# Longitudinal Data Structures

Andrew Zieffler

# Longitudinal Data Structures

Structure of longitudinal data refers to the format of the dataset. There are two primary formats for longitudinal data: (1) Wide data; (2) Long data.

You will need to utilize both structures depending on the analysis you are doing. For example, graphing longitudinal data using ggplot requires the long format, while computing correlations between measurments requires the wide format.

**Example 1:** A researcher wants to examine whether graduation success rate (GSR) has changed over time for football teams in the NCAA Big Ten conference. She has GSR data for the Big Ten teams from 2008, 2011, 2014, and 2015 (based on the 2002, 2005, 2008, and 2009 freshmen cohorts).

# Wide Formatted Data

It is the standard structure for entering longitudinal data into a spreadsheet. It is also referred to as the **subjects-by-variables** format or the **multivariate** format.

**Number of Rows**

*N* rows

where

*N* = Number of clusters (cases / individuals)

| School | GSR2008 | GSR2011 | GSR2014 | GSR2015 |
|---|---|---|---|---|
| Illinois | 69 | 75 | 70 | 70 |
| Indiana | 67 | 70 | 76 | 79 |
| Iowa | 74 | 82 | 71 | 74 |
| Maryland |  |  | 75 | 78 |
| Michigan | 71 | 69 | 72 | 79 |
| Michigan State | 56 | 64 | 66 | 71 |
| Minnesota | 54 | 68 | 69 | 71 |
| Nebraska |  | 68 | 85 | 86 |
| Northwestern | 92 | 97 | 97 | 97 |
| Penn State | 85 | 91 | 81 | 80 |
| Purdue | 59 | 59 | 76 | 81 |
| Rutgers |  |  | 83 | 82 |
| The Ohio State | 62 | 74 | 81 | 74 |
| Wisconsin | 65 | 63 | 71 | 73 |

# Long Formatted Data

The repeated/longitudinal measurements (outcomes) are in a single column. Each predictor, including time is also in a single column. It is also referred to as the **univariate** format.

**<u>Number of Rows</u>**

$N(K)$ rows

where

$N$ = Number of clusters (cases/individuals) and $K$ = Number of measurement occasions

| School | Time | GSR |
|---|---|---|
| Illinois | 2008 | 69 |
| Illinois | 2011 | 75 |
| Illinois | 2014 | 70 |
| Illinois | 2015 | 70 |
| Indiana | 2008 | 67 |
| Indiana | 2011 | 70 |
| Indiana | 2014 | 76 |
| Indiana | 2015 | 79 |
| Iowa | 2008 | 74 |
| Iowa | 2011 | 82 |
| Iowa | 2014 | 71 |
| Iowa | 2015 | 74 |
| ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ |
| Wisconsin | 2008 | 65 |
| Wisconsin | 2011 | 63 |
| Wisconsin | 2014 | 71 |
| Wisconsin | 2015 | 73 |

# Switching Between Formats

Converting data between the wide and long format is quite common and different tools use different language/terminology to describe these conversions. Below is a table that helps you think about these conversions (adapted from the online **tidyr** tutorial; http://tidyr.tidyverse.org/).

| Tool | Wide —> Long | Long —> Wide |
|---|---|---|
| Spreadsheets | Unpivot | Pivot |
| Databases | Fold | Unfold |
| | | |
| **R Packages** | | |
| tidyr | gather | spread |
| reshape2 | melt | cast |

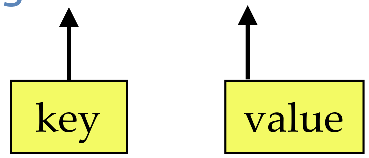
You can use whatever tool you want. I will be using the `gather()` and `spread()` functions from the **tidyr** package.

