AN INTRODUCTION TO MISSING DATA ANALYSES

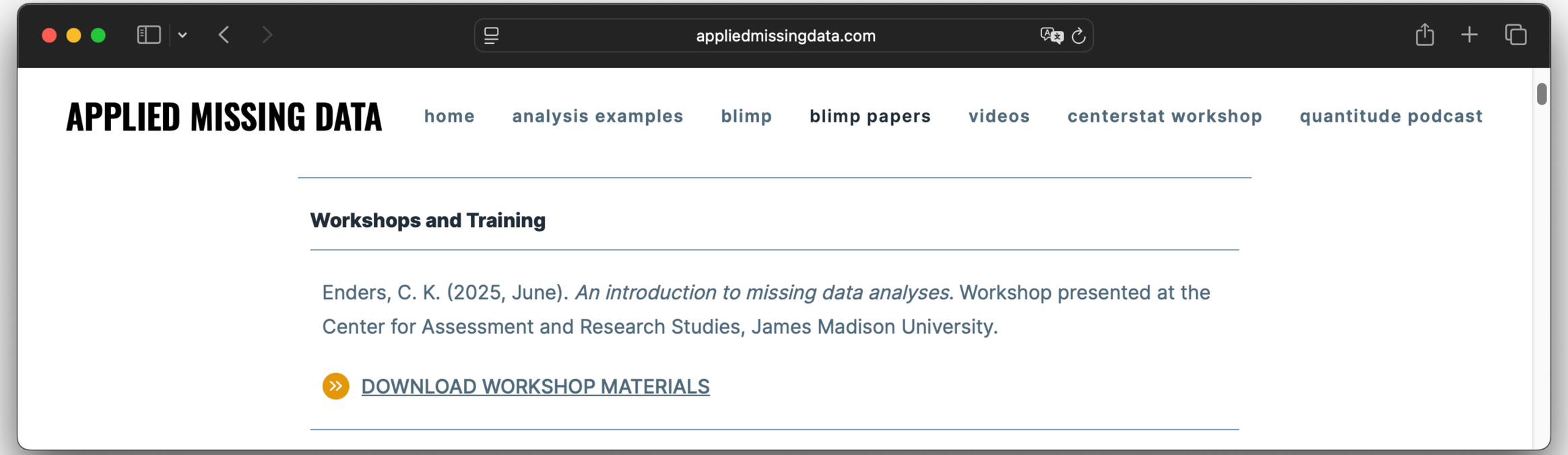
Workshop Sponsored by the Center for Assessment and Research Studies

Craig Enders

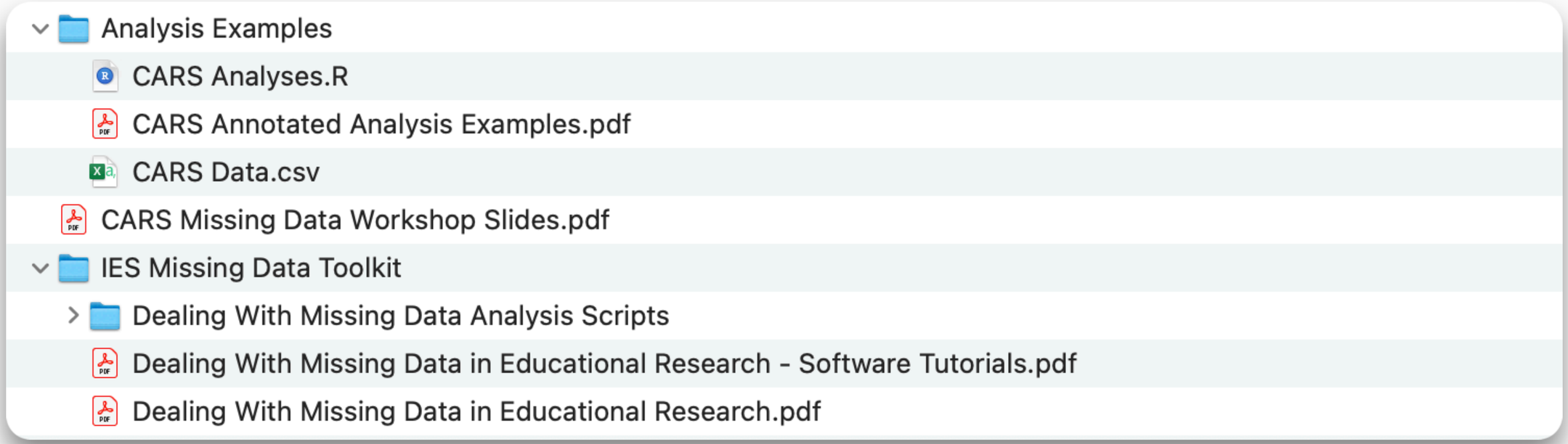
UCLA Department of Psychology

COURSE MATERIAL DOWNLOAD

WWW.APPLIEDMISSINGDATA.COM/BLIMP-PAPERS

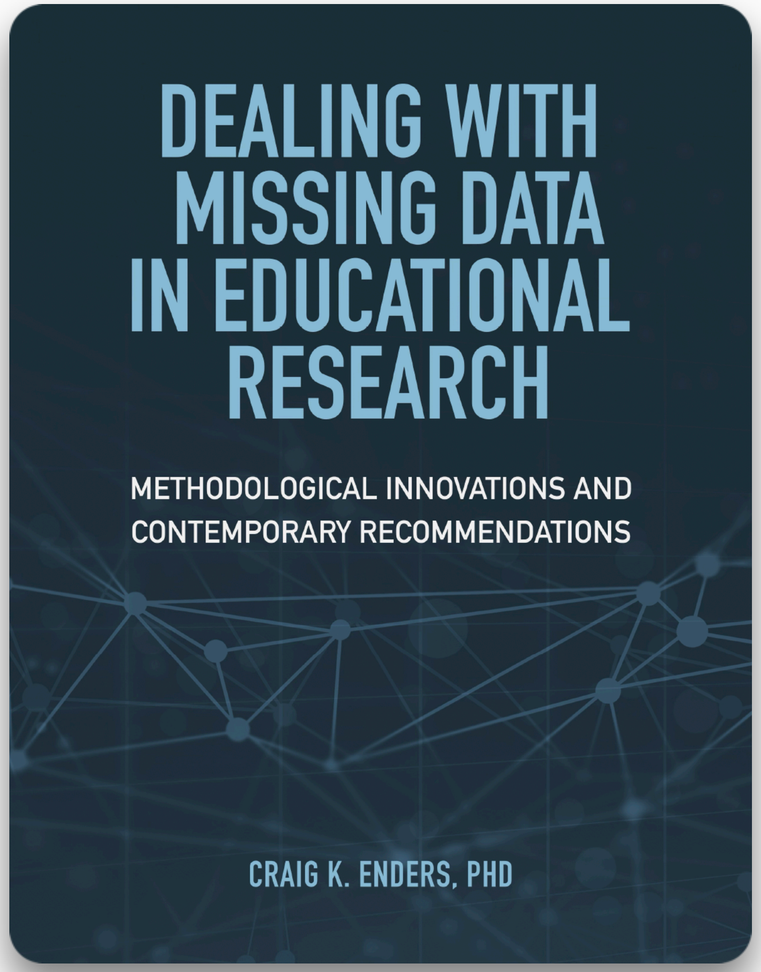


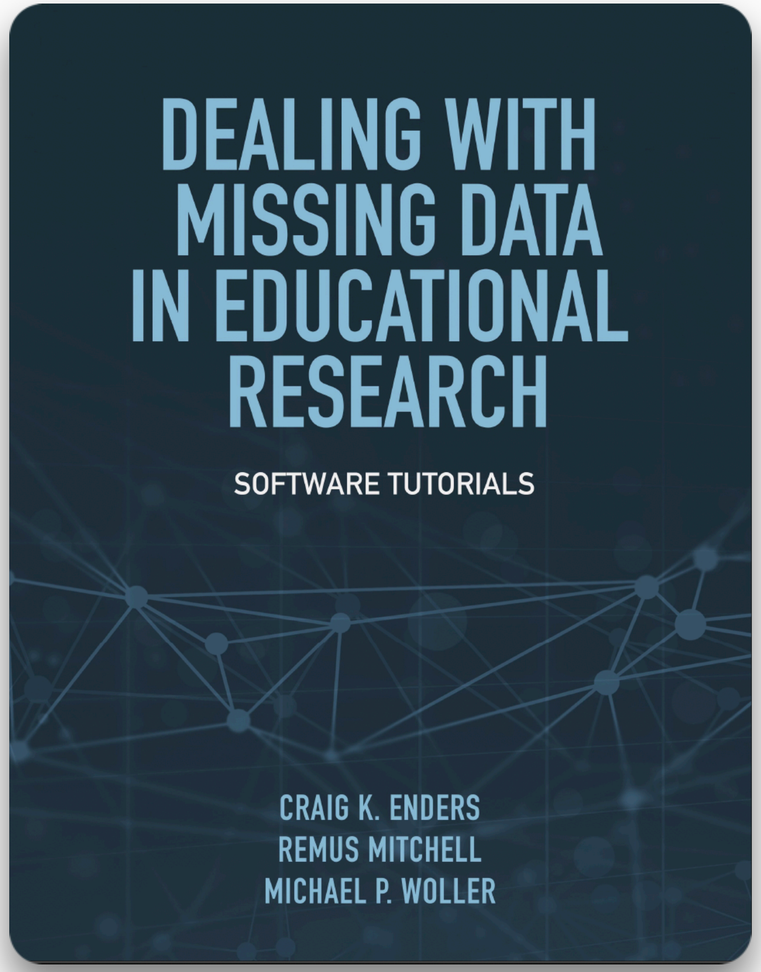
WORKSHOP MATERIALS



IES MISSING DATA TOOLKIT

IES Institute of Education Sciences

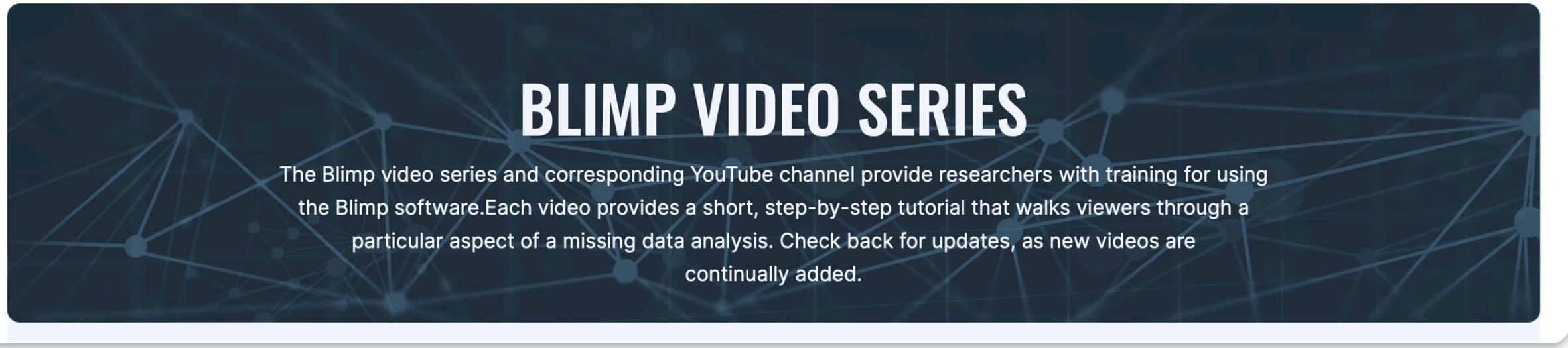




IES MISSING DATA TOOLKIT

IES Institute of Education Sciences

WWW.APPLIEDMISSINGDATA.COM/VIDEOS



INSTALLING BLIMP

IES Institute of Education Sciences

WWW.APPLIEDMISSINGDATA.COM/BLIMP

BLIMP 3.0

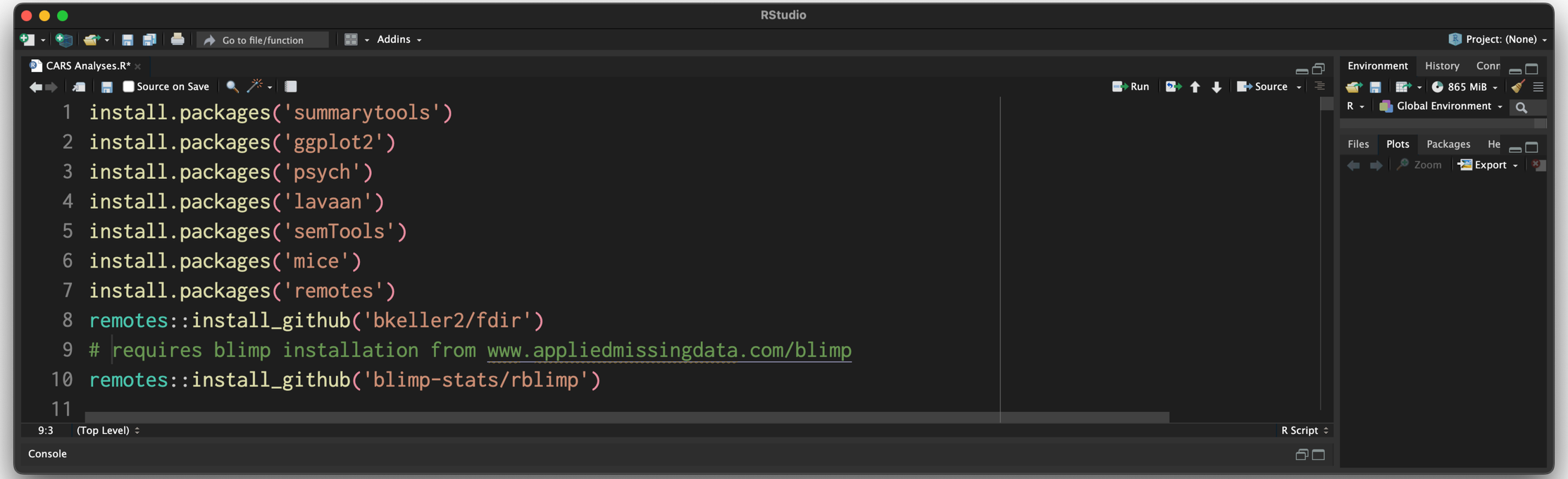
Blimp 3 offers powerful latent variable modeling and imputation for incomplete data sets with up to three levels. Blimp's unique Bayesian computational architecture allows easy specification of complex analyses that are difficult or impossible to fit in other software packages.

Download Now

User's Guide

INSTALLING R PACKAGES

- The CARS Analysis.R script includes package installation commands at the top of the file



MORNING OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

MORNING OUTLINE

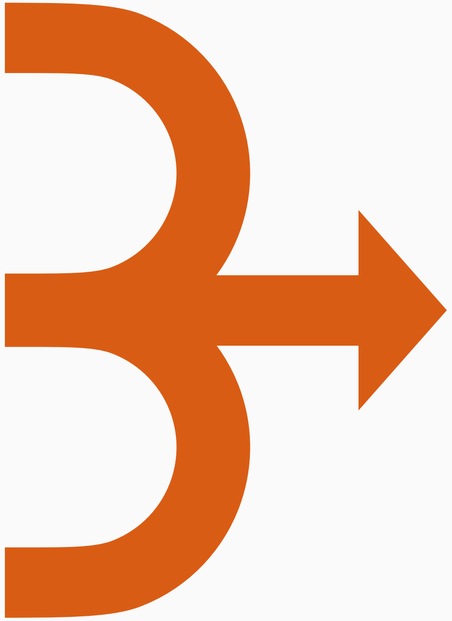
1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

MODERN MISSING DATA METHODS

Maximum likelihood

Bayesian MCMC estimation

Multiple imputation



the Big Three

KEY ADVANTAGES OF BIG THREE

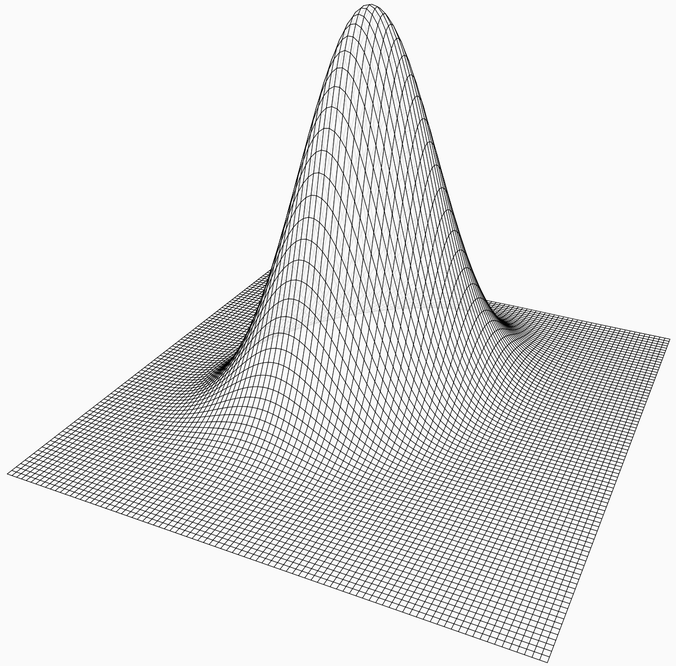
- Achieve unbiasedness with a more realistic assumption about the missing data process
- Allow for alternate assumptions about nonresponse process
- Use all available data, no wasted resources
- Maximize power

CHOOSING A MISSING DATA METHOD

- All things being equal—same data, same variables, same assumptions—the Big Three rarely produce different results
- Missing data analyses require distributional assumptions for variables that wouldn't usually require them (e.g., predictors)
- How we represent those distributions is what matters

MODELING FRAMEWORKS

Multivariate modeling

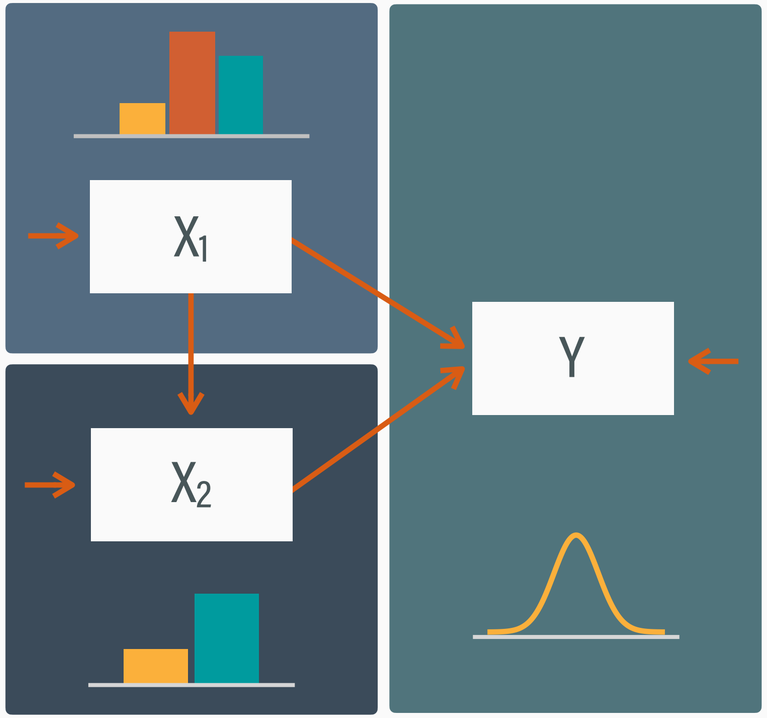


- Classic versions of the Big Three typically assume multivariate normality
- This includes most applications of maximum likelihood and multiple imputation

MODELING FRAMEWORKS

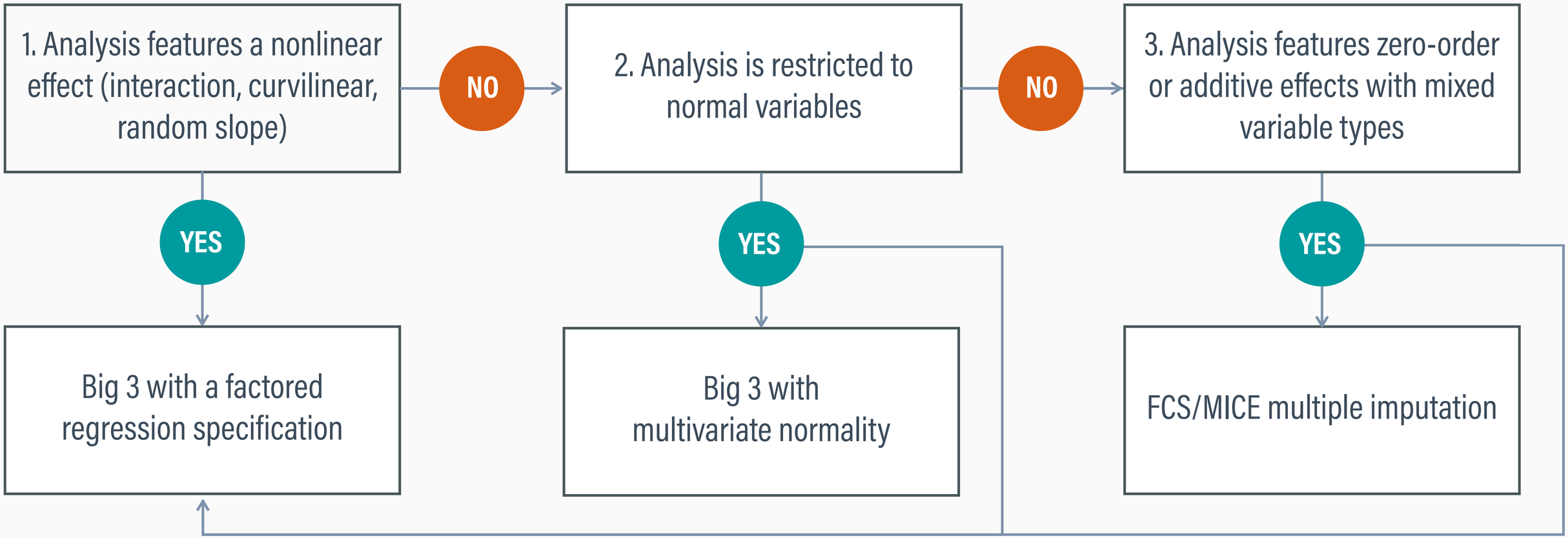
Multivariate modeling

Factored regression specification



- Factored specifications invoke a unique submodel and distribution for each variable
- Submodels can include terms that are at odds with multivariate normality (e.g., discrete variables, interactions, random slopes)

MISSING DATA DECISION TREE



MORNING OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

HOW MUCH MISSING DATA IS TOO MUCH?

