Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 20

9-2-2020

Almost All Missing Data Are MNAR

Thomas R. Knapp
University of Rochester, tomknapp829@gmail.com

Recommended Citation

Knapp, T. R. (2019). Almost all missing data are MNAR. Journal of Modern Applied Statistical Methods, 18(2), eP3523. doi: 10.22237/jmasm/1594045320

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 2, eP3523.
doi: 10.22237/jmasm/1594045320

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

Almost All Missing Data Are MNAR

Thomas R. Knapp
University of Rochester
Rochester, NY

Rubin (1976, and elsewhere) claimed that there are three kinds of “missingness”: missing completely at random; missing at random; and missing not at random. He gave examples of each. The article that now follows takes an opposing view by arguing that almost all missing data are missing not at random.

Keywords: Donald Rubin, missing data, interesting examples

Introduction

One of the most frustrating problems in data analysis is the absence of one or more pieces of data. Researchers usually go to great lengths in designing their studies, choosing the appropriate measuring instruments, drawing their samples, and collecting their data, only to discover that some of the observations are missing, due to a variety of reasons (an item on a questionnaire left blank, a clerk's neglect to enter an important piece of information, a data-recording instrument's failure, etc.). How do you cope with such a problem? The literature suggests that there are essentially two strategies--deletion or imputation. You can delete all or some of the non-missing data for the entities for which any data are missing; or you can try to impute (estimate) what the missing data "would have been".

When are data missing at random? My opinion is "almost never." Nevertheless, Rubin (1976) defined three kinds of "missingness" (see also Little & Rubin, 2002):

1. Missing at random (MAR). Data are said to be MAR if the distribution of missingness does not depend upon the actual values of the missing data (i.e, what they would have been).

doi: 10.22237/jmasm/1594045320 | Accepted: April 30, 2018; Published: September 2, 2020.
Correspondence: Thomas R. Knapp, tomknapp829@gmail.com

2. Missing completely at random (MCAR). Data are said to be MCAR if the distribution of missingness also does not depend upon the actual values of the other data that are not missing.
3. Missing not at random (MNAR). Data are said to be MNAR if the distribution of missingness does depend upon the actual values of the missing data.

What is the meaning of the word "random"? Kac (1983) argued most scientists never define random. It is taken for granted everyone knows what it means, even though something being "random" is actually complicated. May (1997) referred to Kac, and so too Beltrami (1999) noted Kac (1983) "prompted the title of this book." (p. 146)

Product vs. Process

Rubin's definitions of types of missingness are "product-oriented" rather than "process-oriented", i.e., one needs to make certain assumptions and/or analyze some actual evidence in order to determine whether or not, or the extent to which, data might be missing "at random". That view is contrary to mine. It also appears to be contrary to the concept of a random phenomenon, where chance and chance alone plays the essential role. People don't flip coins, roll dice, or draw cards in order to determine whether or not they will respond to a particular item on a questionnaire. Data entry clerks don't employ such devices in order to determine whether or not they will enter a participant's response in a data file. And recording instruments don't choose random times to break down. Do they? And how can you analyze non-missing data to draw conclusions regarding the randomness or non-randomness of missing data?

Whether randomness pertains to a process or a product is a controversial matter, as Bennett (1998) explained (see p. 165-172); also see Eagle (2005) for a discussion of process vs. product. Those who claim something is random if it has been determined by a chance process appeal to some property of the object used in the process (e.g., the balance of a coin and a die) and/or the mixing mechanism for the process (e.g., the shuffling of a deck of cards). Others treat randomness as a characteristic of a product, and they claim that it is the product that must be judged to be random or non-random. Beltrami (1999, p. xiii) stated there was a general shift in the last several decades from randomness-as-process to randomness-as-product. That may be true, but some very strong cases have been made for randomness-as-process. See, for example, Keren and Lewis (1993).

A middle-ground approach is to argue that randomness pertains to the basic product of the alleged randomness-generating process. It is those processes that must be judged to be random or non-random by virtue of the basic products they generate. Once a process has been deemed to be random, any data gathered as a result of the use of such a process need not be judged to be random or non-random.