# Wide —> Long Format

```
# Read in wide data
> bigten_wide = read.csv(file = "~/epsy-8282/data/big-ten-wide.csv")
> head(bigten_wide)

          school gsr2008 gsr2011 gsr2014 gsr2015
1       Illinois      69      75      70      70
2        Indiana      67      70      76      79
3           Iowa      74      82      71      74
4       Maryland      NA      NA      75      78
5       Michigan      71      69      72      79
6 Michigan State      56      64      66      71


# Load the tidyr library (you may need to install it first)
> library(tidyr)

# Use the gather() function
> bigten_long = bigten_wide %>% gather(year, gsr, gsr2008:gsr2015)
```

The `gather()` function takes three arguments: The first two specify a **key-value pair**: year is the key and `gsr` the value. The third argument specifies which variables in the original data to convert into the key-value combination (in this case, all variables from `gsr2008` to `gsr2015`).

```
# Look at the long data (only the first 15 cases shown)
> bigten_long

            school    year gsr
1         Illinois gsr2008  69
2          Indiana gsr2008  67
3             Iowa gsr2008  74
4         Maryland gsr2008  NA
5         Michigan gsr2008  71
6   Michigan State gsr2008  56
7        Minnesota gsr2008  54
8         Nebraska gsr2008  NA
9     Northwestern gsr2008  92
10      Penn State gsr2008  85
11          Purdue gsr2008  59
12         Rutgers gsr2008  NA
13 The Ohio State gsr2008  62
14       Wisconsin gsr2008  65
15        Illinois gsr2011  75
```



key

value

For further examples of using `gather()` see this tutorial:

https://stanford.edu/~ejdemyr/r-tutorials/wide-and-long/

It can be useful (although not necessary) to sort the data by cluster; in our case, sort by school. To do this, we can pipe the long formatted data into the arrange() function. Or the formatting and sorting can be done in one step as shown below.

```
# Use the gather() function
> bigten_long = bigten_wide %>%
     gather(year, gsr, gsr2008:gsr2015) %>%
     arrange(school)

> bigten_long

           school    year gsr
1        Illinois gsr2008  69
2        Illinois gsr2011  75
3        Illinois gsr2014  70
4        Illinois gsr2015  70
5         Indiana gsr2008  67
6         Indiana gsr2011  70
7         Indiana gsr2014  76
8         Indiana gsr2015  79
9            Iowa gsr2008  74
10           Iowa gsr2011  82
11           Iowa gsr2014  71
12           Iowa gsr2015  74
13       Maryland gsr2008  NA
14       Maryland gsr2011  NA
15       Maryland gsr2014  75
16       Maryland gsr2015  78
```

The `arrange()` function works like Excel's Sort function.

Note that Maryland does not have data for 2008 nor 2011. R will automatically denote this with `NA`.

# Long —> Wide Format

```
# Use the spread() function
> bigten_wide = bigten_long %>% spread(year, gsr)
> bigten_wide

           school gsr2008 gsr2011 gsr2014 gsr2015
1        Illinois      69      75      70      70
2         Indiana      67      70      76      79
3            Iowa      74      82      71      74
4        Maryland      NA      NA      75      78
5        Michigan      71      69      72      79
6  Michigan State      56      64      66      71
7       Minnesota      54      68      69      71
8        Nebraska      NA      68      85      86
9    Northwestern      92      97      97      97
10     Penn State      85      91      81      80
11         Purdue      59      59      76      81
12        Rutgers      NA      NA      83      82
13 The Ohio State      62      74      81      74
14      Wisconsin      65      63      71      73
```

The `spread()` function takes two arguments: The first specifies a **key**: and the second specifies the **value**: `year` is the key and `gsr` the value.

# Balanced Design

A **balanced design** refers to a design in which participants are measured at the same time points.

An **unbalanced design** occurs when not all participants are measured at the same time points by design.

| School | GSR2008 | GSR2011 | GSR2014 | GSR2015 |
|---|---|---|---|---|
| Illinois | 69 | 75 | 70 | 70 |
| Indiana | 67 | 70 | 76 | 79 |
| Iowa | 74 | 82 | 71 | 74 |
| Maryland |  |  | 75 | 78 |
| Michigan | 71 | 69 | 72 | 79 |
| Michigan State | 56 | 64 | 66 | 71 |
| Minnesota | 54 | 68 | 69 | 71 |
| Nebraska |  | 68 | 85 | 86 |
| Northwestern | 92 | 97 | 97 | 97 |
| Penn State | 85 | 91 | 81 | 80 |
| Purdue | 59 | 59 | 76 | 81 |
| Rutgers |  |  | 83 | 82 |
| The Ohio State | 62 | 74 | 81 | 74 |
| Wisconsin | 65 | 63 | 71 | 73 |

This constitutes a **balanced design**, as the measurement occasions across all schools were the same.