- The Big Three can tolerate substantial amounts of missing data
- The Big Three are increasingly better than ad hoc methods (e.g., deleting incomplete records) as missingness increases
- The amount of missing data is less important than why the data are missing (the missingness process or mechanism)

RUBIN'S MISSING DATA MECHANISMS

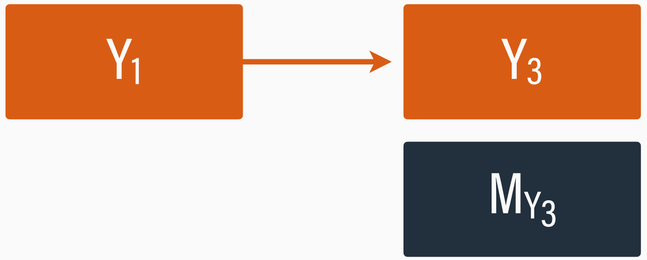
- Missing data mechanisms (processes) describe different ways in which the data relate to nonresponse
- Missingness may be completely random or systematically related to different parts of the data
- Mechanisms function as statistical assumptions that determine our ability to accurately recover the parameter values

PARTITIONING THE DATA

Complete			=	Observed			+	Missing			Indicators		
Y_1	Y_2	Y_3		Y_1	Y_2	Y_3		Y_1	Y_2	Y_3	M_1	M_2	M_3
4	4	3		4	4	3					0	0	0
3	3	5		3	NA	5			3		0	1	0
7	1	6		7	1	6					0	0	0
2	1	6		NA	1	6		2			1	0	0
5	9	3	=	5	9	3	+				0	0	0
3	2	2		3	NA	NA			2	2	0	1	1
1	6	7		1	6	7					0	0	0
9	4	9		9	4	9					0	0	0
2	5	6		2	NA	6			5		0	1	0

* see Rubin (1976) in Biometrica

MISSING COMPLETELY AT RANDOM



- The probability of missing values is completely unrelated to the data

$$f(M = 1 \mid data_{obs}, data_{mis}) = f(M = 1)$$

- MCAR is purely random missingness
- We don't care about this process or testing for it (e.g., Little's MCAR test)

Missingness Indicators

M_1	M_2	M_3
0	0	0
0	1	0
0	0	0
1	0	0
0	0	0
0	1	1
0	0	0
0	0	0
0	1	0

Predictors of nonresponse

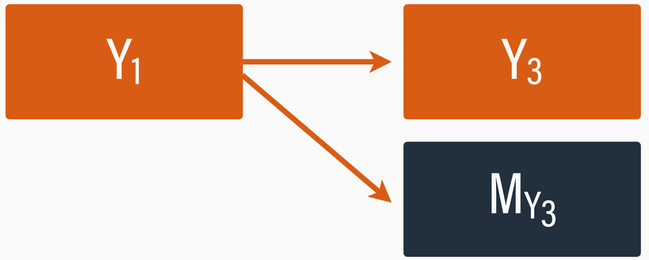
Observed

Y_1	Y_2	Y_3
4	4	3
3	NA	5
7	1	6
NA		6
5		3
3	NA	NA
1	6	7
9	4	9
2	NA	6

Missing

Y_1	Y_2	Y_3
	3	
2		
	2	2
	5	

(CONDITIONALLY) MISSING AT RANDOM



- Systematic missingness related to the observed data but unrelated to the unseen latent data

$$f(M = 1 \mid data_{obs}, data_{mis}) = f(M = 1 \mid data_{obs})$$

- Most Big Three applications assume CMAR

Missingness

Indicators

M_1	M_2	M_3
0	0	0
0	**1**	0
0	0	0
1	0	0
0	0	0
0	**1**	**1**
0	0	0
0	0	0
0	**1**	0

Predictors of nonresponse

Observed

Y_1	Y_2	Y_3
4	4	3
3	**NA**	5
7	1	6
NA	1	6
5	9	3
3	**NA**	**NA**
1	6	7
9	4	9
2	**NA**	6

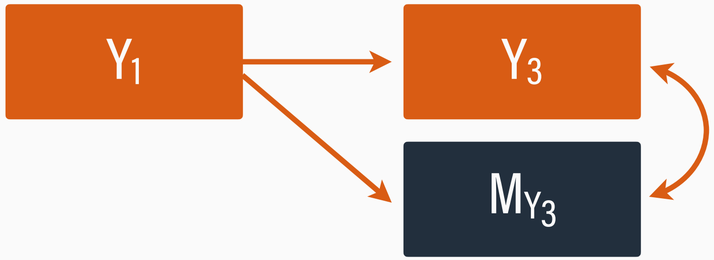
Missing

Y_1	Y_2	Y_3
	3	
2		
	2	2
	5	

CONDITIONALLY MAR EXAMPLE

- CARS is interested in assessing whether information literacy scores differ across the three time points.
- Transfer students are missing the first-year assessment
- Students who drop out are missing later assessments
- These examples could classify as CMAR if the reasons for missingness are unrelated to the unseen literacy scores

MISSING NOT AT RANDOM



- Systematic missingness related to the observed data and the unseen latent data

$$f(M = 1 \mid data_{obs}, data_{mis})$$

- The Big Three also allow MNAR processes (selection and pattern mixture models)

Missingness

Indicators

M_1	M_2	M_3
0	0	0
0	**1**	0
0	0	0
1	0	0
0	0	0
0	**1**	**1**
0	0	0
0	0	0
0	**1**	0

Predictors of nonresponse

Observed

Y_1	Y_2	Y_3
4	4	3
3	**NA**	5
7	1	6
NA	1	6
5	9	3
3	**NA**	**NA**
1	6	7
9	4	9
2	**NA**	6

Missing

Y_1	Y_2	Y_3
	3	
2		
	2	**2**
	5	

CARS is interested in assessing whether information literacy scores differ across the three time points. A missing not at random process would occur if a student's unseen literacy score relates to whether they have a score (missingness). In small groups, discuss whether you think this process is plausible.

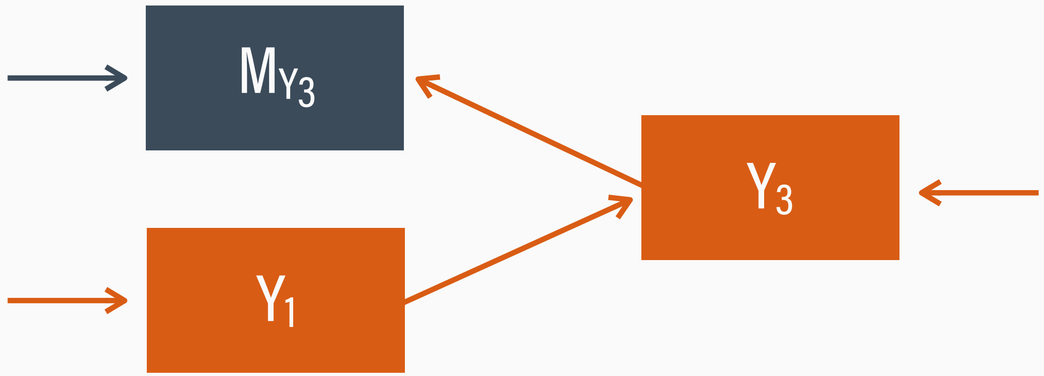
TESTING THE CMAR ASSUMPTION

- The CMAR assumption is untestable because it stipulates no relation between missingness and the unseen scores
- We must rely on logical arguments about why the unseen scores should not be related to missingness
- When in doubt, conduct sensitivity analyses that compare the estimates from CMAR and MNAR assumptions

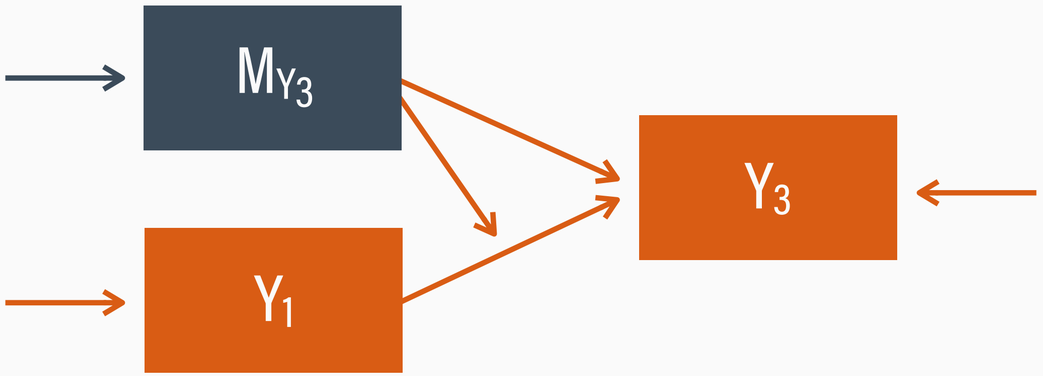
MNAR MODELING

- Missing not at random processes require an explicit model that incorporates missing data indicators into the analysis

Selection Model



Pattern Mixture Model



MNAR-BY-OMISSION PROCESS

= analysis variables
= unused variable

- CMAR is satisfied when Y and M_Y are uncorrelated
- Analyzing the data without A induces a spurious correlation between Y and its indicator M_Y (an MNAR process)
- Variable A must be in the missing data analysis to avoid nonresponse bias

Data-Generating Model

MNAR-by-Omission Process

AUXILIARY VARIABLES

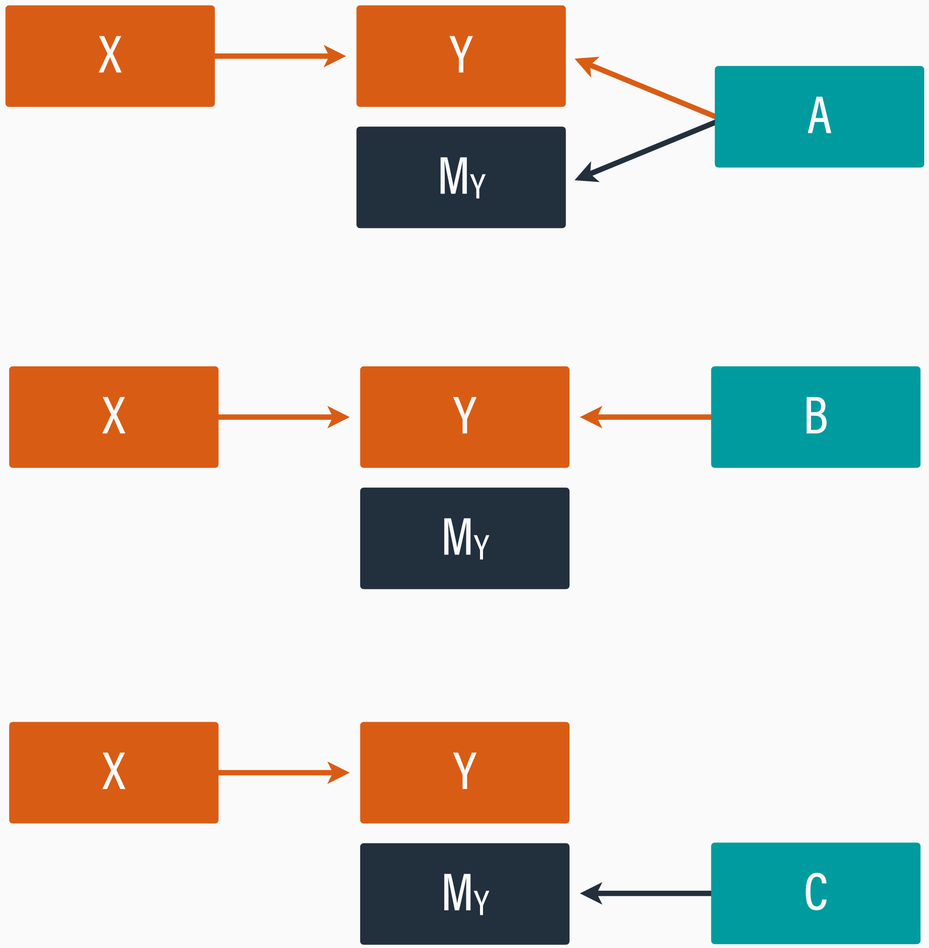
- The CMAR assumption holds when the observed data in a particular analysis model completely explain missingness
- The analysis variables usually constitute a small subset of the available variables in a data set
- The literature often recommends an inclusive strategy that includes auxiliary variables that are not in the focal analysis

AUXILIARY VARIABLE TYPOLOGY

$M_Y = 1$ if missing

$M_Y = 0$ if complete

- Auxiliary variables may correlate with only the residuals of Y (Type B), only the missingness of Y (Type C), or both (Type A)
- Type A variables are most important because they can induce bias if ignored
- Type B variables can improve power, and type C variables are unhelpful



* see Collins et al. (2001) in Psychological Methods

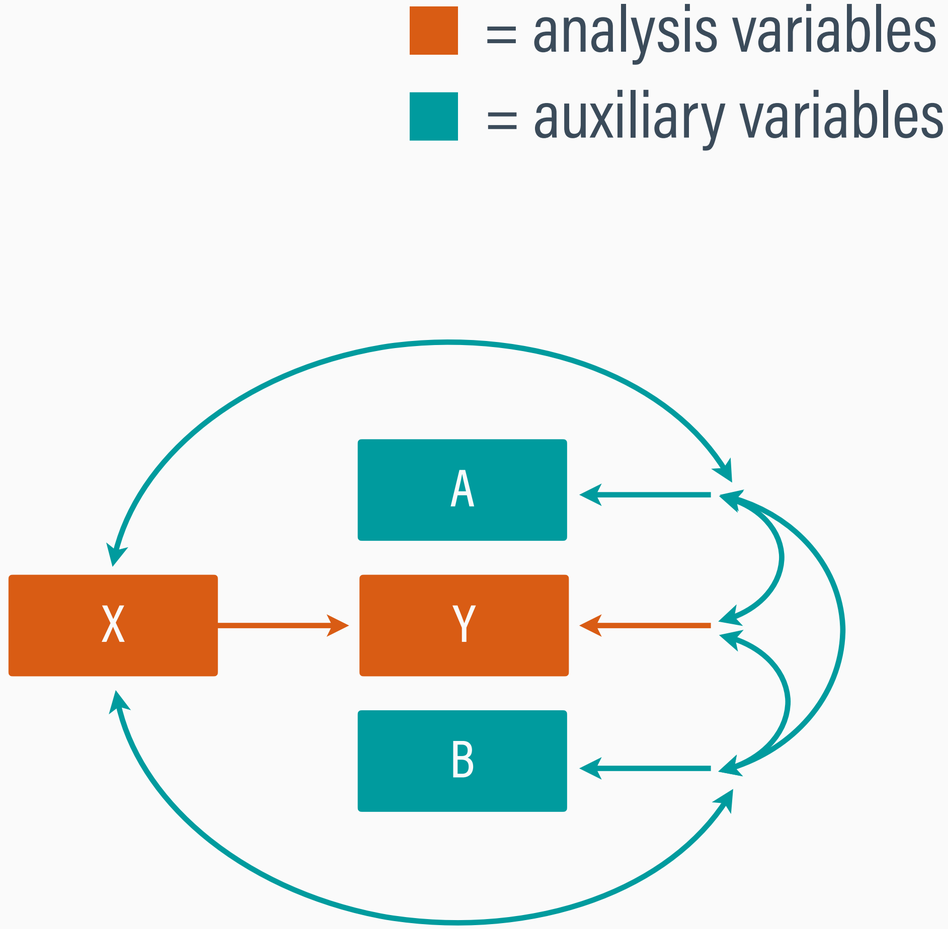
SATURATED CORRELATES MODEL

- Auxiliary variables enter a model via correlations and residual correlations

- Auxiliary variables correlate with …

  1. Each other
  2. Exogenous predictors
  3. The residual(s) of any outcomes

- Available in the R semTools package



* see Graham (2003) in Structural Equation Modeling

SEQUENTIAL SPECIFICATION

= analysis variables

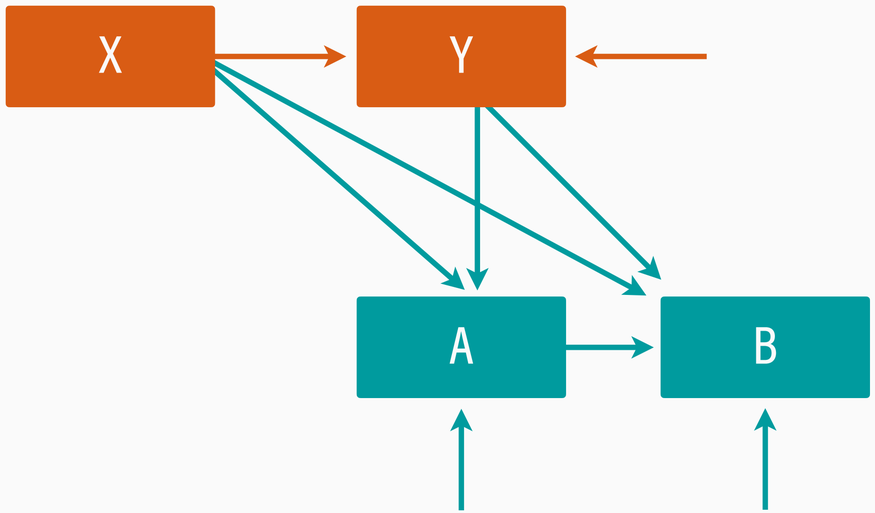
= auxiliary variables

- Auxiliary variables enter model as extra dependent variables

- Auxiliary variables are regressed on …

  1. Analysis variables
  2. Each other in a sequence

- Flexible and simple to implement



* see Lüdtke et al. (2020) in Psychological Methods

MORNING OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

MAXIMUM LIKELIHOOD

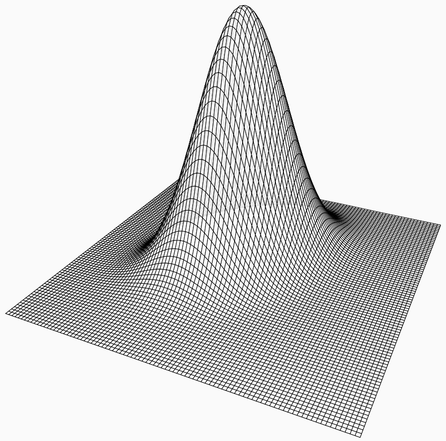
- Maximum likelihood estimation identifies that parameter values that are most likely to have produced the sample data
- Like OLS in the sense that estimation minimizes squared distances between a model's predictions and the data
- The normal distribution quantifies the size of the residuals

MOTIVATING EXAMPLE

- CARS staff want to assess whether information literacy changes between the first and third assessment
- Descriptive statistics (means, variances, and the covariance) are estimated from incomplete data
- For now, we focus on the mechanics of maximum likelihood with complete data

MULTIVARIATE NORMAL LIKELIHOOD

- The multivariate normal distribution function provides the mathematical machinery for maximum likelihood estimation



Scaling term that makes area under curve equal 1

Standardized distances between the data and parameters

$$\text{likelihood}_i = (2\pi)^{(-.5V)}|\boldsymbol{\Sigma}|^{-.5} \times \exp\left(-\frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})\right)$$

$$\text{height coordinate}_i = \text{scaling terms} \times \exp\left(-\frac{1}{2}\text{sum of squared z-scores}\right)$$

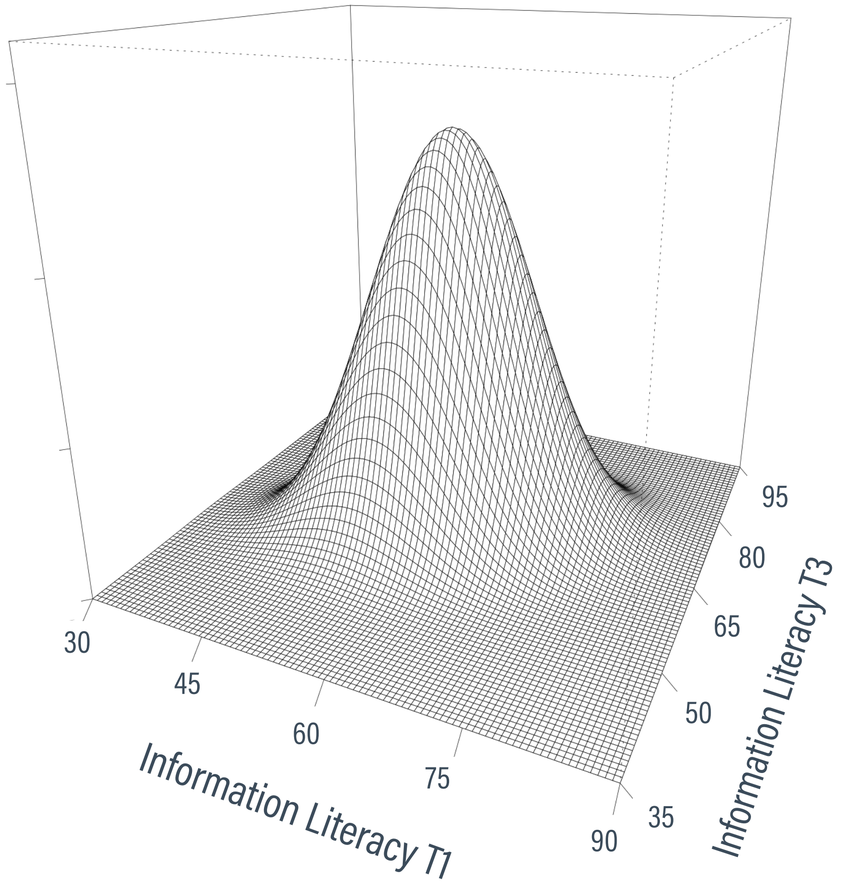
DISTRIBUTION KERNEL

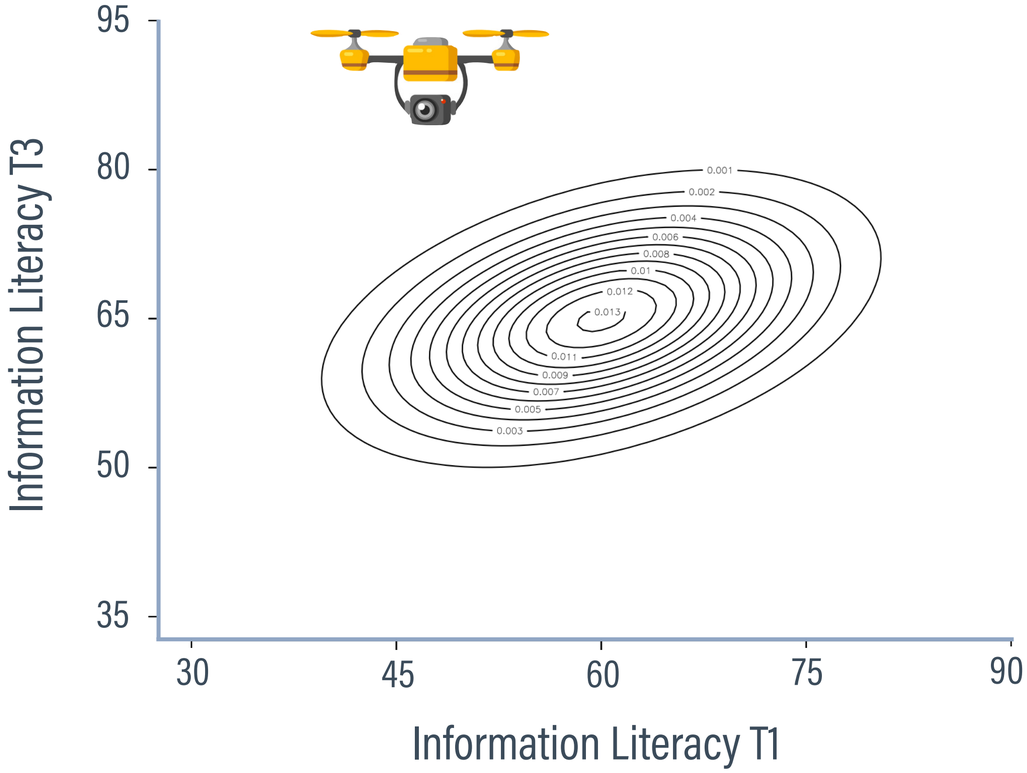
- The kernel of the distribution is key, as it defines a person's data-model fit as the sum of squared standardized distances

$$\exp\left(-\frac{1}{2}\text{sum of squared z-scores}\right) = \exp\left(-\frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})\right)$$

$$= \exp\left(-\frac{1}{2}\begin{pmatrix}\text{info}_{1i} - \mu_1 \\ \text{info}_{3i} - \mu_3\end{pmatrix}'\begin{pmatrix}\sigma_1^2 & \sigma_{13} \\ \sigma_{31} & \sigma_3^2\end{pmatrix}^{-1}\begin{pmatrix}\text{info}_{1i} - \mu_1 \\ \text{info}_{3i} - \mu_3\end{pmatrix}\right)$$

