

Characterizing Patterns of Longitudinal Data Completeness through Successive Refinement

Lawrence Rasouliyan and David J. Pasta
Ovation Research Group, San Francisco, California

ABSTRACT

The analysis of longitudinal data with multiple assessment periods provides many challenges with respect to completeness and usability of data. Patterns of data completeness over time can vary greatly, and the number of possible longitudinal patterns increases exponentially with the number of assessment periods. With a great number of possible completeness patterns, questions arise as to which patterns have enough complete data to be usable. Some patterns occur much more commonly than others, and through the process of successive refinement, rules can be developed to accommodate these more common patterns of data completeness. Common patterns of complete data include “pure dropout” cases, “a little missing” cases, and “barely any data” cases. These cases will be explored and examples provided. Possible rules will be described to help determine where to draw the line with respect to the completeness of data and its subsequent usability.

INTRODUCTION

Longitudinal study designs are extremely important in numerous areas of interest, particularly in the analysis of health-related data. From observing patients to see if they experience an adverse event after surgery to determining the incidence of a particular disease in a particular population to monitoring adherence to a therapeutic regimen, longitudinal methods are applicable in any study where subjects are followed through time.

Longitudinal studies can be as simple as the assessment of subjects at baseline (t_0) and at some later point in time (t_1). In these situations, when data at t_1 are missing, the subject is usually classified as being lost to followup, and the observation is typically thrown out of all analyses that require followup data. However, suppose that, instead of having two assessment periods, the study has 6, or 12, or 20 assessment periods. How should we treat an observation that has complete data at some assessments but has missing data at others? The situation becomes a little more complex.

If we assume that at any particular assessment, data can either be complete or missing and denoted by values of 0 or 1 respectively, then we can use binary notation to construct descriptions of data completeness patterns. For instance, a subject with 6 assessment periods who only has complete data at the first and third assessments would be described by a completeness pattern of 101000. For every additional assessment period present, the number of possible longitudinal completeness combinations increases exponentially. Assuming that baseline data is always present, for n assessment periods, there are 2^{n-1} possible combinations of completeness.

Since such a large number of combinations are possible in studies with several assessment periods, decisions need to be made with respect to where to draw the line in terms of data usability. Exactly how much data needs to be present for an observation to be included? Should the order of complete data in time matter? Through the process of successive refinement, we can classify certain patterns of completeness that occur more frequently than others, and we can develop possible rules to accommodate these more common patterns.

SUCCESSIVE REFINEMENT

Successive refinement describes the process by which a crude set of information is gradually explored and its components classified until it reaches a more developed and understandable state. This outcome is often achieved by starting with the most extreme cases of data, classifying them into groups, and gradually working towards the middle. This logic can be applied to classifying longitudinal patterns of data

completeness. We could start with cases where data at all followup assessments are either 100% complete or 100% missing, then work our way towards the cases that have a mix of complete and missing data at followup assessments.

EXAMPLE

Consider a study designed to collect 6 quarters of data for each subject. Assuming that each subject must have complete baseline data to be included in the study, there are $2^{(6-1)}$, or 32, possible combinations of longitudinal completeness patterns over the 6 assessment periods. Below is a list of all possible combinations along with preliminary classifications:

100000	Immediate dropout		
100001		Barely any data	
100010		Barely any data	
100011			ROD
100100		Barely any data	
100101			ROD
100110			ROD
100111			ROD
101000		Barely any data	
101001			ROD
101010			ROD
101011			ROD
101100			ROD
101101			ROD
101110			ROD
101111			A little missing
110000			Pure dropout
110001			ROD
110010			ROD
110011			ROD
110100			ROD
110101			ROD
110110			ROD
110111			A little missing
111000			Pure dropout
111001			ROD
111010			ROD
111011			A little missing
111100			Pure dropout
111101			A little missing
111110			Pure dropout
111111			Complete data

Beginning with the most extreme cases, we can start classifying patterns from the “complete” end and from the “missing” end. That is, data can be complete at all followup assessments corresponding to a longitudinal pattern of 111111 (complete data), or they can be missing at all followup assessments corresponding to a pattern of 100000 (immediate dropout). These two extremes belong in classes of their own. For almost all purposes, we would want to keep the observations exhibiting the complete data pattern, and we would want to omit the observations exhibiting the immediate dropout pattern, as they have no usability with respect to longitudinal comparisons.

