in an analysis of variance or correlation coefficient greater than one. Statistical software is of little value in this way so the author advocates a model of teaching statistical software to reinforce students' understanding of how to use statistical software properly and efficiently. The model so developed follows the process of statistical analysis of data and consists of six major steps: examining data characteristics, selecting statistical tools, understanding the strengths and weaknesses of statistical software, checking the accuracy of statistical output, interpreting statistical output and presenting statistical results (as illustrated in Fig. 1).

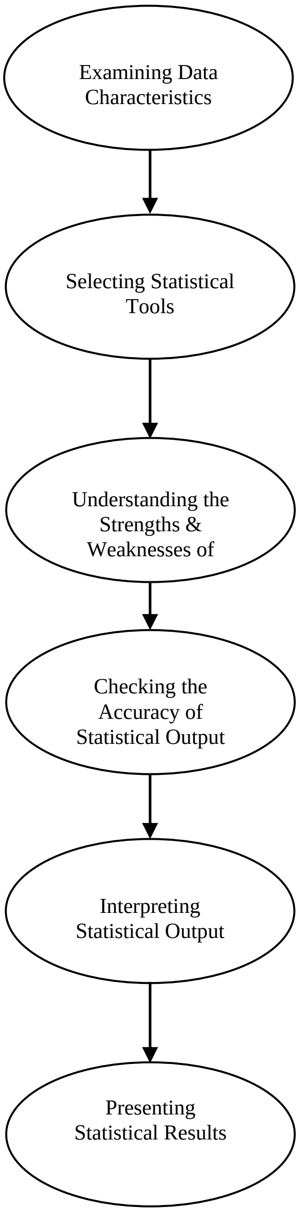
Examining Data Characteristics
Selecting Statistical Tools
Understanding the Strengths & Weaknesses of
Checking the Accuracy of Statistical Output
Interpreting Statistical Output
Presenting Statistical Results

Fig. 1. Flowchart of data analysis using statistical software.

II. A Model of Teaching Software

Statistical software does not address non-statistical questions inherent in a statistical problem. Addressing these questions is beyond its capability but is left to software users to attempt [5]. Hence, the users should possess extensive

Manuscript received August 8, 2014; revised October 9, 2014.

Ken W. Li is with the Department of Information and Communications Technology, Hong Kong Institute of Vocational Education (Tsing Yi), 20 Tsing Yi Road, New Territories, Hong Kong (e-mail: kenli@vtc.edu.hk).

DOI: 10.7763/IJIET.2016.V6.674

statistical knowledge as well as an understanding of a discipline from which the problem arises in order to complete a meaningful data analysis.

A. Examining Data Characteristics

To commence a statistical analysis of data, statistical practitioners should conduct a preliminary study on the characteristics of given data in terms of the types, format, content, context, measurement and measurement units of data. This preliminary study discloses which of the given data are relevant and useful for solving a data analysis problem and also provides general clues for which statistical tools ought to be chosen.

B. Selecting Statistical Tools

Understanding data characteristics assists in selecting correct statistical tools to a certain extent because different statistical tools are developed for different statistical purposes and/or data characteristics. For example, using paired t-test for the testing of equal means between two independent samples of data is a semantic error arising from misuse of a statistical tool. As far as statistical context is concerned, a t-test is used when comparing the means between two independent samples of data, whereas a paired t-test is used to non-independent samples of data.

The following example illustrates how correct selection of a statistical tool is reliant on data characteristics. A histogram and a bar chart are commonly used as statistical tools for presenting and examining the distribution of data. These two different statistical tools are used for the same statistical purposes of exhibiting the measures of central tendency, dispersion, skewness and peakedness of data, and identifying extreme data values/outliers if any. However, these two tools cannot be used interchangeably because the former is suitable for displaying the distribution of continuous data, whereas the latter is appropriate for discrete data [6], [7].

Obviously, a wrong selection or misuse of a statistical tool can generate incorrect statistical output and most probably will result in non-meaningful or even incorrect interpretation. Apart from checking these technical aspects, correct selection of statistical tools should also involve checking whether or not an interpretation of statistical results derived from a specific statistical tool reflects the contextual meaning. Prior to using statistical software, its users must therefore select correct statistical tools [1].

