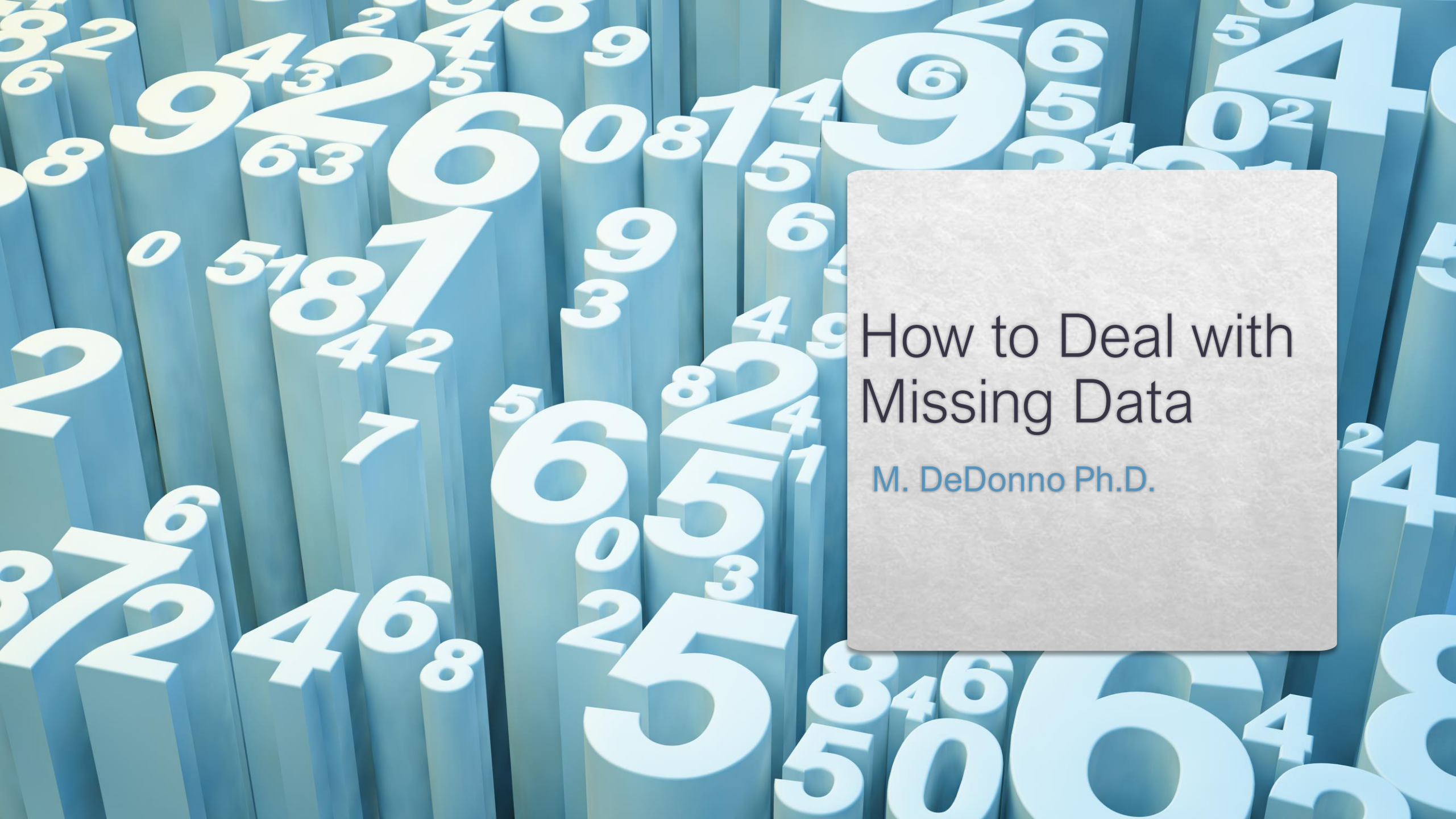

How to Deal with Missing Data

M. DeDonno Ph.D.

Missing Data

- Absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false
- Lost data can cause bias in the estimation of parameters.
- Can reduce the representativeness of the samples
- May complicate the analysis of the study

Missing Data

- Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest (Kang, 2013).

Types of Missing Data

- Before deciding which approach to employ, statisticians must understand why the data is missing
- **Missing completely at random (MCAR).** When data are MCAR, the fact that the data are missing is not related to either the specific value which is supposed to be obtained, or the set of observed responses.
  - Data missing due to equipment failure or samples are lost in transit
  - *The statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data*.

Types of Missing Data

- **Missing at random (MAR).** When data are MAR, the fact that the data are missing is systematically related to the **observed** but not the unobserved data (Little & Rubin, 1987).
  - Example: Male participants are less likely to complete a survey about depression severity than female participants.
  - Example: if blood pressure data are missing at random, conditional on age and sex, then the distributions of missing and observed blood pressures will be similar among people of the same age and sex.

