**What can JMP do for you if hypothesis testing is banned?**[1]

**Chong Ho Yu, Anna Yu, & Samantha Douglas**
**Azusa Pacific University, Azusa, CA**

## To *p* or not to *p* - that is the question

The value of hypothesis testing has been challenged by researchers for several decades (Harlow, Mulaik, & Steiger, 1997; Nuzzo, 2014). Flaws of hypothesis testing and *p*-values have been widely documented. For example, by definition a null hypothesis denotes no difference (zero effect). Loftus (1996) mockingly noted that, "Rejecting a typical null hypothesis is like rejecting the proposition that the moon is made of green cheese. The appropriate response would be 'Well, yes, okay ... but so what?'" It is a well-known fact that the outcomes of many conventional statistics, such as Pearson's *r* and Chi-square, are subject to sample size. Specifically, with very large sample sizes, statistical power is close to .999, resulting in small *p*-values and erroneous conclusions (Type I error).

In 1989, when Kenneth Rothman started the *Journal of Epidemiology*, he discouraged overreliance on *p*-values. However, when Rothman left his position in 2001, the journal reverted to the *p*-value tradition (Nuzzo, 2014). In 2015 the journal *Basic and Applied Social Psychology* attempted to resolve this issue once and for all, strictly prohibiting the publication of articles that involved hypothesis testing or related statistical procedures (Woolston, 2015).

Throughout the past several decades different alternatives have been proposed; these include reportage of confidence intervals (CIs) and effect sizes (Cumming, 2011), and usage of Bayesian statistics (Novella, 2015), exploratory data analysis, data visualization, and data mining (e.g. Behrens, & Yu, 2003; Yu, 2010, 2014). However, many 'gate keepers' - including a large number of journal editors and dissertation advisors – are both reluctant to completely overthrow the classic regime, and are skeptical of fully embracing its alternatives.

It is important to point out - that in addition to hypothesis testing - JMP has rich features on data visualization, exploratory data analysis, and data mining. Therefore, even if hypothesis testing fades away from data analysis, JMP will still be useful. Contrary to popular belief, although conventional and alternative approaches are different in many ways, these approaches share many commonalities. In this presentation two examples are used for illustration: the Bayesian approach to confidence interval and LogWorth in data mining. Since these procedures often work hand in hand, JMP users do not have to choose to exclusively use one of the two.

## The Bayesian approach and the usage of confidence intervals

In 1996, the American Psychological Association Task Force on Statistical Inference endorsed the usage of confidence intervals (CIs) as a supplement to the usage of *p*-values (Wilkinson & the Task Force on Statistical Inference, 1996). Recently, Cumming (2015) mocked the continued use of *p*-values by saying, "psychology struggles out of the *p*-swamp into the beautiful garden of confidence intervals."

The usage of CIs has certain advantages over hypothesis testing. First, hypothesis testing - as a method of point-estimate - yields a dichotomous answer only (i.e. to 'reject' or to 'not reject' the null). In contrast, the usage of CIs returns a possible range of population parameters. Second, in hypothesis

[1] Poster presented at 2015 JMP Discovery Summit, San Diego, CA.

testing, a *p*-value is defined as the 'probability of observing a particular statistic, given that a null hypothesis is true.' In usage of CIs, there is no reliance upon assumption of the truth of the null hypotheses.

Although JMP reports CIs in almost every statistical output, many researchers still report *p*-values only. *P*-values and CIs are complimentary. Some authors state that CIs express margins of error, rather than probabilities. Specific CIs either do or do not contain specific population parameters (i.e. '1' or '0') (Frost, 2015). According to this dichotomous view, both *p*-values and CIs look for on single "true" answer. If exact *p*-values are reported, the function of CIs and hypothesis testing is almost identical.

Consider the following example: Professor Yu offered the same class in three different pedagogical approaches: conventional classroom, online, and hybrid. He wanted to know which teaching method could yield better learning outcomes, as reflected on test scores. The typical approach to address this question would be to first run an ANOVA, and to next run a multiple comparisons procedure. JMP users, however, can explore this question by using a diamond plot - a visual presentation of CIs (see Figure 1).

