International Journal of Applied Science and Mathematics
Volume 8, Issue 1, ISSN (Online): 2394-2894

A Note on Statistical Methods for Most-Frequently Seen 3-Type of Data

Dewi Rahardja
U.S. Department of Defense, Fort Meade, MD 20755, USA.
(Disclaimer Statement – This research represents the author's own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.)
Corresponding author email id: rahardja@gmail.com

Date of publication (dd/mm/yyyy): 02/02/2021

***Abstract* – Data types are important concept when designing and planning a study or analysis because the existing statistical methods in the literature can only be used with certain data types along with their assumptions. The basic/ majority 3-type of (response or outcome measure) data are continuous, categorical/discrete, and Time-To-Event (TTE). One has to analyze continuous data differently than categorical and also differently than TTE data; and vice versa. Otherwise, inappropriate choice of statistical methods would result in improper analyses and very biased results; yield to misleading conclusions and recommendations. Therefore, knowing the types of data one encounters would enable one to pre-specify the appropriate method of analysis. The purpose of this (integrated) short-note article is for teaching and training: to point-out and note the most-frequently seen 3-type of data in the real-world applications with the well-established statistical methodologies in the literature where the computing is doable via popular statistical software.**

***Keywords* – Types of Data, Continuous, Categorical, Time-To-Event, Descriptive Statistics, Inferential Statistics.**

I. INTRODUCTION

In the real-world applications, data analysis, conclusions, and recommendations of a study requirements should be driven by data type (obtained by the purpose and reasons of a study), as opposed to by arbitrary/non-suitable methods. In other words, many common errors are to forcefully utilize a specific statistics method, and then insert the data to the method. Rather, the reverse is a more sensible/proper order – first, determine what data type we encounter [or to design and collect], then select the appropriate method accordingly. In the later way, the method chosen would be statistically valid, has a logical rationale, sound and justifiable to the objective, design and data needs for the purpose of a study.

In this short-note paper, we first present the most-frequently encountered basic/majority 3-type of data (as opposed to advance/minority), in terms of the response variable (i.e., the outcome measure or dependent variable). Such outcome measure or response variable can either be continuous, categorical/discrete, or Time-To-Event (TTE) type of data; and the assumptions of Large-Sample (Asymptotic) Normal Theory and independent-and-identically distributed (iid) cases must be checked and satisfied. In other words, if any of these assumptions are violated severely, then the results of the analyses can be very biased/inaccurate, imprecise, and/ or invalid. Note that the word "majority" indicates more than half (or the greater number). Meaning, more than 50% of cases one encountered would typically fall under these 3-type of data. Note also that the word "basic" implies these are not the only types of data exist but there are more (advance) types beyond the basic. However, those advance/minority data types, such as Time-Series data [1], Spatial data, Spatio-Temporal data, not independent but identically distributed (non-iid data), with one/more covariates inclusion, etc., are beyond the scope of this short-note article. Next, we integrate the roadmaps to emphasize each of these basic/majority 3-type of data (i.e., the outcome measure or response variable or dependent variable). Such (integrated) review of the basic roadmaps (basic/majority 3-type of data) will be helpful for practitioners, investigators, educators,

students, and researchers in various fields (such as clinical trials, epidemiology, finance, business, economics, education, sociology, psychology, human health, along with many others) to determine which methods are suitable for their particular data type. We also highly recommend to well-train these topics in the Statistics and/or Data Science graduate programs so that their graduates will be ready to properly handle the majority of the problems in the real world, adequately. Hence such roadmaps need to be studied/trained sequentially, with the following recommended order: continuous, categorical, and TTE sequence.

On the most-recent 2020 World Statistics Day [2] webinar, a virtual conversation of the perspectives on the future national statistics occurred. In the conversation, presenters pointed out that besides accuracy and precision of the results, timeliness of results are also of a higher-priority rank. Therefore, a well-balanced management of all these important factors are needed to be taken into accounts. Then, in order to reach timeliness goal, on top of accuracy and precision goals, such easy (ready) and quick (fast) access to these prescriptions of the basic/majority 3-type of data will be very helpful to the practitioners, investigators, educators, students, and researchers in the real-world applications.