C. Understanding the Strengths and Weaknesses of Statistical Software

The introduction of user-friendly features to statistical software seems to bypass most software users' thought processes because they seldom make enquiries about computational quality of the software [8]. In fact, they need to be aware that statistical software cannot detect semantic errors; nor are all statistical outputs generated by statistical software accurate [9]-[12]. Inaccurate results are yielded for various reasons, rounding errors, truncation errors, computer platforms, computing architecture and algorithms as well as computer memory registry.

Sawitzki [13] conducted exhaustive tests on the numerical accuracy of most popular statistical software packages namely, BMDP, Data Desk, Excel, GLIM, ISP, SAS, S-PLUS, SPSS and STATGRAPHICS and found none of them could pass all the tests. In addition, the same version of SAS yielded different results when running on different computer platforms. In particular, results produced by mainframe computers are more accurate than those produced by personal computers because the former has powerful computer compilers and libraries.

McCullough [10] reported that two different statistical software packages might end up with different results after performing the same statistical calculation because they did not necessarily adopt the same statistical algorithm. Sawitzki [12] compared three SAS computer modules, REG, GLM and ORTHOREG which could be used for regression modelling. Each of them uses different statistical algorithms and produces statistical results in varying accuracy accordingly. Furthermore, some statistical software even outputs a negative value for a variance and a value greater than one for a correlation coefficient [10]. With statistical logic borne in mind, these two resulting values are impossible. These two errors are not a matter of accuracy but the concern is whether these values make any sense in a statistical context. The first incorrect result can be disproved by the following statistical algorithm. A sample variance is a measure of dispersion of data that gives the average squared deviation of each item in the sample from the sample mean. The formula, $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is generally used to compute a sample variance, where x_i represents the real value of i^{th} data point in a set of data, $\bar{x}$ stands for the sample mean of a set of data and n is the total number of data points in the set of data. To obtain a meaningful sample variance, the restriction for the value of n is greater than one so that the denominator $n-1$ cannot be negative. For the numerator, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is to sum up the squared difference between each data point and the mean of the data set, starting from the first term to the last term of the data set. The square of a real value can never be negative, for the same reason, the square of the difference between two real values can never be negative either. This means each term of the squared difference cannot be negative. The sum of all non-negative values cannot be negative. This implies that the numerator is a non-negative value. The ratio of two non-negative values cannot be negative. As non-negative numerator and non-negative denominator are obtained, the sample variance cannot be negative. These results in a statistical algorithm which is a critical issue and data analysts should acknowledge.

These errors that uncover the problems are twofold. The first fold is errors in numerical computation existing in the forms of binary representation and finite precision, rounding error and truncation and algorithmic error [10]. The second fold is the quality of computer programming. McCullough pointed out that data being kept by computers in single precision would not yield numerical results as accurate as the data being kept in double precision provided that the computer programming is in good quality. On the other hand, poor programming would result in less accurate numerical results irrespective of whether single or double data precision had been used.

Students should be conscious of statistical logic when using statistical software. That is, they must understand what the goal of statistical computation is; what statistical computation is to achieve; how computers register numerical data; and what the logic of the strategy is. Intuitively, the variance of three given values, 90000001, 90000002 and 90000003 is one because when someone looks up the definition of variance which governs statistical logic it assists in checking the meaningfulness of a statistical result. Nevertheless, a result other than 1 is yielded by statistical software mainly because how data being processed and stored by computers is different from the way of data is processed by human brain. A number is translated into binary representation of a computer memory register. With representation and algorithm, someone may ask how this computational theory can be implemented; what the representation for the input and output is and what the computer algorithm for the statistical procedures are being used. And at the program implementation level someone may also ask how the representation and algorithm can be realized physically at the computational level, for students' knowledge of statistics is that of statistical principles; the representation and algorithm level is that of formulae and tables; and the hardware implementation is calculator, computer, and so on.