Types of Missing Data

- **Missing not at random (MNAR).** When data are MNAR, the fact that the data are missing is systematically related to the **unobserved** data, (i.e., the missingness is related to events or factors which are not measured by the researcher).
  - Example: Participants with severe depression (not assessed) are more likely to refuse to complete the survey about depression severity.
  - Example: On a health survey, illicit drug users are less likely to respond to a question about illicit drug use.

Dealing with Missing Data

- Two primary methods
  - Deletion / removal of data where related data is deleted
    - Most common approach
    - When data do not fulfil assumption of MCAR, may cause bias in estimates (Doner, 1982)
  - Imputation method develops reasonable guesses for missing data
    - Most useful when the percentage of missing data is low
    - When the portion of missing data is too high, the results lack natural variation that could result in an inappropriate model

Dealing with Missing Data

- Deletion
  - Dropping variables - If data is missing for more than 60% of variable, it may be wise to discard it if the variable is insignificant.
  - Do nothing – SPSS will include the variable, but ignore missing values.

Dealing with Missing Data

- Deletion
  - Listwise deletion - all data for a case (e.g., a participant) that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data.
    - Look at the entire case to explore potential issues
  - Pairwise deletion - Cases that contain some missing data are maintained.
    - A case may contain 3 variables: VAR1, VAR2, and VAR3. A case may have a missing value for VAR1, but this does not prevent some statistical procedures from using the same case to analyze variables VAR2 and VAR3.
    - Pairwise deletion allows you to use more of your data. However, each computed statistic may be based on a different subset of cases.
  - Dropping variables - If data is missing for more than 60% of variable, it may be wise to discard it if the variable is insignificant.

Dealing with Missing Data

- Imputation
  - vs deletion - When data is missing, it may make sense to delete data. However, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.
  - Instead of deletion, statisticians have multiple solutions to impute the value of missing data.

Imputation

- Mean Substitution - The mean value of a variable is used in place of the missing data value for that same variable.
  - Theoretical background - The mean is a reasonable estimate for a randomly selected observation from a normal distribution.
  - However, with missing values that are not strictly random, the mean substitution method may lead to inconsistent bias.
  - Furthermore, this approach adds no new information but only increases the sample size and leads to an underestimate of the errors. Thus, mean substitution is not generally accepted (Malhotra, 1987).

Imputation

- Most frequent value – Impute the most frequently occurring value
  - Best with categorical data
- Hot deck imputation - A randomly chosen value from an individual in the sample who has similar values on other variables.
  - Find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.
  - One advantage is you are constrained to only possible values. (i.e., if Age in your study is restricted to being between 5 and 10, you will always get a value between 5 and 10 this way).
  - Another is the random component, which adds in some variability.

Imputation

- Regression Imputation - The information of other variables is used to predict the missing values in a variable by using a regression model.
  - Two options for regression imputation
    - Regression option - SPSS has some flaws in the estimation of the regression parameters (Hippel 2004).
    - Expectation Maximization (EM) option. Recommended option
      - SPSS: Analyze → Missing Value Analysis → Estimation: EM

Imputation

- Multiple Imputation - Allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each data set.
  - Multiple datasets with imputed values
  - Conduct statistical analysis on each of the imputed data sets and results are combined
  - Available in SPSS – See website for video instructions

Imputation

- K Nearest Neighbors (KNN) – Statistician selects a distance measure for k neighbors, and the average is used to impute an estimate (Beretta & Santaniello, 2016).
  - Must select the number of nearest neighbors and the distance metric.
  - KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

Imputation

- Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB) - These options are used to analyze longitudinal repeated measures data where follow-up observations may be missing.
  - Every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline.
  - Easy to understand and implement.
  - However, may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

Final Word

- Be sure to clearly explain why and how you handled missing data – particularly if you used some technique.
- Remember, minimizing missing data starts during the design of your study.
  - Appropriately identifying a sample
  - Format, order, complexity, and personal nature of questions
  - Time to complete survey / assessment / task
  - Where and when participation is required
    - E.g., college students less responsive at the end of a semester
  - Management and maintaining the data

Final Word

- Be careful when creating a value from multiple variables (SPSS –Compute a variable)

Var1	Var2	Var3	Total (new variable)
4	2	3	9
3	3	4	10
5	-	4	9
3	3	3	9

References

- Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making, 16*(74), 197-208.
- Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician, 36*, 378-81.
- Hippel, P. T. von. (2004). Biases in SPSS 12.0 Missing Values Analysis. *The American Statistician 58* (2), 160–64.
- Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology, 64*(5), 402-406.
- Little R. J., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.
- Malhotra, N. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Market Research, 24*, 74-84.
- Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.