MULTIVARIATE NORMAL CONTOURS

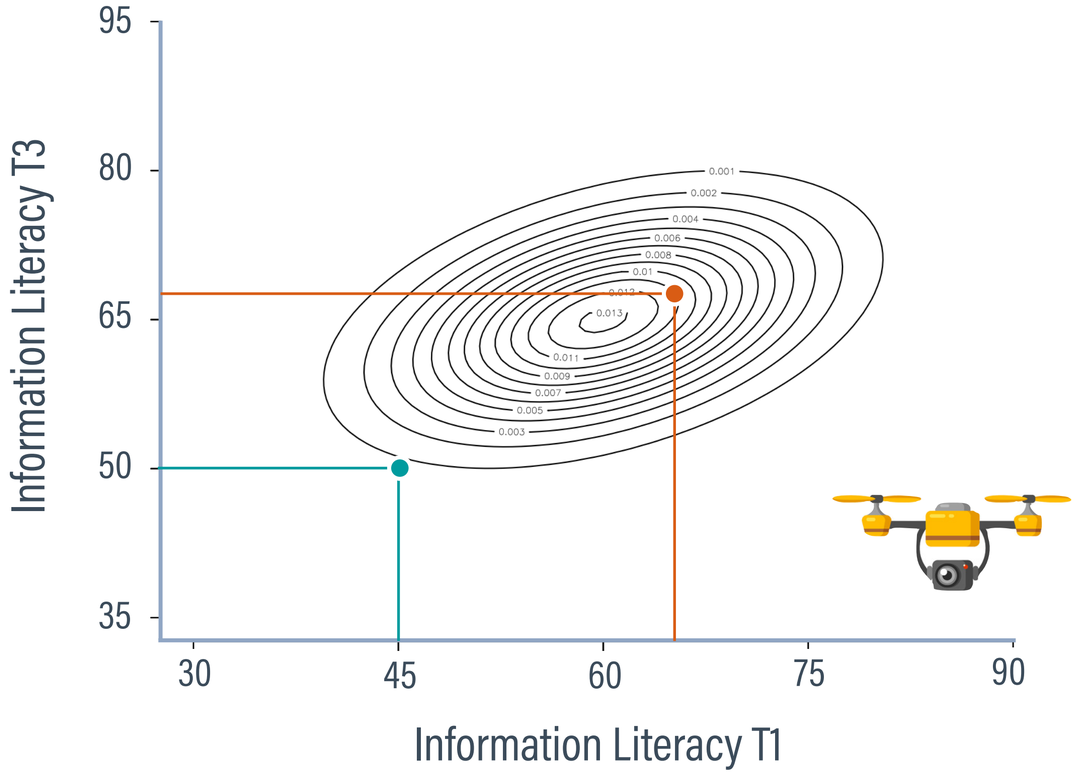




LIKELIHOOD IS A HEIGHT COORDINATE

- Likelihood(**Y** = 65 and 67 | μ, σ, ρ) = .0007
- Likelihood(**Y** = 45 and 50 | μ, σ, ρ) = .0003

- Assume μ = (60, 65), σ = 15, and ρ = .40
- Substituting parameters and data into the function returns height coordinates
- Likelihood = size of residual distance (fit) expressed as a height coordinate



INDIVIDUAL LOG-LIKELIHOOD

- Taking the natural log of each person's likelihood expresses probability-like fit quantities on a more tractable metric

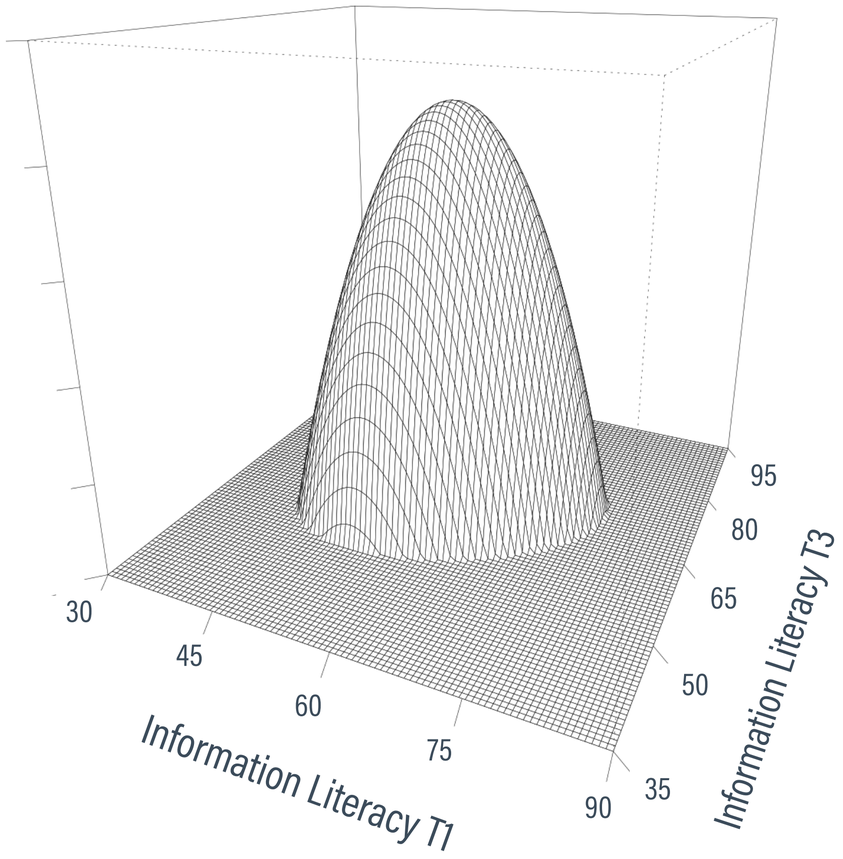
Scaling term that makes area under curve equal 1

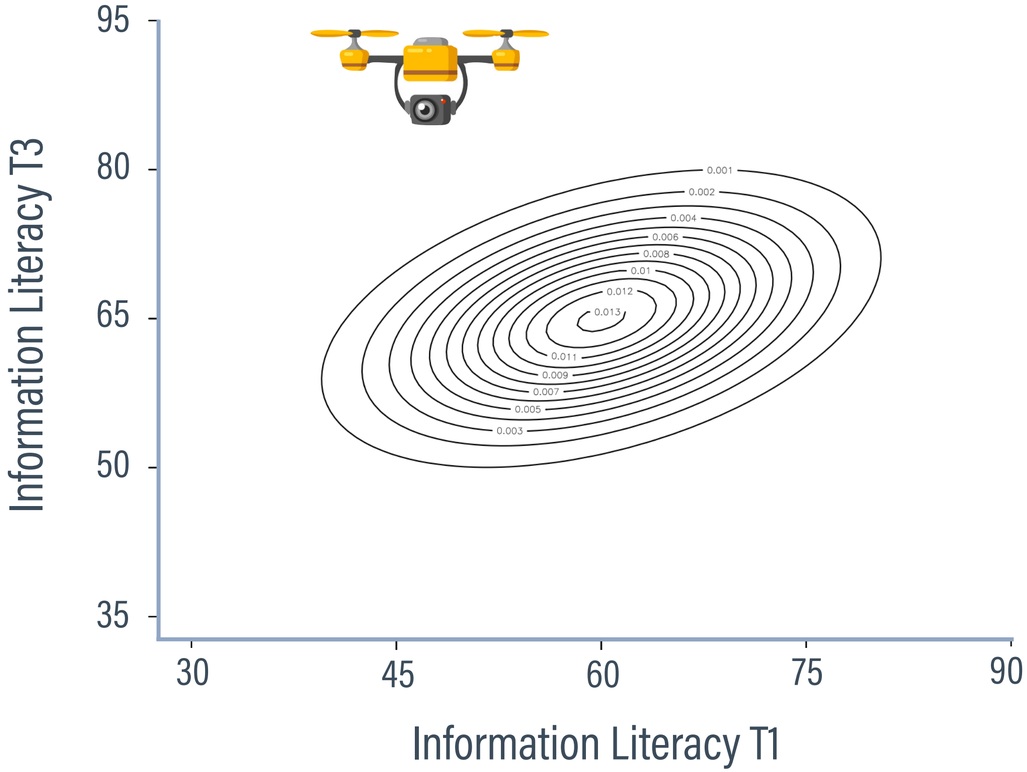
Standardized distances between the data and parameters

$$\text{log-likelihood}_i = -\frac{V}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})$$

$$\text{height coordinate}_i = \text{scaling terms} + \left(-\frac{1}{2}\text{sum of squared z-scores}\right)$$

LOG-LIKELIHOOD CONTOURS





LOG-LIKELIHOOD IS ALSO A HEIGHT COORDINATE

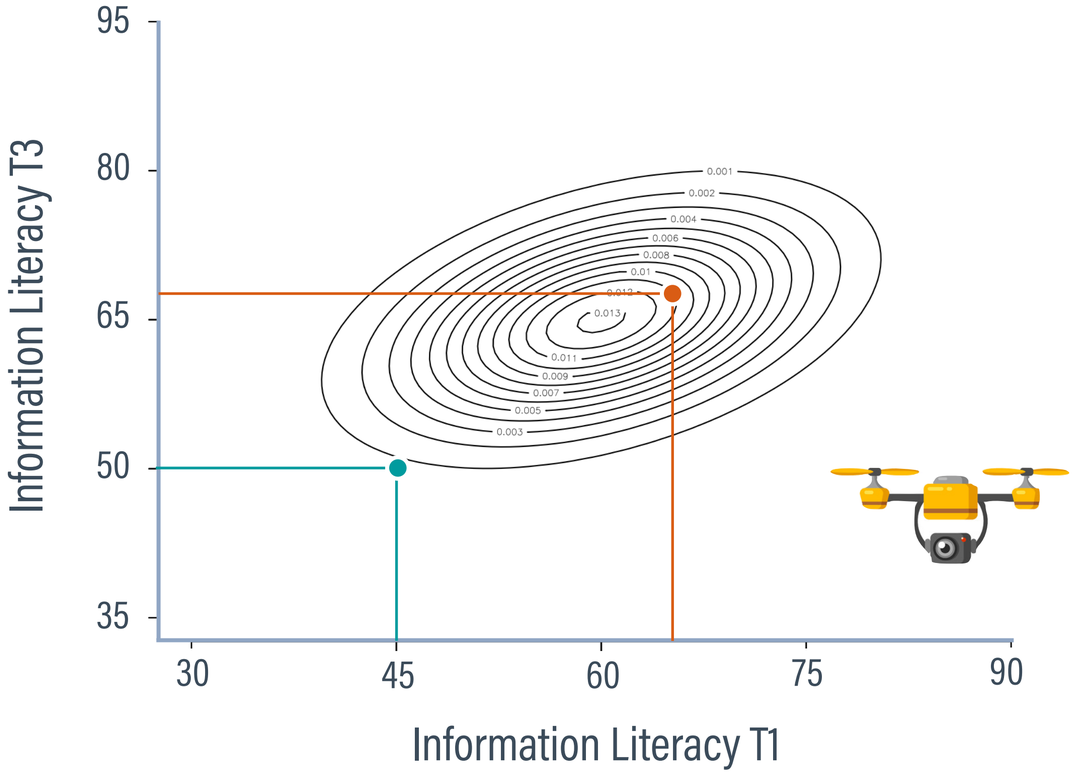
- Assume $\boldsymbol{\mu} = (60, 65)$, $\sigma = 15$, and $\rho = .40$
- Substituting parameters and data into the function returns height coordinates
- log-Likelihood = size of residual (fit) expressed as a height coordinate

Likelihood($\mathbf{Y}$ = 65 and 67 | $\boldsymbol{\mu}$, σ, ρ) = .0007

- log-Likelihood($\mathbf{Y}$ = 65 and 67 | $\boldsymbol{\mu}$, σ, ρ) = −7.22

Likelihood($\mathbf{Y}$ = 45 and 50 | $\boldsymbol{\mu}$, σ, ρ) = .0003

- log-Likelihood($\mathbf{Y}$ = 45 and 50 | $\boldsymbol{\mu}$, σ, ρ) = −7.88



SAMPLE LOG-LIKELIHOOD

- The sample log-likelihood is the sum of the individual fits

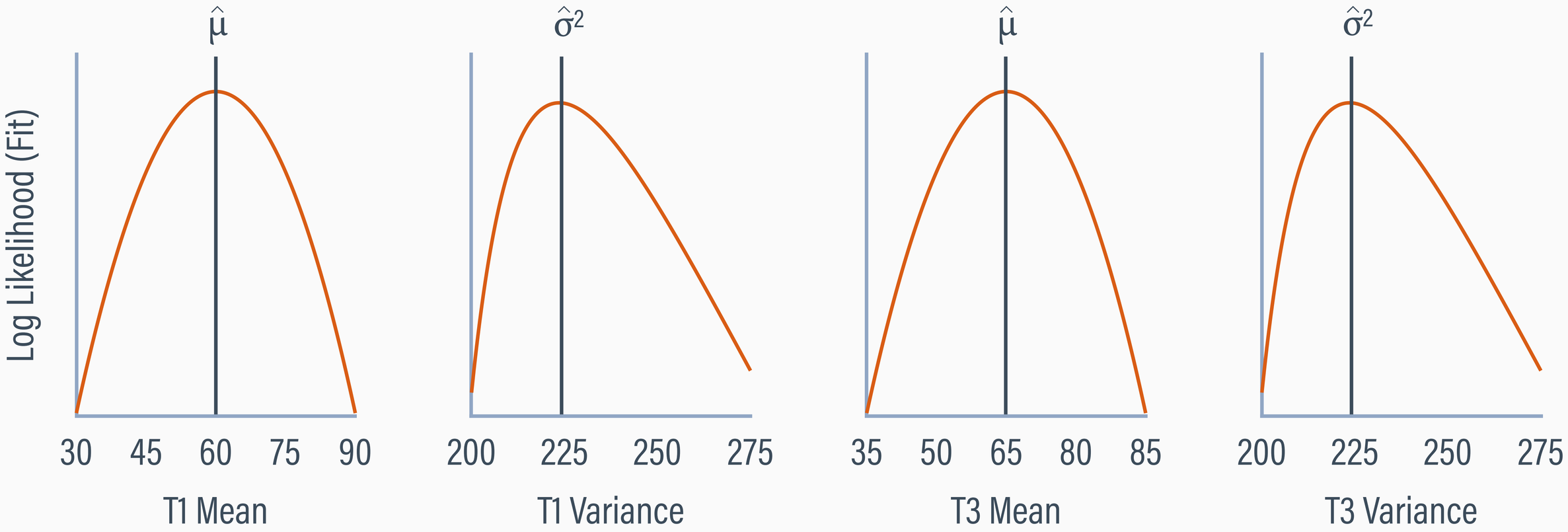
$$\text{log-likelihood} = \sum_{i=1}^{N} \ln(\text{likelihood}_i) = \sum_{i=1}^{N} (\text{individual data-model fit})$$

- Higher (less negative) values imply smaller residuals

- The goal is to find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize the log-likelihood (minimize standardized residual distances)

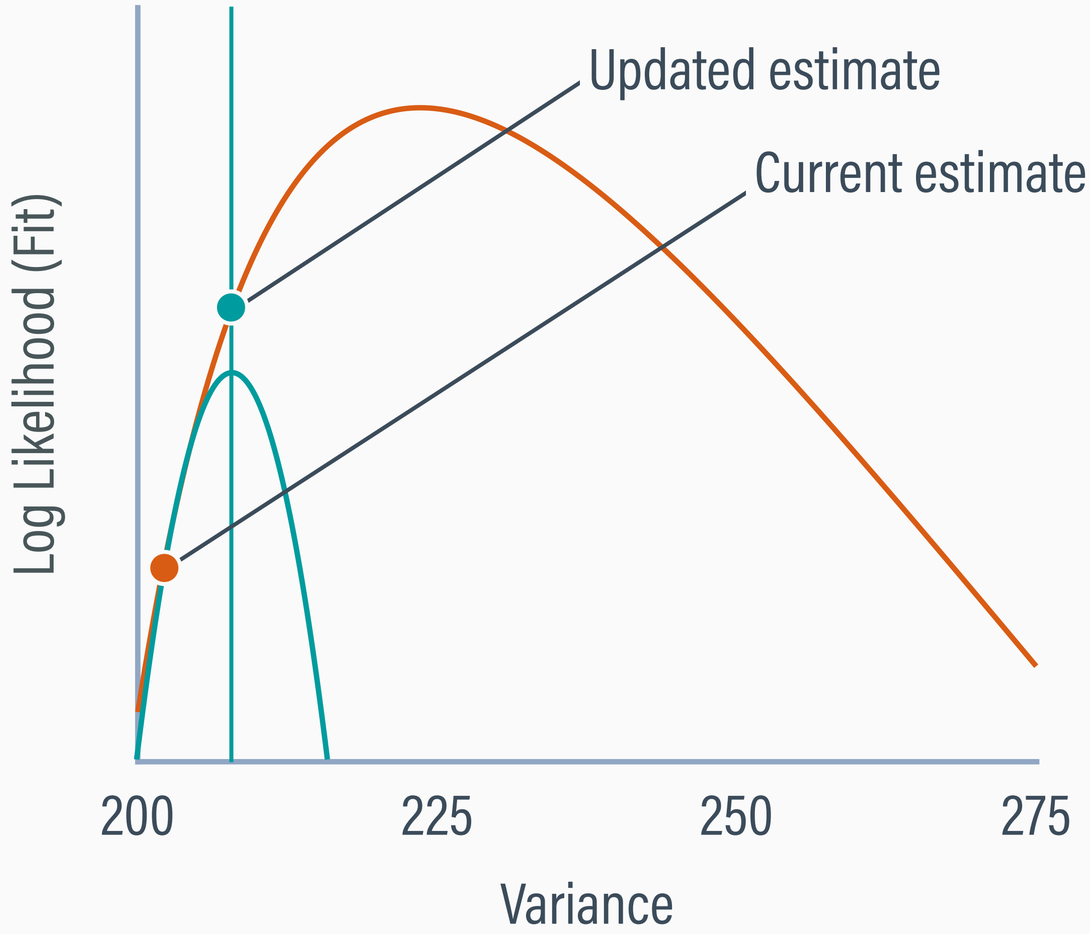
LOG-LIKELIHOOD SURFACES

- The log-likelihood surface shows changes to fit as different parameters are substituted into the normal curve function



NEWTON'S ALGORITHM

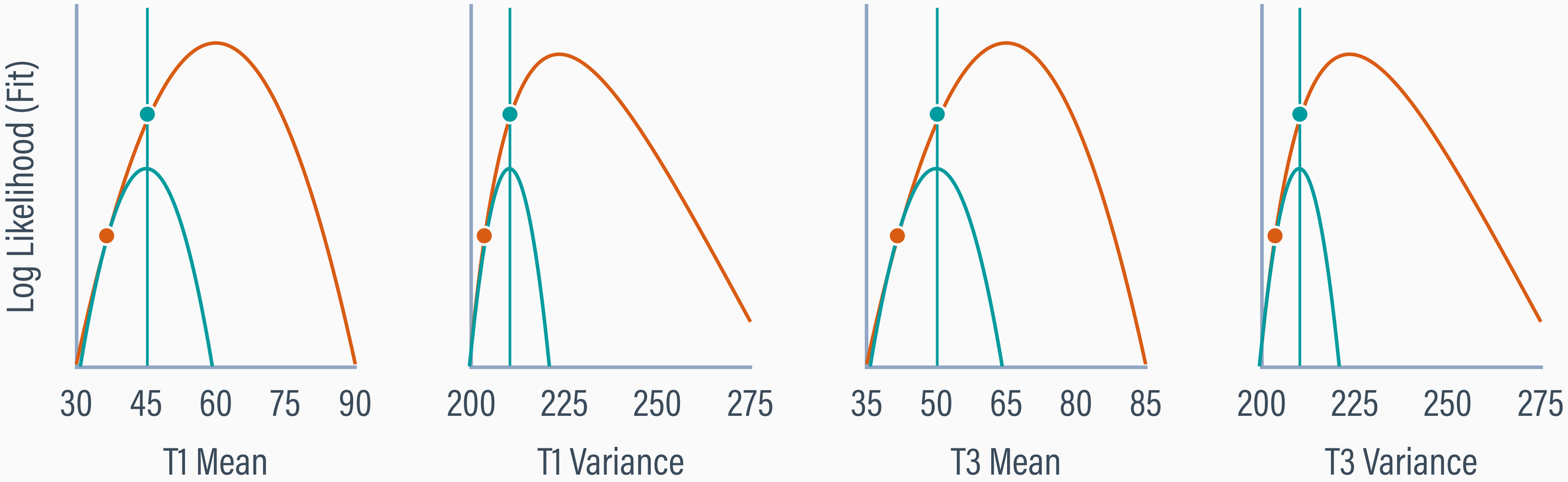
- The peak of the log-likelihood function resembles the peak of a parabola
- Newton's algorithm iteratively projects a parabola through the current estimate
- The point below the peak of the parabola is the updated estimate



UPDATING STEP

- Updated
- Current
- Previous

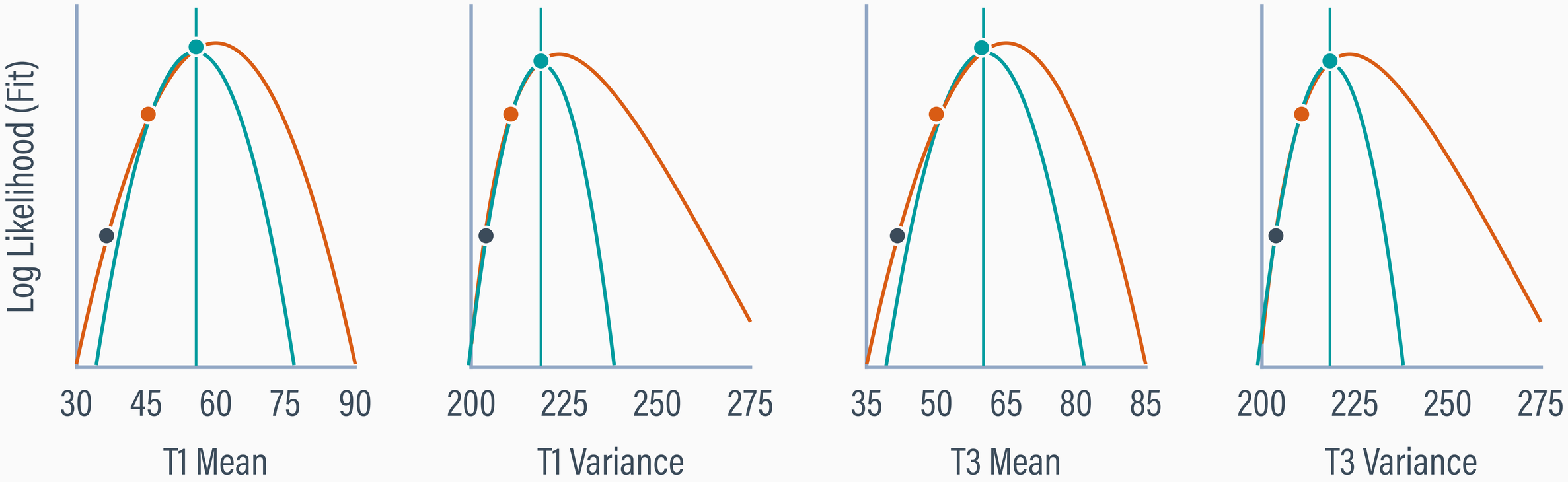
- The updated parameter estimates improve fit and move closer to the maximum of the log-likelihood surfaces



UPDATING STEP

- Updated
- Current
- Previous

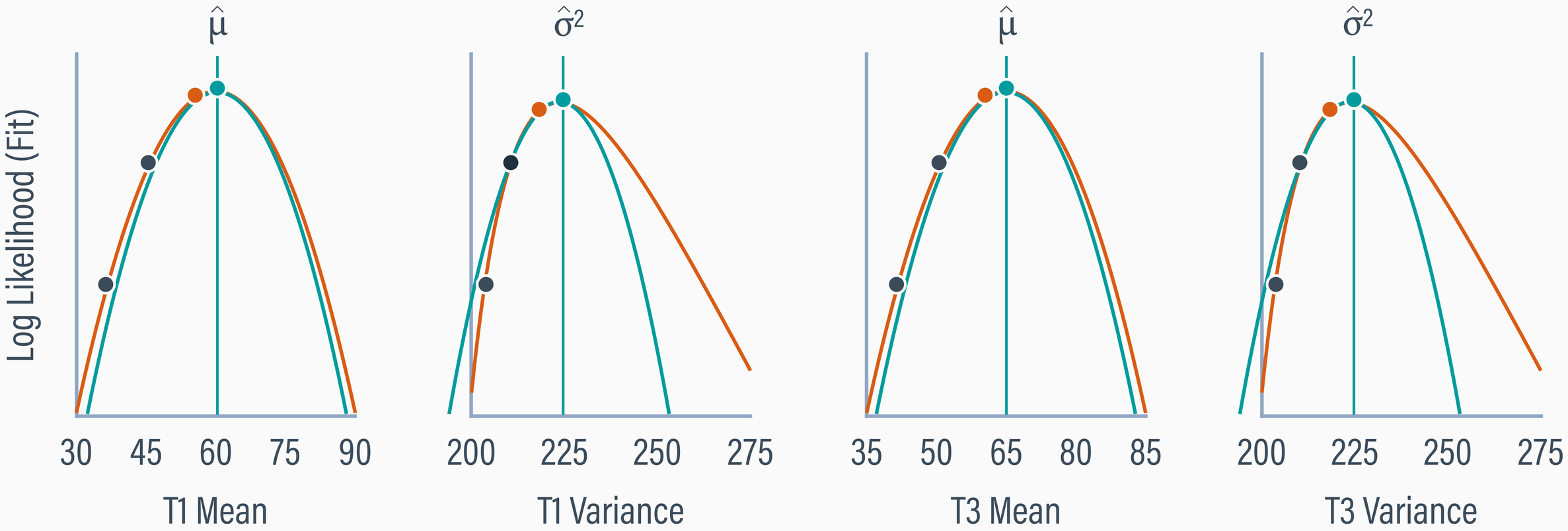
- The updated parameter estimates improve fit and move closer to the maximum of the log-likelihood surfaces