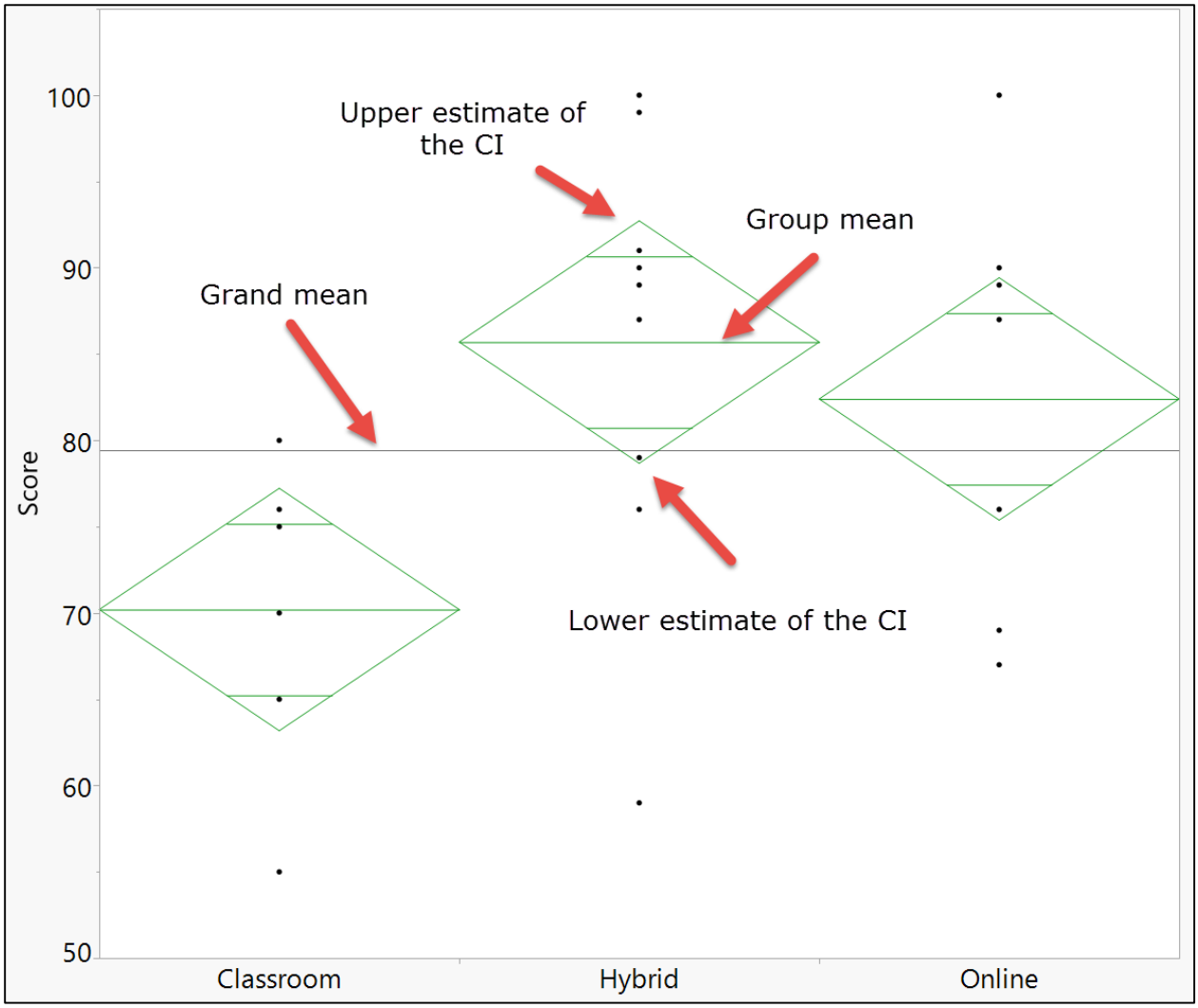