Thus, there are risks in using statistical software without extensive knowledge of data handling and interpretation. Software users must therefore have considerable statistical knowledge and concepts as well as an awareness of the capability and incapability of statistical software [8]. Actually, they should explore what underlying limitations about the data range the software can take in their computer platform. Software users should also think about how modifications or adjustments that may be needed. For example, ill-conditioned (linearly dependent) data must be re-scaled prior to using regression SAS REG module to build a regression model [12].

D. Checking the Accuracy of Statistical Output

Sawitzki [12] and Knusel [9] reminded users of statistical software not to totally accept statistical output from the software because proper use of the software does not necessarily to assure an absolutely correct or accurate statistical output although correctness and accuracy are a prerequisite.

On some occasions, the software users must know what underlying assumptions about data have to be made. When the distribution of data is not known, assuming the data follows a normal distribution is a common practice provided that the sample size of data is large. In regression modelling, it is common to fit data to a regression model after a preliminary study of data characteristics and data relationships but without prior checking of the data distribution. As such, software users should bear in mind that the regression output is based on the normality assumption of data unless a subsequent normality assumption check confirms the assumption valid.

Reading statistical output without a well-understood statistical context is a risky undertaking. For example, using the Kruskal-Wallis Test for testing of equal means between three or more independent groups of continuous data is inappropriate for two reasons. First, this statistical tool is developed for testing ordinal data that cannot be measured on a quantitative scale. Otherwise, One-way Analysis-of-Variance (ANOVA) is used provided that the data follow a normal distribution. Second, the actual values of the continuous data are converted to data rankings, so that parts of the data values are distorted. This test becomes insensitive and its consequence can be as severe as producing a misleading statistical output. That is, the output may show the test statistically significant but in fact it may not be, or vice versa. Hence, when reading statistical output one needs to take careful account of what statistical tool is being used and what the implications for data characteristics are.

Of course, beyond checking computational performance and accuracy, software users must study whether or not statistical output makes sense in terms of statistical logic and reasoning [14], [15].

E. Interpreting Statistical Output

Royall [16] stated the core objective of interpretation of statistical output as "what the data say", that is, "to interpret data as evidence". To interpret statistical output meaningfully, students must understand how the results are derived and what they are representing. Using graph comprehension as an example, students must read the labels of a statistical graph; relate the labels and graphical display to the context of given data; describe the graphical appearance of the display; and deduce the implications for the graphical display for instance, data relationship, data projection, data discrepancy and so on. However, this interpretation process needs to be developed.

Royall [16] stressed an inclusion of probabilistic view was vital for evidential interpretation of observations. For example, 95% and 99% confidence intervals for the population parameter, θ are (a,b) and (c,d) respectively. For given information, 99% confidence interval has a wider interval and a higher probability of enclosing θ than that of 95%. There is a need for organizing the teaching of statistical software in line with evaluation of statistical evidence.

Interpretation is more than a superficial explanation of a statistics in a numerical context; instead it takes a deeper look at what it really means; how it measures; and what practical meaning and/or implications can be deduced from given statistics. An example can be quoted from Truran's study [17] on interpretation of correlation in which he pointed out that the superficial interpretation of coefficient of determination, $R^2 = \frac{\text{unexplaine d variation}}{\text{total variation}}$ (where unexplained variation and total variation mean the sum of the squared deviations between predicted and observed values and the sum of the squared deviations of observed data from their mean respectively) as a ratio of these figures was insufficient. Interpretation of R^2 only takes a look at what amount of the variation in Y is explained by the regression model and is not enough either. Specifically, the value of R^2 lying between 0 and 1 indicates to what the extent the linear regression model fits given data. R^2 close to 0 indicates that the given data do not fit the regression model at all and signifies not to use the model for making any prediction. If R^2 is about 0.5, it indicates that given data do not fit the model well, thus making prediction by using the model is not persuasive. This implies that such a model needs to be refined. R^2 near 1 implies that given data fit the regression model very well and there is a strong linear relationship between two variables, but says nothing about the direction of data relationship. Using this model for making a prediction is reliable but no further induction or deduction can be made, such as the cause-and-effect relationship and exact relationship between two variables, that is, how much of the change in a dependent variable is caused by an independent variable.

