Missing Data

Some degree of missing data is nearly inevitable in social and behavioral sciences. Missing values can occur for a variety of different reasons and can occur in a variety of different patterns. Individuals may fail to respond to some questions on a scale or test, some participants may be missing an entire assessment (e.g., participant drops out), such as when a participant completes the pretest but not the posttest, or data may be missing in a planned fashion, as when a random selection of participants are given a special module in a survey.

Before discussing some classic definitions and remedies for dealing with missing data, I will first describe a few computational details about how missing data are routinely handled in SPSS and R.

Computational Details and Missing Data in SPSS

There are several somewhat *small* details with how missing data are handled by some common operations in SPSS that *may* have *major* implications for how a computed variable may be interpreted. These details involve very common circumstances for researchers, and I'm not sure all researchers are always fully cognizant of them.

MEAN function. One can compute a new variable by averaging several variables using a `COMPUTE` statement and the `MEAN` function. For each person, the `MEAN` function takes an average of the values for each variable that is listed. For example, if you want to create a composite measure by averaging variables `X1` through `X4`, you would do the following:

```
COMPUTE NEWVAR=MEAN(X1,X2,X3,X4).
```

An alternative way to compute the same thing is by using the following:

```
COMPUTE NEWVAR=(X1+X2+X3+X4)/4.
```

If there are any data missing for variables `X1`, `X2`, `X3`, or `X4`, however, these two methods will not give equivalent results. Assume we have a small data set of only 5 subjects. A period '.' indicates a missing value.

Subject	X1	X2	X3	X4	NEWVAR (MEAN Function)	NEWVAR (adding and dividing)
1	2	3	1	2	2	2
2	5	4	6	.	5	.
3	10	10	14	14	12	12
4	7		5		6	.
5	8	.	.	.	8	.

When the `MEAN` function is used, SPSS computes the average of the variables in the list for a subject based on whatever data are available, even if only one of the variables is present for a subject. The second method in which the variables are added and then they are divided by the number of variables will not produce a result for the variable `NEWVAR` unless all of the variables are present.

Notice that one way to think about how the `MEAN` function computes the average is that the mean of the items for a case is assumed to be the value for the missing values for that case. For example, for subject 4 in above example, the value for `NEWVAR` using the `MEAN` function is the same as if the scores were 7,*6*,5,*6*, where the missing values are assumed to be equal to the average score in the list for that person.

SUM Function. Analogous methods can be used to compute the sum of a set of variables when creating a composite score.

```
COMPUTE NEWVAR=SUM(X1,X2,X3,X4).
```

Alternatively, one could specify it this way:

```
COMPUTE NEWVAR=X1+X2+X3+X4.
```

Let's see what the NEWVAR scores look like with these two methods, using the data from the previous example.

Subject	X1	X2	X3	X4	NEWVAR (SUM Function)	NEWVAR (adding)
1	2	3	1	2	8	8
2	5	4	6	.	15	.
3	10	10	14	14	48	48
4	7		5		12	.
5	8	.	.	.	8	.

NEWVAR is not computed using the adding method unless all the data are present. The SUM function, however, computes a sum based on the data available. This means that the sums will tend to be lower for individuals who have any missing data on the variables in the list. Notice that this is equivalent to assuming that the missing data are equal to '0's. For example, using the SUM function, the data for subject 4 is really assumed to be: 7,*0*,5,*0*.

Requiring a certain number of present cases of the SUM or MEAN function. One can require that all or some of the values in the list be nonmissing when either the SUM or the MEAN function are used by specifying a number after the MEAN or SUM function. This is best explained with an example:

```
COMPUTE NEWVAR=MEAN.3(X1,X2,X3,X4).
```

or

```
COMPUTE NEWVAR=SUM.3(X1,X2,X3,X4).
```

The above command requires that at least three of the values be present or NEWVAR will not be computed. In the above examples, subject 3 would get a score but subject 4 and 5 would not. Any value can be used in place of '.3' up to the number of variables in the list. Note that .1 does not do much good, because any subject has to have at least one variable present to get a value of NEWVAR anyway. Note also that using .4 in this instance is the same as using the adding method.

