1 Introduction to Missing Data

Missing data is a common problem in all kinds of research. The way you deal with missing depends on how much data is missing, and why it is missing. Handling missing data is an important step in several phases of a scientific study. Missing values means reduced sample size and loss of data. You conduct research to measure empirical reality so missing values thwart the very purpose of research. The less data collected, the less data that can be analyzed, and reducing the data that can be analyzed reduces statistical power, which is the ability to detect real relationships in the data.

Why do missing values occur?

- Missing values are either random or non-random.

- In survey-based studies, random missing values may through data entry mistakes, and the respondent accidentally skipped the question. When using sensor devices, missing data may occur because the devices malfunctioned.

- Random missing values may also occur in survey-based studies because the subject inadvertently did not answer some questions. The study may be overly complex and/or long, or the subject may be tired and/or not paying attention. You may noticed that many questions at the end of the survey that are not answered.

- Non-random missing values may occur because the subject deliberately did not answer some questions. The question may be confusing, so many subjects do not answer the question. Also, the question may not provide appropriate answer choices, such as ***no opinion*** or ***not applicable***, so the subject chooses not to answer the question.

- Another reason for non-random missing values is that respondents may be reluctant to answer some questions because of social desirability concerns about the content of the question, such as questions about sensitive topics like Criminal record, addiction issues, personal and political beliefs etc.

- Missing values may also indicate bias in the data. If the missing values are non-random, then the study is not accurately measuring the intended constructs. The results of your study may have been different if the missing data was not missing.

Remark: This lesson plan mostly deals with Missing Data. We will discuss short topics such as complete data, truncated data and censored data before progressing to Missing Data

Complete Data

Complete data means that the value of each sample unit is observed or known. Complete data is much easiest to work with, and basic statistical analysis techniques assume that we have complete data. Unfortunately it is a situation that does not always happen in practice.

Truncated Data

- Truncation results in values that are limited above or below, resulting in a truncated sample.
- Truncation is similar to but distinct from the concept of statistical censoring. A truncated sample can be thought of as being equivalent to an underlying sample with all values outside the bounds entirely omitted, with not even a count of those omitted being kept.
- With statistical censoring, a note would be recorded documenting which bound (upper or lower) had been exceeded and the value of that bound.

1.1 Examples of Analysis with Truncated Data

Example 1. One example of truncated samples come from historical military height records. Many armies imposed a minimum height requirement (MHR) on soldiers. This implies that men shorter than the MHR are not included in the sample. This implies that samples drawn from such records are perforce deficient i.e., incomplete, inasmuch as a substantial portion of the underlying population's height distribution is unavailable for analysis.

Example 2. A study of students in a special GATE (gifted and talented education) program wishes to model achievement as a function of language skills and the type of program in which the student is currently enrolled. A major concern is that students are required to have a minimum achievement score of 40 to enter the special program. Thus, the sample is truncated at an achievement score of 40.

Example 3. A researcher has data for a sample of Americans whose income is above the poverty line. Hence, the lower part of the distribution of income is truncated. If the researcher had a sample of Americans whose income was at or below the poverty line, then the upper part of the income distribution would be truncated. In other words, truncation is a result of sampling only part of the distribution of the outcome variable.

Censoring and Censored Data

There are three types of possible censoring schemes, right censored data (also called suspended data), interval censored data, and left censored data.

Right Censored (Suspended)

These are data for which we know only its minimum value. In reliability testing, for example, not all of the tested units will necessarily fail within the testing period. Then all we know is that the failure time exceeds the testing time. In microbiology, there is a practical threshold above which we cannot count colonies on a Petri dish. In sequential sifting, we known only the minimum diameter of the largest particles that don't pass through the first sieve. This type of data is commonly called ***right-censored*** or ***suspended data***.

Left Censored

These are data for which we know only its maximum value. In scientific experiments, for example, we may not be able to measure some quantity because it is below the threshold of detection (e.g. chemical concentration).

1.2 Examples of Analysis with Censored Data

Example 1. In the 1980s there was a federal law restricting speedometer readings to no more than 85 mph. So if you wanted to try and predict a vehicle's top-speed from a combination of horse-power and engine size, you would get a reading no higher than 85, regardless of how fast the vehicle was really traveling. This is a classic case of right-censoring (censoring from above) of the data. The only thing we are certain of is that those vehicles were traveling at least 85 mph.

Example 2. A research project is studying the level of lead in home drinking water as a function of the age of a house and family income. The water testing kit cannot detect lead concentrations below 5 parts per billion (ppb). The EPA considers levels above 15 ppb to be dangerous. These data are an example of left-censoring (censoring from below).

