Tourism Management 62 (2017) 89–96

Contents lists available at ScienceDirect

Tourism Management

journal homepage: www.elsevier.com/locate/tourman

Research note

Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research

Jalayer Khalilzadeh[*], Asli D.A. Tasci

Rosen College of Hospitality Management, University of Central Florida, 9907 Universal Blvd, Orlando, FL 32819, USA

HIGHLIGHTS

- Discusses the issue of significance in statistical versus practical sense when using big data.
- Summarizes different effect size metrics for practical significance in social sciences.
- Exemplifies effect size metrics commonly used when comparing two groups (*t*-test).
- Lists commonly used effect size metrics for test of associations (correlation and regression).
- Explains commonly used effect size measures for association of nominal variables (chi-square test).

ARTICLE INFO

Article history:
Received 21 October 2016
Received in revised form
28 March 2017
Accepted 30 March 2017

> *"The world is one big data problem."*
>
> Andrew McAfee

1. Increasing popularity of big data

The origins of big data started with companies collecting mass amounts of information related to their metrics in order to make strategic decisions related to market intelligence. Following advancements in information technologies in the early 2000s, the size, speed, and variety of stored information has increased exponentially in recent years, hailing the terminology of *big data* as another asset for companies and businesses in different industries. As the importance of information for competitive advantage has been realized, not only businesses, but also non-profit organizations, destination marketing organizations, associations, and governments have become interested in collecting big data to allow for better informed strategies and actions. Big data accumulated over a period of time can be harnessed to help in better branding, gaining a sustainable competitive advantage, increasing returns on investment, innovation, new product development, avoiding risks, assuring customer satisfaction and loyalty, adjusting to the volatile market environment that changes by the minute, and identifying issues and their causes, thereby increasing the overall effectiveness, efficiency, and productivity of an entity.

Gathering big data mainly appears to be treated as an end in itself, however, rather than as a means to an end, since much of the available data has not been mined or analyzed for different reasons, such as proprietary concerns, a lack of understanding of the importance of information, or a lack of analysis skills. Proprietary concerns are the typical reasons given for the lack of academic access to big data related to different fields of research. Academicians typically do not have access to the big data that have been streamed by companies' IT systems through the web of diverse connected devices. For example, Fuchs, Hopken, and Lexhagen (2014) are some of the few academics privileged with access to the Destination Management Information System Åre, a knowledge infrastructure implemented by a Swedish mountain tourism destination.

Nevertheless, all academicians have access to big data available on social media platforms and other publicly-available sources, such as governments (e.g. CIA World Fact book), regional organizations (e.g. European Union Open Data Portal), consumer rating sites (e.g. TripAdvisor), meta search engines (e.g. Kayak, Trivago), and Online Travel Agents (e.g. Expedia). Much of such data is qualitative, but can be handled quantitatively by applying content analysis. For example, Xiang, Schwartz, Gerdes, and Uysal (2015) used publicly-available consumer feedback and ratings on Expedia to analyze guest experience and satisfaction. In addition to such publicly-available sources, academicians also create their own

[*] Corresponding author.
E-mail addresses: Jalayer.Khalilzadeh@ucf.edu (J. Khalilzadeh), Asli.Tasci@ucf.edu (A.D.A. Tasci).

http://dx.doi.org/10.1016/j.tourman.2017.03.026

systems of big data related to their expertise, sometimes through their partnerships with industry sponsors and funding agencies.

2. Issues of using big data

One of the most important issues when using big data is the large sample size. There are different ways of analyzing big data, some of which are conducted by taking small samples from the data; however, even these so-called 'small samples' can be very large from a statistical point of view. Nonetheless, the concept of large sample size appears to be relative. Lin, Lucas, and Shmueli (2013) considered sample sizes over 10,000 cases to be large. In a study where the authors used approximately 2500 cases, however, they received reviewer requests for effect size measures along with the p-values of statistical tests. A conservative approach may consider sample sizes of over 300 to be large enough to demand effect size measures for practical significance, along with statistical significance. It is therefore apparent that different researchers have different levels of awareness and approaches to the threshold of a large sample.

Using large samples offers both advantages and disadvantages. One advantage is that large samples allow the discovery of rare associations or rare events that cannot be revealed by small samples, as they are sufficient for discovering only average behavior (Lin et al., 2013, p. 914). Identifying such rare associations by using big data may reveal important insights for theoretical and managerial purposes. Conversely, a disadvantage is the issue of p-values approaching zero, guaranteeing statistical significance. To be more specific, as the sample size grows and become closer to the real-world population size, the power of the test also increases, identifying small, impractical effects. Effect size reports the practical significance, whereas the p-value reports the statistical significance (Chatfield, 1995). There are many different effect size tests for validating statistical significance of results obtained by different tests. The purpose of this paper is to provide a guide for researchers concerning appropriate effect size measures for different types of tests. The following section summarizes the most commonly-used effect size measures for validating statistically significant results of different tests, along with analysis software with the appropriate test tools for effect size tests, and cautions concerning interpreting effect size test results.

