Lecture 20

1. Types of missing values
2. Making example missing-value datasets: MCAR, MAR, and MNAR
3. Common methods for missing data
4. Compare results on example MCAR, MAR, MNAR data

Missing Data Methods

Clinical trial randomly assigned 100 patients with major depression to an experimental drug (D) or to placebo (P). (Source: Dmitrienko *et.al.* (2005).

Participants completed the Hamilton depression rating scale (HAMD) at baseline and again after 9-week treatment. Study outcome was HAMD at end; higher scores mean worse depression. Participants at 5 centers:

```
drug      center

Frequency|       1|       2|       3|       4|       5|  Total
---------+--------+--------+--------+--------+--------+
Drug     |     11 |      7 |     16 |      9 |      7 |     50
---------+--------+--------+--------+--------+--------+
Placebo  |     13 |      7 |     14 |     10 |      6 |     50
---------+--------+--------+--------+--------+--------+
Total          24       14       30       19       13      100
```

The first 5 observations from the Depression Study data

ID	baseline	final	drug	center
1	27	4	D	1
2	27	9	D	1
3	26	8	D	1
4	27	5	D	1
5	36	8	D	1

Model(s) to compare final HAMD between treatments, adjusted for baseline and center:

We'll return to these models to analyze this data.

Missing Values

Suppose that some final surveys were missing—not completed.

What happens to these participants' data in the fitting the adjusted model?

What if patients with the worst side-effects to the experimental drug (D) dropped out and didn't complete the final survey?

Types of Missing Data

Missing completely at random (MCAR): data are missing independently of both observed and unobserved data.

Example: a participant flips a coin to decide whether to complete the depression survey.

Missing at random (MAR): given the observed data, data are missing independently of unobserved data.

Example: male participants are more likely to refuse to fill out the depression survey, but it does not depend on the level of their depression.

MCAR implies MAR, but not the other way round. Most methods assume MAR.

We can ignore missing data (= omit missing observations) if we have MAR or MCAR.

Missing Not at Random (MNAR): missing observations related to values of unobserved data.

Example: participants with severe depression, or side-effects from the medication, were more likely to be missing at end.

Informative missingness: the fact that data is missing contains information about the response.

Observed data is biased sample. Missing data cannot be ignored.

Cannot distinguish MAR from MNAR without additional information.

SAS default is to omit cases with missing data = ignore missing data.

With MNAR, you get a non-representative sample and biased estimates.

References:

Dmitrienko *et. al.* (2005) *Analysis of Clinical Trials Using SAS*, Chapter 5

R Little and D Rubin (2002) *Statistical Analysis with Missing Data, Second Edition*

Plan:

1. Delete observations from HAMD data to make an example of each type of missing data.

2. Discuss approaches to handling missing data.

3. Compare these approaches on our constructed examples from HAMD.

Make missing completely at random (MCAR) example

MCAR: data are missing independently of both observed and unobserved data.

Example: participant flips a coin to decide whether to complete final survey.

Randomly select 30% of the observations in HAMD, set to missing.

```
data MCAR;
  set ph6470.hamd2;
  missing = 0;
  if (ranuni(457392) < .3) then do;  select 30% random sample
       final =. ;
       missing=1;   label missing values
  end;
```

MCAR example, first 10 observations.

```
Obs    missing    baseline    final    drug    center
  1       0          27          4      D         1
  2       0          27          9      D         1
  3       0          26          8      D         1
  4       0          27          5      D         1
  5       0          36          8      D         1
  6       0          39         18      D         1
  7       0          25         14      D         1
  8       0          33          8      D         1
  9       0          38          9      D         1
 10       1          39          .      D         1
```

```
proc freq data=MCAR;
   tables missing;

  missing    Frequency      Percent      Frequency       Percent
        0           67      67.00               67         67.00
        1           33      33.00              100        100.00
```

What percent are actually missing?

Missing at random (MAR) example

Missing at random (MAR): given the observed data, data are missing independently of unobserved data.

Example: male participants more likely to refuse to fill out final survey, independent of their level of their depression.

Data does not include gender. Missing values related to observed data: only at centers 1, 2, and 3.