Working towards the middle from the complete end, we have cases where data are complete at all assessments, except for the last. Exhibiting a pattern of 111110, subjects with this longitudinal combination may have dropped out of the study or been lost to followup before the last assessment period. Since the data are complete at consecutive assessments up until a point in time after which data are missing, this pattern is classified as a “pure dropout.” The other pure dropout patterns in this example are 111100, 111000, and 110000. These patterns can be further sub-classified corresponding to the first assessment at which data become missing (i.e. late pure dropout, early pure dropout, intermediate pure drop out).

From the missing end, cases arise where data are missing at all but one of the followup assessments. Subjects exhibiting these completeness patterns barely have any followup data; hence, they are classified as “barely any data.” Barely any data patterns include 100001, 100010, 100100, and 101000. Of course, in other examples where the number of assessments may be very large, other rules of thumb can be developed to make the barely any data pattern classifications more inclusive. For instance, a rule could be made to classify the barely any data pattern as having complete data at up to 20% of followup assessments.

Analogous to the barely any data completeness patterns on the other end of the spectrum are the “a little missing” patterns. With these patterns, data are missing at only one of the followup assessments, while the rest of the assessments have complete data. Patterns classified as a little missing include 111101, 111011, 110111, and 101111. As with the barely any data classification, an “up to 20%” rule of missing assessments can be implemented to define a little missing pattern for studies with very large numbers of assessments.

Almost half of all completeness patterns in this example have been classified into one of the preceding categories. The rest of the data (or ROD) represent a more generic class of patterns that do not fit within any of these categories. The ROD patterns are characterized through inspection of the five followup assessments; they have either two 1s and three 0s or two 0s and three 1s. These followup assessments can also be used to further distinguish different combinations from one another. ROD patterns can be classified by the number of runs in the followup period, that is, the number of times a 1 precedes a 0 and vice versa. The below table classifies the followup ROD completeness patterns according to corresponding number of runs for the followup assessment periods.

Number of runs in followup period	Longitudinal ROD completeness patterns in followup period
2	00011, 00111
3	00110, 01100, 10001, 01110, 10011, 11001
4	00101, 01001, 10010, 10100, 01011, 01101, 10110, 11010
5	01010, 10101

Hence, beginning with 32 possible combinations of longitudinal completeness, we were able to employ successive refinement to classify these patterns into more structured groups. After characterization of the patterns, group frequencies can be examined and rules developed to determine where to draw the line in terms of data usability.

CHARACTERIZING LONGITUDINAL DATA WITH SEVERAL ASSESSMENT PERIODS

The principles presented in the preceding example are also applicable to study designs with many more assessments. Suppose a study is designed to collect weekly data on subjects for six months, yielding 26 assessment periods. Assuming complete data at baseline for inclusion in the study, the number of possible completeness patterns is enormous: $2^{(26-1)}$, or over 33.5 million combinations. It would take several lifetimes to list and classify all possible combinations. We could begin with the basic concepts of successive refinement and characterize complete data, immediate dropouts, pure dropouts, etc.; however, more generalized approaches would be necessary to characterize the bulk of longitudinal patterns. Furthermore, the classification and determination of subsequent usability would depend on the objectives of the study.

One approach would be to examine the longitudinal pattern in smaller sections and to classify these sections in terms of the previous concepts. For instance, consider the following possible completeness pattern: 11111011110011001000000000. The section from the 1st to the 10th assessment follows the “a little

missing” pattern, while the section from the 17th assessment onward follows the “immediate drop out” pattern. The section between these two appears as a ROD and could be examined more closely:

1 1 1 1 0 1 1 1 1	0 0 1 1 0 0 1	0 0 0 0 0 0 0 0 0
A little missing	ROD	Immediate dropout

Another approach would be starting with the last assessment and looking backwards in time towards baseline. Similar concepts could be applied in backwards examination as they were in forward examination to classify patterns. For instance, with the preceding pattern, we observe that the first time a 1 appears going backwards is at the 10th from the last visit. We could also calculate the percent missing assessments from this point towards baseline (in this case, that value would be 5/17 = 29%). After examining the distribution of this percent missing until last complete visit, further groups could be formed.