To date, there is no literature that clearly and concisely emphasize, note, and summarizes the [statistical methods] review of such these basic/majority 3-type of data: continuous, categorical/discrete, and TTE. Such inadequacy is evidenced from questions and lengthy discussions raised on the American Statistician Association (ASA) group emails or webinar or postings, various different workplaces, and a professional subject-matter experts (SME) research association such as Research Gate (RG), regarding how to analyze and what are the appropriate methods to use when they encounter such basic/ majority 3-type of data. Therefore, in our standard [statistical] line of practice, we frequently find practitioners, investigators, and researchers get confused about the data type and the appropriate methods available to use, and post questions on ASA, workplaces, and/or RG platforms. To close such confusion gap, this short-note article will be very practical and helpful to various practitioners, investigators, educators, students, and researchers, in various fields of study, for both prospective and retrospective studies, and in both the statistical design and data analysis plan.

II. Most Frequently Seen 3-Type of Data (Outcome Measure / Response / Independent Variable)

2.1. *Continuous Data*

The first type of data or response variable or outcome measure that we most encounter in the real-world applications is continuous type of data. Continuous Data response or outcome is very common in real-data applications such as clinical trials, finance, business, economics, epidemiology, sociology, etc. The reviews of the analysis of such Continuous [3] outcome measure (or response variable) is shown to have a long history, beginning with the two-sample t-test with s-pooled [4], One-Way ANOVA [4], two-sample t-test with Satterth-Waite Exact degrees-of-freedom [4], Welch ANOVA test [5], Wilcoxon Rank-Sum test [6] or Mann-Whitney test [7], Kruskal-Wallis test [8], Paired t-test [9], up to the multiple comparison procedure (MCP) [10]. These literature presents the hypothesis testing procedures that are available for various types [one-sample, two-sample, and multiple-sample tests] of continuous-data type of outcome measure (or response variable) with one grouping variable (factor) of multiple levels to multiple factors with multiple levels. For master-degree analysts, they need to [at least] master all these Continuous [3] data-type of analysis methods, in order to thrive well in the real-world jobs.

2.2. *Categorical Data*

The second type of most commonly seen data or response variable or outcome measure in the real-world application is categorical (binary outcome) type of data. Categorical Data or most particularly binary or dichotomous outcome (i.e., success vs. failure, dead vs. alive, 1 vs. 0) is very common in real-data applications. The analysis of such categorical outcomes has a long history, beginning with the single 2 × 2 table, multiple/stratified 2 × 2 tables, matched/paired 2 × 2 tables, to big table such as K × K tables. Rahardja *et. al*. in 2016 [11] provide a comprehensive review of the hypothesis testing procedures that are available in the literature for various types of categorical data, particularly on the two independent or dependent (matched pair) samples with binary data, with or without stratum effects. The review includes classical methods such as Fisher's Exact [4], Pearson's Chi-Square [4], McNemar [12], Bowker [13], Stuart-Maxwell [14–15], Breslow-Day [16] and, Cochran-Mantel-Haenszel [17–18], as well as newly developed ones such as Homogeneous Stratum Effect (HSE) test [19] and Common Risk Difference (CRD) test [20]. In order to be reasonably ready in the workplaces, any Master-Degree graduate programs (of the aforementioned majors) should master [at least] half of these Categorical data-type of analysis methods prescribed in the Rahardja *et. al*. [11] paper; while PhD-Degree graduates should master all these methods.

2.3. *Time-To-Event (TTE) Data*

The third type of most commonly seen data in the real-world application is TTE type of data. The analysis of such TTE Data or outcome measure (or response variable) has a long literature, beginning with the Kaplan-Meier (K-M) [21] or Product Limit (PL) Estimator, Life Table (Actuarial Estimator) [22], Log-Rank (Cox-Mantel) Test [23–25], up to the Cox (Proportional Hazard) Model [26]. Rahardja and Wu in 2018 [27] summarize both descriptive and inferential methods available in the TTE literature. Similarly, any graduate programs (of the aforementioned majors) would be much beneficial to master such TTE data-type of analysis methods [27]. In the next Section, we summarize the roadmaps of the above basic/majority 3-type of data.

III. ROADMAPS

3.1. *Overall (Big Picture) Roadmap*

In Figure 1, we display the roadmap for practitioners and researchers to select the most suitable method available in the literature for their type of data. As described in Section 2, the corresponding roadmap method in the Figure 1 is organized by whether the data (outcome or response) type is continuous, categorical, or TTE.