3. The root cause of the 'large sample size' issue

As in many other disciplines and fields of scientific inquiry, null-hypothesis (H_0) significance testing (NHST) is the most commonly used and abused approach to statistical analysis in tourism and hospitality. Much criticism has been directed toward the utilization of NHST (e.g. Nickerson, 2011). In the NHST process, the test yields a statistic and a probability (p-value). The obtained p-value denotes the probability of which the statistic's value and its larger values has been obtained by mere chance. Conventionally, the H_0 is rejected when the p-value is less than or equal to the Type I error probability's (known as α or alpha) criterion value, typically set at 0.05 or 0.01 in social sciences (Nickerson, 2011). Strictly speaking, the decision to 'accept or support' the H_0 is not made in most practical situations since data favoring H_0 may take on other values, thus making H_0 false. As a result, the appropriate language is 'to reject H_0' or 'fail to reject H_0'. Theoretically, four outcomes are possible: (1) rejecting a non-false H_0, (2) failing to reject a false H_0, (3) failing to reject a non-false H_0, and (4) rejecting a false H_0 in favor of the alternative hypothesis. The probabilities of each of these decisions are estimated, respectively, by (1) the probability of Type I error (also known as α error), (2) the probability of Type II error (also known as β error), 3) $1-\alpha$, whose interpretation is similar to 'confidence' in a 'confidence interval', and 4) $1-\beta$, also known as the power of the test. Note that the decisions described in 1 and 2 are errors, and the decisions described in 3 and 4 are correct decisions. Hence, Type I error is commonly considered to be more dangerous than Type II error (Nickerson, 2011). There are many misconceptions about NHST (e.g. Nickerson, 2011); one common misconception is that a small p-value means there is a strong treatment effect. A p-value does not provide any information on the magnitude of the effect, the practical significance of the relationship, or any differences identified by the statistical analysis. In fact, nearly any H_0 can be rejected if appropriate precision and a large enough sample size are selected.

For example, in comparing the averages of two independent groups (*t*-test), if researchers want to test with a power $(1-\beta)$ of 0.80, a confidence level of 0.95, and equal group sizes, a sample size of 46 (23 per group) would be large enough if the effect size is approximately 0.85 (strong effect; assuming μ_1 and μ_2 to be the means of two groups and σ to be the common standard deviation in the population underlying the groups, the effect size is the difference in true means adjusted for the standard deviation, or $|\mu_{1-}\mu_2|/\sigma$). Assuming the exact same scenario, but this time with an effect size of 0.15 (weak effect), a sample size of 1398 (699 per group) is needed to achieve a test power of 0.80. In other words, if the effect size is 0.85 with a sample size of 23 in each group, there is enough test power $(1-\beta)$ to reject H_0 (see Fig. 1a). If, however, the effect size is 0.15 with the same sample size ($n = 46$), the power of the test will decline significantly (see Fig. 1b). Simply put, even when the impact of an independent variable is negligible, by using a large enough sample size, researchers can achieve enough power to reject H_0 in favor of the alternative hypothesis (H_A). The relationship between sample size and the required effect size for different power levels is displayed in Fig. 2. For example, with a sample size of 50, if we investigate a relationship with an effect size of 0.80, the test power $(1-\beta)$ to reject the null hypothesis will be approximately 0.85, whereas with the same sample size, if we investigate a relationship with an effect size of 0.20, the test power $(1-\beta)$ to reject the null hypothesis will be approximately 0.10.

The example above shows the importance of considering effect size when dealing with big data. When analyzing or mining big data, a large enough sample size compels researchers to ensure that the statistically significant effect inferred from the sample is substantial enough to be considered practically significant. This issue may not apply to a situation where a company's big data on its entire consumer population (or any other unit of analysis) is being analyzed in order to generate insights specifically for the consumers of this specific company. The effect size issue does apply, however, when big data are analyzed or mined to be generalized to a larger population beyond the sample at hand.

4. Measures to avoid the 'large sample size' issue

In the behavioral sciences, effect size, or standardized mean differences, is one of the measures of the magnitude of the effect (Kirk, 2005). Effect size has now, however, been substituted for the magnitude of the effect. Hence, in this manuscript, effect size is used in this broad sense, covering all effect magnitude measures. In contrast to p-value, or the statistical significance, effect size, or the practical significance (Ellis & Steyn, 2003), provides the magnitude of the effect identified in a statistical test (Grissom & Kim, 2005). Effect size is less biased toward sample size; hence, it is more reliable in the case of a large sample size, where the likelihood of finding a statistically significant result increases enormously. There are more than 65 different measures for identifying effect size. Table 1 provides a comprehensive list of common effect sizes used in the behavioral sciences.