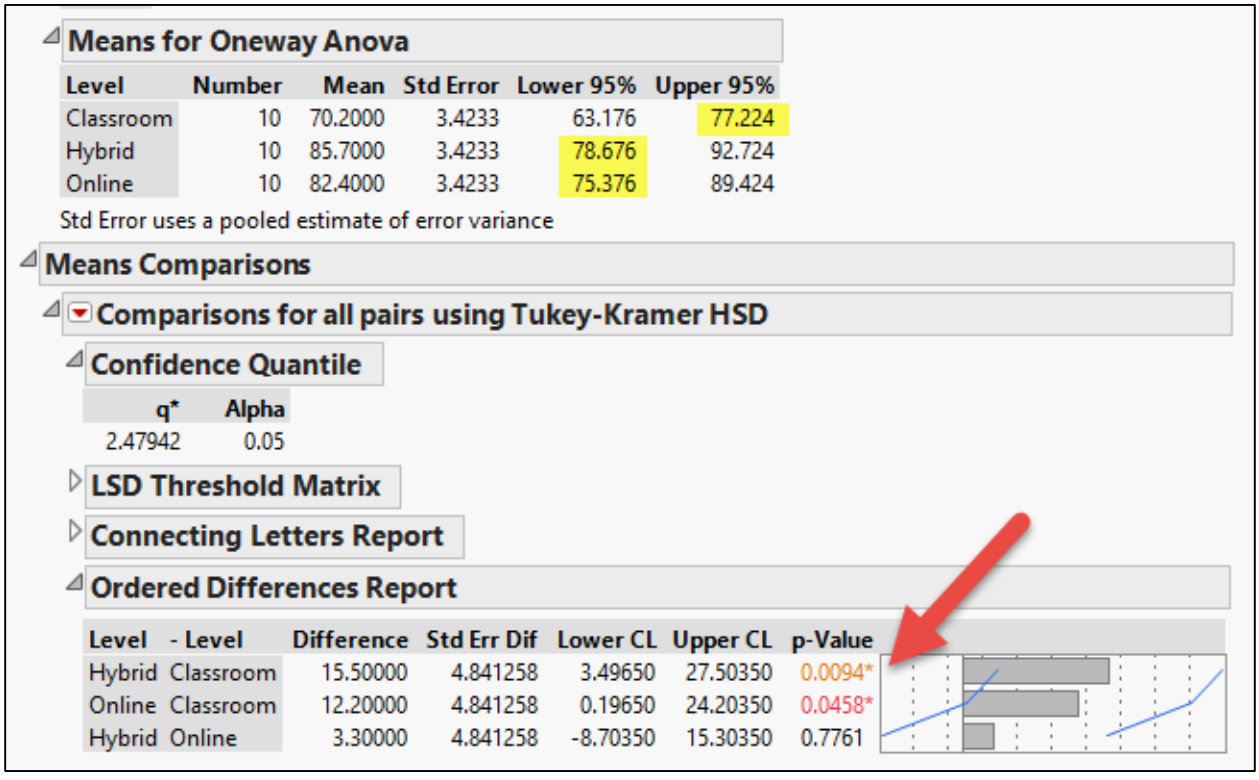