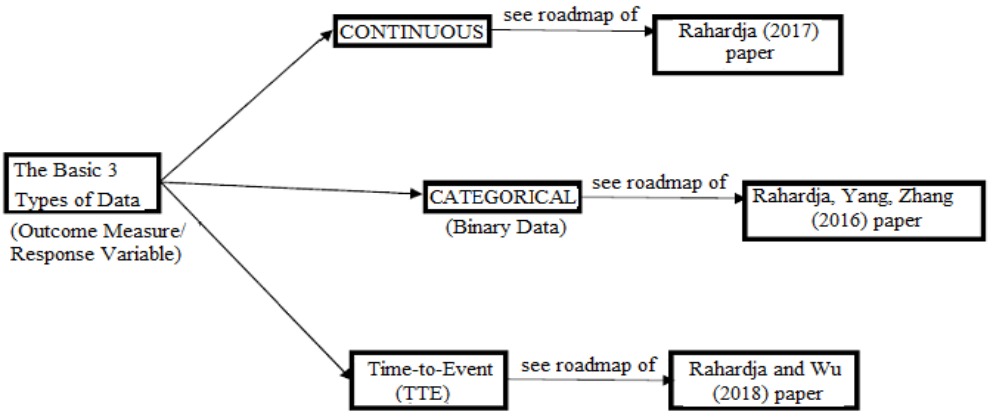

Fig. 1. The Most-Frequently Seen Basic/Majority 3-Type of Data (Response Variable or Outcome Measure or Dependent Variable).

Next, if the data (outcome or response) type is continuous, the roadmap will lead to Figure 2; whereas if the data (outcome or response) type is categorical, the roadmap will guide to Figure 3; and finally, if the data (outcome or response) type is TTE, the roadmap will direct to Figure 4.

3.2. *Continuous-Data Roadmap*

In Figure 2, for the continuous data (outcome or response), the roadmap method is provided by whether or not the response variable (outcome measure) is independent, then by whether or not the outcome is normally distributed data, and finally, by whether or not the outcome variable has homogeneous variance. Then either yes/no response variable (in each of the 3-sequential aforementioned questions) will lead to whether the grouping variable (or factor) is two-sample for a two-level factor or is multiple-sample or k-sample (where k is greater than 2) for a multiple-level factor.