FINAL UPDATING STEP

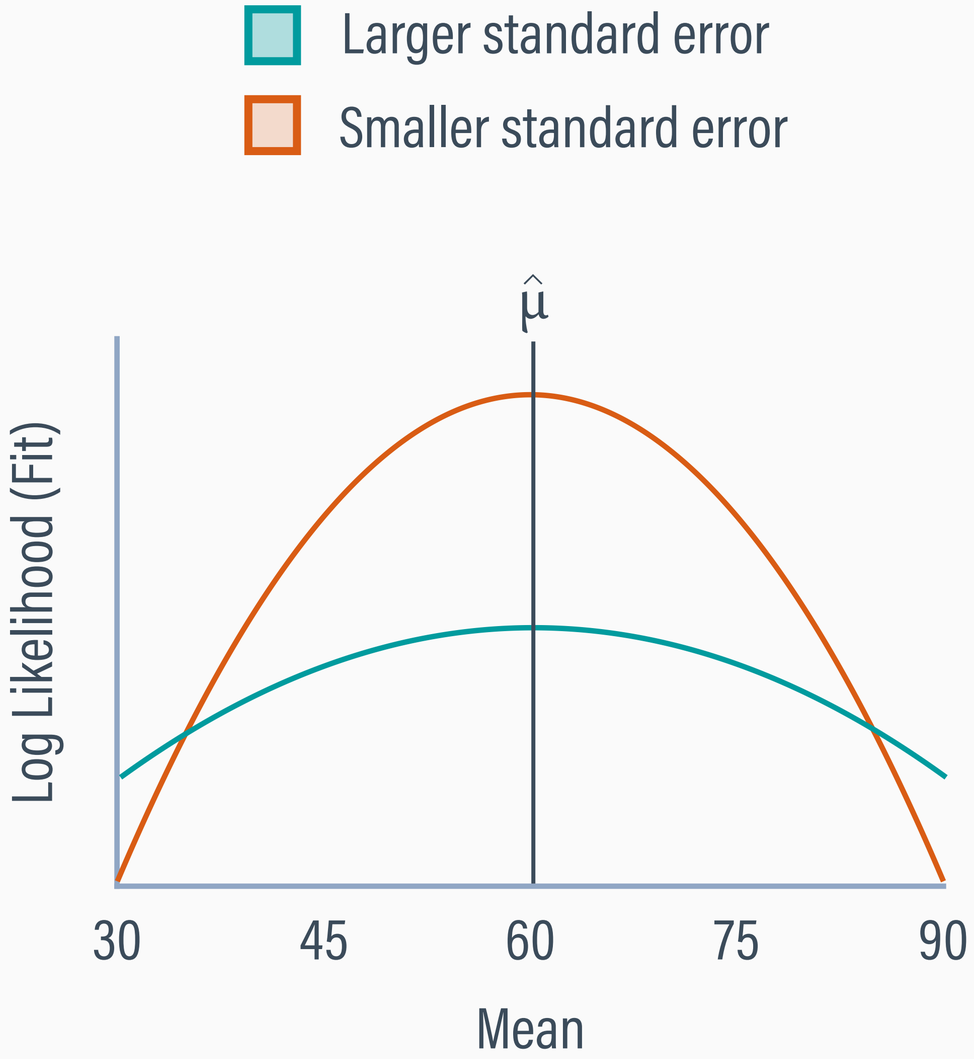
- Updated
- Current
- Previous

- The final updating step occurs when the peak of the parabola is at the same location as the peak of the log-likelihood



STANDARD ERRORS

- Standard errors depend on the curvature at the peak of the log-likelihood
- Steeper functions have smaller standard errors (the summit is more obvious), and flatter functions imply more uncertainty
- Large second derivatives = more peaked

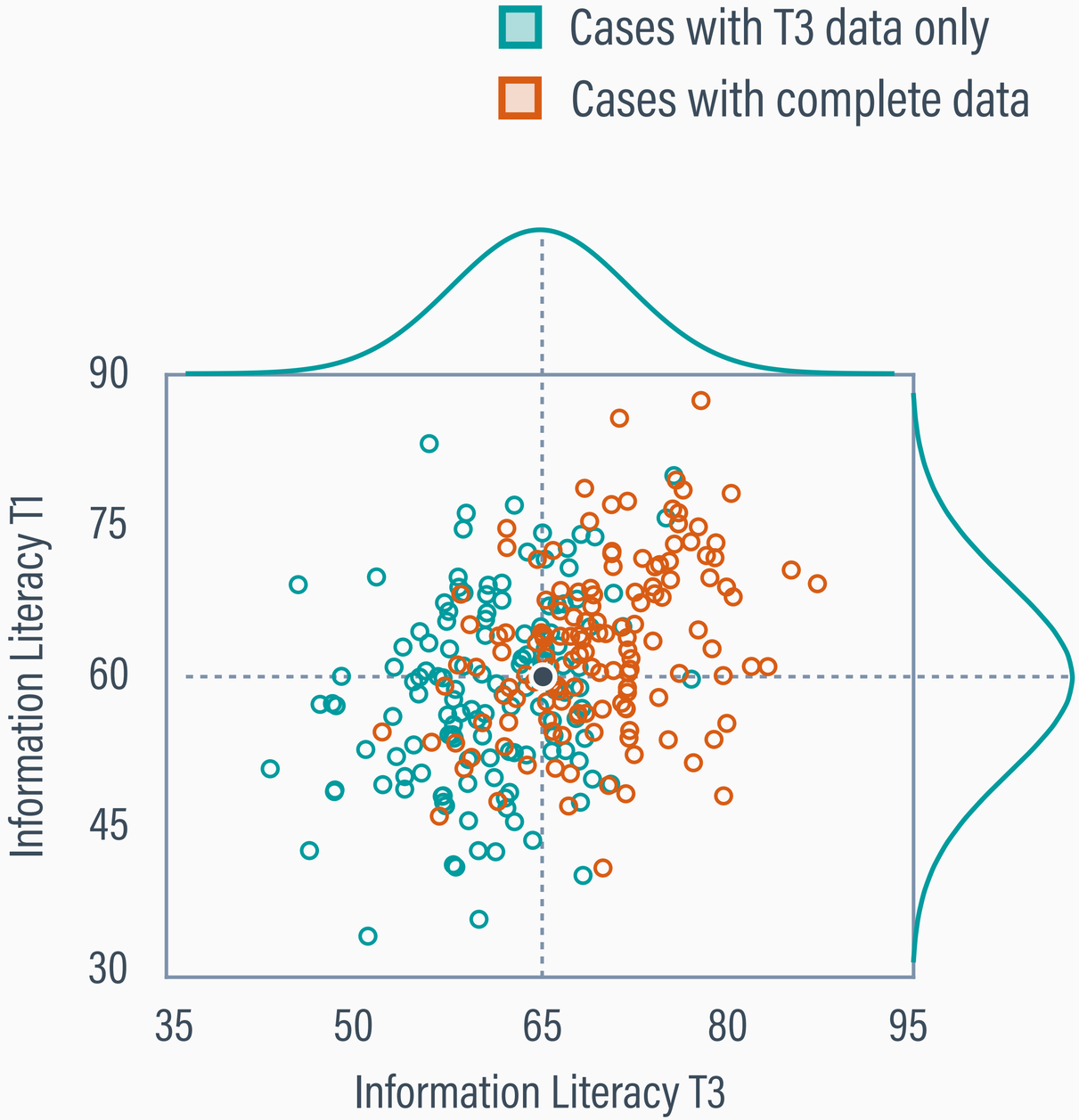


MISSING DATA PREVIEW

- ML identifies the parameter values that minimize squared distances between a model's predictions and the data
- Estimation uses incomplete data, no imputation performed
- People contribute different numbers of data points
- Each observation's contribution to estimation is restricted to the subset of parameters for which there is data

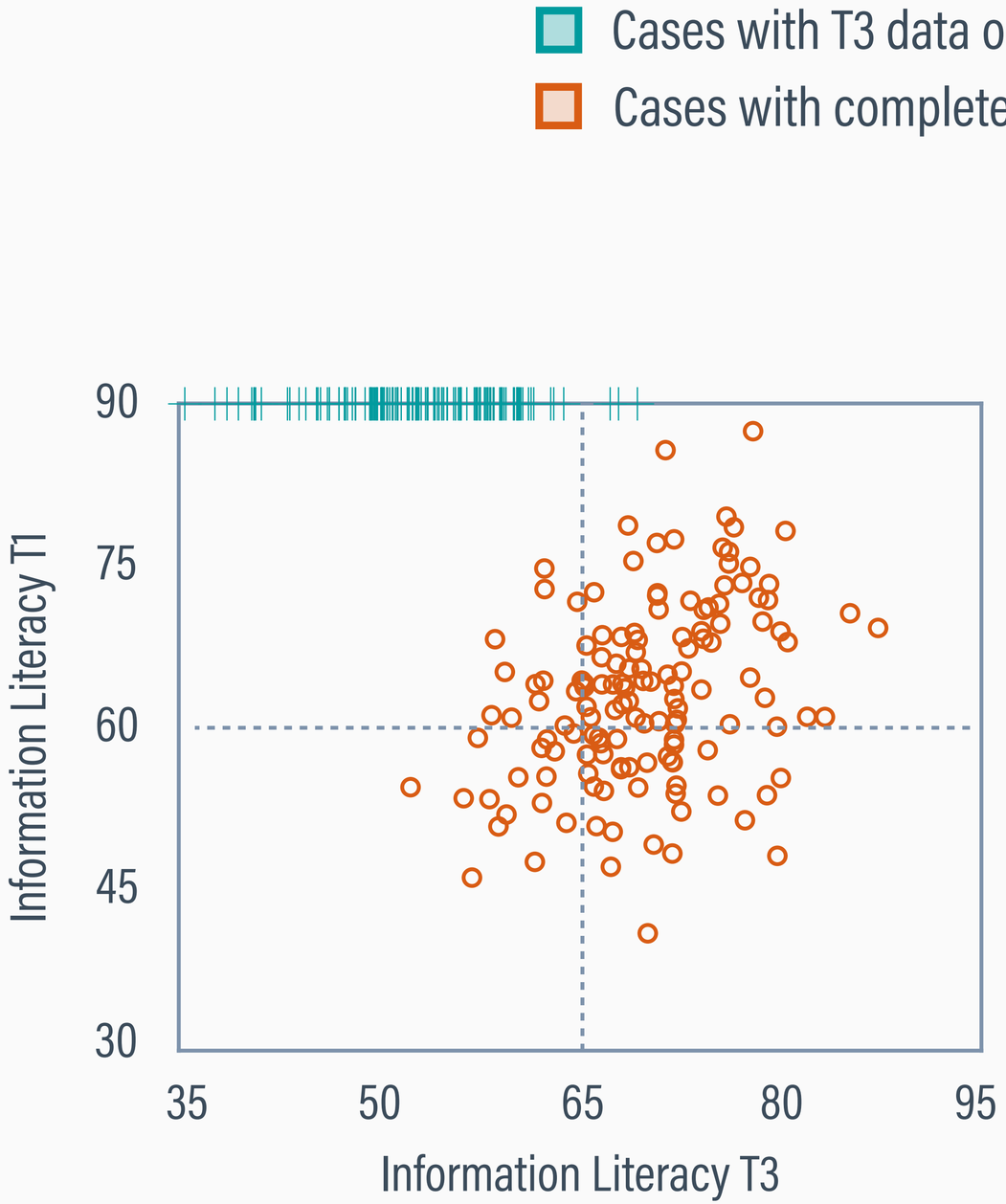
BIVARIATE CARS ILLUSTRATION

- CARS staff want to assess whether information literacy changes between the first and third assessment
- T1 scores are incomplete
- The scatterplot shows the hypothetically complete data



OBSERVED DATA

- A subset of students has information literacy scores observed at T1 and T3
- T1 scores are missing for another subset
- The partial T3 scores tend to be located in the lower tail of the distribution



OBSERVED-DATA LOG-LIKELIHOOD

- The complete- and observed-data log-likelihood equations are the same except for the i subscripts on the parameters

- The parameter subscripts convey that each person's fit is computed using only the parameters for which there is data

$$\text{complete-data log-likelihood}_i = -\frac{V}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})$$

$$\text{observed-data log-likelihood}_i = -\frac{V_i}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

LOG-LIKELIHOOD WITH MISSING DATA

- The log-likelihood (data-model fit) for incomplete records includes only parameters for which there is observed data

$$\text{log-likelihood}_i = -\frac{V_i}{2}\ln(2\pi) - \frac{1}{2}\ln\begin{vmatrix}\sigma_1^2 & \sigma_{13}\\ \sigma_{31} & \sigma_3^2\end{vmatrix} - \frac{1}{2}\begin{pmatrix}? - \mu_1\\ \text{info}_{3i} - \mu_3\end{pmatrix}'\begin{pmatrix}\sigma_1^2 & \sigma_{13}\\ \sigma_{31} & \sigma_3^2\end{pmatrix}^{-1}\begin{pmatrix}? - \mu_1\\ \text{info}_{3i} - \mu_3\end{pmatrix}$$

SAMPLE LOG-LIKELIHOOD

- The sample log-likelihood is the sum of the individual fits

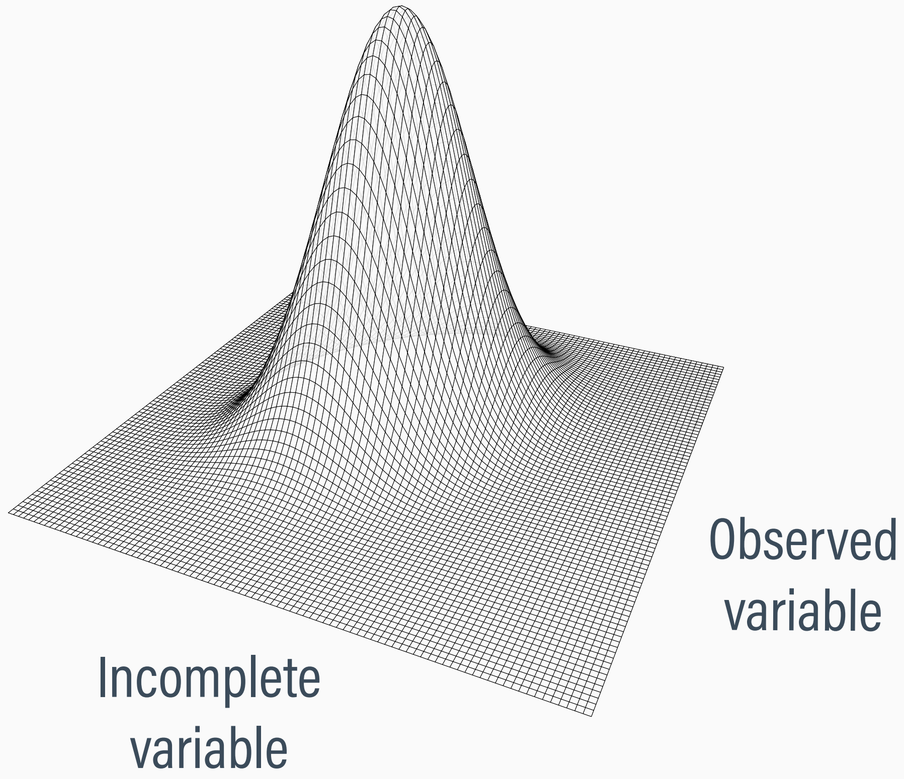
$$\text{log-likelihood} = \sum_{i=1}^{n_{com}} (\text{individual fit with complete data}) + \sum_{j=1}^{n_{mis}} (\text{individual fit with partial data})$$

- The goal is to find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that maximize fit

- The observed data contain more information about some parameters than others

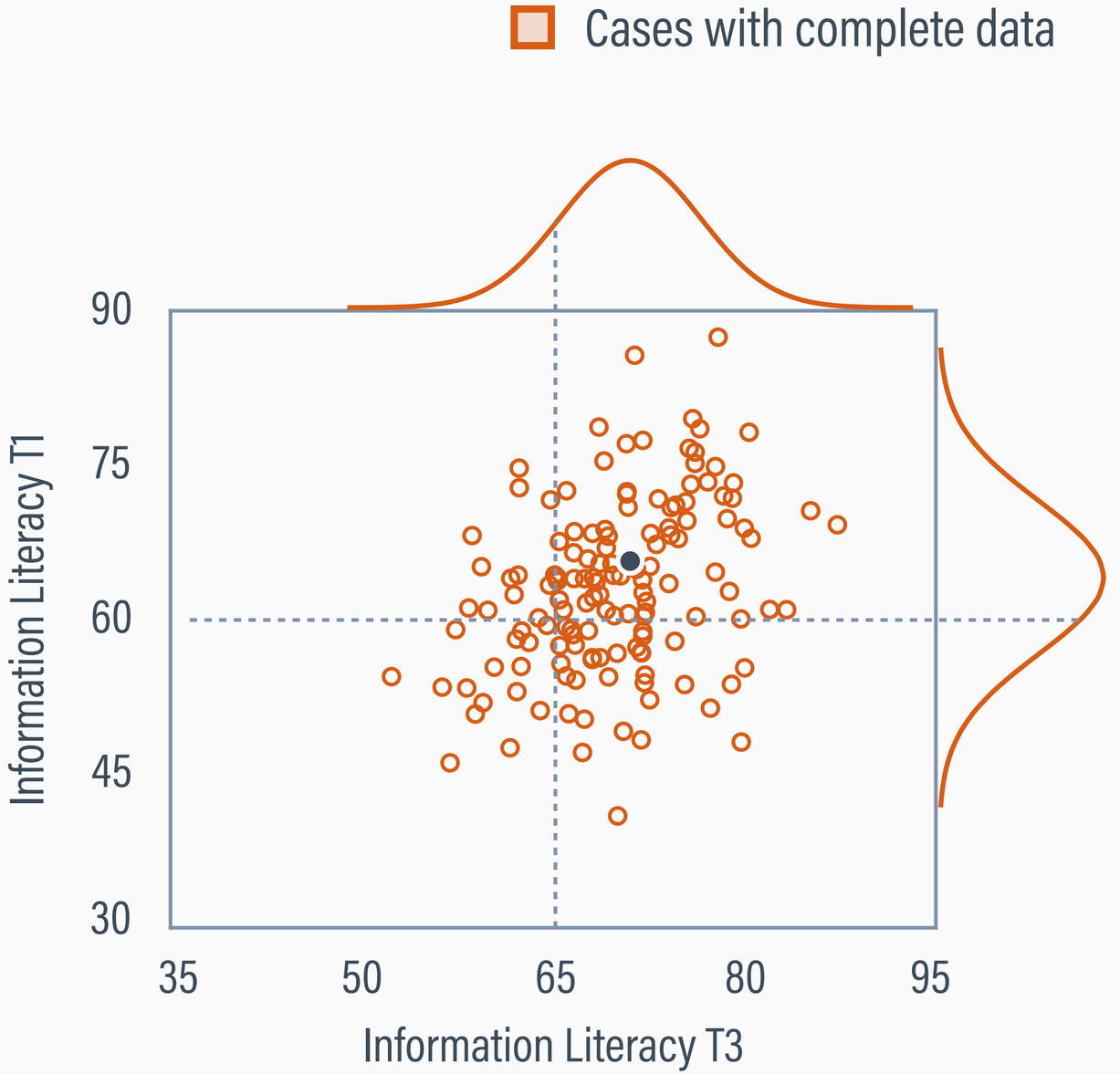
HOW DO PARTIAL DATA RECORDS HELP?

- Data are not filled in, but the multivariate normal distribution acts like an imputation machine
- Given the observed data, the normal curve implies probable locations of the unseen data
- Estimates adjust to account for missing scores



DELETING INCOMPLETE DATA

- Deleting cases with missing scores produces a non-representative sample
- Scores are systematically missing from the lower tails of both distributions
- Both means are too high

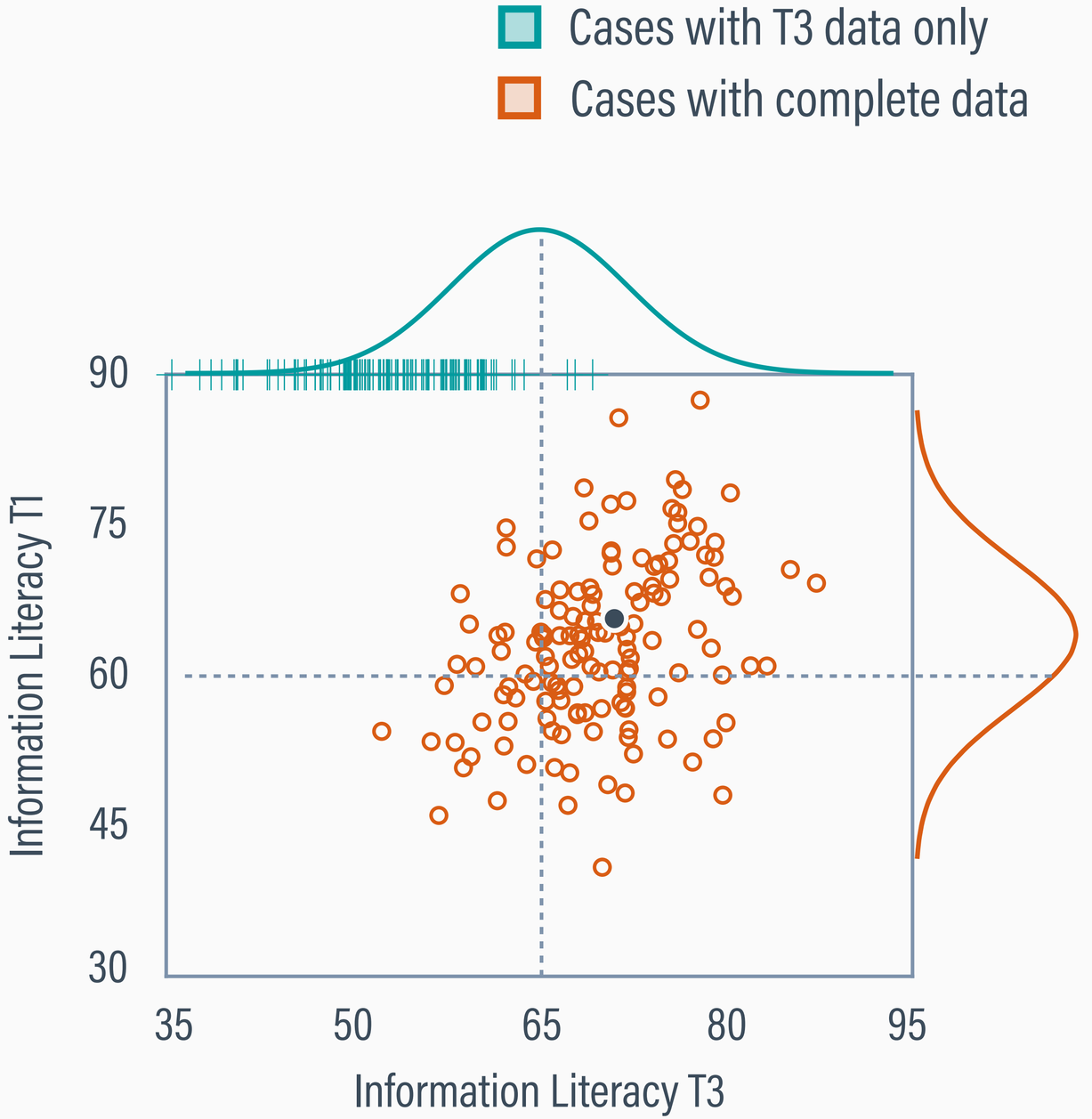


PARTIAL DATA RECORDS

- Maximum likelihood uses the partial data for students with missing T1 scores
- The observed T3 scores tend to be located in the lower tail of the distribution