Need to get ≈ 33 missing cases. Centers 1, 2, 3 together have 64/100 patients in study. What proportion p should be missing?

$$p * 64 = 33 \quad \text{gives} \quad x = .516$$

```
 data MAR;
   set ph6470.hamd2;
   missing = 0;
   if (ranuni(457392) < .516  and center IN (1, 2, 3))  then do;
        final =. ;
        missing=1;
   end;

 proc freq data=MAR;
   tables missing;
```

```
                                      Cumulative    Cumulative
missing    Frequency      Percent     Frequency      Percent
-----------------------------------------------------------
      0          63       63.00             63        63.00
      1          37        37.00            100       100.00
```

Adjusting the cutoff for the uniform random number gives:

```
 data MAR;
   set ph6470.hamd2;
   missing = 0;
   if (ranuni(457392) < .435  and center IN (1, 2, 3)) then do;
        final =. ;
        missing=1;
   end;
```

This produces 34 missing values, nearly the same number as the MCAR example.

MAR example, first 10 observations.

```
Obs   missing   baseline   final   change   drug   center

  1      1         27        .       23      D        1
  2      0         27        9       18      D        1
  3      0         26        8       18      D        1
  4      0         27        5       22      D        1
  5      0         36        8       28      D        1
  6      0         39       18       21      D        1
  7      1         25        .       11      D        1
  8      0         33        8       25      D        1
  9      1         38        .       29      D        1
 10      1         39        .       18      D        1
```

Missing not at random (MNAR) example

MNAR: missing observations related to values of unobserved data.

Example: participants with most severe depression were less likely to complete final HAMD survey.

Identify “high” final values.

Randomly select 33 among these to delete—want same amount of missing data as other examples.

How do we identify top 50% of baseline values?

```
Proc univariate data=ph6470.hamd2;
   var final;

  Quantile      Estimate
  100% Max          35.0
  99%               34.0
  95%               28.0
  90%               23.5
  75% Q3            19.0
  50% Median        14.5
  25% Q1             8.0
  10%                4.0
  5%                 2.0
  1%                 1.0
  0% Min             1.0
```

What proportion do we remove? $p * 50 = 33$ gives $p = .66$

```
data MNAR;
  set ph6470.hamd2;
  missing=0;
  if (final GE 14.5  and ranuni(884739) < .66 ) then do;
       final =. ;
       missing=1;
  end;
proc freq data=MNAR;
  tables missing;
```

This gives only 30 missing values, and we want 33 or 34.

What do we adjust to get a few more missing values?

Trial and error leads to:

```
data MNAR;
  set ph6470.hamd2;
  missing=0;
  if (final GE 14.5 and ranuni(884739) <  .69  ) then do;
      final =. ;
      missing=1;
  end;
```

which gives 33 missing values.

MNAR example, first 10 observations:

```
Obs    missing    baseline    final    change    drug    center
  1       0          27         4        23       D         1
  2       0          27         9        18       D         1
  3       0          26         8        18       D         1
  4       0          27         5        22       D         1
  5       0          36         8        28       D         1
  6       1          39         .        21       D         1
  7       0          25        14        11       D         1
  8       0          33         8        25       D         1
  9       0          38         9        29       D         1
 10       1          39         .        18       D         1
```

Review the plan:

1. Delete observations from HAMD data to make an example of each type of missing data: MCAR, MAR, MNAR.

   All data sets have 33% missing data.

2. Overview: approaches to handling missing data.

3. Compare these approaches on our constructed examples from HAMD.

   Results will depend on type of missingness, not amount of missing data.

Common methods for MAR data

MAR property: missing-ness related only to observed data, not the missing data.

1. **Complete case analysis.** Omit observations missing any part of the data.

   SAS default for many procedures.

   Requires MCAR to be unbiased.

2. **Last observation carried forward (LOCF).** Longitudinal data collection where early measurements are not missing but final measurements are missing. Use each subject's last non-missing measurement to fill in later missing values. Reduces apparent change in response.

   Requires strong assumptions about response; does not account for uncertainty of missing data.

   Better approach: use Proc Mixed which can handle missing values in longitudinal data.

3. **Imputation.** This means filling in each missing value with a guess. Many ways to impute, such as:

   - Use mean of individual's other values.
   - Replace missing value in a group with group mean.
   - Predict missing values in a variable V from regression of V on other variables.