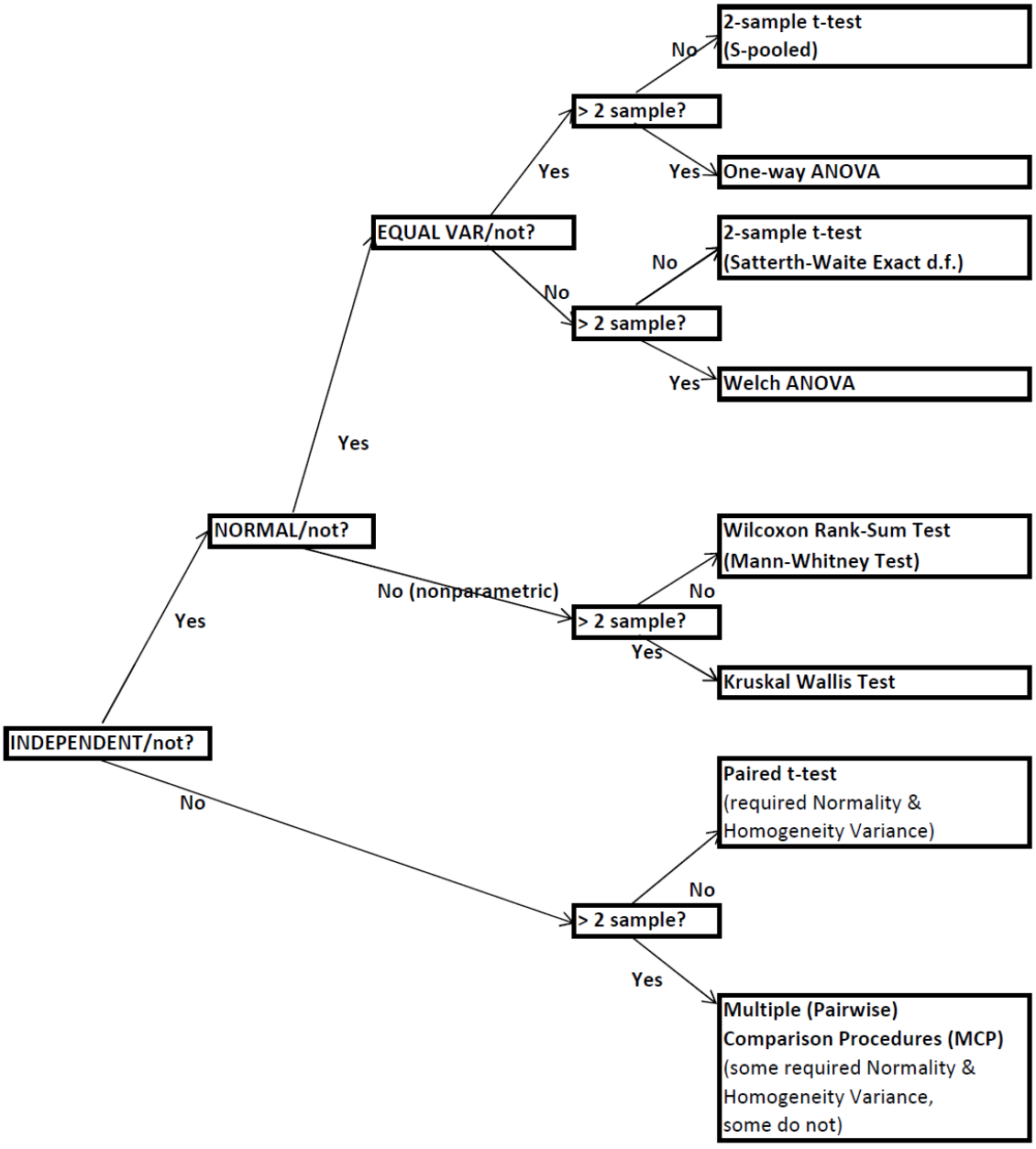

Fig. 2. Continuous-Data Roadmap for Two-Sample and Multiple-Sample Testing.

3.3. *Categorical-Data Roadmap*

In Figure 3, for the categorical data (outcome or response), the roadmap is presented by whether or not stratification table (multiple contingency table) is needed, then by whether the two groups are independent or dependent (paired) data, and finally whether or not equality/commonality/homogeneity exists (intuitive, depending on each of the scenario asked in the roadmap).