The **incomplete data** is a function of missingness, not the design.

An example of an **unbalanced design**: A researcher plans to measure subjects from School A annually, but subjects from School B every two years.

|  |  | Age |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **Student** | **School** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| 2 | A | 9.3 | 10.7 | 11.9 | 14.2 | 14.1 | 15.0 | 14.8 | 14.7 |
| 3 | A | 10.3 | 11.1 | 11.7 | 12.6 | 13.1 | 12.8 | 13.1 | 14.1 |
| 4 | B | 10.1 |  | 11.6 |  | 12.4 |  | 13.9 |  |
| 5 | B | 11.1 |  | 12.8 |  | 15.3 |  | 12.2 |  |
| 6 | B | 9.5 |  | 14.1 |  | 12.6 |  | 13.2 |  |

Here the **incomplete data** is a function of the design.

Same **unbalanced design**: A researcher plans to measure subjects from School A annually, but subjects from School B every two years.

|  |  | Age |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Student | School | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| 2 | A | 9.3 |  |  | 14.2 | 14.1 | 15.0 | 14.8 |  |
| 3 | A |  | 11.1 | 11.7 | 12.6 | 13.1 | 12.8 |  | 14.1 |
| 4 | B | 10.1 |  | 11.6 |  |  |  | 13.9 |  |
| 5 | B | 11.1 |  | 12.8 |  | 15.3 |  | 12.2 |  |
| 6 | B | 9.5 |  | 14.1 |  | 12.6 |  | 13.2 |  |

Here the **incomplete data** is a function of the design and missingness.

# Imbalance, Incompleteness and Analytic Method

Classical methods of analyzing longitudinal data (Repeated-Measures ANOVA and MANOVA) require **balanced and complete** data.

Classical methods work on the WIDE data. If there is any missing data, that row is eliminated from the analysis (**listwise deletion**).

| | | Age | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Student** | **School** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| ~~2~~ | ~~A~~ | ~~9.3~~ | | | ~~14.2~~ | ~~14.1~~ | ~~15.0~~ | ~~14.8~~ | |
| ~~3~~ | ~~A~~ | | ~~11.1~~ | ~~11.7~~ | ~~12.6~~ | ~~13.1~~ | ~~12.8~~ | | ~~14.1~~ |
| ~~4~~ | ~~B~~ | ~~10.1~~ | | ~~11.6~~ | | | | ~~13.9~~ | |
| ~~5~~ | ~~B~~ | ~~11.1~~ | | ~~12.8~~ | | ~~15.3~~ | | ~~12.2~~ | |
| ~~6~~ | ~~B~~ | ~~9.5~~ | | ~~14.1~~ | | ~~12.6~~ | | ~~13.2~~ | |

To avoid the problem of eliminating a great deal of cases, sometimes researchers will **create complete and balanced data from unbalanced or incomplete data**. To do this, the researcher must delete data to force complete data and an equal number of waves for all participants.

|  |  | Age |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Student | School | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| 2 | A | 9.3 |  |  | 14.2 | 14.1 | 15.0 | 14.8 |  |
| 3 | A |  | 11.1 | 11.7 | 12.6 | 13.1 | 12.8 |  | 14.1 |
| 4 | B | 10.1 |  | 11.6 |  |  |  | 13.9 |  |
| 5 | B | 11.1 |  | 12.8 |  | 15.3 |  | 12.2 |  |
| 6 | B | 9.5 |  | 14.1 |  | 12.6 |  | 13.2 |  |

For example, rather than focus on age, a researcher might choose to focus on **successive measurement occasions**.

|  |  | Successive Measurement Occassions |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Student | School | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| 2 | A | 9.3 | 14.2 | 14.1 | 15.0 | 14.8 |  |  |  |
| 3 | A | 11.1 | 11.7 | 12.6 | 13.1 | 12.8 | 14.1 |  |  |
| 4 | B | 10.1 | 11.6 | 13.9 |  |  |  |  |  |
| 5 | B | 11.1 | 12.8 | 15.3 | 12.2 |  |  |  |  |
| 6 | B | 9.5 | 14.1 | 12.6 | 13.2 |  |  |  |  |

There is complete data for three **successive measurement occasions**. Everything else would be deleted.