Example 3. Consider the situation in which we have a measure of academic aptitude (scaled 200-800) which we want to model using reading and math test scores, as well as, the type of program the student is enrolled in (academic, general, or vocational). The problem here is that students who answer all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude. The same is true of students who answer all of the questions incorrectly. All such students would have a score of 200, although they may not all be of equal aptitude.

Censored Data vs Truncated Data

- **IMPORTANT** There is sometimes confusion about the difference between truncated data and censored data. With censored variables, all of the observations are in the dataset, but we don't know the "true" values of some of them. With truncation some of the observations are not included in the analysis because of the value of the variable.

Interval Censoring

- These are data for which we know only that they lie between a certain minimum and maximum. Interval censoring arises commonly when we assign measurements into categories or intervals.
- For example, a survey may ask people which income range they have, and offer several contiguous intervals, rather than ask their exact income.
- In reliability testing, for example, we may only be inspecting the units every T hours, so can only record that a unit failed between nT and (n+1)T hours. This is sometimes called ***inspection data***.

1.3 Examples of Analysis with Interval Data

Example 1. We wish to model annual income using years of education and marital status. However, we do not have access to the precise values for income. Rather, we only have data on the income ranges: ¡15, 000,15,000-25, 000,25,000-50, 000,50,000-75, 000,75,000-100, 000, *and* >100,000. Note that the extreme values of the categories on either end of the range are either left-censored or right-censored. The other categories are interval censored, that is, each interval is both left- and right-censored. Analyses of this type require a generalization of censored regression known as interval regression.

Example 2. We wish to predict GPA from teacher ratings of effort and from reading and writing test scores. The measure of GPA is a self-report response to the following item:

```
Select the category that best represents your overall GPA.

[1]- less than 2.0      [5]- 3.4 to 3.8
[2]- 2.0 to 2.5         [6]- 3.8 to 3.9
[3]- 2.5 to 3.0         [7]- 4.0 or greater
[4]- 3.0 to 3.4
```

Missing Data

Missing completely at random

- There are several reasons why the data may be missing. They may be missing because equipment malfunctioned, the weather was terrible, or people got sick, or the data were not entered correctly. Here the data are missing completely at random (MCAR).
- When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of Xi or to the value of any other variables. Thus data on family income would not be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes.

- Similarly, if Whites were more likely to omit reporting income than African Americans, we again would not have data that were MCAR because "missingness" would be correlated with ethnicity. However if a participant's data were missing because he was stopped for a traffic violation and missed the data collection session, his data would presumably be missing completely at random. Another way to think of MCAR is to note that in that case any piece of data is just as likely to be missing as any other piece of data.

- Notice that it is the value of the observation, and not its "missingness," that is important. If people who refused to report personal income were also likely to refuse to report family income, the data could still be considered MCAR, so long as neither of these had any relation to the income value itself.

- This nice feature of data that are MCAR is that the analysis remains unbiased. We may lose power for our design, but the estimated parameters are not biased by the absence of data.

Missing at random

- Often data are not missing completely at random, but they may be classifiable as **missing at random** (MAR). For data to be missing completely at random, the probability that Xi is missing is unrelated to the value of Xi or other variables in the analysis. But the data can be considered as missing at random if the data meet the requirement that "missingness" does not depend on the value of X_i after controlling for another variable.

- For example, people who are depressed might be less inclined to report their income, and thus reported income will be related to depression. Depressed people might also have a lower income in general, and thus when we have a high rate of missing data among depressed individuals, the existing mean income might be lower than it would be without missing data. However, if, within depressed patients the probability of reported income was unrelated to income level, then the data would be considered MAR, though not MCAR.

- The terminology is a bit awkward here because we tend to think of randomness as not producing bias, and thus might well think that Missing at Random is not a problem. Unfortunately is is a problem, although in this case we have ways of dealing with the issue so as to produce meaningful and relatively unbiased estimates. But just because a variable is MAR does not mean that you can just forget about the problem.

Missing Not at random

- If data are not missing at random or completely at random then they are classed as **Missing Not at Random** (MNAR). For example, if we are studying mental health and people who have been diagnosed as depressed are less likely than others to report their mental status, the data are not missing at random.

- Clearly the mean mental status score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data. The same thing happens when people with low income are less likely to report their income on a data collection form.

- When we have data that are MNAR we have a problem. The only way to obtain an unbiased estimate of parameters is to model missingness. In other words we would need to write a model that accounts for the missing data. That model could then be incorporated into a more complex model for estimating missing values. This is not a task anyone would take on lightly.