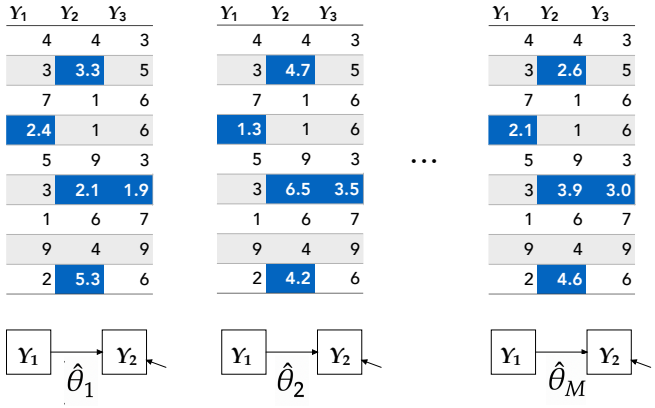
11. Multiple Imputation: Analysis And Pooling Phases

Multiple Imputation Step 2: Perform Analysis on Each Data Set



Paired-Samples t Test

Regression parameterization for a dependent-samples t test evaluating change between the two assessments, Y_1 and Y_2

Mean change

$$(Y_{2i} - Y_{1i}) = \beta_0 + \varepsilon_i = E(Y_\Delta) + \varepsilon_i$$

$$Y_{\Delta i} \sim N\left(E(Y_\Delta), \sigma^2_\varepsilon\right)$$

Change score ($Y_{\Delta i}$)

Math Achievement Data

Math achievement data for 250 students

The data set includes pre-test and post-test math achievement scores and academic-related variables such as math self-efficacy, math anxiety, standardized reading scores, socio-demographic variables

math.dat

Variable	Name	Missing %	Scaling
Identifier variable	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Free or reduced lunch	LUNCHASST	4.3	0 = none, 1 = assistance
Achievement group	ACHIEVEGRP	2.0	1 = learning disability, 2 = low achieving, 3 = average achieving
Standardized reading	STANREAD	10.0	Continuous
Math self-efficacy	EFFICACY	9.7	6-point ordinal scale
Math anxiety	ANXIETY	9.3	Continuous
Pre-test math achievement	MATHPRE	0	Continuous
Post-test math achievement	MATHPOST	18.0	Continuous

Substantive Example

Do math scores improve between the pre-test and post-test assessments?

$$(MATHPOST_i - MATHPRE_i) = \beta_0 + \varepsilon_i$$

Pre-test scores are complete, 18% of the post-test scores are missing and need to be imputed

Paired-Samples t Test With Imputed Data

1. Compute change scores, $Y_\Delta = Y_2 - Y_1$, within each filled-in data set

2. Fit the regression model to each data set

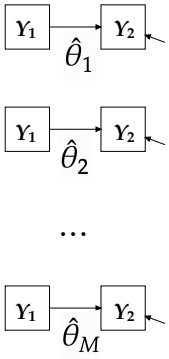
3. Pool (average) estimates and standard errors

Dataset-Specific Estimates

The average change score and its standard error vary across imputed data sets

Imputation	$\beta_{0(m)}$	SE
1	6.577	0.576
2	6.841	0.573
3	6.154	0.570
4	6.623	0.572
5	6.671	0.574
…	…	…
20	6.698	0.576

Multiple Imputation Step 3: Combine Estimates and Standard Errors



$$\hat{\theta} = \left(\hat{\theta}_1 + \hat{\theta}_2 + \cdots + \hat{\theta}_M\right)/M$$

Pooling Parameter Estimates

The multiple imputation regression coefficient is the arithmetic average of the M estimates

$$\hat{\beta}_0 = \left(\beta_{0(1)} + \beta_{0(2)} + \cdots + \beta_{0(M)}\right) \div M$$

$$\hat{\beta}_0 = (6.577 + 6.841 + \cdots + 6.698)/20 = 6.546$$

On average, scores improved by 6.55 points

Pooling Standard Errors

Each standard error estimates sampling variation that would have resulted from a complete data set