DISTINGUISHING POTENTIAL DROPOUTS

A potential dropout is defined as a subject who has missing data at the last assessment. That is, if a subject does not have complete data at the final assessment period, he potentially could have left the study permanently. However, the absence of data at the last assessment does not necessarily mean that the subject definitely dropped out of the study. The likelihood of such a subject being a drop out depends on the pattern of completeness preceding the final assessment. How can we further distinguish a potential dropout as a “likely dropout” or an “unlikely dropout?”

EXAMPLE

Consider a study designed to collect quarterly data on subjects for 5 years. The following SAS® code was used to simulate a random sample of 10,000 subjects with 21 assessment periods:

```
data simulate;
  attrib pattern length=21;
  array x (0:20) x0-x20;
  retain p1 0.05;
  do k = 1 to 10000;
    p2 = 0.1+ranuni(2342)*0.3;
    x(0) = 1;
    drop=0;
    pattern=put(x(0),6.0);
    do i = 1 to 20;
      if (ranuni(321) le p1) then drop=1;
      if (drop eq 1) then x(i)=0;
      else if (ranuni(528) le p2) then x(i)=0;
      else x(i) = 1;
      pattern = compress(pattern) || compress(put(x(i),6.0));
    end;
    output;
  end;
run;
```

For each simulated study subject, the above code generates a complete pattern for which there exists a probability (p1) of 5% that any specific assessment and all subsequent assessments are missing; that is, the subject drops completely from the study. The value p2, ranging from 10% to 40%, is randomly generated for each simulated study subject and corresponds to the probability that any particular assessment itself is missing. An indicator variable named drop, which is initialized at 0, is created for each observation and denotes whether a subject has truly dropped out of the study.

Based on the number of 1s and 0s in any given pattern, we can calculate the expected number of runs and compare that to the observed number of runs. Through the runs test (Dudley 1997, Siegal 1956,

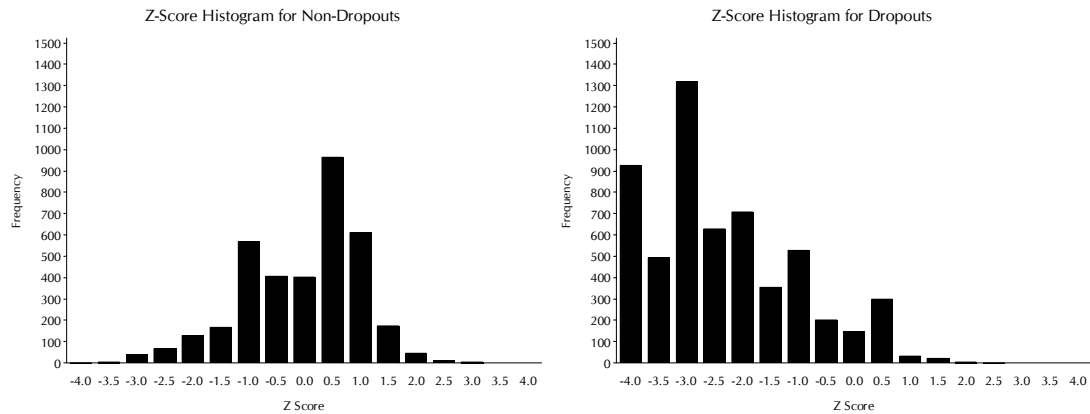
Mendenhall et al. 1986), a z statistic can be calculated, which we can use to help classify likely versus unlikely dropouts. The following code was used to calculate the runs test z scores for the simulated data:

```

data runsdata;
  set simulate;
  array x (0:20) x0-x20;
  array r (0:20) r0-r20;
  r(1) = 1;
  do i = 2 to 20;
    if x(i) ne x(i-1) then r(i)=1;
    else if x(i) eq x(i-1) then r(i)=0;
  end;
  runs = sum(of r0-r20);
  n1 = sum(of x1-x20);
  n0 = 20 - n1;
  if (n1+n0) ne 0 then mur = ((2*n1*n0) / (n1+n0)) + 1;
  if (((n1+n0)**2)*(n1+n0-1)) gt 0 then
    sigmar = ((2*n1*n0*(2*n1*n0-n1-n0)) / (((n1+n0)**2)*(n1+n0-1)))**0.5;
  if sigmar ne 0 then z = (runs-mur) / sigmar;
run;