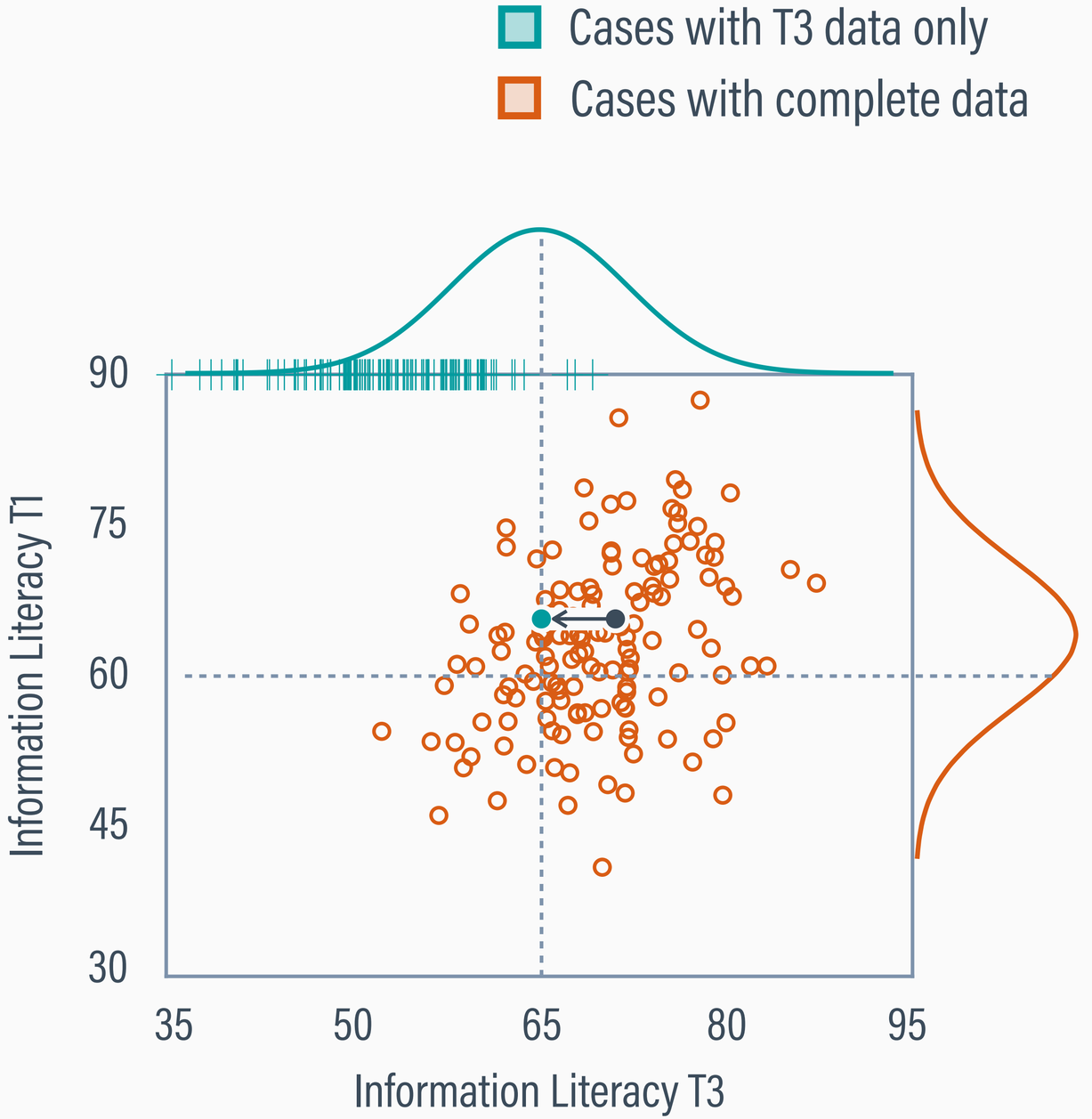
ADJUSTING THE T3 VARIANCE

- The distribution stretches out to accommodate a wider range of scores
- Adding lower T3 values increases the variance relative to deletion
- The variance is no longer biased



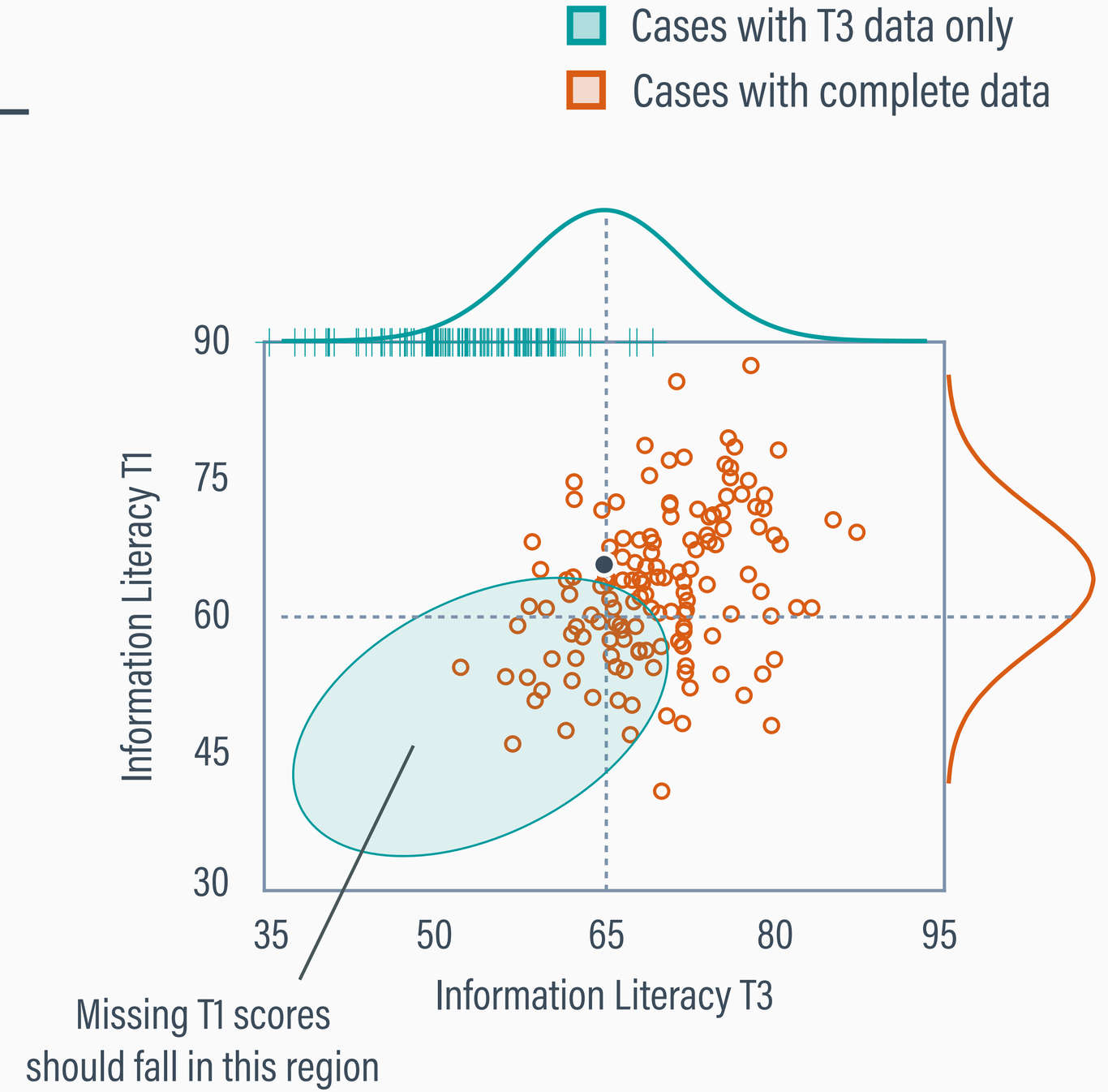
ADJUSTING THE T3 MEAN

- The partial data in the lower tail pull the T3 mean lower than the deletion average
- The T3 mean is no longer biased



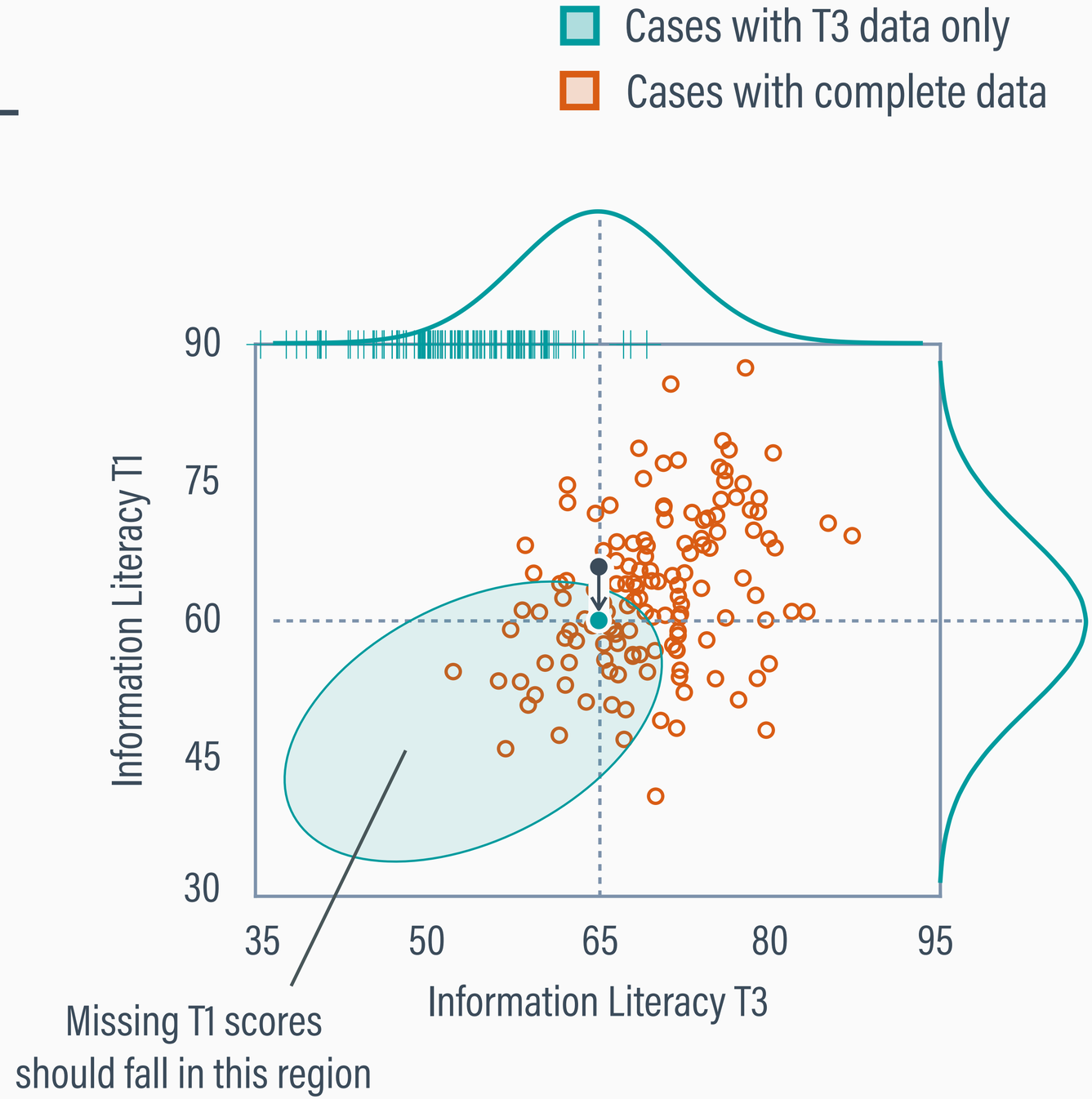
IMPLICIT IMPUTATION

- The normal curve implies probable values for the missing scores
- In a normal curve with a positive correlation, lower T3 scores should pair with lower (missing) T1 scores
- ML implicitly imputes missing values

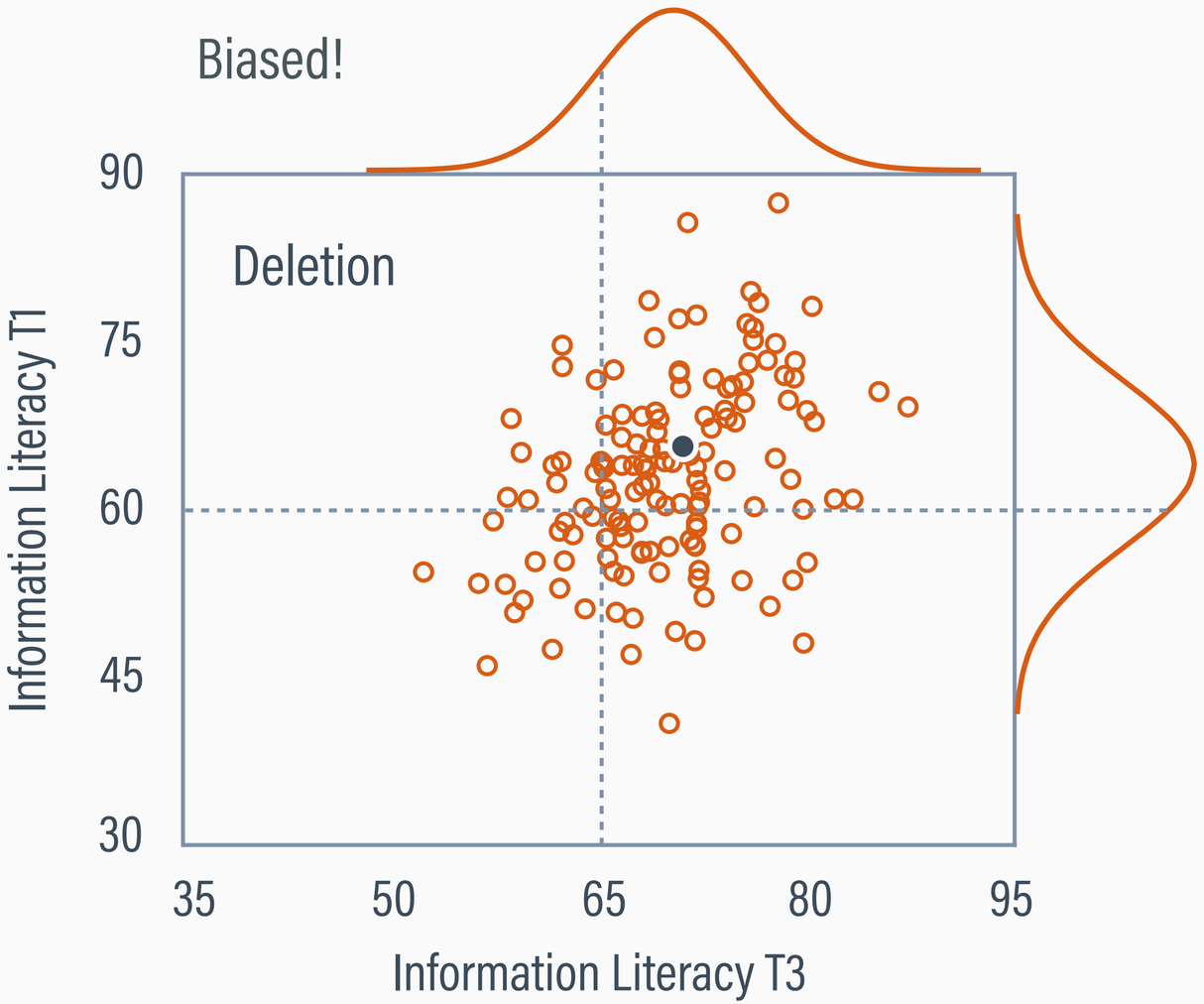


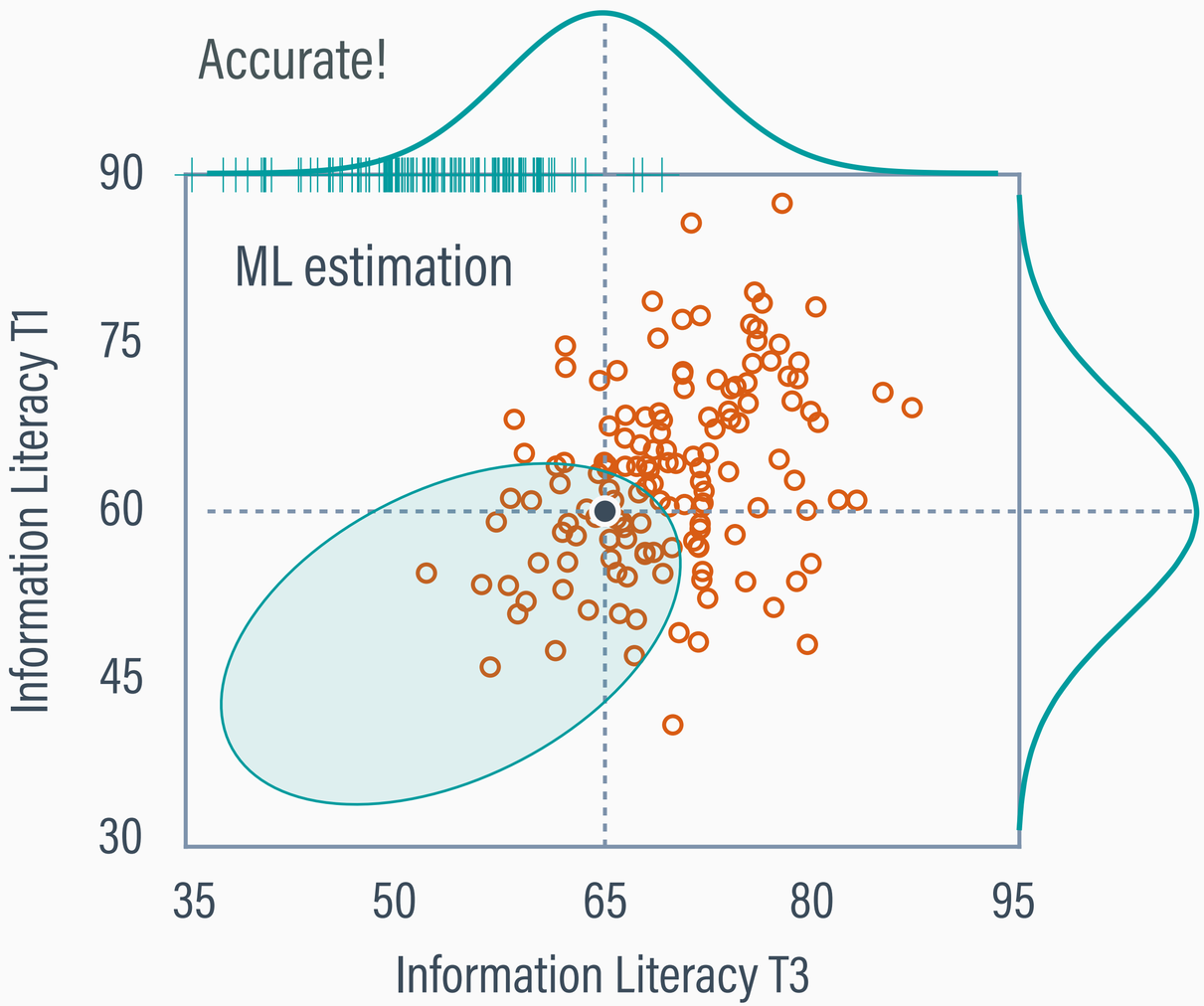
ADJUSTING THE T1 DISTRIBUTION

- From the positive correlation, ML can intuit the presence of unseen T1 scores in the lower tail of the distribution
- The estimated T1 variance increases
- Unseen scores in the lower tail imply a lower T1 mean, eliminating bias



ESTIMATION SUMMARY





MAXIMUM LIKELIHOOD PROS AND CONS

Pros

- Direct estimation for a wide range of analysis models
- Widely available in software packages (any SEM program)
- Easy to use, missing data handling occurs behind the scenes

Cons

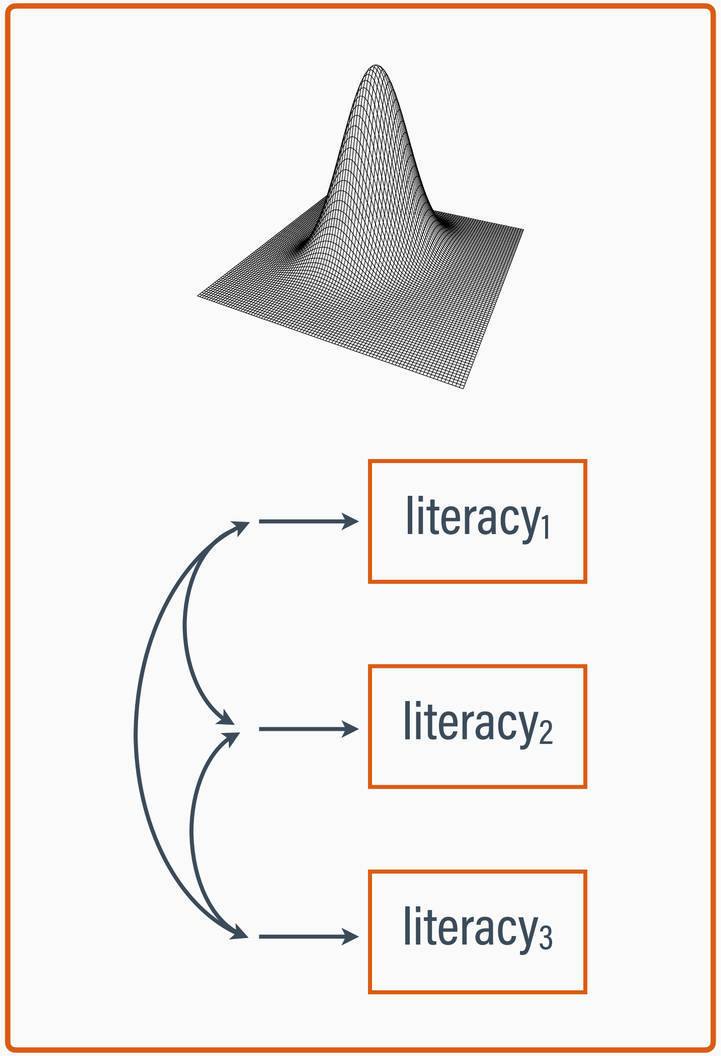
- Generally limited to normal data, options for mixed metrics are less common
- Normal-theory methods are biased with interactions and non-linear terms
- MLM software usually discards observations with missing predictors

MORNING OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

CARS ANALYSIS EXAMPLE

- CARS wants to assess whether information literacy changes over time
- Descriptive statistics are obtained by specifying a saturated model with all possible means, variances, and covariances
- lavaan's ML estimator assumes that all variables are multivariate normal



LAVAAN SCRIPT

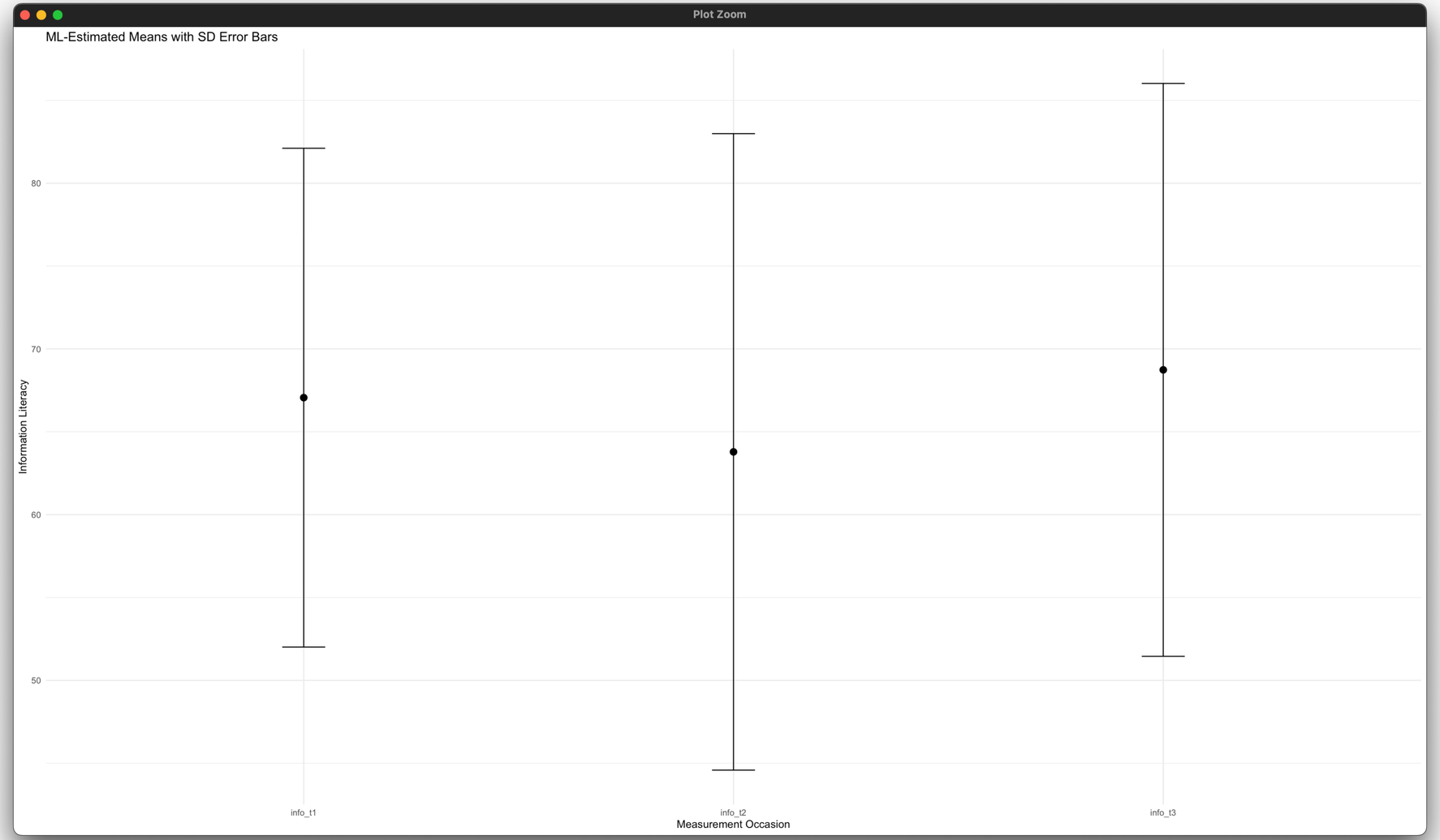
```
# all possible means, variances, and covariances
syntax <- '
  info_t1 ~ 1
  info_t2 ~ 1
  info_t3 ~ 1
  info_t1 ~~ info_t1
  info_t2 ~~ info_t2
  info_t3 ~~ info_t3
  info_t1 ~~ info_t2
  info_t1 ~~ info_t3
  info_t2 ~~ info_t3'

mymodel <- sem.auxiliary(  # semTools package for auxiliary variables in lavaan
  model = syntax,
  data = carsdat,
  fixed.x = FALSE,  # ml missing data requires distribution for all variables
  aux = c('extra_t3','cont_t3','om_t3','ag_t3','ne_t3','male'))  # saturated correlates model
summary(mymodel, standardized = TRUE)  # summarize results
```

LAVAAN OUTPUT

The lavaan output also includes the means, variances, and covariances for the auxiliary variables

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  info_t1 ~~
    info_t2         124.738    5.831   21.392    0.000  124.738    0.432
    info_t3         105.597    4.793   22.030    0.000  105.597    0.406
  info_t2 ~~
    info_t3         160.350    6.460   24.823    0.000  160.350    0.483
...