*Figure 1.* Diamond plots of CIs of three groups receiving different treatments

Figure 1 reveals the information about the data as follows:

- Grand sample mean: the overall mean is represented by the horizontal line across the three groups.
- Group means: the horizontal line inside each diamond is the group mean.
- Sample size: The width of the diamond reflects the sample size. In this example the sizes of all three groups are the same.
- Confidence intervals: The diamond represents the CI of each group.

Obviously, there is a large amount of overlap between the diamond of the Hybrid group and the diamond of the Online group; it is therefore concluded that - at the population level, there is no difference between these two groups. In addition, there is no overlap between the Classroom group and the Hybrid group. Using the sample mean as a basis for inference, the best estimate of the Classroom group mean is 77.22 while the worst estimate of the Hybrid group mean is 78.676 (see the yellow highlight in Figure 2). In other words, even the highest test score mean of the Classroom group (by estimation) is lower than the worst test score mean of the Hybrid group (by estimation). In this example, the usage of CIs and hypothesis testing is in agreement. Things become tricky, however, when the Classroom group and the Online group are compared against one another. The lowest estimate of the Online group is 75.76; this number is lower than the highest estimate of the Classroom group (77.22). Since these numbers slightly overlap, this conclusion is not clear-cut.



**Means for Oneway Anova**

| Level | Number | Mean | Std Error | Lower 95% | Upper 95% |
|---|---|---|---|---|---|
| Classroom | 10 | 70.2000 | 3.4233 | 63.176 | 77.224 |
| Hybrid | 10 | 85.7000 | 3.4233 | 78.676 | 92.724 |
| Online | 10 | 82.4000 | 3.4233 | 75.376 | 89.424 |

Std Error uses a pooled estimate of error variance

**Means Comparisons**

**Comparisons for all pairs using Tukey-Kramer HSD**

**Confidence Quantile**

| q* | Alpha |
|---|---|
| 2.47942 | 0.05 |

**LSD Threshold Matrix**

**Connecting Letters Report**

**Ordered Differences Report**

| Level | - Level | Difference | Std Err Dif | Lower CL | Upper CL | p-Value |
|---|---|---|---|---|---|---|
| Hybrid | Classroom | 15.50000 | 4.841258 | 3.49650 | 27.50350 | 0.0094* |
| Online | Classroom | 12.20000 | 4.841258 | 0.19650 | 24.20350 | 0.0458* |
| Hybrid | Online | 3.30000 | 4.841258 | -8.70350 | 15.30350 | 0.7761 |

*Figure 2.* CIs and multiple comparison procedures of the three groups receiving different treatments.

If we look at the Tukey test result (one of the multiple comparison procedures), it is clear that both the Hybrid group and the Online group significantly outperform the Classroom group ($p$ = 0.0094, $p$ =

0.0458, respectively; see the orange and red numbers in Figure 2). In this example, it seems that using CIs alone cannot answer the research question; as a result, reporting *p*-values become necessary. Further, Payton, Greenstone and Schenker (2003) warned researchers that inferring from non-overlapping CIs to significant mean differences is a dangerous practice, because the error rates associated with these types of comparisons tend to be quite large. The probability of overlap is a function of the standard error. As the standard errors become less homogeneous, the probability of overlap decreases. Results of simulations indicated that when standard errors are approximately equal, using 83% or 84% size for the intervals will yield an approximate alpha = 0.05 test, while using 95% confidence intervals (which is a common practice), will yield very conservative results.

This is a problem if and only if a researcher wants to obtain a clear-cut conclusion (e.g. 'yes' or 'no') based on the frequentist approach of probability (e.g. How likely can we observe the test statistics at hand in the long run, assuming that the same study is replicated multiple times?) Nevertheless, the very reason of using CI is to yield a range of possible answers, rather than a single answer. Second, the notion of repeating a same multiple times is questionable. The results yielded from hypothesis testing vary from study to study; scientists are concerned with the inability of replication. A common misinterpretation of significance is that *p*-values tells you how "right" conclusions are (i.e. If *a p*-value is .01, there is only a 1% chance that a conclusion is wrong). Actually, a *p*-value of .01 corresponds to a Type I error rate (false alarm) of 11%, and a *p*-value of .05 is equivalent to an error rate of 29%. Given the fact that the result of one particular study could be wrong, it is not surprising to see that researchers are often unable to replicate original results (Nuzzo, 2014). Indeed, CIs can be interpreted according to both the Bayesian and frequentist approaches.

A Fisherian who subscribes to the objective, frequentist philosophy interprets a CI as, "Given the choice of $z = \pm 1.96$, for every 100 samples drawn, 95 of them will capture the population parameter within the bracket." According to the objectivist school of probability, the population parameter is constant, and there is only one true value in the population.

However, in the view of Bayesians, the same CI can be interpreted as "given the choice of $z = \pm 1.96$, a researcher is 95% confident that the population parameter is bracketed by the CI." It is important to note that in the second interpretation "confidence" becomes a subjective, psychological property. In addition, Bayesians do not treat the population parameter as a constant or true value.

As mentioned before, hypothesis testing is based upon probability, which is defined as a relative frequency in the long run. However, it is problematic to apply this frequentist view of probability to a single event that is not repeatable or to a new event that has no comparable events (Carver, 1978). In the real world, the subjective approach usually makes more sense. For example, if someone asks a man whether his wife loves him, he can employ a subjective approach to say, "I am 95% sure that my wife really loves me." If the objective approach is used, then the answer is: "If I marry a woman similar to my wife 100 times, 95 of them would really love me."