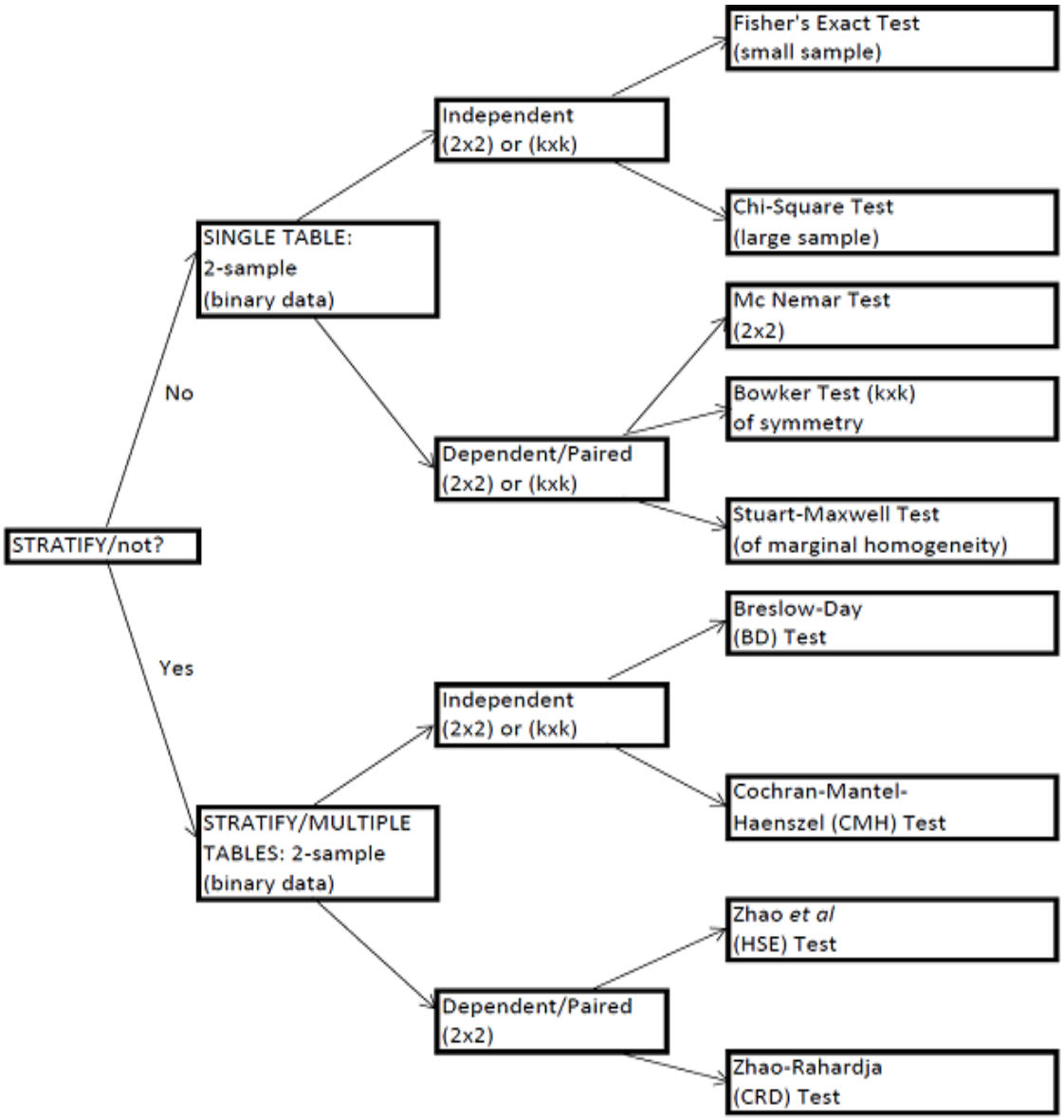

Fig. 3. Categorical-Data Roadmap by Stratify or Not.

3.4. *TTE-Data Roadmap*

In Figure 4, for the TTE data (outcome or response), the roadmap is prescribed by whether the statistics type desired is descriptive or inferential statistics. Then, if the statistics type is descriptive, the roadmap will lead to whether the estimates displayed is graphical (K-M Curve or Product Limit Estimator) [21] or tabular (Life Table or Actuarial Estimator) [22]; whereas if the statistics type is inferential, then the roadmap will lead to whether there is only one covariate with multiple levels (Log-Rank or Cox-Mantel Test) [23–25] or multiple covariates/predictors (Cox or Proportional Hazard Model) [26].