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    info_t1           67.060    0.220  305.189    0.000   67.060    4.456
    info_t2           63.787    0.309  206.355    0.000   63.787    3.322
    info_t3           68.737    0.254  270.706    0.000   68.737    3.977
...

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    info_t1         226.500    4.766   47.528    0.000  226.500    1.000
    info_t2         368.763    8.877   41.544    0.000  368.763    1.000
    info_t3         298.772    6.446   46.350    0.000  298.772    1.000
...
```

MEANS WITH STD. DEV. ERROR BARS

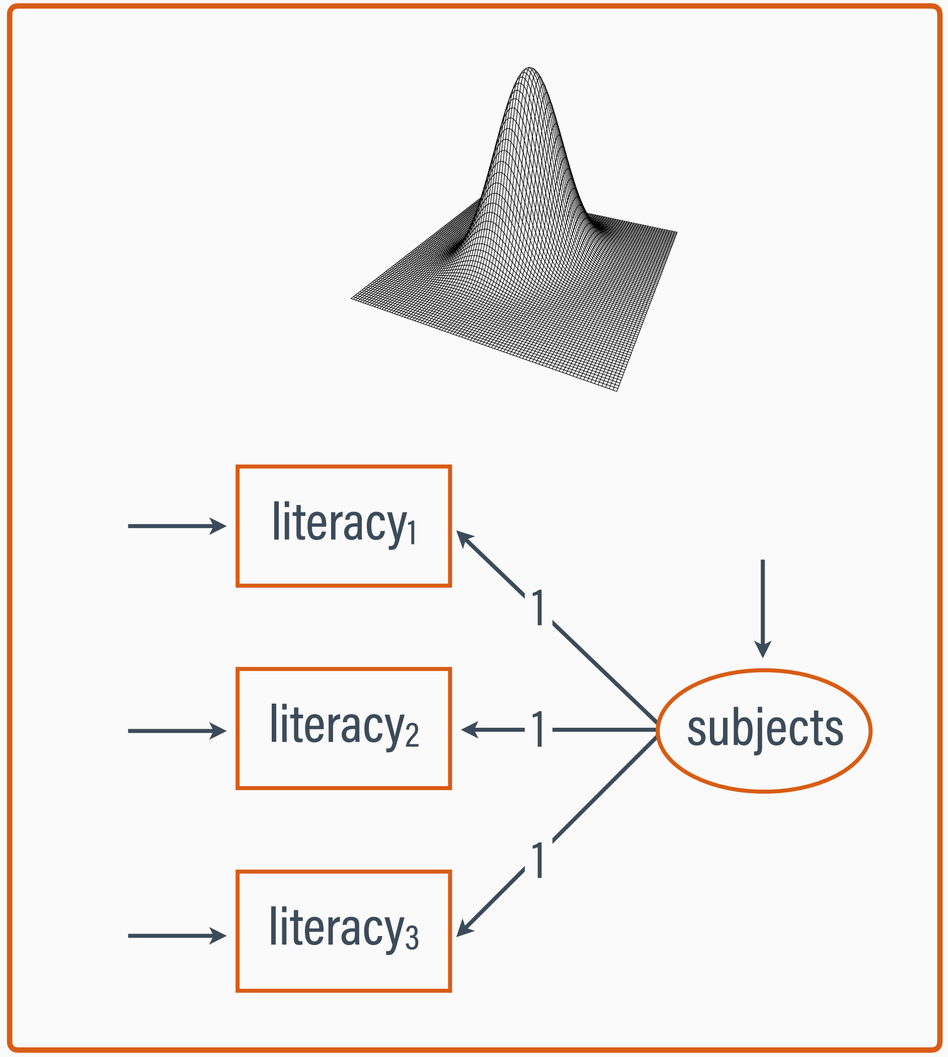


INTERPRETATIONS

- Means exhibit a nonlinear pattern with a decrease at T2
- Standard deviations are unequal with greater variation at T2
- Estimates assume a CMAR process that depends on the observed repeated measures data and auxiliary variables

REPEATED MEASURES ANALYSIS

- Repeated measures ANOVA features a random “subjects factor” that quantifies differences among person means
- The subjects latent variable has a mean of zero and loadings fixed to one
- Each outcome has an intercept (mean)
- lavaan’s ML estimator assumes that all variables are multivariate normal



LAVAAN SCRIPT

```
syntax <- '
  subjects =~ 1*info_t1 + 1*info_t2 + 1*info_t3  # random subjects factor with loadings = 1
  subjects ~ 0*1  # random subjects factor with mean = 0
  info_t1 ~ mu1*1  # label the means
  info_t2 ~ mu2*1
  info_t3 ~ mu3*1
  info_t1 ~~ res*info_t1  # label res sets equal residual variances for compound symmetry assumption
  info_t2 ~~ res*info_t2
  info_t3 ~~ res*info_t3'

mymodel <- sem.auxiliary(  # semTools package for auxiliary variables in lavaan
  model = syntax,
  data = carsdat,
  fixed.x = FALSE,  # ml missing data requires distribution for predictors
  aux = c('extra_t3','cont_t3','om_t3','ag_t3','ne_t3','male'))  # saturated correlates auxiliary model
summary(mymodel, standardized = TRUE)  # summarize results

lavTestWald(mymodel, 'mu1 == mu2; mu2 == mu3')  # null that all means are equal
lavTestWald(mymodel, 'mu1 == mu2')  # pairwise comparison null
lavTestWald(mymodel, 'mu1 == mu3')
lavTestWald(mymodel, 'mu2 == mu3')
```

LAVAAN OUTPUT

The lavaan output also includes the means, variances, and covariances for the auxiliary variables

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  subjects =~
    info_t1           1.000                              11.115    0.652
    info_t2           1.000                              11.115    0.652
    info_t3           1.000                              11.115    0.652

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  extra_t3 ~~
    cont_t3           0.139    0.007   20.226    0.000    0.139    0.324
    om_t3             0.101    0.006   15.961    0.000    0.101    0.251
...

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    subjects          0.000                               0.000    0.000
   .info_t1  (mu1)   67.046    0.249  269.753    0.000   67.046    3.932
   .info_t2  (mu2)   63.933    0.276  231.307    0.000   63.933    3.750
   .info_t3  (mu3)   68.749    0.251  273.848    0.000   68.749    4.032
...

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .info_t1  (res)  167.178    3.088   54.142    0.000  167.178    0.575
   .info_t2  (res)  167.178    3.088   54.142    0.000  167.178    0.575
   .info_t3  (res)  167.178    3.088   54.142    0.000  167.178    0.575
    subjects        123.538    4.210   29.345    0.000    1.000    1.000
...
```

WALD SIGNIFICANCE TEST OUTPUT

```
> lavTestWald(fiml_repeated_aux, 'mu1 == mu2; mu2 == mu3’)

$stat
[1] 229.2187

$df
[1] 2

$p.value
[1] 0
```

Null that all means are equal

SIGNIFICANCE TEST OUTPUT, CONT.

```
> lavTestWald(fiml_repeated_aux, 'mu1 == mu2')
$stat
[1] 94.87408
$df
[1] 1
$p.value
[1] 0
```

Pairwise comparison

```
> lavTestWald(fiml_repeated_aux, 'mu1 == mu3')
$stat
[1] 32.48125
$df
[1] 1
$p.value
[1] 1.20348e-08
```

Pairwise comparison

```
> lavTestWald(fiml_repeated_aux, 'mu2 == mu3')
$stat
[1] 226.9077
$df
[1] 1
$p.value
[1] 0
```

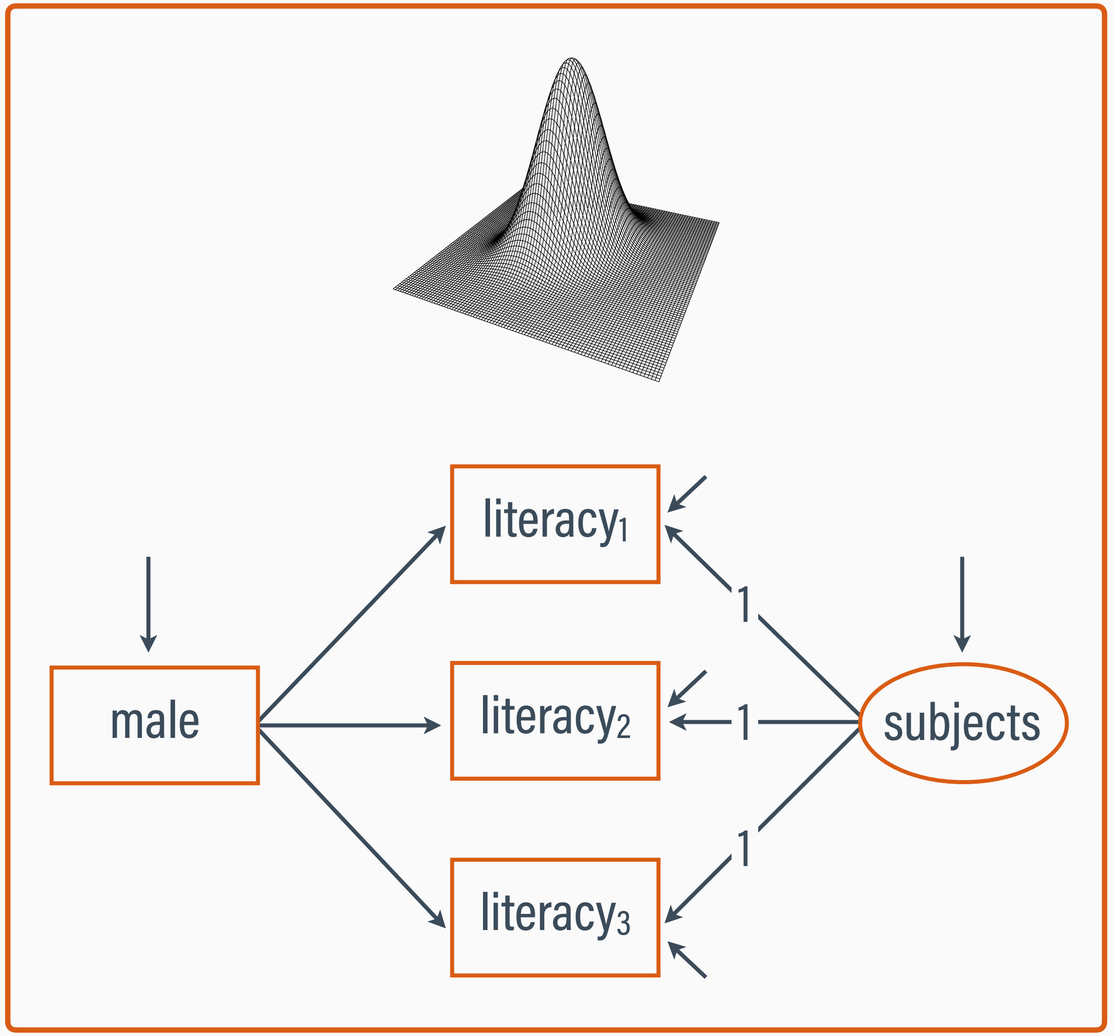
Pairwise comparison

MORNING OUTLINE

1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

CARS ANALYSIS EXAMPLE

- Incomplete predictors require distributional assumptions for missing data handling
- WLS for categorical data assumes MCAR (same bias-inducing assumption as deletion)
- lavaan's ML estimator treats the binary predictor as normally distributed
- Conceptually, this "imputes" the dummy code with decimals instead of 0s and 1s



LAVAAN SCRIPT

```
syntax <- '
  subjects =~ 1*info_t1 + 1*info_t2 + 1*info_t3  # random subjects factor with loadings = 1
  subjects ~ 0*1  # random subjects factor with mean = 0
  info_t1 ~ mu1*1 + dif1*male  # label means and differences
  info_t2 ~ mu2*1 + dif2*male
  info_t3 ~ mu3*1 + dif3*male
  info_t1 ~~ res*info_t1  # label res sets equal residual variances for compound symmetry assumption
  info_t2 ~~ res*info_t2
  info_t3 ~~ res*info_t3
  fem_mu1 := mu1  # define group means
  fem_mu2 := mu2
  fem_mu3 := mu3
  male_mu1 := mu1 + dif1
  male_mu2 := mu2 + dif2
  male_mu3 := mu3 + dif3'

mymodel <- sem.auxiliary(  # semTools package for auxiliary variables in lavaan
  model = syntax,
  data = carsdat,
  fixed.x = FALSE,  # ml missing data requires normal distribution for predictors (not optimal for binary predictor)
  aux = c('extra_t3','cont_t3','om_t3','ag_t3','ne_t3'))  # saturated correlates auxiliary model
summary(mymodel, standardized = TRUE)  # summarize results

lavTestWald(mymodel, 'dif1 == dif2; dif2 == dif3')  # wald test that group-by-time interaction = 0
```

LAVAAN OUTPUT

The lavaan output also includes the means, variances, and covariances for the auxiliary variables

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
...

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  info_t1 ~
    male    (dif1)   -2.030    0.501   -4.054    0.000   -2.030   -0.059
  info_t2 ~
    male    (dif2)   -5.301    0.560   -9.474    0.000   -5.301   -0.153
  info_t3 ~
    male    (dif3)   -4.896    0.510   -9.606    0.000   -4.896   -0.142

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  extra_t3 ~~
    cont_t3            0.139    0.007   20.212    0.000    0.139    0.324
    om_t3              0.101    0.006   15.962    0.000    0.101    0.251
...

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    subjects           0.000                              0.000    0.000
   .info_t1  (mu1)   67.946    0.324  209.452    0.000   67.946    4.008
   .info_t2  (mu2)   66.218    0.361  183.582    0.000   66.218    3.867
   .info_t3  (mu3)   70.879    0.328  216.146    0.000   70.879    4.146
    male               0.436    0.006   71.720    0.000    0.436    0.879
...

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .info_t1  (res)  166.359    3.079   54.037    0.000  166.359    0.579
   .info_t2  (res)  166.359    3.079   54.037    0.000  166.359    0.567
   .info_t3  (res)  166.359    3.079   54.037    0.000  166.359    0.569
    subjects        120.032    4.147   28.941    0.000    1.000    1.000
    male              0.246    0.004   57.642    0.000    0.246    1.000
...
```

LAVAAN OUTPUT, CONTINUED

```
Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    fem_mu1           67.946    0.324  209.452    0.000   67.946    4.008
    fem_mu2           66.218    0.361  183.582    0.000   66.218    3.867
    fem_mu3           70.879    0.328  216.146    0.000   70.879    4.146
    male_mu1          65.916    0.381  173.052    0.000   65.916    3.949
    male_mu2          60.917    0.427  142.766    0.000   60.917    3.713
    male_mu3          65.983    0.388  169.984    0.000   65.983    4.004
```

WALD SIGNIFICANCE TEST OUTPUT

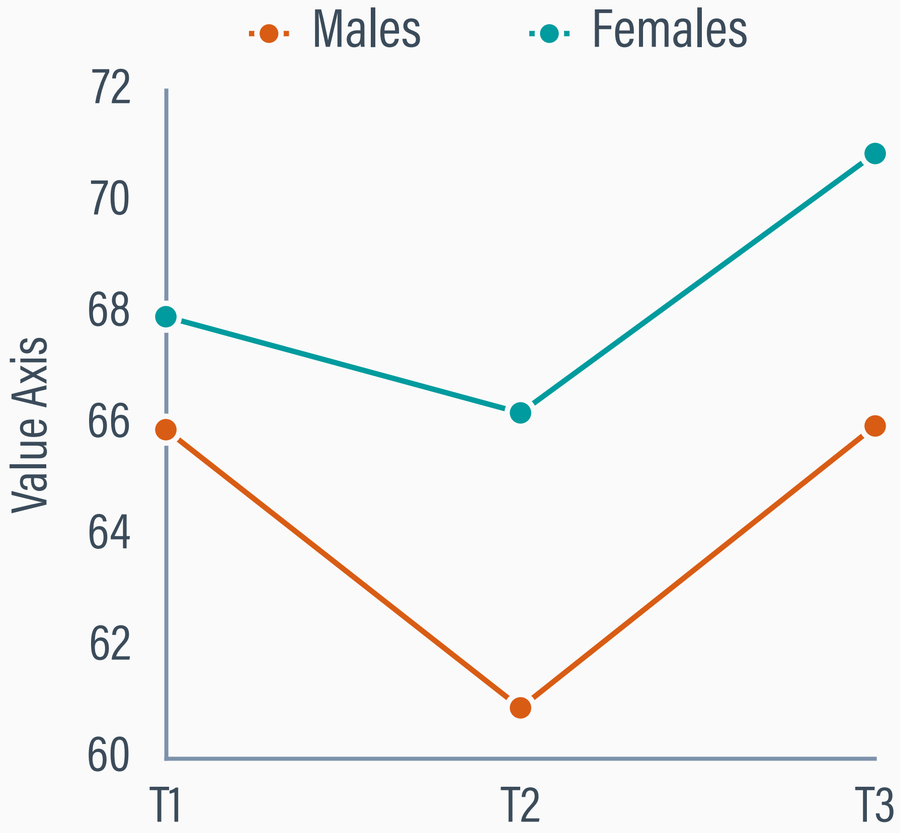
```
> lavTestWald(fiml_repeated_gender_aux, 'dif1 == dif2; dif2 == dif3')
```

Null hypothesis that group-by-time interaction = 0 (time-specific mean differences are equal)

```
$stat
[1] 32.65378

$df
[1] 2

$p.value
[1] 8.115618e-08
```

INTERPRETATIONS

- The group-by-time interaction was significant ($\chi^2 = 32.65$, $p < .001$)
- Females decreased at T2 then rebounded to a higher mean at T3
- Males decreased by a larger amount at T2 (about five points versus less than two), and their T3 mean is the same as T1



MORNING OUTLINE

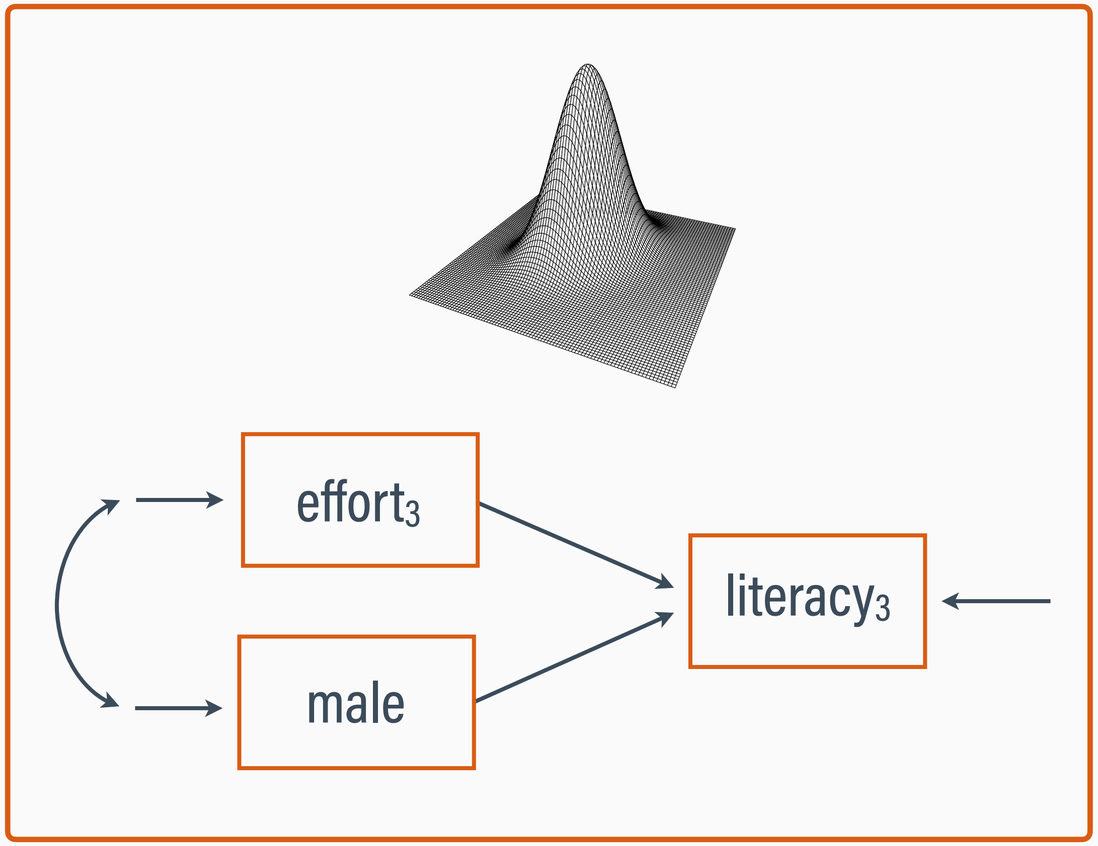
1. Modern Missing Data Methods
2. Missing Data Mechanisms
3. Maximum Likelihood Estimation
4. Analysis Example 1: Descriptive Statistics and Repeated Measures
5. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
6. Analysis Example 3: Multiple Regression

CARS ANALYSIS EXAMPLE

- Are there gender differences in T3 information literacy, controlling for T3 effort?

$$literacy_3 = \beta_0 + \beta_1(effort_3) + \beta_2(male) + \varepsilon$$

- WLS for categorical data assumes MCAR (same bias-inducing assumption as deletion)
- lavaan's ML estimator treats the binary predictor as normally distributed



LAVAAN SCRIPT

```
# auxiliary variables for use with sem.auxiliary function
auxvars <- c('extra_t3','cont_t3','om_t3','ag_t3','ne_t3','admit_type_num')

# regression model syntax
syntax <- 'info_t3 ~ effort_t3 + male'

mymodel <- sem.auxiliary(
  model = syntax,
  data = carsdat,
  fixed.x = FALSE,  # missing data requires normal distribution for predictors (not optimal for binary)
  aux = auxvars)  # saturated correlates auxiliary model
summary(mymodel, standardized = TRUE)  # summarize results
```

REGRESSION SUMMARY TABLE

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  info_t3 ~
    effort_t3 (b1)   10.010    0.337   29.699    0.000   10.010    0.413
    male      (b2)   -2.497    0.488   -5.114    0.000   -2.497   -0.072

Covariances:
                    Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  effort_t3 ~~
    male               -0.049    0.005   -8.996    0.000   -0.049   -0.139
...

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .info_t3          33.077    1.312   25.210    0.000   33.077    1.924
    effort_t3         3.694    0.011  341.451    0.000    3.694    5.206
    male              0.436    0.006   71.741    0.000    0.436    0.880
...

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .info_t3         241.188    5.201   46.378    0.000  241.188    0.816
    effort_t3         0.504    0.011   46.367    0.000    0.504    1.000
    male              0.246    0.004   57.643    0.000    0.246    1.000
...
```

INTERPRETATIONS

- $\beta_0 = 33.08$ is the mean for a female with zero effort (an extrapolation because effort has no zero point)
- For two students with the same gender, scoring one point higher on the effort measure was associated with a $\beta_1 = 10.01$ increase in information literacy
- For two students with the same effort score, males scored $\beta_2 = -2.50$ points lower than females

AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

WHY CHOOSE MCMC?