   Requires strong assumptions about response; does not account for uncertainty of missing data.

4. **Multiple imputation**:
   (a) Impute observations for all missing values in a variable V: use random samples from normal distribution with mean and SD of V.
   (Or use regression to predict mean and SD, then sample from this normal distribution.)
   (b) Do the imputation M times, creating M complete data sets.
   (c) Analyze each of the M complete data sets.
   (d) Combine the results of the M analyses to draw conclusions.

   Requires MAR to be unbiased. Partially accounts for uncertainty of missing data.

Compare these approaches on our constructed missing-data examples from HAMD

Estimate treatment means, test treatment×center interaction from full data, and constructed examples of MCAR, MAR, and MNAR.

For MCAR, MAR, and MNAR, apply

1. complete case analysis
2. last observation carried forward (LOCF)
3. multiple imputation

Full data analysis

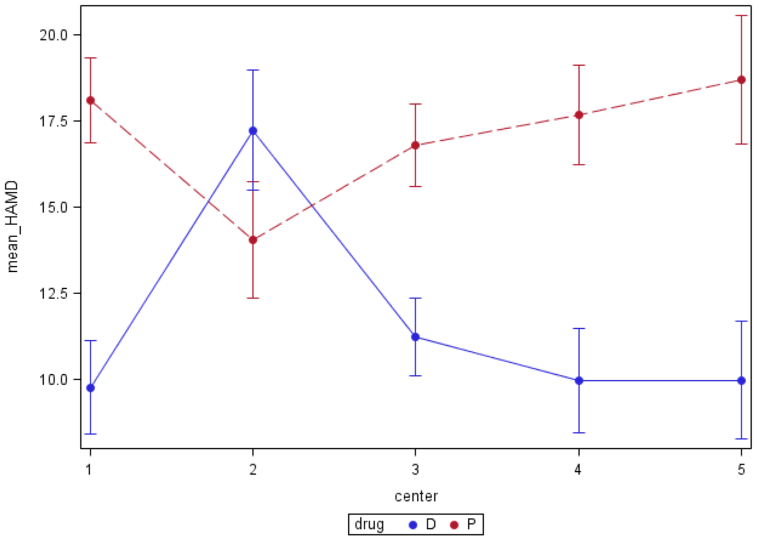
Test interaction between treatments and centers:

```
Proc GLM   data=ph6470.hamd2;
    class drug center;
    model final =  baseline drug center drug*center;
```

Estimate treatment means using main-effect model:

```
Proc GLM   data=ph6470.hamd2;
    class drug center;
    model final =  baseline drug center;
    LSmeans drug / stderr;
```

```
Source                    DF     Type III SS     Mean Square     F Value     Pr > F
baseline                   1     2937.733457     2937.733457      145.94     <.0001
drug                       1      665.325939      665.325939       33.05     <.0001
center                     4       33.166939        8.291735        0.41     0.7996
drug*center                4      348.644165       87.161041        4.33     0.0030
```



From the main-effects model (parallel lines), which assumes no interaction:

```
                                              Standard
Parameter              Estimate                  Error    t Value    Pr > |t|

Intercept        -5.029306548 B         2.40541540      -2.09      0.0393
baseline          0.732392169           0.06750514      10.85      <.0001
drug       D     -5.780464986 B         0.96279506      -6.00      <.0001
drug       P      0.000000000 B          .                .          .
center     1     -0.140497046 B         1.65460640      -0.08      0.9325
center     2      1.282776802 B         1.85476395       0.69      0.4909
center     3     -0.125739959 B         1.59580285      -0.08      0.9374
center     4     -0.092401374 B         1.75244924      -0.05      0.9581
center     5      0.000000000 B          .                .          .

                         Least Squares Means

                                                                H0:LSMean1=
                                      Standard     H0:LSMEAN=0     LSMean2
  drug    final LSMEAN                   Error          Pr > |t|       Pr > |t|

  D         11.4640040           0.6960627            <.0001          <.0001
  P         17.2444690           0.6983823            <.0001
```

Where is estimate of treatment difference?

Higher scores on HAMD mean worse depression.

Drug effect shown by *higher* final scores with placebo than with drug.