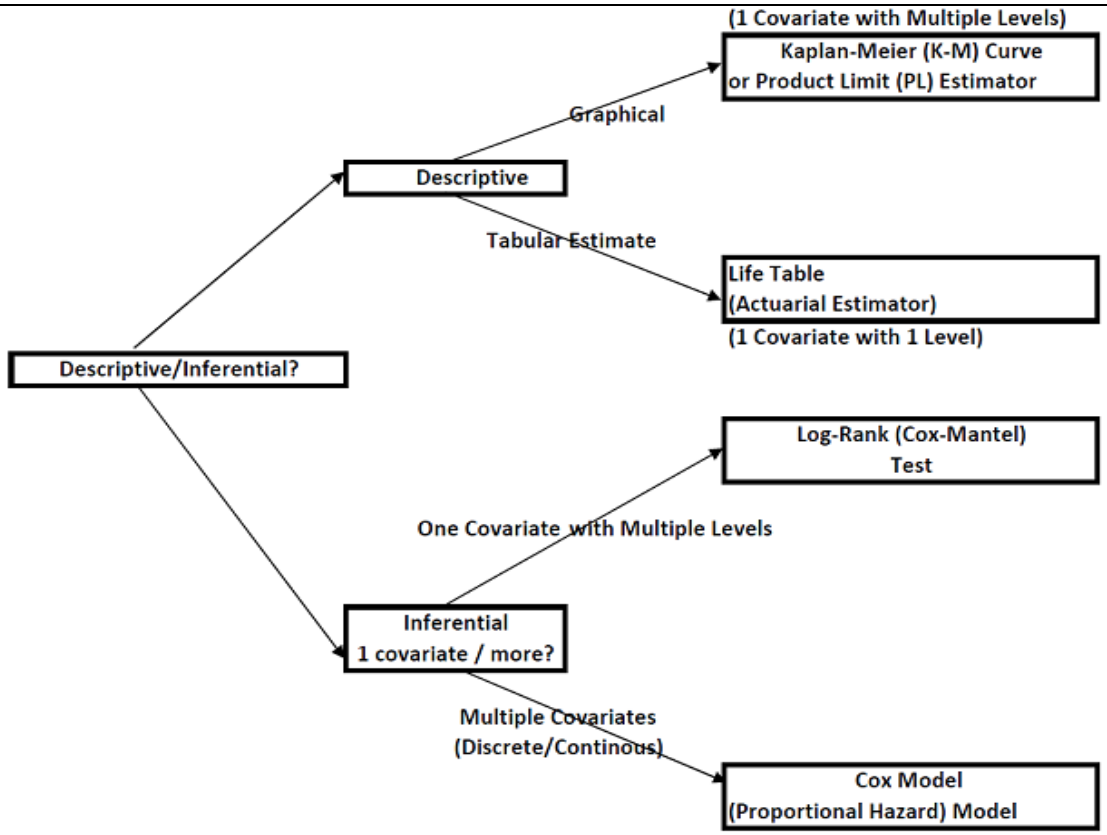

Fig. 4. TTE-Data Roadmap for Descriptive and Inferential Methods.

IV. Summary

In real-world applications, the above roadmaps of the basic/majority 3-type of data response or outcome are very frequently seen. Those 3-type of data (i.e., outcome measure or response variable or dependent variable) are continuous, categorical, or Time-To-Event (TTE) type of data. All three types of data have long history of literature. In this short-note (integrated) paper, we point out a review of those three basic/majority types of data in the provided roadmap, as shown in Figures 1-4. As pointed out in the Introduction (Section 1), there are (more advanced/minority) data types beyond the scope of these basic/majority 3-type of data, presented here.

In summary, this short-note (integrated) paper will be helpful for the practitioners, investigators, educators, students, and researchers in the various fields of study (such as clinical trials, oncology, finance, business, economics, education, epidemiology, sociology, psychology, etc.) to determine the appropriate method suitable for their data types, according to the provided roadmap in Figures 1–4, for both prospective and retrospective studies, and in all the statistical design phase, data collection, and analysis plan phase. It is also highly recommended for statistics-related graduate programs (i.e., those majors such as Statistics, Biostatistics, Data Science, Economics, Finance, etc.) to teach and train their graduate students to master these methods in order to be properly well-prepared to handle/solve real-world problems, adequately.

V. Conclusion

The prescribed methodologies and roadmaps (Figures 1–4) for the (basic/majority) 3-type of data will be useful for teaching-and-training and helpful for the benefits of many others (practitioners, investigators, educators, students, and researchers) in the various fields of study.

VI. Dedication

The author is grateful and thankful to her widowed father, Djohan Rahardja, for his continual prayers and encouragements; and to her mother (https://sites.google.com/site/statistiks/in-loving-memory), Ismajati Kiswojo, who passed away on 24-May-2019 at 8:15 AM (EST). They both have been raising her very responsibly, to provide-and-protect, and lovingly to care-and-nurture. There would not be the best-version of her today, without her parents.

The author would like to thank, praise, and dedicate this manuscript to her God and Savior, Lord Jesus Christ, who always inspires, guides step-by-step, loves, and blesses her abundantly. *Soli Deo Gloria*.

VII. Disclaimer

This research represents the author's own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.

Acknowledgement

The author would like to thank the two anonymous referees for their constructive comments and suggestions which helped improved the manuscript.