**LogWorth in data mining**

Today, researchers are often swamped by large-scale data sets. As mentioned previously, when conventional statistical analyses are used, even trivial differences may mistakenly be reported as significant. Take the Programme for International Student Assessment (PISA) as an example. The sample from North America only (US and Canada) consists of over twenty thousand students. Needless to say,

using regression analysis (to identify factors contributing to PISA science test performance) would be problematic. In this case, the recursive partition tree - also known as the 'classification tree' or 'decision tree,' would be a better alternative.

In this example, students were classified into two groups (i.e. 'proficient' or 'not proficient') based on their abilities, as estimated by Item Response Theory. Thirty-five independent variables were used to predict the preceding outcome variable. The partition tree indicated that the 'degree to which student enjoyed science' was the most important predictor of PISA science test performances, whereas the 'number of books at home' was the second most important predictor. In this data set the degree of science enjoyment was represented by a composite score of many Likert-scaled items; therefore, the variable was treated as continuous, while the variable 'number of books' was treated as categorical (0-10, 11-50, 51-100, 101-200…etc.).

In partition trees the splitting criterion is LogWorth statistics. It may be surprising to some readers that LogWorth statistics is based on *p*-values! The classification tree examines each independent variable, in order to identify ones that can decisively split a sample, with reference to the dependent variable. If the input is continuous (e.g. the degree of enjoyment of science), every value of the variable could be a potential split point. If the input is categorical (e.g. the number of books), then the average value of the outcome variable is taken for each level of the predictor variable. Afterwards, two sub-groups demarcated by the split point are generated, and a 2x2 crosstab table is formed (e.g. ['proficient' or 'not proficient'] x ['enjoy science more' or 'enjoy science less']). Next, Pearson's Chi-square is used for examining the association between the two variables. As mentioned before, the result of Chi-square is dependent on the sample size. When the sample size is extremely large, the *p*-value is close to 0 and virtually everything appears to be 'significant.' As a remedy, the quality of the split is reported by LogWorth, which is defined as –log10(p). Because the LogWorth statistics is the inverse of the *p* value, a bigger LogWorth is considered better. If the outcome variable is categorical, G^2, (the likelihood ratio of chi-square), is also reported (Klimberg & McCullough, 2013).

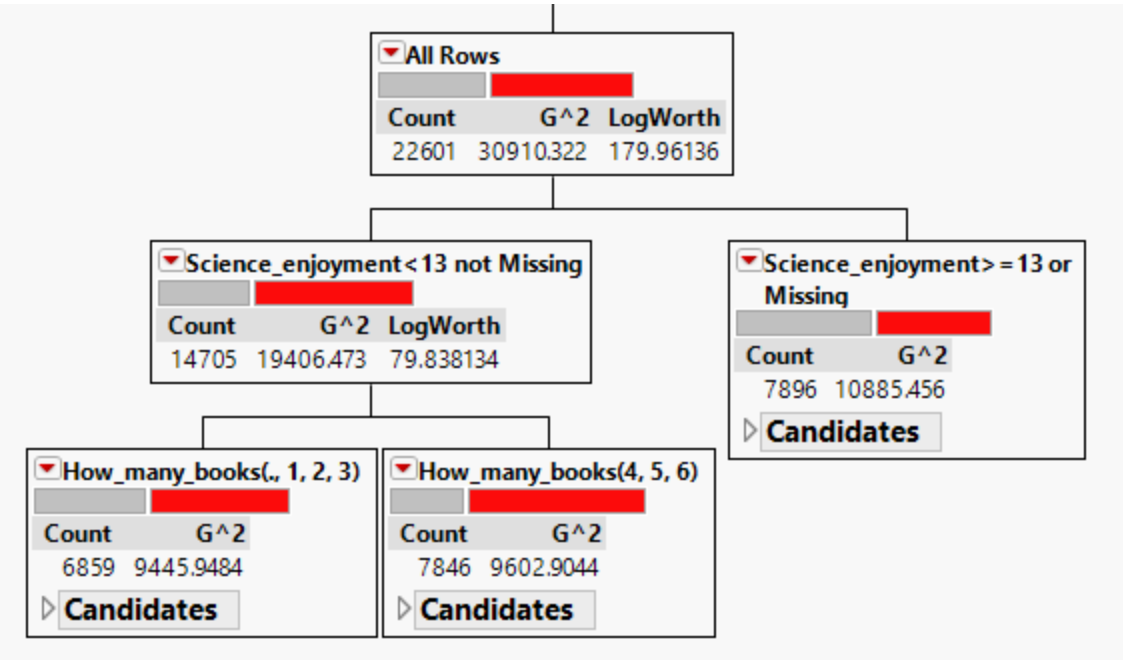


*Figure 3.* Partition tree for identifying factors contributing to better PISA scores.