Data	*Method*	*Interaction* *P-value*	*Drug Effect* *(Pbo − Drug) ± SE*	*Drug Effect* *P-value*
Full		.003	5.8 ± 1	< .0001
MCAR				
MAR				
MNAR				

Complete Case (CC)

Proc GLM omits any observations with missing values for the response or any predictors in the model *or class statement.*

Apply interaction and main-effects Proc GLM to MCAR, MAR, MNAR data sets.

```
                         MCAR complete case

                         The GLM Procedure

                      Class Level Information

                  Class         Levels    Values
                  drug               2    D P
                  center             5    1 2 3 4 5

               Number of Observations Read         100
               Number of Observations Used          67
```

```
Source                    DF     Type III SS     Mean Square    F Value    Pr > F
baseline                   1     1859.138751     1859.138751      81.53    <.0001
drug                       1      402.153831      402.153831      17.63    <.0001
center                     4       81.206129       20.301532       0.89    0.4759
drug*center                4      292.270566       73.067641       3.20    0.0194

                                             Standard
Parameter           Estimate                    Error    t Value    Pr > |t|
Intercept       -6.020340071 B             3.15225778      -1.91      0.0609
baseline         0.764130773               0.09024947       8.47      <.0001
drug       D    -7.111619538 B             1.29954422      -5.47      <.0001
drug       P     0.000000000 B              .                .         .
center     1     0.477570916 B             2.28410457       0.21      0.8351
center     2     0.027583323 B             2.65767011       0.01      0.9918
center     3     1.384149384 B             2.26772684       0.61      0.5439
center     4    -0.437788226 B             2.41622554      -0.18      0.8568
center     5     0.000000000 B              .                .         .

                                                          H0:LSMean1=
                              Standard    H0:LSMEAN=0       LSMean2
drug    final LSMEAN             Error       Pr > |t|       Pr > |t|
D          9.8085482         0.9973089         <.0001         <.0001
P         16.9201677         0.8911077         <.0001
```

Repeat this analysis with MAR and MNAR examples.

		Interaction	*Drug Effect*	*Drug Effect*
Data	*Method*	*P-value*	*(Pbo − Drug) ± SE*	*P-value*
Full		.003	5.8 ± 1	< .0001
MCAR	CC	.019	7.1 ± 1	< .0001
MAR	CC	.071	7.4 ± 1	< .0001
MNAR	CC	.001	5.7 ± 1	< .0001

Last observation carried forward (LOCF)

Fill in the missing final values with baseline in a data step.

```
data MCAR_lcf;
  set MCAR;
  final_lcf =final;          create a new response variable
  if final=. then final_lcf=baseline;   fill in missing with baseline
data MAR_lcf;
  set MAR;
  final_lcf=final;
  if final=. then final_lcf=baseline;
data MNAR_lcf;
  set MNAR;
  final_lcf=final;
  if final=. then final_lcf=baseline;
```

```
                        MCAR last value carried forward

                               The GLM Procedure

                        Class         Levels    Values

                        drug               2    D P
                        center             5    1 2 3 4 5


                     Number of Observations Read         100
                     Number of Observations Used         100

                               The GLM Procedure

Dependent Variable: final_lcf
```

No missing data now, because we have filled all the holes.

```
Source                      DF     Type III SS    Mean Square   F Value   Pr > F
baseline                     1     5248.650484    5248.650484     74.77   <.0001
drug                         1      114.620844     114.620844      1.63   0.2046
center                       4      313.984729      78.496182      1.12   0.3531
drug*center                  4     1387.792987     346.948247      4.94   0.0012

                                              Standard
Parameter             Estimate                   Error    t Value    Pr > |t|
Intercept         -5.502658173 B        4.54351199      -1.21      0.2289
baseline           0.974066666          0.12750829       7.64      <.0001
drug       D      -3.951806237 B        1.81859271      -2.17      0.0323
drug       P       0.000000000 B         .                .          .
center     1      -3.996833406 B        3.12533295      -1.28      0.2041
center     2      -0.042933940 B        3.50340413      -0.01      0.9902
center     3      -2.298745160 B        3.01426080      -0.76      0.4476
center     4      -4.075868689 B        3.31014514      -1.23      0.2213
center     5       0.000000000 B         .                .          .

                                                       H0:LSMean1=
          final_lcf        Standard    H0:LSMEAN=0      LSMean2
drug         LSMEAN           Error       Pr > |t|       Pr > |t|

D        17.8405100       1.3147705         <.0001         0.0323
P        21.7923162       1.3191518         <.0001
```