- MCMC readily handles complex missing data problems, including:
  - Mixed metrics (normal, ordinal, nominal, skewed, count, latent)
  - Nonlinear effects (interactions, curvilinear effects)
  - Multilevel data (random coefficients, interactions)
  - Latent variable modeling (interactions)
- FIML estimators for these scenarios are far more limited

FREQUENTIST VS. BAYESIAN PARADIGMS

Frequentist

- The parameter is a fixed quantity, estimates vary across different samples
- Statements about probability, precision, and confidence refer to estimates
- Probability = long run frequency of outcomes across many samples

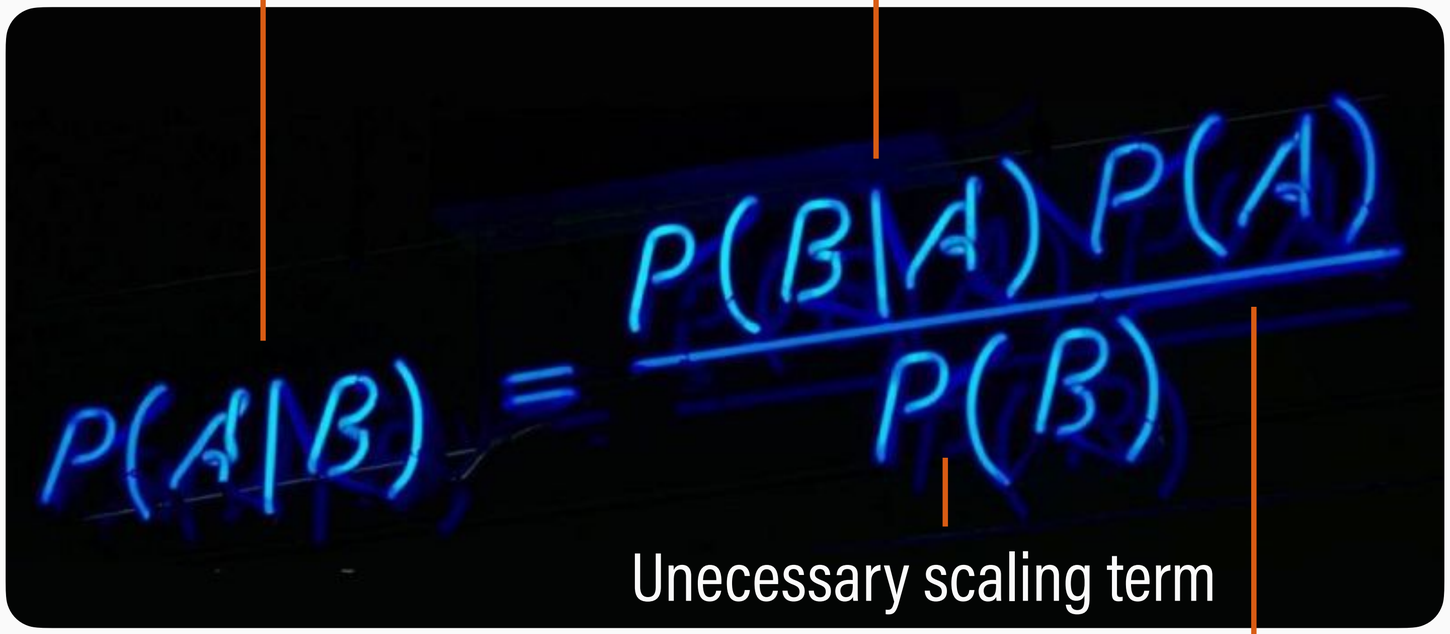
Bayesian

- Parameters are random variables with a distribution of plausible realizations
- Statements about probability, precision, and intervals refer to the parameter
- Probability = our degree of certainty about a parameter after analyzing data

BAYES' THEOREM

Posterior = parameters (A) given the data (B)

Frequentist likelihood = data (B) given the parameters (A)

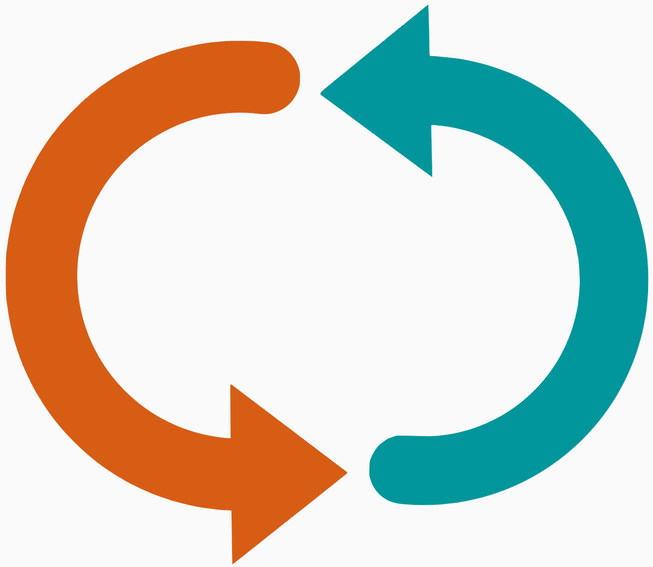


Unecessary scaling term

Prior = a priori belief about parameters (A)

MCMC ESTIMATION

Estimate model parameters



Impute missing values

Do for t = 1 to 10,000 iterations

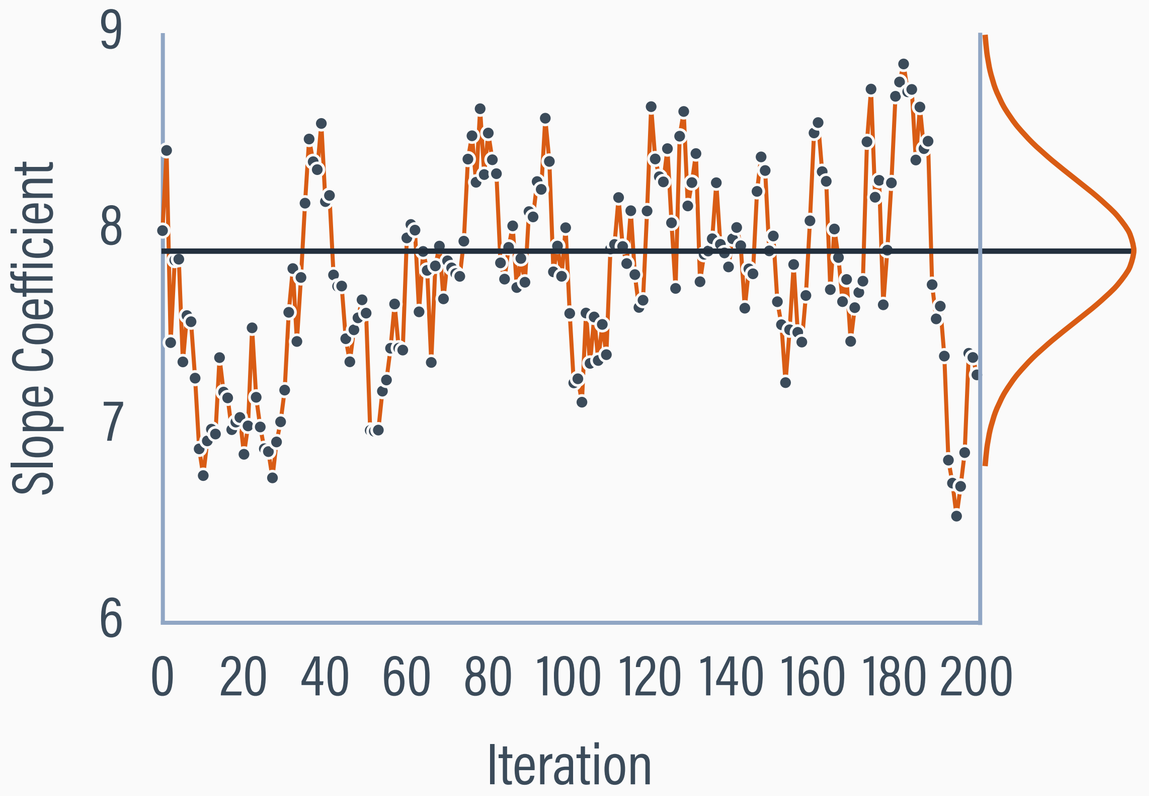
- Estimate model parameters, conditional on the filled-in data
- Impute missing values, conditional on the model parameters

Repeat

Summarize model parameters

MEANING OF ESTIMATION

- MCMC uses computer simulation to “sample” parameters from a distribution
- Estimates continually vary across iterations in a random pattern
- Each iteration gives plausible parameter values that could have produced our data



SIMPLE REGRESSION ILLUSTRATION

- Information literacy at T3 regressed on effort at that occasion

$$info_3 = \beta_0 + \beta_1(effort_3) + \varepsilon$$

- Each iteration yields plausible model parameters and unique imputations based on those parameters

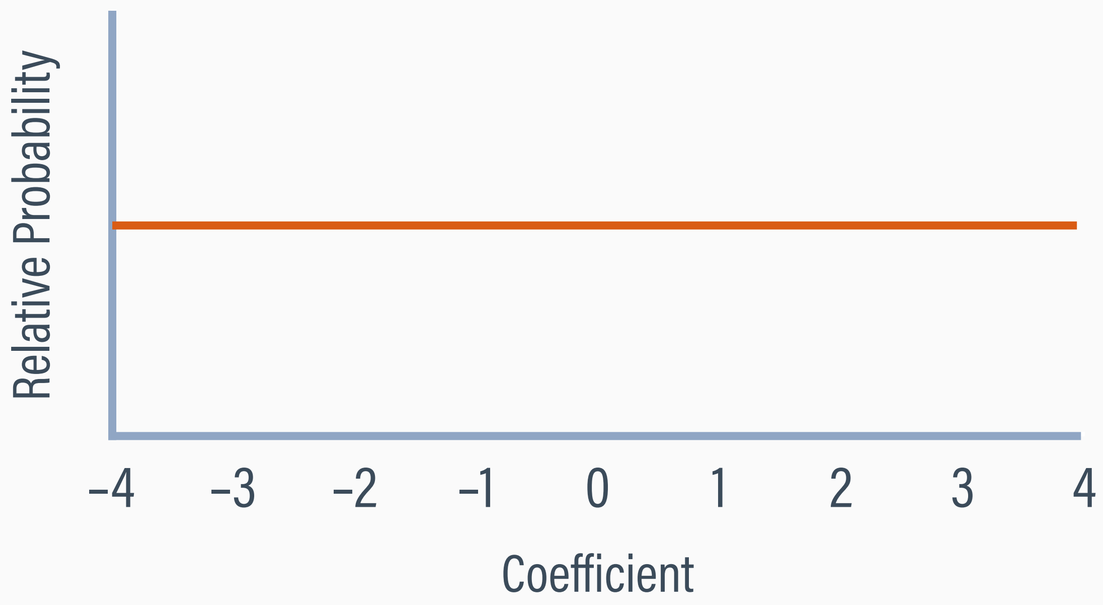
- The goal is to summarize the parameter distributions

PRIOR DISTRIBUTIONS

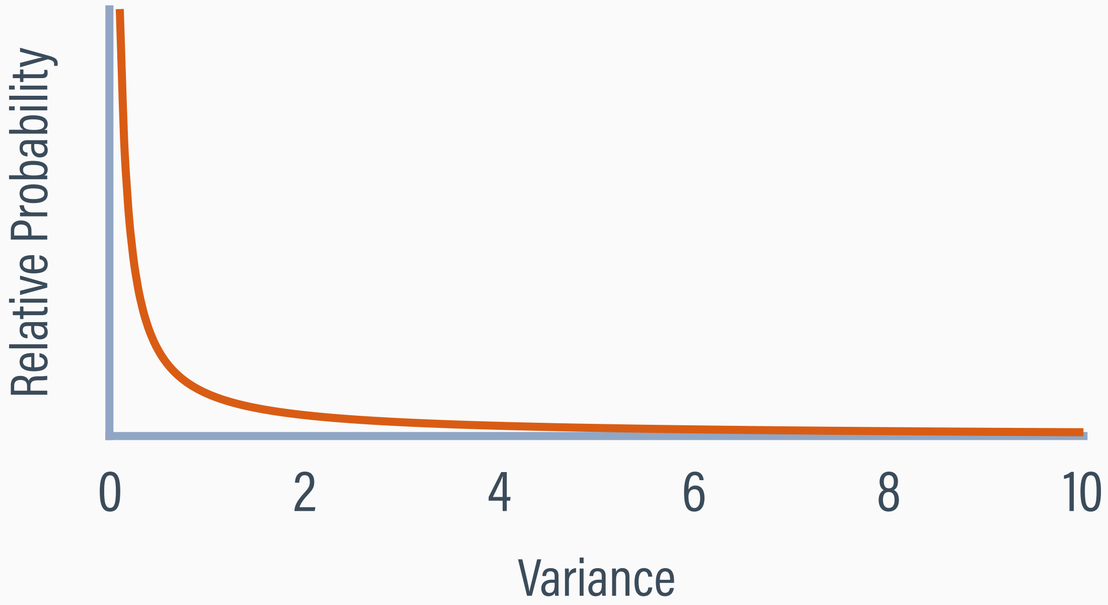
- Bayesian analyses require prior distributions that encode our beliefs about the parameter values prior to analyzing the data
- Conceptually, prior distributions function like secondary inputs that augment the data during estimation
- It is common to non-informative (diffuse) priors that impart as little information as possible (let the data do the talking)

PRIOR DISTRIBUTIONS

- A diffuse prior for means and coefficients conveys that all possible parameter values are equally likely a priori

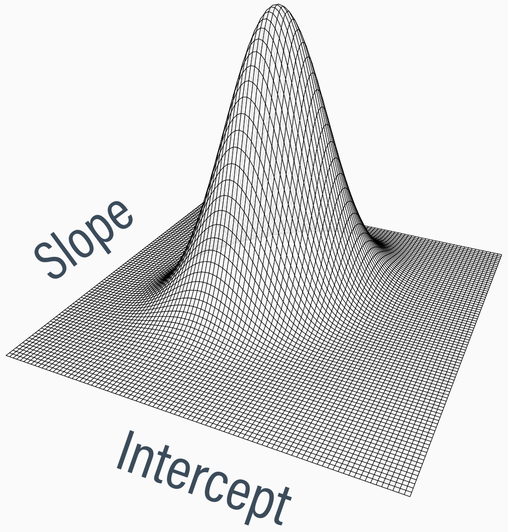


- Diffuse priors for variances are slightly informative, and different options function like df adjustments in regression

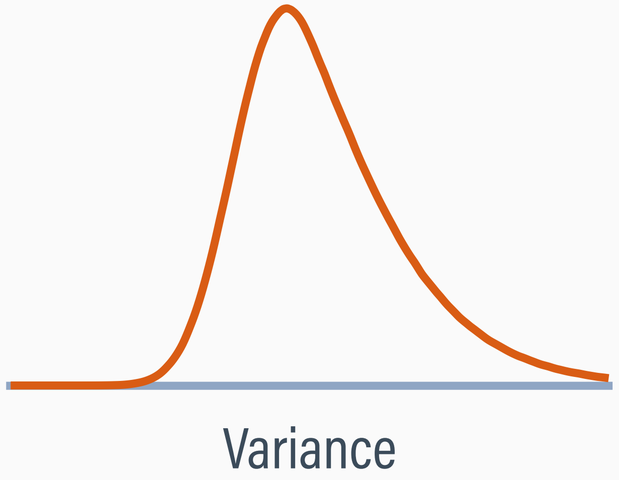


PARAMETER-GENERATING DISTRIBUTIONS

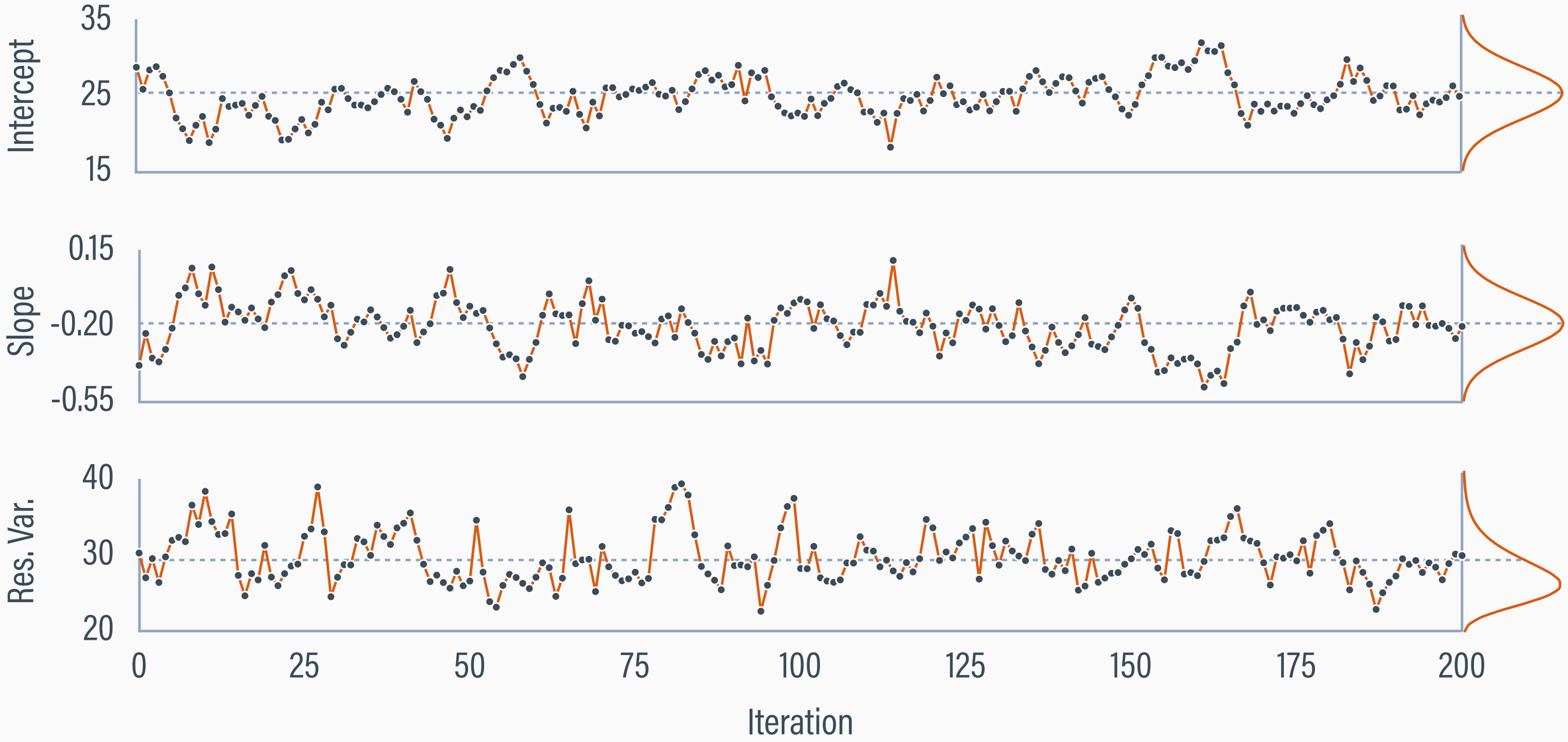
- MCMC draws coefficients from a multivariate normal distribution, with least-squares estimates defining shape



- MCMC draws variances from an inverse gamma distribution with its shape determined by the df and residual SS



PARAMETERS FROM 200 MCMC CYCLES

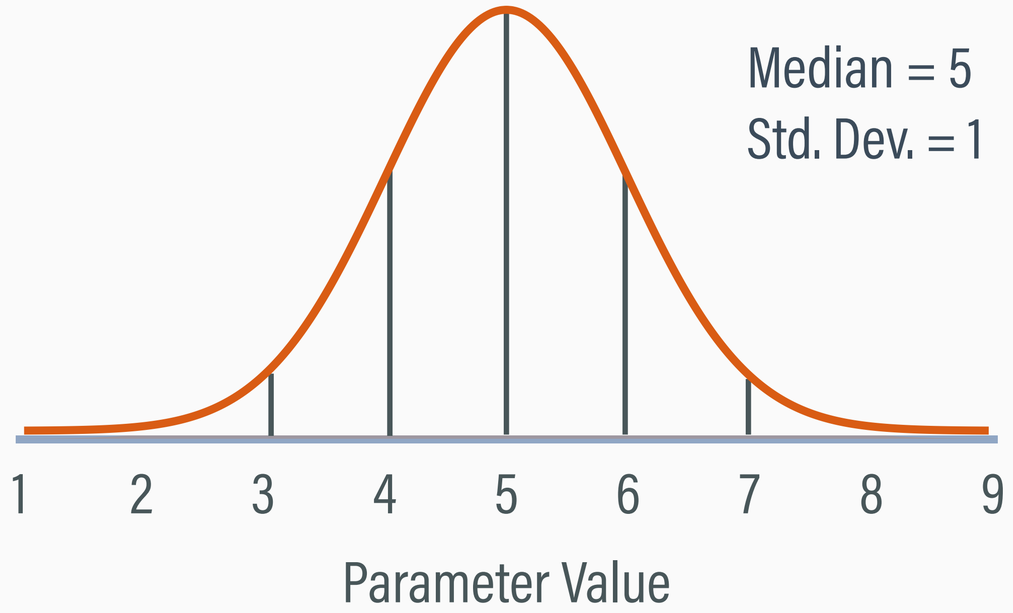


SUMMARIZING MCMC ESTIMATES

- MCMC iterates for thousands of cycles, and each cycle produces estimates based on one filed-in data set
- Bayesian estimation yields a distribution of parameters—called a posterior—that averages over thousands of imputations
- The posterior is a distribution of plausible parameter values that could have produced our particular data

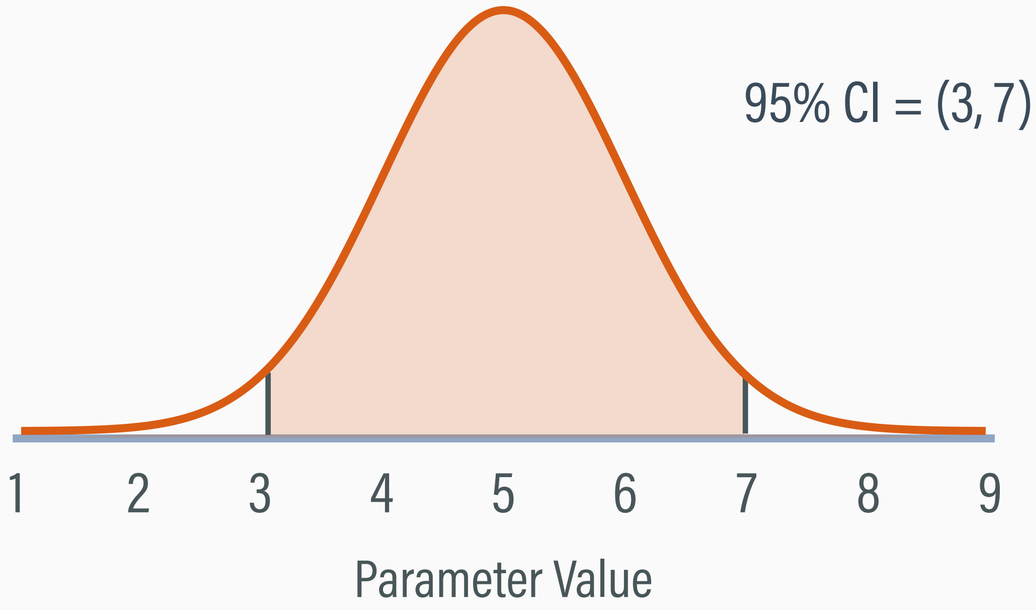
POSTERIOR MEDIAN AND STD. DEV.