```

In the code, the variable n1 represents the number of 1s in a pattern, the variable n0 represents the number of 0s, and the variable runs represents the number of runs. Since we are assuming that complete baseline data is necessary for inclusion in the study, we are only interested in examining the runs among the followup assessments. Therefore, we ignore the baseline assessment when calculating the number of runs. Once z scores have been calculated, we can calculate statistics and construct histograms by true dropout status (the variable drop) to compare the distribution between groups:



The mean z scores for non-dropouts and dropouts are -0.02 and -2.37, respectively. This is because the non-dropouts have the expected number or runs on average. In contrast, the dropouts have fewer total runs than one would have expected had the missing assessments been missing at random rather than missing in a patterned way. The overall distribution of z scores for dropouts (median = -2.59) is shifted more towards the left compared to that of non-dropouts (median = 0.312). Furthermore, the 5th percentile z score for non-dropouts is -1.88, which is larger than the median for dropouts. Hence, if we were blinded to true dropout status and calculated a z score for a subject that was below -2, we could conclude that the subject was a likely dropout. Similarly, the 95th percentile z score for dropouts is 0.33, suggesting that subjects with z scores above a threshold close to zero would be unlikely dropouts. The threshold values from these simulated data could be applied to similar actual datasets where true dropout status is unknown to classify subjects as likely or unlikely dropouts.

To further illustrate this point, two true dropouts (drop=1) and two true non-dropouts (drop=0) were selected from the simulated data. The first dropout / non-dropout comparison represents more extreme cases; that is, the patterns appear to be more “obvious” with respect to dropout status. The true dropout pattern of

10011111110000000000 and true non-dropout pattern of 11101011010111111110 have z scores of -3.60 and 0.94, respectively. The z score for the true dropout pattern is well below the threshold of -2, and the z score for the true non-dropout pattern is above the threshold of 0. Therefore, if we had been blinded to the true dropout status of these two patterns, we would have classified the true dropout as a likely dropout and the true non-dropout as an unlikely dropout, based on their respective z scores.

The next comparison represents combinations with misleading patterns with respect to dropout status. Upon initial inspection, a true non-dropout pattern of 101101111011110100000 and a true dropout pattern of 11111111110111111110 would appear as if the true dropout were a non-dropout and vice versa. The z scores for these patterns are -0.88 and -0.85, respectively; both of which are in between the threshold values for likely and unlikely dropouts. Although the z scores suggest inconclusive evidence to classify either pattern as a likely or unlikely dropout, neither of the patterns would have been misclassified if we were blinded to true dropout status.

CONCLUSION

Working with longitudinal data is often very challenging, especially when multiple assessment periods are involved. Longitudinal patterns of data completeness can widely vary, and questions arise with respect to usability. Although the processes by which longitudinal data are examined and classified depend on the nature and objectives of the particular study, generalized rules can be developed to assist with data classification. Through the process of successive refinement, cases at extreme ends are initially classified, and by working towards the middle, we can gradually chip away at the data until it becomes more structured and understandable.

As the number of assessment periods increase, we can apply the same concepts of successive refinement to multiple times to the combinations of data. With very high numbers of assessments other statistical techniques can be applied, such as calculating the z score for the runs test to further classify data and distinguish potential dropouts as likely or unlikely dropouts.

REFERENCES

Dudely, The Runs Test: Nonparametric Testing for Randomness in a Series of Runs. http://www.jmp.com/news/jmpcable/04/sprint1997/runs_test.html

Siegel (1956), *Nonparametric Statistics*, McGraw-Hill Book Company, NY

Mendenhall, Schaeffer, and Wackerly (1986), *Mathematical Statistics with Applications*, 3rd Ed., Duxbury Press, CA

CONTACT INFORMATION

The authors welcome questions and comments. Please direct inquiries to:

Lawrence Rasouliyan
Senior Statistical Analyst
Ovation Research Group
120 Howard Street, Suite 600
San Francisco, CA 94105
lrasouliyan@ovation.org

David J. Pasta
Vice President, Statistics & Data Operations
Ovation Research Group
120 Howard Street, Suite 600
San Francisco, CA 94105
dpasta@ovation.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.