Moreover, mis-interpretation of "significance" which is probably due to misconception about, or incomplete understanding of, the rationale for significance tests has drawn numerous researchers' attention (e.g., [18]-[20]). The superficial interpretation of a significance test which says whether or not to reject the null hypothesis, H_0 is neither convinced nor is related to a practical context. The rejection decision is merely made by contrasting the p-value with a pre-determined level of significance α . This accept-reject decision-making perspective without taking into account a Type I or Type II error is not proper [18]. Falk and Greembaum [6] believed that the p-value had been misinterpreted as $\Pr\left(H_0 \middle| \text{data passes a statistica l test}\right)$. In fact, p-value informs data analysts how likely it is that a wrong decision is made in rejecting H_0 but H_0 is universally true. The smaller the p-value (a smaller chance of committing Type I error that is a true null hypothesis, H_0 is rejected), the greater the weight of evidence favors rejection of H_0. The implication is what is being stated in H_0 is most likely untrue when turning this statistical evidence into a practical context.

F. Interpreting Statistical Output

Statistics students need to be familiar with database management, statistical computing [8], [21] and word processing techniques [13]. This includes computing aspects, data handling, programming and the tailoring of statistical software packages which are essential for statistical analysis of data.

Prior to statistical analysis of data, data must be well organized and managed by a computer database system that can store data, edit data and retrieve data efficiently. Setting up a new database system, or using an existing database system requires creative, critical, and complex thinking skills [22]. Giving field names to data is an example of creative thinking because names given should be meaningful and connected with the data content, possibly facilitating database users to read the data without the need for referencing to a data dictionary within the database. Critical thinking identifies the characteristics of data, that is field type (numeric, character, date, logical or memo) and field length (memory space for each individual piece of data) and prioritizes field in the context of database and aiming at maintaining data storage space to a minimum and utilizing data storage for efficient data processing. Looking more closely at the hidden relation of data requires complex thinking skills to integrate data, reduce data redundancy and build data security for various levels of authorized access.

The dissemination of statistical results through written reports is a vital component of statistical work. Statistical practitioners need to incorporate statistical results with other knowledge to make a decision, such as checking whether or not the quality of a certain product is satisfactory based on the testing of a sample of products when drafting a report.

In the era of information, word processing software is a very popular and sophisticated IT tool used by students to prepare and present their written assignments [13]. The software offers a variety of facilities at various stages of writing: planning, drafting and revising. In the planning stage of writing, students think and determine what to write. Loose but creative ideas, notes or outlines of writing are easily recorded and stored into computer files which can be quickly retrieved for subsequent drafting. Students need to articulate their thoughts and have control over the computer keyboard when drafting. For instance, students present their thoughts into written words in which they state the theme of a section and indicate the direction and development of ideas at the same time of utilizing desktop publishing features: headings, paragraphing, etc. so as to enhance the presentation layout of their writing. When students revise their own writing, they generally need to re-organize their texts so as to ensure the logic and coherence of their viewpoint and so forth. Perhaps, blocks of text may need to be moved, copied or deleted from a draft by using word processing software. Scrimshaw [13] reported that word processing software did not improve the semantic quality of students' writing but assisted them in overseeing and reorganizing writing tasks and composing more succinct and error-free texts.

III. Conclusion

Students have illusions of the power and capability of statistical software because they lack statistical logic about how reasonable or unreasonable a statistical answer is [23]. To teach students how to use statistical software properly, instruction must be structured to illustrate the importance of examining data characteristics, selecting statistical tools, understanding the strengths and weaknesses of statistical software, checking the accuracy of statistical output, interpreting statistical output and presenting statistical results that emphasize the close connections between them.

Acknowledgment

The author would like to thank Dr. Barbara Jack for her valuable comments on an earlier version of the manuscript.

References

[1] J. M. Chambers, "Greater or lesser statistics: a choice for future research," *Statistics and Computing*, vol. 3, no. 4, pp. 182-184, 1993.

[2] K. W. Li, "Enhancing students' graphic communication," in *Proc. Science and Technology Education Conference 2002*, Hong Kong, 2002, pp. 411-422.

[3] S. M. Kosslyn, *Elements of Graph Design*, New York: W.H. Freeman and Company, 1994.

[4] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed., Connecticut: Graphics Press, 2001.