In this example, the **chronology metric** (i.e., time scale) is ignored and so is the variability in timing of observations. Ignoring time scale (e.g., age) may be indefensible, especially if the scores reflect some type of developmental phenomenon that is naturally tied to time scale.

A second way to deal with missing data is to create complete and balanced data via **imputation**. Imputation is the process of replacing missing data with substituted values.

| | | Age | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Student** | **School** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 1 | A | 9.2 | 10.5 | 9.8 | 12.6 | 14.1 | 13.7 | 13.7 | 15.2 |
| 2 | A | 9.3 | **10.8** | **12** | 14.2 | 14.1 | 15.0 | 14.8 | **14.65** |
| 3 | A | **9.84** | 11.1 | 11.7 | 12.6 | 13.1 | 12.8 | **13.56** | 14.1 |
| 4 | B | 10.1 | **10.8** | 11.6 | **13.13** | **13.84** | **13.83** | 13.9 | **14.65** |
| 5 | B | 11.1 | **10.8** | 12.8 | **13.13** | 15.3 | **13.83** | 12.2 | **14.65** |
| 6 | B | 9.5 | **10.8** | 14.1 | **13.13** | 12.6 | **13.83** | 13.2 | **14.65** |

Here mean imputation has been used to impute the values of each missing data point.

One major problem with ignoring time or with imputation (regardless of the method of imputation), is that we **ignore or truncate variation**.

- When creating data by only using the sequential measurement occasions, we ignore the variability in timing of observations.
- When imputing data, we truncate the variation in the measurements themselves.

Variation is key to data analysis, and truncating variation is something you do not want to do.

Modern methods of analyzing longitudinal data (e.g., linear mixed-effects models) **do NOT require** balanced and complete data. This allows for much more flexibility in the design and structure of the data.

Modern methods work on the LONG data. Using listwise deletion on LONG data only results in the elimination of a single measurement occasion for a case…not the entire case.

| Student | School | Age | Outcome |
|---|---|---|---|
| 1 | A | 15 | 13.7 |
| 1 | A | 16 | 15.2 |
| 2 | A | 9 | 9.3 |
| ~~2~~ | ~~A~~ | ~~10~~ | |
| ~~2~~ | ~~A~~ | ~~11~~ | |
| 2 | A | 12 | 14.2 |
| 2 | A | 13 | 14.1 |
| 2 | A | 14 | 15.0 |
| 2 | A | 15 | 14.8 |
| ~~2~~ | ~~A~~ | ~~16~~ | |
| ~~3~~ | ~~A~~ | ~~9~~ | |
| 3 | A | 10 | 11.1 |

This allows us to keep the time scale (Age) and use any available information, even in cases with partial data.

Although we will cover classical methods, it is recommended that you **always use modern methods in practice**—even when you have complete data!

Unfortunately, listwise deletion omits cases even when the missing data occurs only in covariates. Consider this example where a researcher is interested in examining students' reading development.

| Student | Read | Grade | Sex |
|---|---|---|---|
| 1 | 172 | 5 | |
| 1 | 185 | 6 | |
| 1 | 179 | 7 | |
| 1 | 194 | 8 | |
| 2 | 200 | 5 | F |
| 2 | 210 | 6 | F |
| 2 | 209 | 7 | F |
| ~~2~~ | | ~~8~~ | ~~F~~ |
| 3 | 191 | 5 | M |
| 3 | 199 | 6 | M |
| 3 | 203 | 7 | M |
| 3 | 215 | 8 | M |

All of Student 1's data would be deleted because of the missing data on the Sex covariate. **Should we delete these rows?**

This row should be deleted since Student 3 is missing a reading score (outcome) at Grade 8.