Consider the following example: You would like to estimate the standard deviation of the heights of adult males. You have identified a target population of 1000 adult males, with associated ID numbers of 000 to 999, and you plan to draw a sample of size 50. You decide to use Minitab's random sampling routine (the process). You give it the proper commands and ask it to print the 50 sampled ID numbers. You indicate whether each of those numbers is less than or equal to 499 (and call those 0) or greater than 499 (and call those 1). You subject that string of 1s and 0s (the product) to one or more "tests of randomness". Let's say that the string passes that (those) test(s), i.e., it is declared "random". You then measure the heights of the 50 adult males, record them, calculate their standard deviation, and make whatever inference to the population of 1000 adult males is warranted. The heights of the 50 men in the sample need not be subject to any test of randomness (The argument of Siegel & Castellan, 1988; to the contrary notwithstanding)--neither the randomness of their actual magnitudes nor the randomness of the order in which they were drawn--because the process has already been judged to be random. You may wind up with a poor estimate of the standard deviation of the heights of all 1000 adult males in the sampled population, but that is a separate matter.

Rubin's Reflections on the Publication of His 1976 Article

It doesn't seem to be well-known that Rubin had considerable difficulty in getting that article published, as he noted

> “This article is extremely well known because it established the basic terminology for missing data situations, which is now so standard that this paper often isn’t cited for originating the ideas, although often the definitions are summarized somewhat incorrectly. As Molenberghs (2007) wrote: “... it is fair to say that the advent of missing data methodology as a genuine field within statistics, with its proper terminology, taxonomy, notation and body of results, was initiated by Rubin’s (1976) landmark paper.” But was this a bear to get published! It was rejected, I think twice, from both sides of JASA; also from JRSS

> B and I believe JRSS A. I then decided to make it more “mathy,” and I put in all this measure theory “window dressing” (a.s., a.e., both with respect to different measures because I was doing Bayesian repeated sampling and likelihood inference). Then it got rejected twice from The Annals of Statistics, where I thought I had a chance because I knew the Editor — knowing important people doesn’t always help. But when I told him my woes after the second and final rejection from The Annals, and I asked his advice on where I should send it next, he suggested “Yiddish Weekly” — what a great guy! But I did not give up even though all the comments I received were very negative; but to me, these comments were also very confused and very wrong. So I tried Biometrika — home run! David Cox liked it very much, and he gave it to his PhD student, Rod Little, to read and to contribute a formal comment. All those prior rejections created not only a wonderful publication, but lead [sic] to two wonderful friendships. The only real comment David had as the Editor was to eliminate all that measure theory noise, not because it was wrong but rather because it just added clutter to important ideas. Two important messages: First, persevere if you think that you have something important to say, especially if the current reviewers seem not up to speed. Second, try to find a sympathetic audience, and do not give up." (Rubin, 2014)

An Example of a Situation where Data Might be Missing at Random

Consider a high school student taking a multiple-choice examination which is extremely difficult for him and for which there is a penalty for guessing. When he comes to an item for which he has no idea of the correct answer he has two possible strategies: (1) guess, hope he has guessed correctly, and therefore doesn't suffer the penalty; or (2) omit the item. He decides to omit the item. Datum missing at random!

References

Beltrami, E. (1999). *What is random? Chance and order in mathematics and life*. New York: Springer-Verlag. doi: 10.1007/978-1-4612-1472-4

Bennett, D. J. (1998). *Randomness*. Cambridge, MA: Harvard University Press.

Eagle, A. (2005). Randomness is unpredictability. *The British Journal for the Philosophy of Science, 56*(4), 749-790. doi: 10.1093/bjps/axi138

Kac, M. (1983). Marginalia: What is random? *American Scientist, 71*(4), 405-406.

Keren, G., & Lewis, C. (Eds.). (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum. doi: 10.4324/9781315799582

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd edition). New York: Wiley. doi: 10.1002/9781119013563

May, M. (1997). What is random? *American Scientist, 85*(3), 222-223.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (2014). Converting rejections into positive stimuli. In X. Lin, C. Genest, D. L. Bamnks, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds), *Past, present, and future of statistical science* (pp. 593-603). New York: CRC Press.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd edition). New York: McGraw-Hill.