References

[1] Rahardja, D. (2020) "Statistical Methodological Review for Time-Series Data," *Journal of Statistics and Management Systems*, Vol. 23, No. 8, pp. 1445–1461.
[2] World Statistics Day Webinar. (2020) "A Virtual Conversation on National Statistics – Perspectives on the Future of National Statistics" American Statistical Association – AMSTAT Videos, Webinar on 10 October 2020. Website: https://worldstatistics day.org/. YouTube accessed 23 October 2020. https://www.youtube.com/watch?v=3rAZS3I3Vh8&feature=youtu.be.
[3] Rahardja, D. (2017), "A Review of the Multiple-Sample Tests for the Continuous-Data Type," *Journal of Modern Applied Statistical Methods*, Vol. 16, No. 1, pp. 127–136.
[4] Casella, G. (2008). *Statistical design*. New York, NY: Springer.
[5] Welch, B.L. (1947), "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, Vol. 34, No.1/2, 28-35.
[6] Wilcoxon, F. (1945) ,"Individual comparisons by ranking methods," *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80-83.
[7] Mann, H.B., & Whitney, D.R. (1947), "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, Vol. 18, No. 1, pp. 50-60.
[8] Kruskal, W., & Wallis, W.A. (1952), "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 583-621.
[9] Zimmerman, D. W. (2004), "A note on preliminary tests of equality of variances," *British Journal of Mathematical and Statistical Psychology*, Vol. 57, No. 1, pp.173-181.
[10] Cao, J., & Zhang, S. (2014), "Multiple comparison procedures," *The Journal of the American Medical Association*, Vol. 312, No. 5, pp. 543-544.
[11] Rahardja, D., Yang, Y., and Zhang, Z. (2016) "A Comprehensive Review of the Two-Sample Independent or Paired Binary Data – with or without Stratum Effects," *Journal of Modern Applied Statistical Methods*, Vol. 15, No. 2, pp. 215–223.
[12] McNemar, Q. (1947), "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, Vol. 12, No. 2, pp. 153-157.
[13] Bowker, A.H. (1948), "A test for symmetry in contingency tables," *Journal of the American Statistical Association*, Vol. 43, No. 244, pp. 572-574.
[14] Maxwell, A.E. (1970), "Comparing the classification of subjects by two independent judges," *British Journal of Psychiatry*, Vol. 116, No. 535, pp. 651-655.
[15] Stuart, A. (1955), "A test for homogeneity of the marginal distributions in a two-way classification," *Biometrika*, Vol. 42, No. 3/4, pp. 412-416.
[16] Breslow, N.E., & Day, N.E. (1980). *Statistical methods in cancer research: The analysis of case-control studies*. Lyon, France: International Agency for Research on Cancer.
[17] Cochran, W.G. (1954), "Some methods for strengthening the common $\chi 2$ tests," *Biometrics*, Vol. 10, No. 4, pp. 417-451.
[18] Mantel, N., & Haenszel, W. (1959), "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, Vol. 22, No. 4, pp. 719-748.
[19] Zhao, Y.D., Rahardja, D., Wang, D.-H., & Shen, H. (2014), "Testing homogeneity of stratum effects in stratified paired binary data," *Journal of Biopharmaceutical Statistics*, Vol. 24, No. 3, pp. 600-607.
[20] Zhao, Y.D., & Rahardja, D. (2013), "Estimation of the common risk difference in stratified paired binary data with homogeneous stratum effect," *Journal of Biopharmaceutical Statistics*, Vol. 23, No. 4, pp. 848-855.
[21] Kaplan, E.L., Meier, P. (1958) "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, Vol.53, No.282, pp. 457–481.
[22] Lee, E.T. (1992), *Statistical Methods for Survival Data Analysis*, Second Edition, New York: John Wiley & Sons.

[23] Harrington, David (2005). Linear Rank Tests in Survival Analysis. *Encyclopedia of Biostatistics*. Wiley Interscience.
[24] Mantel, Nathan (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, Vol. 50, No.3, pp. 163–70.
[25] Peto, Richard, and Peto, Julian (1972). Asymptotically Efficient Rank Invariant Test Procedures. Journal of the Royal Statistical Society, Series A (Blackwell Publishing), Vol. 135, No. 2, pp. 185–207.
[26] Cox, David R. (1972) "Regression Models and Life-Tables." *Journal of the Royal Statistical Society*, Series B, Vol.34, No. 2, pp. 187–220.
[27] Rahardja, D., and Wu, H. (2018) "Statistical Methodological Review for Time-To-Event Data Type," *Journal of Statistics and Management Systems*, Vol. 21, No. 1, pp. 189–199.

AUTHOR'S PROFILE

Dewi Rahardja, PhD, MM, PhD is a Statistician with the U.S. Department of Defense. She obtained her Ph.D. degree in Statistics from the Baylor University, another Ph.D. degree in Industrial Engineering from the Iowa State University, and an M.M. degree in Piano Performance from the Butler University. Previously, she was a Mathematical Statistician at the U.S. Food and Drug Administration; Biostatistician at the University of Texas Southwestern Medical Center & the NCI designated Simmons Cancer Center; and Senior Statistician at Children's Hospital of Philadelphia. She has been actively serving as an Associate Editor in 2 refereed journals and a Statistical Referee/Reviewer in 16 refereed journals; presenting in conferences and invited talks.