Pooling incorporates an adjustment that reflects the additional error from missing data

Standard error = complete-data sampling error + additional variation due to missing values

Within-Imputation Variance

var_W is the average squared standard error from the complete data sets

$$var_W = \sum_{m=1}^{M} SE_{(m)}^2 \div M$$

var_W estimates the squared standard error that would result had there been no missing data

Within-Imputation Variance Example

$$var_W = (.332 + .329 + \cdots + .332)/20 = .318$$

Imputation	$\beta_{0(m)}$	SE	SE^2
1	6.577	0.576	0.332
2	6.841	0.573	0.329
3	6.154	0.570	0.324
4	6.623	0.572	0.327
5	6.671	0.574	0.330
…	…	…	
20	6.698	0.576	0.332

$var_W = .318$

Between-Imputation Variance

Between-imputation variance is the sample variance formula applied to the M estimates

$$var_B = \sum_{m=1}^{M} \left(\beta_{0(m)} - \hat{\beta}_0\right)^2 \div (M-1)$$

var_B reflects additional variation in the estimates that arises from using different imputations

Between-Imputation Variance Example

$$var_B = \frac{(6.577\text{-}6.546)^2 + (6.841\text{-}6.546)^2 + \ldots + (6.577\text{-}6.546)^2}{20} = .041$$

Imputation	$\beta_{0(m)}$	$\hat{\beta}_0$	$\beta_{0(m)} - \hat{\beta}_0$
1	6.577	6.546	0.031
2	6.841	6.546	0.295
3	6.154	6.546	-0.392
4	6.623	6.546	0.077
5	6.671	6.546	0.125
…	…	…	
20	6.698	6.546	0.152

$var_B = .041$

Pooled Standard Error

The pooled standard error combines complete-data sampling error and missing data uncertainty

$$SE = \sqrt{var_W + var_B + \frac{var_B}{M}}$$

$$= \sqrt{.318 + .041 + \frac{.041}{20}} = .601$$

Significance Testing

A single-*df* test statistic is based on the pooled estimate and standard error

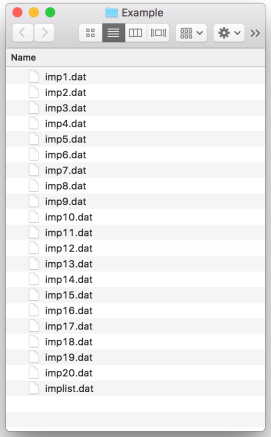
$$t = \frac{\hat{\beta}_0 - \theta_0}{SE} = \frac{6.546 - 0}{.601} = 10.889$$

Software packages use different reference distributions (e.g., t with df adjustments, z)

Mplus Imputation Format

Mplus requires imputed data sets as separate files

Blimp creates a text file containing the names of the data sets, and this file serves as the input data for subsequent analyses



Mplus Imputation Analysis Script

```
DATA:
file = imps_list.dat;
type = imputation;
VARIABLE:
names = id male lunchasst achievegrp stanread efficacy
    anxiety mathpre mathpost;
usevariables = change;
DEFINE:
change = mathpost - mathpre;
MODEL:
[change];
change;
```


Mplus Output

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value
 Means
    CHANGE           6.578      0.598     10.992      0.000
 Variances
    CHANGE          79.512      7.470     10.645      0.000
```

R Imputation Analysis Script

```
library(mitml)

# read stacked data and compute change scores.
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
impdata <- read.csv(paste0(getwd(), "/imps.csv"), header = F)
names(impdata) <- c("imputation", "id", "male", "lunchasst", "achievegrp", "stanread",
  "efficacy", "anxiety", "mathpre", "mathpost")
impdata$change <- impdata$mathpost - impdata$mathpre

# analysis and pooling.
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(change ~ 1))
estimates <- testEstimates(analysis, var.comp = T)
estimates
```


R Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

             Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)     6.546     0.601    10.889  1348.470     0.000     0.135     0.120

                 Estimate
Residual~~Residual   79.621

Unadjusted hypothesis test as appropriate in larger samples.
```