The statistical recommendations about when to delete data are not clear. For example:

- When data is missing (NA values occur) in the **response variable**, it is recommended the missing value rows be omitted and the resulting data used in all analyses.

- When data is missing (NA values occur) in **static predictors** (predictors that are the same across all measurement occasions; e.g., sex), it is unclear whether these cases should be omitted. This is because cases might then be retained/excluded depending on the covariates used in the analysis.

To carry out listwise deletion, we pipe the data into the `drop_na()` function. (To use this function the **tidyr** package will need to be loaded.)

```
# Minneapolis reading data
> mpls

   student read grade  sex
1        1  172     5 <NA>
2        1  185     6 <NA>
3        1  179     7 <NA>
4        1  194     8 <NA>
5        2  200     5    F
6        2  210     6    F
7        2  209     7    F
8        2   NA     8    F
9        3  191     5    M
10       3  199     6    M
11       3  203     7    M
12       3  215     8    M
```

```
# Use drop_na()
> mpls2 = mpls %>% drop_na()
> mpls2

   student read grade sex
5        2  200     5   F
6        2  210     6   F
7        2  209     7   F
9        3  191     5   M
10       3  199     6   M
11       3  203     7   M
12       3  215     8   M
```

Without any arguments, the `drop_na()` function carries out listwise deletion.

```
# Delete only cases missing reading score
> mpls3 = mpls %>% drop_na(read)
> mpls3

   student read grade  sex
1        1  172     5 <NA>
2        1  185     6 <NA>
3        1  179     7 <NA>
4        1  194     8 <NA>
5        2  200     5    F
6        2  210     6    F
7        2  209     7    F
9        3  191     5    M
10       3  199     6    M
11       3  203     7    M
12       3  215     8    M
```

When variable names are included in the `drop_na()` function, it will carry out listwise deletion only on those variables. Multiple variables can be separated by commas.

# Missing Data, Validity, and Inference

In longitudinal analysis (as in any other data analysis), the results and inferences drawn are most valid when there is **no missing data**.

When there are missing values the validity of the analysis depends on certain assumptions about the mechanism underlying the missingness. This mechanism can be thought of as a process that acts on the **unobserved complete data** set to produce the incomplete observed data set.

Consider the measurement of a response variable at two time points. Suppose there is no missing data at the first time point but there is missing data at the second time point. How we classify the missing data mechanism is related to **whether the missing data at Time 2 depend on**:

1. The observed data at Time 1.
2. The observed data at Time 2.
3. The missing data at Time 2.
4. None of the above .

# Missing Completely at Random (MCAR)

The missing data in our example can be classified as *Missing Completely at Random* (MCAR) when #4 on the list (None of the above) is the case. Specifically, the missing data at Time 2:

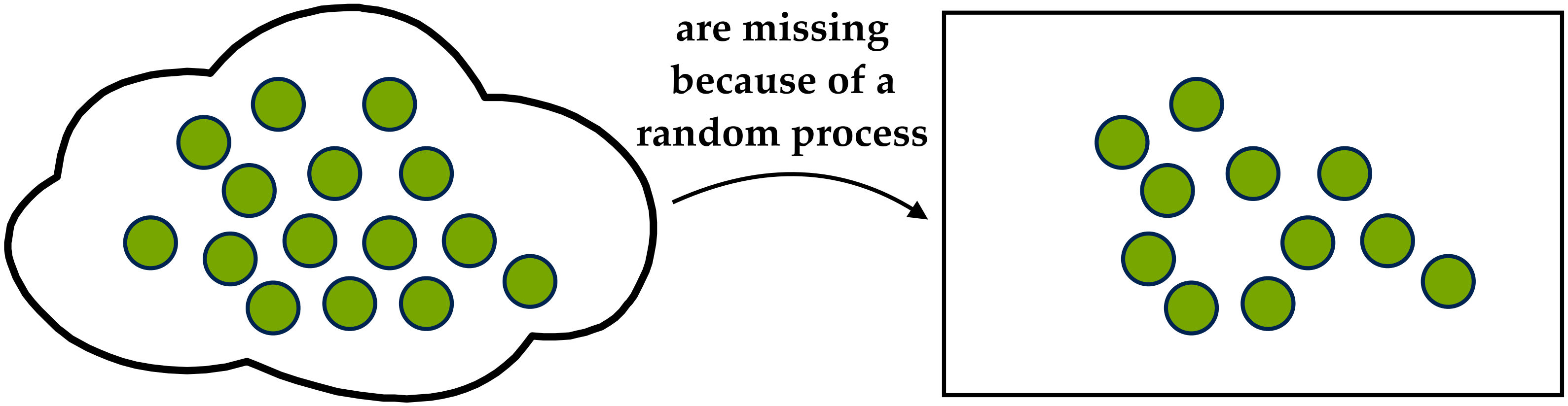
- Does not depend on the observed data at Time 1.
- Does not depend on the observed data at Time 2.
- Does not depend on the missing data at Time 2.