Surprisingly, the idea of inversing the *p*-value was introduced by R. A. Fisher, the inventor of significance testing! The original form used by Fisher was -2log10(p); this is known as the Fisher's method or the Fisher's combined probability test (Elston, 1991). This method is commonly used in meta-analysis, in which results from several independent studies are combined, in order to obtain an overall picture of the issue under study. Unlike the *p* value in which certain alpha levels ($p < .1$, .05, or .01) are used, there is no cutoff in LogWorth. JMP automatically checks all possible split points of all predictors in order to maximize the LogWorth statistics (Klimberg & McCullough, 2013). Simply put, bigger is better. Nonetheless, there is no contradiction between LogWorth and *p*-values. As Nuzzo (2014) states, "The irony is that when UK statistician Ronald Fisher introduced the *p* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look. The idea was to run an experiment, then see if the results were consistent with what random chance might produce" (pp.150-151). If Fisher were alive today, he would not have rejected the exploratory principle of data mining.

**Conclusion**

As demonstrated in the preceding sections, there are both incompatibility and continuity between conventional and alternate procedures. Usage of confidence intervals occasionally yields the same conclusions as hypothesis testing. Nonetheless, CI can be interpreted in terms of the frequentist approach or in terms of the Bayesian approach. Conventionalists are happy to embrace the former, while revolutionists tend to accept the latter. In the partition tree the same pattern is observed. On the one hand, the data-driven decision tree that focuses on pattern-recognition seems to be at odds with hypothesis-driven significance testing. On the other hand, the mechanism driving the splitting process in the decision tree – LogWorth - is based on *p*-values. Reviewers often reject manuscripts simply because data miners do not include *p-values*. However, in the process of data mining, transformed *p*-values are everywhere. What can JMP do for you if hypothesis testing is banned? You can conduct research as usual by - reporting CIs and running data mining.

**References**

Behrens, J. T., & Yu, C. H. (2003). *Exploratory data analysis*. In J. A. Schinka & W. F. Velicer, (Eds.), *Handbook of psychology Volume 2: Research methods in psychology* (pp. 33-64). New Jersey: John Wiley & Sons, Inc.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge

Cumming, G. (2015, May). *The new statistics and open science in practice: Estimation, meta-analysis and research integrity*. Workshop at the Western Psychological Association Conference, Las Vegas, NV.

Elston, R. C. (1991). On Fisher's method of combining *p*-values. *Biometrical Journal, 3*, 339-345. doi: 10.1002/bimj.4710330314.

Frost, J. (2015). Understanding hypothesis tests: Confidence intervals and confidence levels. *The Minitab Blog.* Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/understanding-hypothesis-tests%3A-confidence-intervals-and-confidence-levels

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What If there were no significance tests?* New York, NY: Psychology Press.

Nuzzo, R. (2014) Statistical errors: *P*-values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature, 506*, 150-152.

Novella, S. (2015). Psychology journal bans significance testing. *Science-based Medicine*. Retrieved from https://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/

Klimberg, R., & McCullough, B. D. (2013). *Fundamentals of predictive analytics with JMP*. Cary, NC: SAS Institute

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-170.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3*(34). Retrieved from http://insectscience.org/3.34

Wilkinson, L, & the Task Force on Statistical Inference. (1996). *Statistical methods in psychology journals: Guidelines and explanations*. Retrieved from http://www.apa.org/science/leadership/bsa/statistical/tfsi-followup-report.pdf

Woolston, C. (2015). Psychology journal bans *p*-values. *Nature, 519*(9). doi:10.1038/519009f

Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research, 3*(1), 9-22. Retrieved from http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/download/455/460

Yu, C. H. (2014). *Dancing with the data: The art and science of data visualization*. Saarbrucken, Germany: LAP.

**Correspondence:**

Chong Ho Yu, Ph.D., D. Phil.
Associate Professor
Department of Psychology
Azusa Pacific University
chonghoyu@gamil.com