Computational Details and Missing Data in R

Creating composite scale scores is not as convenient in R. Several ways are shown below. The scale score is not computed if any of the items are missing for a case in the first two methods.

```
#simple ways to compute mean composite (no missing data allowed)
#base R just adding and dividing
d$ias = (d$q1+d$q2+d$q3+d$q4r)/4


#more flexible methods
#list-wise deletion approach (matches the base R approach above)
d$ias<-rowMeans(d[, c("q1", "q2", "q3","q4r")], na.rm=F)
#for summing, use rowSums in the same way
d$ias<-rowSums(d[, c("q1", "q2", "q3","q4r")], na.rm=F)

#available information approach (matches SPSS mean method) by adding na.rm=T
d$ias<-rowMeans(d[, c("q1", "q2", "q3","q4r")], na.rm=T)
```

Mechanisms

MAR and MCAR. A distinction about the nature of missing data was made by Rubin (1976; Little, 1995), who classified missing values as missing at random (MAR), missing completely at random (MCAR), or neither. Both MAR and MCAR require that the true values of the variable with missing values be unrelated to whether

or not a person has missing values on that variable. For example, if those with lower incomes are more likely to have missing values on an income question, the data cannot be MAR or MCAR. The difference between MAR and MCAR is whether or not other variables in the data set are associated with whether or not someone has missing values on a particular variable (say y). For example, when income values are missing, are older people more likely to refuse to respond to an income question? If other variables, say x, are related to missingness on y, then the missing values are MAR. If no other variables are related to missingness (not the unknown values of y and also not any x), then missing values are MCAR. The term "missing at random" is confusing because values are not really missing at random—for MAR, missingness seems to depend on some of the variables in the data set. MCAR is more what we think of when we think values are missing at random. For MCAR, it is as if we took a completely random selection of cases, and deleted their values for a variable.

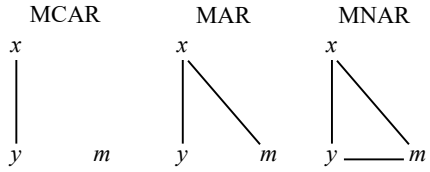

Figure 14.1. Analogue representation of missing data mechanisms. Reprinted from Newsom (2024). Adapted from Schafer and Graham (2003). In the figure, y is the variable of interest that has some missing data, x is another variable with no missing data, and m is the missingness indicator.

Determining Whether Missing Values Are MCAR or MAR

Researchers can investigate whether any variables in the data set are related to missingness on a variable by computing a new variable that indicates (0, 1) whether data are missing or present and then using correlations or group comparisons. Little (1988) developed a simultaneous test along these lines.[1] If none of the variables in the data set are related to missingness, then the results are consistent with the missing completely at random but does not prove that data are MCAR. The result can miss relationships due to lack of power and it does not guarantee that the values for the missing variable are not related to missingness for that variable because there is no way to know this (Allison, 2002). Practically speaking, it is not possible ever to determine whether the true values of a variable are related to the probability of missingness on that variable, because we do not have the missing information. As Schafer and Graham (2002) state: "*When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents or by imposing an unverifiable model.*" (p. 152). With attrition over time, it may be possible to test whether missingness is associated with the value of the variable that has present values at an earlier time point (i.e., usually all cases have mostly complete data at the first time point). For example, in a pretest-posttest design, we could investigate whether the variable at Time 1 (i.e., with complete data) is associated with the missingness for that variable at Time 2 (Little, 1995), which provides some information but is nearly always an imperfect proxy. In other circumstances, one may have to provide a theoretical argument that missingness is not associated with the variable or rely on information in the literature.

Simulation studies illustrate that modeling potential causes or correlates of the variables with missing values has important advantages when values are only MAR, particularly when the association of those "auxiliary" variables with the variable with missing values is high (e.g., > .4) and when the amount of missing data is large (e.g., > 25%; Collins, Schafer, & Cam, 2001; Graham, 2003). So, to the extent that we can incorporate some of the variables or proxies for the variables that may be causally related to the probability of missingness, we may be closer to meeting the MAR assumption. For this reason, there is an argument for always using modern missing data techniques, such as multiple imputation or full maximum likelihood estimation (see below), because there are few if any cases in which listwise deletion would provide better statistical tests.

[1] Little's test for MCAR, [Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association , 83*, 1198-1202] can be conducted in SPSS with the missing data module (must be separately purchased), in SAS using a macro https://communities.sas.com/kntur85557/attachments/kntur85557/programming/162574/1/Little%20Code.pdf, and through other specialty packages such as Mplus.

Dealing with Missing Data: Conventional Methods

There are a variety of less sophisticated methods of handling missing data that are worth mentioning, because several of them are still employed in certain circles. I do not recommend any of these methods.

Listwise deletion. Most analyses that we have studied, including t tests, ANOVA, chi square, and regression, require that all cases have data on all of the variables specified in the analysis. Until more recently, listwise deletion has been the most common way of dealing with missing data. That is, complete data were required on all variables in the analysis—any cases with missing values on one or more of the variables was eliminated from the analysis. Because statistical software programs have not incorporated better approaches in their common statistical test procedures, listwise deletion is still overwhelmingly the most common way of dealing with missing data (e.g., regression and ANOVA use this method by default). When there are small percentages of missing data (5%, 10%, less than 20% of cases are lost with listwise deletion?), there may not be serious harm in this practice (Arbuckle, 1995). In the last few years, however, researchers have begun to use data estimation techniques, such as multiple imputation, more often, particularly when many cases will be lost in the analysis. And simulation studies convincingly show that when there are a lot of missing values, listwise deletion will have biased parameters and standard errors (see Enders, 2001, for an illustration).