SAS Imputation Analysis Script

```
/* read stacked imputation data */
data impdata (where = (_imputation_ gt 0));
infile '/folders/myfolders/imps_stacked.dat';
input _imputation_ id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost;
change = mathpost - mathpre;
run;

/* analyze imputations */
proc reg data = impdata outest = estimates covout noprint;
model change = ;
by _imputation_; run;

/* pool estimates and standard errors */
proc mianalyze data = estimates;
modeleffects Intercept ; run;
```


SAS Analysis Output

The MIANALYZE Procedure

Model Information	
Data Set	WORK.ESTIMATES
Number of Imputations	20

Variance Information (20 Imputations)

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Intercept	0.040854	0.318483	0.361379	1348.5	0.134689	0.120006	0.994036

Parameter Estimates (20 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	6.545988	0.601148	5.366701	7.725274	1348.5	6.154129	6.841431	0	10.89	<.0001

SPSS Imputation Analysis Script

```
* set working directory.
CD "YOUR-FILE-PATH".

* read stacked imputation data and compute change score.
DATA LIST free file = "imps_stacked.dat"
 /imputation_ id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost.
MISSING VALUES all (999).
COMPUTE change = mathpost - mathpre.

* initiate pooling routines.
SORT CASES by imputation_.
SPLIT FILE layered by imputation_.

* analysis and pooling.
MIXED change
 /print = testcov solution
 /method = reml
 /fixed = intercept.
```


SPSS Analysis Output

Estimates of Fixed Effects[a]

imputation_	Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
.00	Intercept	6.451923	.621810	207	10.376	.000	5.226030	7.677816
1.00	Intercept	6.577354	.576471	249	11.410	.000	5.441973	7.712735
2.00	Intercept	6.841431	.573235	249	11.935	.000	5.712423	7.970440
3.00	Intercept	6.154129	.569563	249	10.805	.000	5.032353	7.275904
4.00	Intercept	6.623176	.572236	249	11.574	.000	5.496137	7.750216
14.00	Intercept	6.788940	.559447	249	12.135	.000	5.687089	7.890792
15.00	Intercept	6.230578	.578870	249	10.763	.000	5.090471	7.370684
16.00	Intercept	6.472906	.556788	249	11.625	.000	5.376292	7.569520
17.00	Intercept	6.404010	.554772	249	11.543	.000	5.311367	7.496654
18.00	Intercept	6.595235	.549586	249	12.000	.000	5.512806	7.677664
19.00	Intercept	6.412370	.554625	249	11.562	.000	5.320016	7.504723
20.00	Intercept	6.697848	.576228	249	11.624	.000	5.562945	7.832751
Pooled	Intercept	6.545988	.601148		10.889	.000	5.366701	7.725274

a. Dependent Variable: change.

Stata Imputation Analysis Script

```
// set working directory
cd "YOUR-FILE-PATH"

// read stacked data
clear
infile imp id male lunchasst achievegrp stanread efficacy anxiety mathpre mathpost
   using "imps_stacked.dat"

// recode missing data in original data (imp = 0)
recode male - mathpost (999 = .)

// compute change scores
generate change = mathpost - mathpre

// convert to mi data, analyze and pool
mi import flong, m(imp) id(id) imputed(male - change) clear
mi estimate, cmdok: regress change
```


Stata Output

```
Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =        250
                                                Average RVI       =     0.1347
                                                Largest FMI       =     0.1210
                                                Complete DF       =        249
                                                DF:     min       =     187.44
                                                        avg       =     187.44
DF adjustment:   Small sample                           max       =     187.44
                                                F(   0,      .)   =          .
Within VCE type:          OLS                   Prob > F          =          .

------------------------------------------------------------------------------
      change |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   6.545988   .6011479    10.89   0.000     5.360103    7.731872
------------------------------------------------------------------------------
```