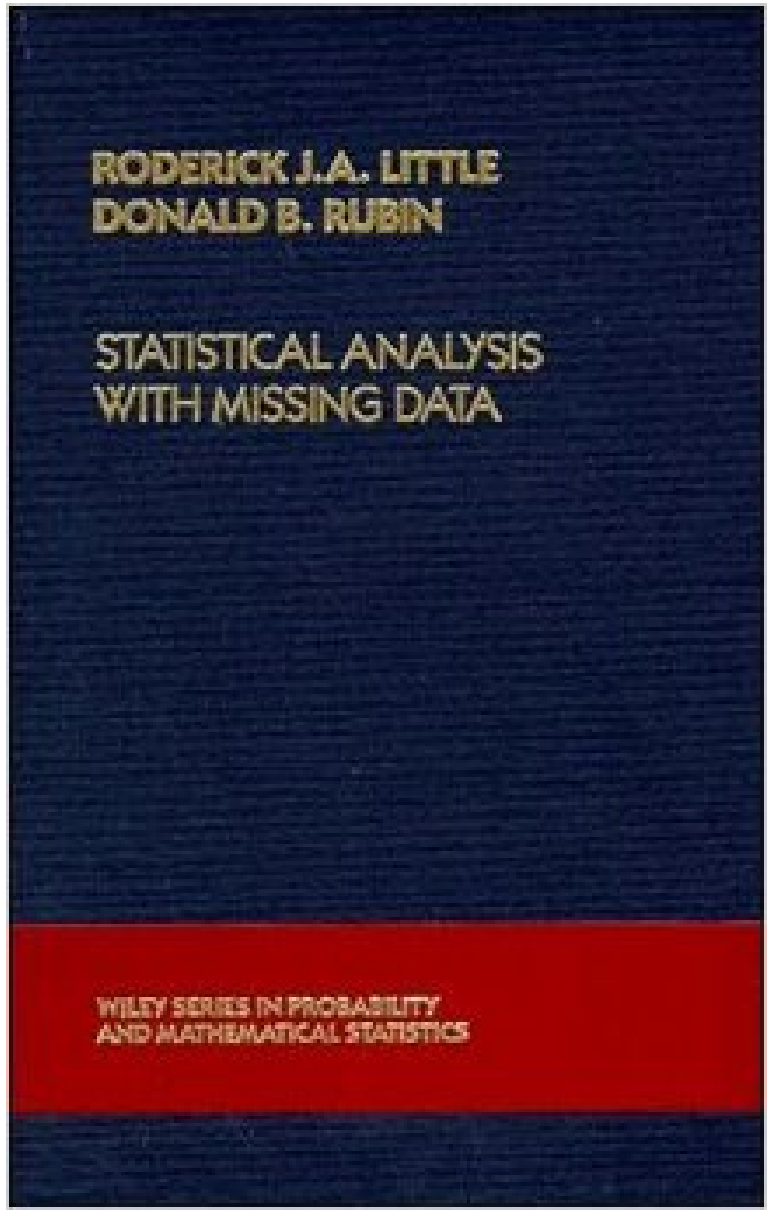
RODERICK J.A. LITTLE
DONALD B. RUBIN
STATISTICAL ANALYSIS
WITH MISSING DATA
WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

What is multiple imputation?

Imputation

- Imputation, the practice of '*filling in*' missing data with plausible values, is an attractive approach to analyzing incomplete data.
- The intention is to solve the missing-data problem at the beginning of the analysis.
- However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others.

Imputation

- For **MNAR**, imputation is not sufficient, because the missing data are totally different from the available data, i.e. your complete data has become a selective group of persons.
- For **MCAR** and **MAR**, there are roughly two kinds of techniques for imputation; Single and Multiple Imputation.

Single Imputation

- Single imputation techniques are based on the idea that in a random sample every person can be replaced by a new person, given that this new person is randomly chosen from the same source population as the original person.
- In that case you can use the observed available data of the other persons to make an estimation of the distribution of the test result in the source population.
- It is called **single imputation**, because each missing is imputed once.

What is multiple imputation?

- There are many methods for single imputation, such as
  - replacement by the mean,
  - regression,
  - expected maximization (EM).
- **Expected maximization** is preferred, because in the other methods the variance and standard error are reduced and the chance for Type II errors increases.

Multiple Imputation

- Multiple imputation is a simulation-based approach to the statistical analysis of incomplete data.
- In multiple imputation, each missing datum is replaced by $m > 1$ simulated values.
- The resulting m versions of the complete data can then be analyzed by standard complete-data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing-data uncertainty.

What is multiple imputation?

- The difference with single imputation is that in MI the value is imputed for several times. There are more imputed datasets created.
- The different imputations are then based on random draws of different estimations of the underlying distribution in the source population.
- In this way, the imputed data comes from different distributions and therefore are less look alike.
- There is more uncertainty created in the dataset. Therefore the standard error increases.

What is multiple imputation?

- The amount of imputations is dependent on the amount of missing data, but mostly 5 to 10 imputations are enough.
- A drawback of this method it that several imputed datasets are created and that the statistical analysis has to be repeated in each dataset.
- Finally, results have to be pooled in a summary measure. Most statistical packages can do this automatically.

Multiple Imputation

- The question of how to obtain valid inferences from imputed data was addressed by Rubin's (1987) book on multiple imputation (MI).
- MI is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (e.g. 3-10).

Multiple Imputation

- In Rubin's method for '**repeated imputation**' inference, each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.
- Rubin (1987) addresses potential uses of MI primarily for large public-use data files from sample surveys and censuses.

Multiple Imputation

- With the advent of new computational methods and software for creating MI's, however, the technique has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences whose investigations are hindered by missing data.
- These methods are documented in a recent book by Schafer (1997) on incomplete multivariate data.