Repeating this analysis with MAR and MNAR examples gives:

		Interaction	*Drug Effect*	*Drug Effect*
Data	*Method*	*P-value*	*(Pbo − Drug) ± SE*	*P-value*
Full		.003	5.8 ± 1	< .0001
MCAR	CC	.019	7.1 ± 1	< .0001
	LOCF	.001	4.0 ± 2	.032
MAR	CC	.071	7.4 ± 1	< .0001
	LOCF	.233	5.9 ± 2	.0003
MNAR	CC	.001	5.7 ± 1	< .0001
	LOCF	.006	8.1 ± 2	< .0001

Multiple Imputation: Proc MI + Proc MIanalyze

We want to estimate a parameter θ (eg. adjusted mean or regression coefficient) from data with missing values.

1. Proc MI For each missing value Y_i, generate M estimates y_{im}, $m = 1, \ldots, M$ using the distribution of observed values.

   Use MAR property: missingness related only to observed data.

   Fill in missing values in the data using each set $\{y_{im}\}$, to produce M complete data sets.

2. Fit a model to each of the M complete data sets to get a parameter estimate $\hat{\theta}_m$ with variance V_m (squared standard error).

3. Proc MIanalyze Combine the results of the M analyses.

   Combined estimate of θ is the average of the M estimates $\{\hat{\theta}_m\}$:

$$\bar{\theta}_M = \frac{1}{M}\sum_{1}^{M}\hat{\theta}_m.$$

   Variance of this estimate comes from the *within-imputation* variance,

   estimated by the mean $\bar{V}_M$ of the variances $\{V_m\}$,

   and the *between-imputation* variance

$$B_M = \frac{1}{M-1}\sum_{1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2,$$

   and so its standard error is:

$$\mathrm{SE}(\bar{\theta}_M) = \sqrt{\bar{V}_M + \frac{M+1}{M}B_M}.$$

Little & Rubin (2002) *Statistical Analysis with Missing Data, Second Edition*

For Depression Study example, imputation code will have 3 steps:

1. Proc MI generates M complete data sets, indexed by _Imputation_

2. Proc GLM fits the model, BY _Imputation_ , and outputs the results as a dataset (use ODS close listing to prevent writing them to the output window)

3. Proc MIanalyze reads output dataset, makes combined estimate $\bar{\theta}_M$ and $\mathrm{SE}(\bar{\theta}_M)$

An additional problem is that drug and center are CLASS variables and MIanalyze has problems with these. Need to add these indicators to data.

Make indicators for CLASS variables in MCAR, MAR, and MNAR data sets:

```
 data ph6470.hamd_MCAR;
   set mar;
   drugD = (drug="D");   logical variables to make indicators
   center1=(center=1);
   center2=(center=2);
   center3=(center=3);
   center4=(center=4);
   drugcenter_1 = drugD * center1;
   drugcenter_2 = drugD * center2;
   drugcenter_3 = drugD * center3;
   drugcenter_4 = drugD * center4;
```

Multiple Imputation SAS code

Step 1. Make 20 complete datasets using imputation

```
Proc MI data=ph6470.hamd_mcar   out=C   output data set
   nimpute=20   number of filled-in datasets
   seed=74950631
   minimum= 0  maximum= 40   reject values outside 0 - 40, range of HAMD
   round=1.0;  round to integer
   var  final;    variables to fill in
```

```
                         The MI Procedure

                        Model Information

        Data Set                             PH6470.HAMD_MCAR
        Method                               MCMC
        Multiple Imputation Chain            Single Chain
        Initial Estimates for MCMC           EM Posterior Mode
        Start                                Starting Value
        Prior                                Jeffreys
        Number of Imputations                20
        Number of Burn-in Iterations         200
        Number of Iterations                 100
        Seed for random number generator     74950631

                        Missing Data Patterns

                                                     --------Group Means--------
Group    baseline    final        Freq     Percent       baseline           final

    1    X           X              67       67.00      29.641791       13.686567
    2    X           .              33       33.00      31.212121               .
```

Step 2. Fit model in Proc GLM to each of the 20 imputed datasets.