In fact, to be classified as MCAR, missing data on a particular variable must be **unrelated to any other variables** of interest in the study.

One way to think about data that are MCAR is that the incomplete observed sample is assumed to be a random sample of the unobserved complete data.

**Example of MCAR Process:** Suppose a researcher wants to study students' reading development between 5th and 9th grade. Her original plans was to administer a reading test each year of the study, but she is worried about the response load. By design, she plans to obtain only four waves of data for two cohorts of individuals. Cohort 1 will be measured over grades 5–8 and Cohort 2 will be measured over grades 6–9.  Students will be randomly assigned to a cohort.

| | Grade | | | | |
|---|---|---|---|---|---|
| Cohort | 5 | 6 | 7 | 8 | 9 |
| 1 | x | x | x | x | |
| 2 | | x | x | x | x |



**Scores are missing because of a random process**

Unobserved complete data at Grade 5 and 9

Observed incomplete data at Grade 5 and 9

In the observed (incomplete) data, students assigned to Cohort 1 will be missing reading measurements at Grade 9, and students assigned to Cohort 2 will be missing measurements at Grade 5. The missingness is a **function of the random process used to assign students to cohorts**, and thus can be considered MCAR.

# Missing at Random (MAR)

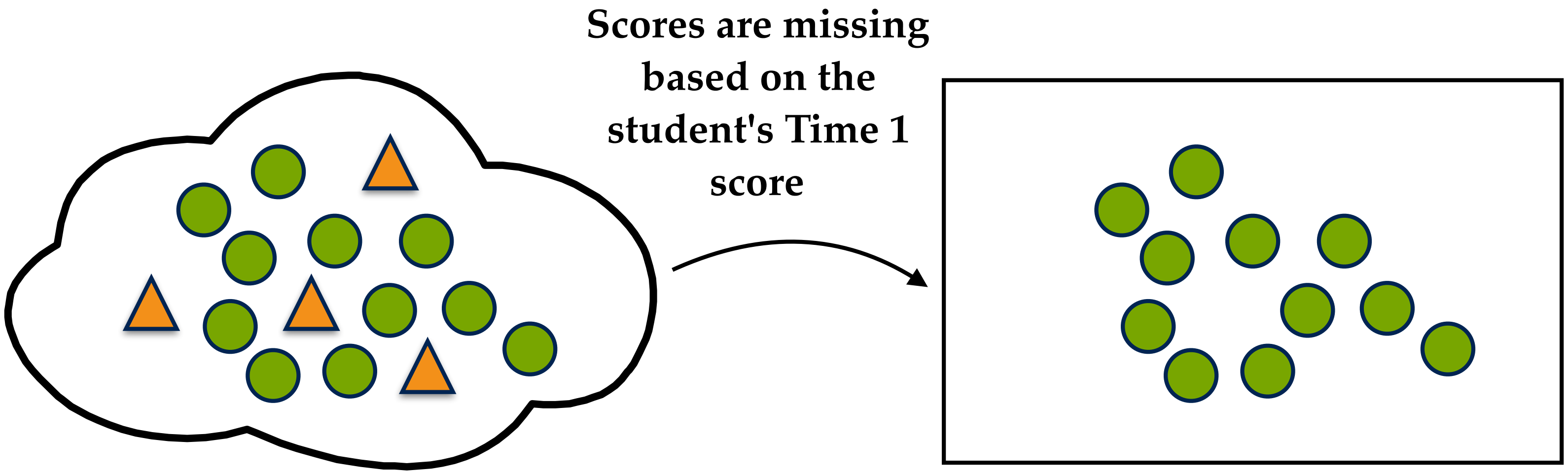
The missing data in our example can be classified as *Missing at Random* (MAR) when #1 or #2 on the list is the case. Specifically, the missing data at Time 2:

- Depends on the observed data at Time 1; OR
- Depends on the observed data at Time 2; BUT
- Does not depend on the missing data at Time 2.

To be classified as MAR, missing data on a particular variable must be **associated with other variables of interest in the data set.**

The term MAR is confusing since data are not really missing at random—missingness depend on some of the variables in the data set.

**Example of MAR Process:** Suppose a researcher wants to study students' reading development over the course of year. She is planning on taking two measurements for each student. However, any student with a perfect score at Time 1 will not be measured a second time.



**Scores are missing based on the student's Time 1 score**

Unobserved complete data at Time 2

Observed incomplete data at Time 2

In the observed (incomplete) data, students with perfect scores at Time 1 will be missing reading measurements at Time 2. The missingness is a **function of another variable in the dataset** (the Time 1 scores), and thus can be considered MAR.

Knowing a subject's score at Time 1, is predictive of a missing value at Time 2. (Note: The prediction of missing values can be based on any variable, not only the response variable.)

# Not Missing at Random (NMAR)

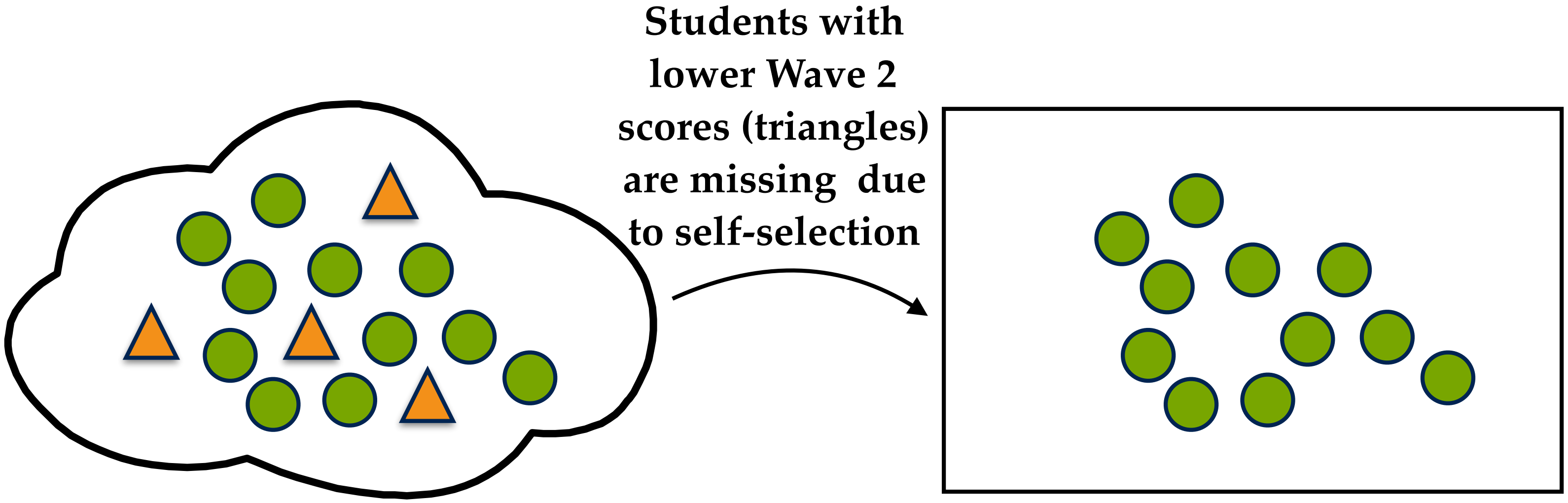
The missing data in our example can be classified as *Not Missing at Random* (NMAR) when #3 on the list is the case. Specifically, the missing data at Time 2:

- Depends on the missing data at Time 2.