Pairwise deletion. Pairwise deletion is what occurs when correlations are required. You get the maximum N for each pair of variables. The correlation of X_1 with Y_1 may have a different N than the correlation of X_1 with Y_2. Pairwise deletion is a possible option in some analyses, such as multiple regression or structural equation modeling, but there are other potential problems with the approach and I do not recommend it (see Enders, 2022, for more information).

Other imputation methods. There are several other estimation approaches in which the data are imputed. That is, a full data set is created based on the imputation method that fills in data based on information from existing data. Older methods, such as mean imputation (in which the average of all scores on a variable for individuals with data is filled in for a case that has a missing value on that variable), regression-based methods (in which a regression is used to predict a score), and resemblance-based “hot-deck imputation” (in which new values are imputed using similar cases) do not perform as well as other methods, and some may produce highly biased coefficients and/or standard errors (Gold & Bentler, 2000).

Dealing with Missing Data: Modern Methods

Modern approaches, in particular multiple imputation (MI; Rubin, 1987) and full maximum likelihood (FIML; Dempster, Laird, & Rubin, 1977, which uses a structural modeling approach), produce superior estimates compared with listwise deletion and the other conventional methods mentioned above as long as data are at least MAR (Enders, 2022; Schafer & Graham, 2002). The standard multiple imputation approach requires an initial step (the I Step) in which multiple data sets are imputed with some degree of uncertainty built into the imputed estimates. There are a number of different methods for doing this (see Enders, 2022, for a nice summary). Current recommendations are for at least 20 imputed data sets (Graham, Olchowski & Gilreath, 2007) but precision seems to continue to improve with more, and Enders (2022) argues that there is generally little time cost and no statistical harm, given modern speed of computers, to using 100 data sets. In the second step analyses (the P step), the multiple, imputed data sets are analyzed and results are combined (or "pooled") using variability across the multiple imputations to better estimate standard errors in the analysis. Special software or special procedures within existing software are needed for multiple imputation, including SPSS Missing Values (which is an add-on with additional cost), several packages, such as `mice` and `mitml` in R, and free software Blimp (Enders, Keller, & Levy, 2018), which also handles multilevel data sets.

Structural equation modeling packages, such as Mplus, AMOS and the lavaan package in R, use FIML that is employed seamlessly in a single step when specifying a model (Mplus and R lavaan also can be used with MI). Regression models can be specified within these packages conveniently by simply requesting FIML estimation (often it is the default). These types of models are regression-based, and, although models can be constructed to test group difference hypotheses, structural equation modeling packages are generally not set up for analysis of variance per se as we are familiar with at this point.

Conclusion

Modern missing data handling, namely multiple imputation and FIML, have been shown to provide better estimates and better power than listwise deletion and older conventional methods of handling missing data. Moreover, the ability of the modern missing data handling approaches to incorporate even a large number of auxiliary variables that may be related to missingness and to the values of the variable with missing values, means that we can reduce or perhaps sometimes eliminate missing data biases when data do not meet the MAR assumption otherwise. So, as long as these methods are available to use, there is no harm and only advantages of using them as a default. Given that listwise deletion has been shown to produce biased estimates and will reduce power unless data meet the MCAR assumption, it seems there is little justification for using it other than greater convenience.

References and Further Reading

Arbuckle, J.L. (1996) Full information estimation in the presence of incomplete data. In G.A. Marcoulides and R.E. Schumacker [Eds.] *Advanced structural equation modeling: Issues and Techniques.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bernaard, C.A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research, 35*, 321-364.

Collins, L. M, Schafer, J.L., & Kam, C-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Structural Equation Modeling, 6*, 330-351.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38

Downey and King (1998). Missing data in likert ratings: A comparison of replacement methods. *The Journal of General Psychology, 125*, 175-191.

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61*, 713-740.

Enders, C. K. (2022). *Applied missing data analysis, second edition.* Guilford Press.

Enders, C.K., Keller, B.T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23,* 298-317. http://dx.doi.org/10.1037/met0000148.

Gold, M.S., & Bentler, P.M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling, 7*, 319-355.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8(*3), 206-213.

Little, R.J.A., & Rubin, D.B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research, 18*, 292-326.

Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data, third edition*. *New York: Wiley.*

Newsom, J.T. (2024). *Longitudinal structural equation modeling: A comprehensive introduction, second edition.* New York: Routledge.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons Inc., New York.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.