Write results to output datasets—*see examples in Help Documentation for* Proc MIanalyze.

```
 ODS listing close;
 Proc GLM data=C;
    model final =  baseline drugD center1 center2 center3 center4
      drugcenter_1 drugcenter_2 drugcenter_3 drugcenter_4
       / inverse solution;
    by _Imputation_;
    ODS output  ParameterEstimates=glmparms  InvXPX=glmxpxi;
run;
ODS listing;
```

Step 3. Combine estimates.

```
Proc MIanalyze  parms=glmparms xpxi=glmxpxi ;
    modeleffects  Intercept baseline drugD
       center1 center2 center3 center4
       drugcenter_1 drugcenter_2 drugcenter_3 drugcenter_4;
```

Very difficult to figure out what output should be passed from procedures (step 2) to Proc MIanalyze.

Follow examples given in documentation for MIanalyze or use Google to look for examples.

MIanalyze: interaction model

```
                          The MIANALYZE Procedure

                            Parameter Estimates

Parameter            Estimate      Std Error    95% Confidence Limits        DF
drugD               -3.860516       3.986366     -11.7520       4.03099   121.86
center1              1.196681       3.582645      -5.9081       8.30146   103.68
center2             -2.553557       3.543349      -9.5269       4.41982   295.71
center3              0.812648       3.467136      -6.0503       7.67560   123.06
center4              1.073320       3.369253      -5.5580       7.70464   289.65
drugcenter_1        -3.860369       4.912036     -13.5889       5.86821   116.38
drugcenter_2         6.867292       5.583987      -4.1853      17.91987   123.66
drugcenter_3        -1.096627       4.830671     -10.6706       8.47731    109.3
drugcenter_4        -2.212305       4.919742     -11.9263       7.50166   164.53

                                                 t for H0:
          Parameter             Theta0   Parameter=Theta0    Pr > |t|
          drugcenter_1               0              -0.79      0.4335
          drugcenter_2               0               1.23      0.2211
          drugcenter_3               0              -0.23      0.8208
          drugcenter_4               0              -0.45      0.6535
```

Interaction significant?

From main-effects model:

```
                          The MIANALYZE Procedure

                            Parameter Estimates

Parameter        Estimate      Std Error    95% Confidence Limits        DF
Intercept       -4.967379       3.421668     -11.7158       1.78101   194.27
baseline         0.709821       0.098281       0.5157       0.90391   160.63
 drugD          -4.540941       1.288706      -7.0748      -2.00704   379.39
center1         -0.600204       2.325569      -5.1838       3.98341   216.81
center2          0.764762       2.596219      -4.3512       5.88071   225.56
center3          0.277993       2.152946      -3.9567       4.51271   341.24
center4          0.143256       2.413208      -4.6081       4.89457   267.26

                                     t for H0:
Parameter          Theta0   Parameter=Theta0   Pr > |t|
Intercept               0              -1.45     0.1482
baseline                0               7.22     <.0001
drugD                   0              -3.52     0.0005
```

Data	*Method*	*Interaction* *P-value*	*Drug Effect* *(Pbo − Drug) ± SE*	*Drug Effect* *P-value*
Full		.003	5.8 ± 1	< .0001
MCAR	CC	.019	7.1 ± 1	< .0001
	LOCF	.001	4.0 ± 2	.032
	MI	NS	4.5 ± 1	.0005
MAR	CC	.071	7.4 ± 1	< .0001
	LOCF	.233	5.9 ± 2	.0003
	MI	NS	4.8 ± 1	.0001
MNAR	CC	.001	5.7 ± 1	< .0001
	LOCF	.006	8.1 ± 2	< .0001
	MI	NS	3.7 ± 1	.0036

Imputing values when data are not missing at random can lead to severe bias.

Difficult questions with missing data imputation:

1. Do you have missing at random? How do you know?
2. How do you choose an imputation method?

   How can you use what you know to improve the process of imputation?

References:

Dmitrienko *et. al.* (2005) *Analysis of Clinical Trials Using SAS*, Chapter 5

R Little and D Rubin (2002) *Statistical Analysis with Missing Data, Second Edition*

Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials* 1999; 20:408–422.