To be classified as NMAR, missing data on a particular variable must be **associated with the variable that has the missing data.**

For example, survey respondents with lower incomes are more likely to have missing data on the income item of a survey.

**Example of NMAR Process:** Consider computer administered reading test with two measurement waves. After taking the test the second time (Wave 2), subjects are allowed to see both scores and can decide to retain or delete their second score. Subjects whose score decreases from Wave 1 to Wave 2, are more likely to delete their Wave 2 score. Only retained scores are ultimately observed by the researcher (deleted scores are not observable).



**Students with lower Wave 2 scores (triangles) are missing due to self-selection**

Unobserved complete data at Wave 2

Observed incomplete data at Wave 2

In the observed (incomplete) data, students with lower scores at Wave 2 will be missing reading measurements at Wave 2. The missingness is a **function of the Wave 2 missing data** (the low Wave 2 scores), and thus can be considered NMAR.

In this case the missing data are also dependent on the observed data at Wave 1. More importantly, they depend on the missing data at Wave 2, as the researcher never sees the deleted subjects' Wave 2 scores.

What do these missing data mechanisms mean for estimation and inference? Well, this depends. If one is using listwise deletion, then only data that is MCAR will produce unbiased estimates of the coefficients and SEs. If likelihood-based estimation methods are utilized, then unbiased estimation also extends to MAR data (Rubin, 1976).

In the statistical literature, for likelihood-based inference the **missingness is ignorable** under MCAR or MAR.

| | Listwise Deletion<br>(Non-Likelihood-Based Inference) | | Listwise Deletion<br>(Likelihood-Based Inference) | |
|---|---|---|---|---|
| MCAR | Unbiased coefficient estimates | Unbiased estimates of SE | Unbiased coefficient estimates | Unbiased estimates of SE |
| MAR | | | Unbiased coefficient estimates | Unbiased estimates of SE |
| NMAR | | | | |

Modern methods for dealing with longitudinal data (mixed-effects models) use likelihood-based inference. Classical methods (RM-ANOVA; MANOVA) do not use likelihood-based inference.

It is up to the researcher to show that any missing data that is NMAR is ignorable. This is a tall order and often means explicitly modeling the missing data mechanism to obtain unbiased estimates of the coefficients and SEs. This, in turn, requires a wealth of substantive knowledge about the underlying missing data mechanism because the data itself contains no information about what models would be appropriate. This means there is no way to test goodness of fit of missing data model. Results, which are often very sensitive to choice of model, are quite uncertain..

There have been many partial solutions for specific types and cases of NMAR. See, for example, Rubin and Little's (2002) book on missing data.

## What about other methods of dealing with missing data?

**Pairwise Deletion:** Generally we can use PD to estimate the means, variances and correlations between variables. These values are then plugged into formulas to estimate the parameters (e.g., coefficients, SEs). If the missing data is MCAR, the coefficient estimates are *approximately* unbiased. The SEs, however, tend to be incorrect (larger sample sizes do not fix this). For missing data that is MAR, both the coefficient estimates and SEs tend to be biased.

**Imputation:** Imputation methods often produce biased parameter estimates, especially, variances. Because the variation in measures is truncated with most imputation methods, SE estimates tend too be biased downward (too small), which in turn produce *p*-values that are too small. Most analysts treat imputed data as actual data; ignoring the inherent uncertainty in producing imputed values. Multiple imputation attempts to account for this uncertainty.

# References and Source Material

- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. New York: Wiley.
- Kohli, N. (2016). *Data structures and longitudinal data analysis (course notes).* Minneapolis, MN: Author.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed). New York: Wiley.
- Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R.* Thousand Oaks, CA: Sage. (Minneapolis reading data)
- National Collegiate Athletic Association. (2016, December 12). *Graduation success rate.* Retrieved July 21, 2017, from http://www.ncaa.org/about/resources/research/graduation-success-rate
- Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

- Graphics used from Open Clip Art Library (https://openclipart.org/)