[5] J. M. Chambers, "Users, programmers, and statistical software," *American Statistical Association Journal of Computational and Graphical Statistics*, vol. 9, no. 3, pp. 404-422, 2000.

[6] M. C. Fleming and J. G. Nellis, *Principles of Applied Statistics,* London: Routledge, 1994.

[7] K. W. Li, "Developing secondary students' through the use of computers," *The Journal of Educators*, vol. 10, no. 1, pp. 1-15, 2000.

[8] D. Nicholls, "Statistics into the 21st century," *Australian & New Zealand Journal of Statistics,* vol. 41, no. 2, pp. 127-139, 1999.

[9] L. Knusel, "On the accuracy of statistical distributions in Microsoft Excel 97," *Computational Statistics and Data Analysis*, vol. 26, pp. 375-377, 1998.

[10] B. D. McCullough, "Assessing the reliability of statistical software: Part I," *American Statistician,* vol. 52, no. 3, pp. 358-366, 1998.

[11] B. D. McCullough, "Assessing the reliability of statistical software: Part II," *American Statistician,* vol. 53, no. 2, pp. 149-159, 1999.

[12] G. Sawtizki, "Report on the numerical reliability of data analysis systems," *Computational Statistics and Data Analysis*, vol. 18, pp. 289-301, 1994.

[13] P. Scrimshaw, "Teachers, learners and computers," in *Language, Classrooms and Computers*, P. Scrimshaw, Ed., New York: Routledge, 1993, pp. 3-10.

[14] J. Garfield, "Assessing statistical reasoning," *Statistics Education Research Journal,* vol. 2, no. 1, pp. 22-38, 2003.

[15] J. Garfield and I. Gal, "Teaching and assessing statistical reasoning," in *Developing Mathematical Reasoning in Grades,* L. V. Stiff and F. R. Curico, Eds., Reston: NCTM, 1999, pp. 207-219.

[16] R. Royall, *Statistical Evidence: A Likelihood Paradigm,* Reprint ed., New York: Chapman & Hall/CRC, 1999.

[17] J. Truran, "Understanding of association and regression by first year economics students from two different countries as revealed in responses to the same examination questions," in *Proc. the 20th Annual Conference of the Mathematics Education Research Group of Australasia*, New Zealand, 1997, pp. 530-537.

[18] S. L. Chow, *Statistical Significance: Rationale, Validity and Utility*, London: SAGE Publications, 1996.

[19] N. R. Falk and C. W. Greenbaum, "Significance tests die hard: The amazing persistence of a probabilistic misconception," *Theory & Psychology,* vol. 5, pp. 75-78, 1995.

[20] A. V. Jimenez, "Students' conceptions of the logic of hypothesis Testing," *Hiroshima Journal of Mathematics Education*, vol. 4, pp. 43-61, 1996.

[21] D. Nicholls, "Future directions for the teaching and learning statistics at the tertiary level," *International Statistical Review,* vol. 69, no. 1, pp. 11-15, 2001.

[22] D. H. Jonassen, *Computers as Mindtools for Schools,* New Jersey: Merrill Prentice Hall, 2000.

[23] P. T. Smith, "Levels of understanding and psychology students' acquisition of statistics," in *Cognitive Processes in Mathematics*, J. A. Sloboda & D. Rogers, Eds., Oxford: Oxford University Press, 1987, pp. 157-168.

Ken W. Li received his BSc, MSc degrees from University of Western Ontario, Canada. He got his Ph.D(PGDipEd) degree from the University of Queensland, Australia. Dr. Li's current research areas are in the use of IT in education and human-computer interaction. He has developed teaching models and instruments for assessing learning outcomes of students. He has published an undergraduate textbook, book chapters, and more than thirty research articles in international journals as well as proceedings of international conferences. Recently, his research paper, "A study on students' attitudes towards teacher's interventions within an IT environment" was awarded "The Best Paper" by International Conference on Technology in Education in 2014. In 2006, his research paper, "A Simulation Study on Intra-cluster Correlation" was awarded Certificate of Merits by the International Association of Engineers. He is associate editors of academic journals, and a member of various paper reviewer panel, program and organizing committees of conferences as well as competitions.