- The posterior median and standard deviation quantify the most likely parameter value and uncertainty
- Analogous to a point estimate and standard error but no reference to other hypothetical samples



95% CREDIBLE INTERVALS

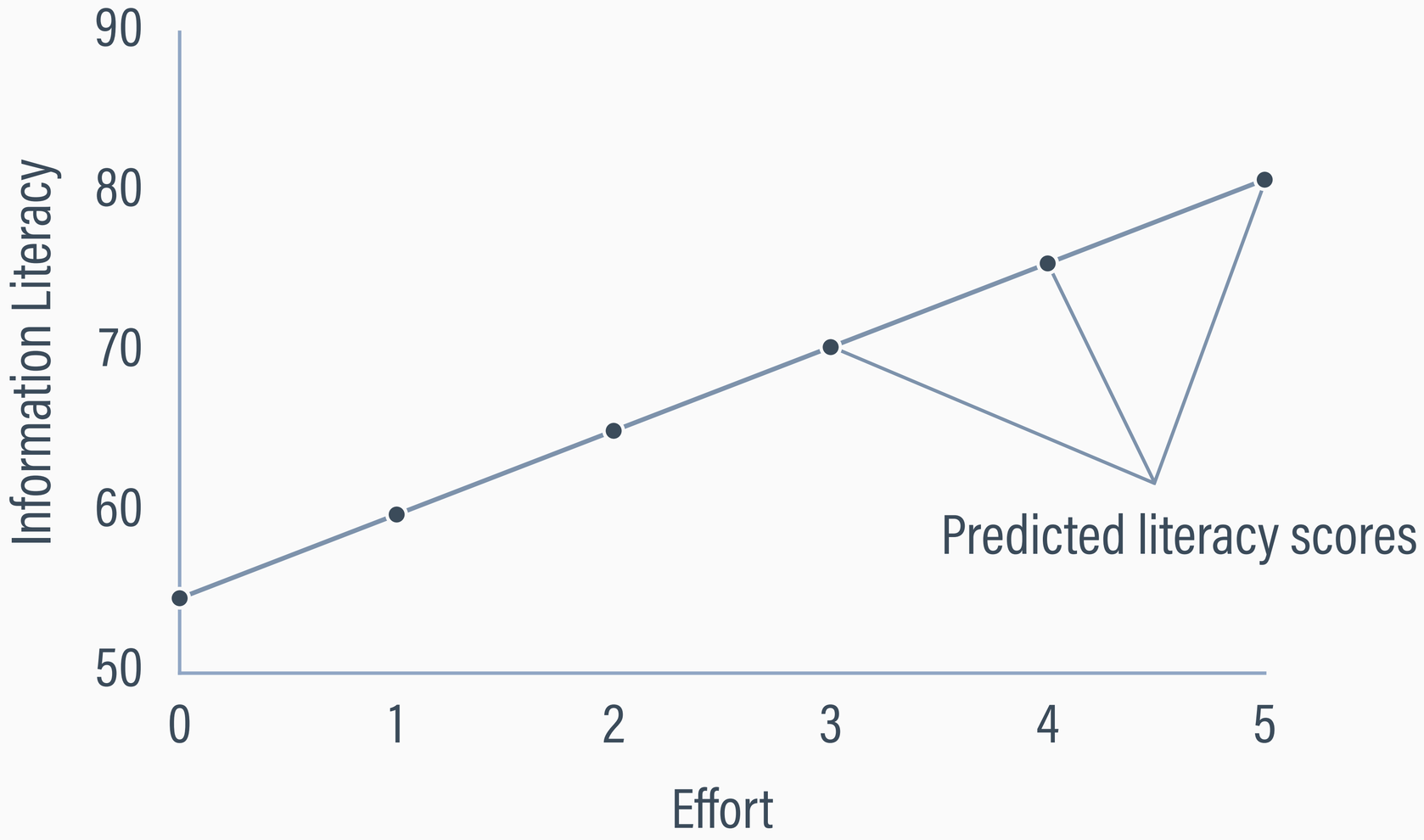
- The 95% credible interval gives limits spanning 95% of the parameter's range
- Akin to a confidence interval, but references a range of highly plausible parameter values for one data set



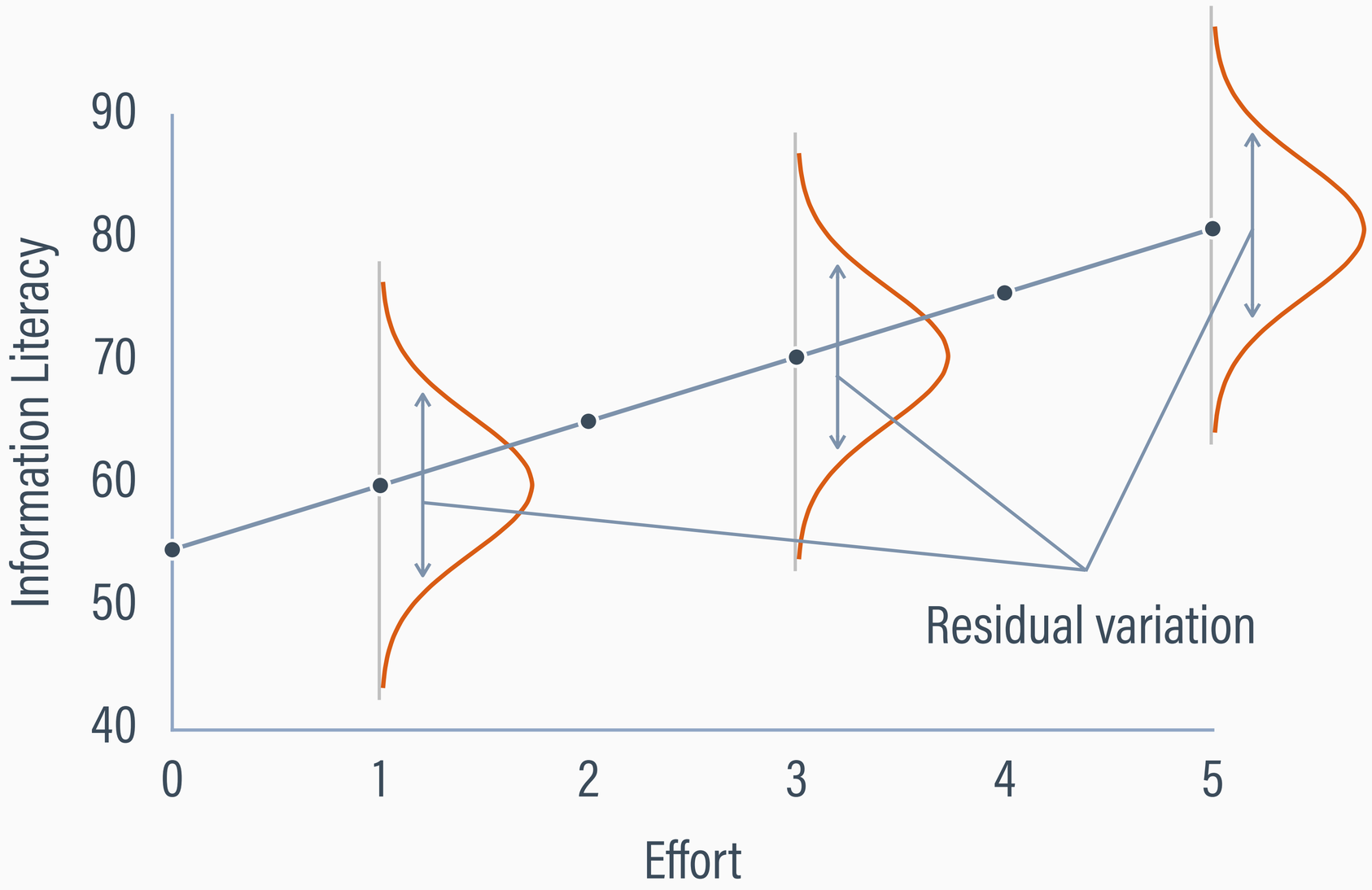
MISSING DATA IMPUTATION STEP

- Missing scores are imputed by drawing replacement scores at random from a distribution of plausible values
- The model parameters at each iteration combine to define the center and spread of the missing data imputations
- Each imputation can be viewed as a predicted score plus a computer-simulated random noise term

PREDICTED VALUES

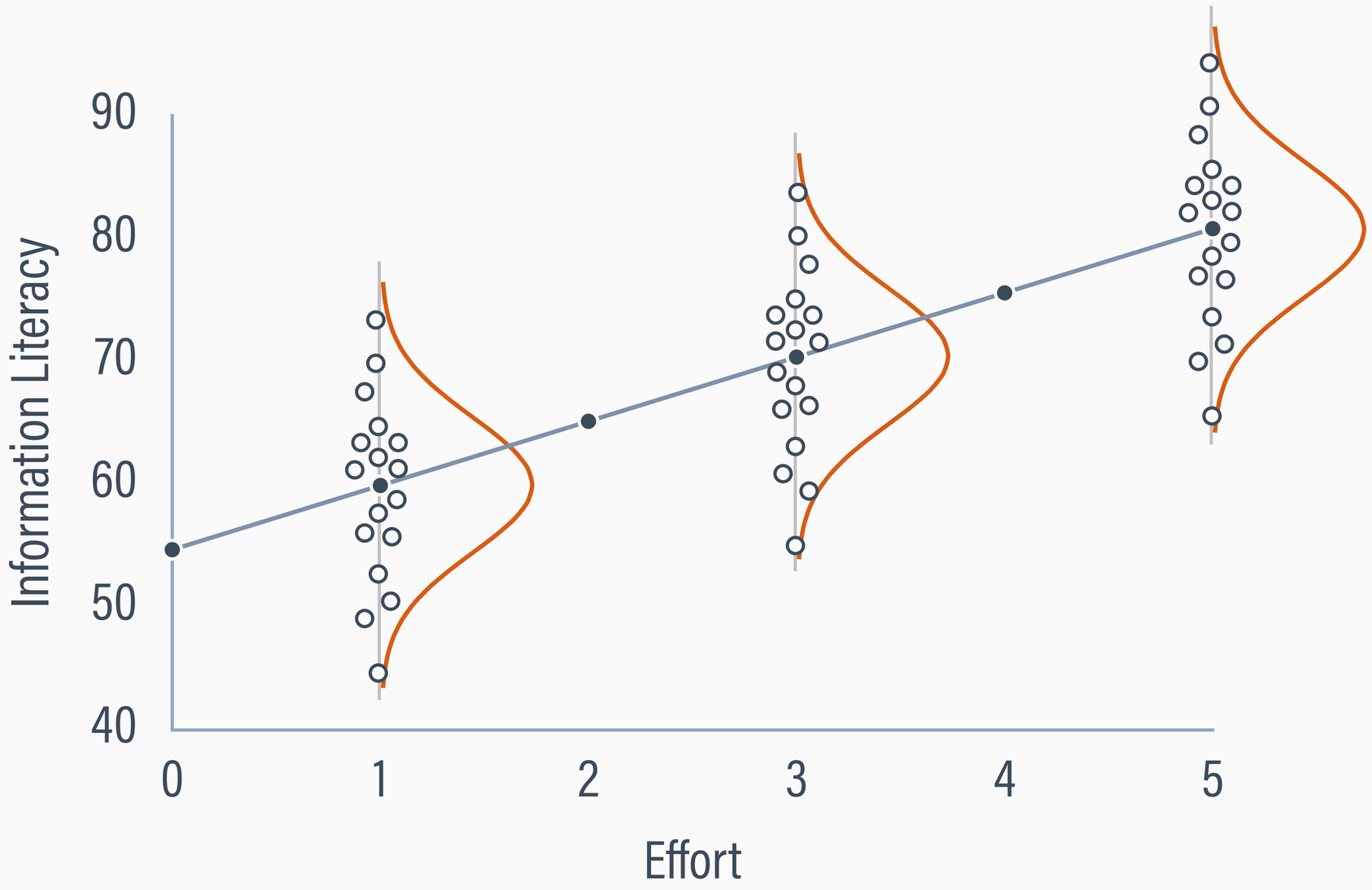


RESIDUAL VARIATION



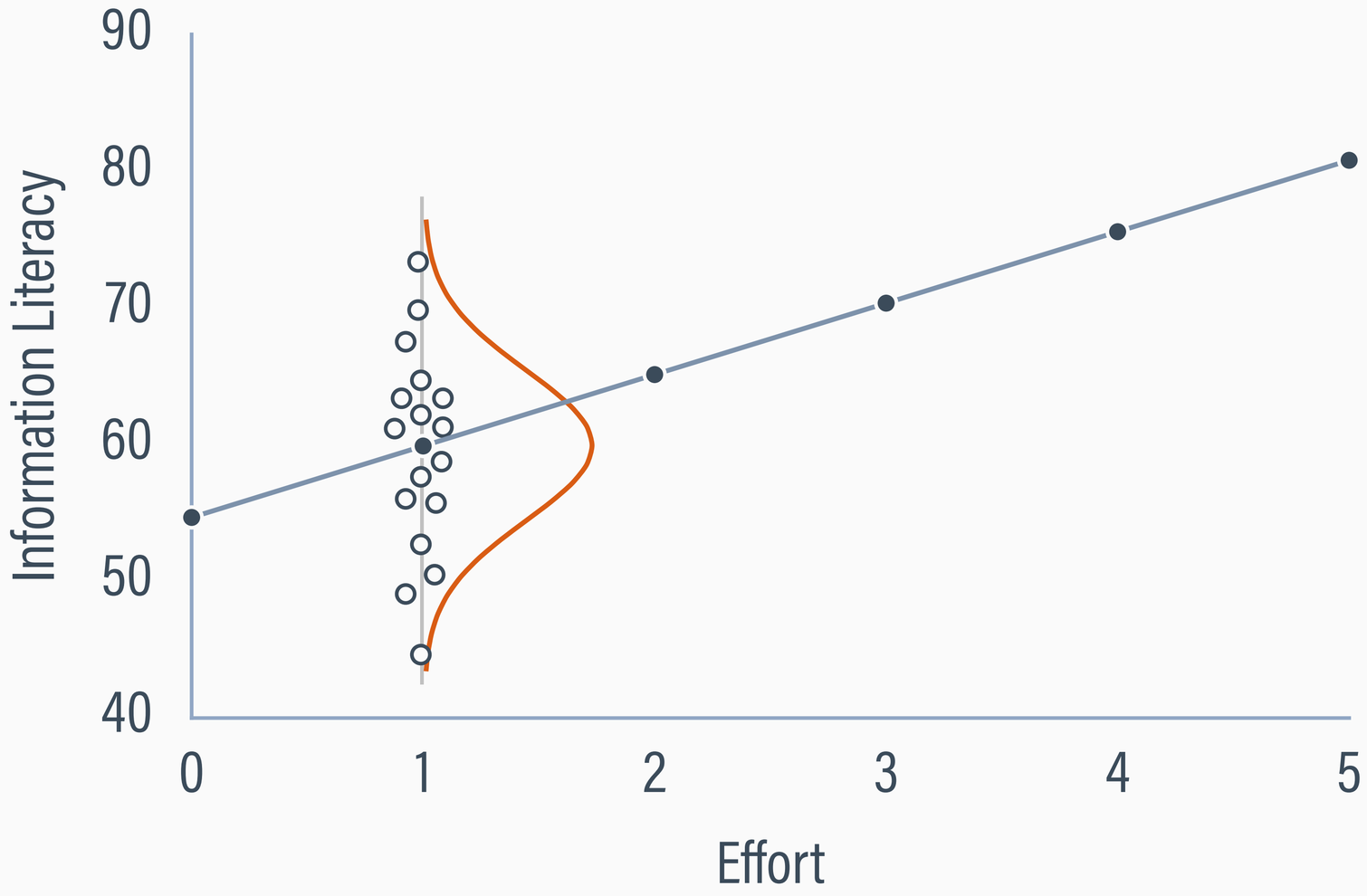
DISTRIBUTIONS OF IMPUTATIONS

○ = plausible literacy imputations



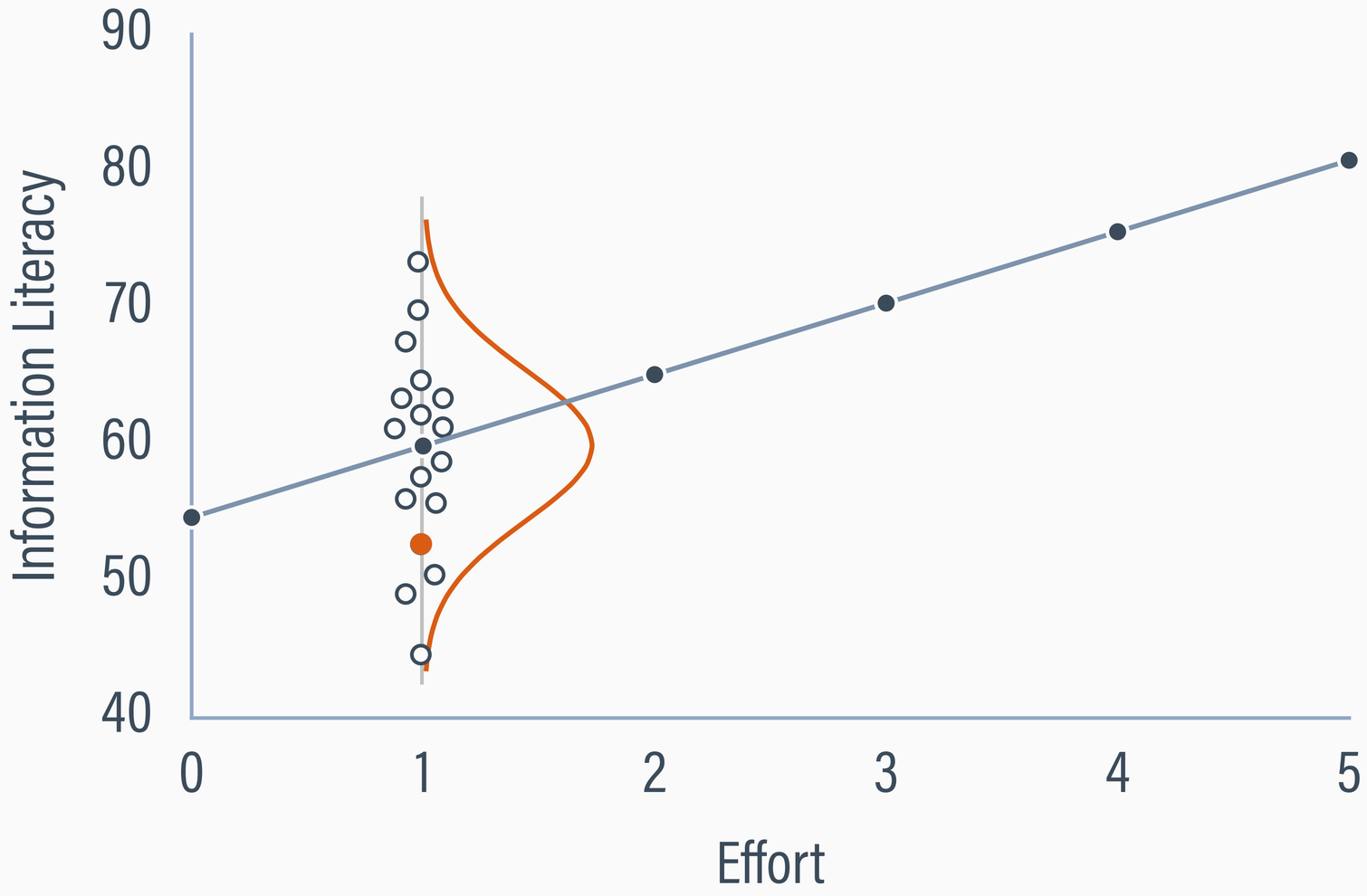
IMPUTATIONS FOR LOW EFFORT

○ = plausible literacy imputations

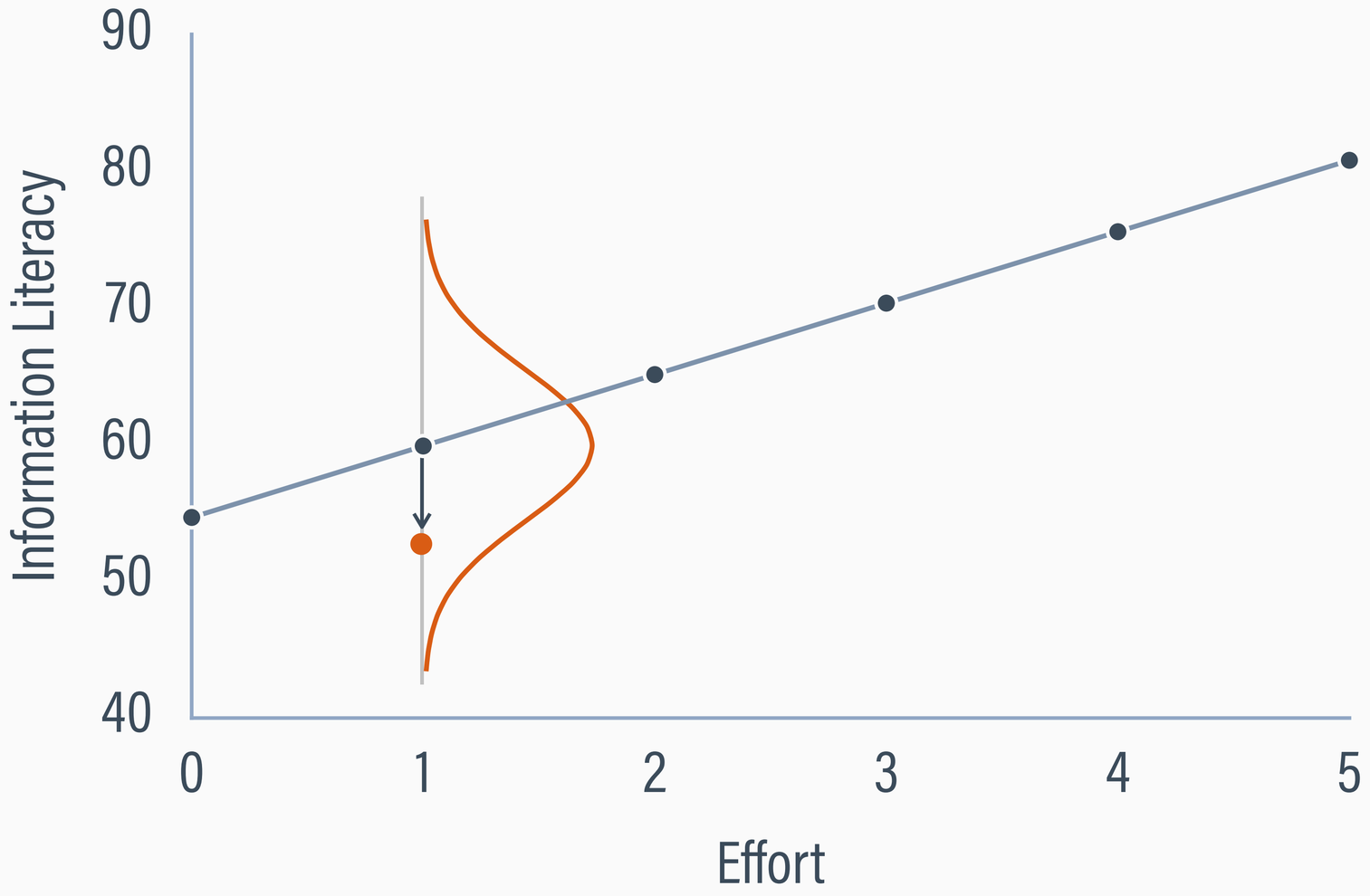


DRAW IMPUTATION AT RANDOM

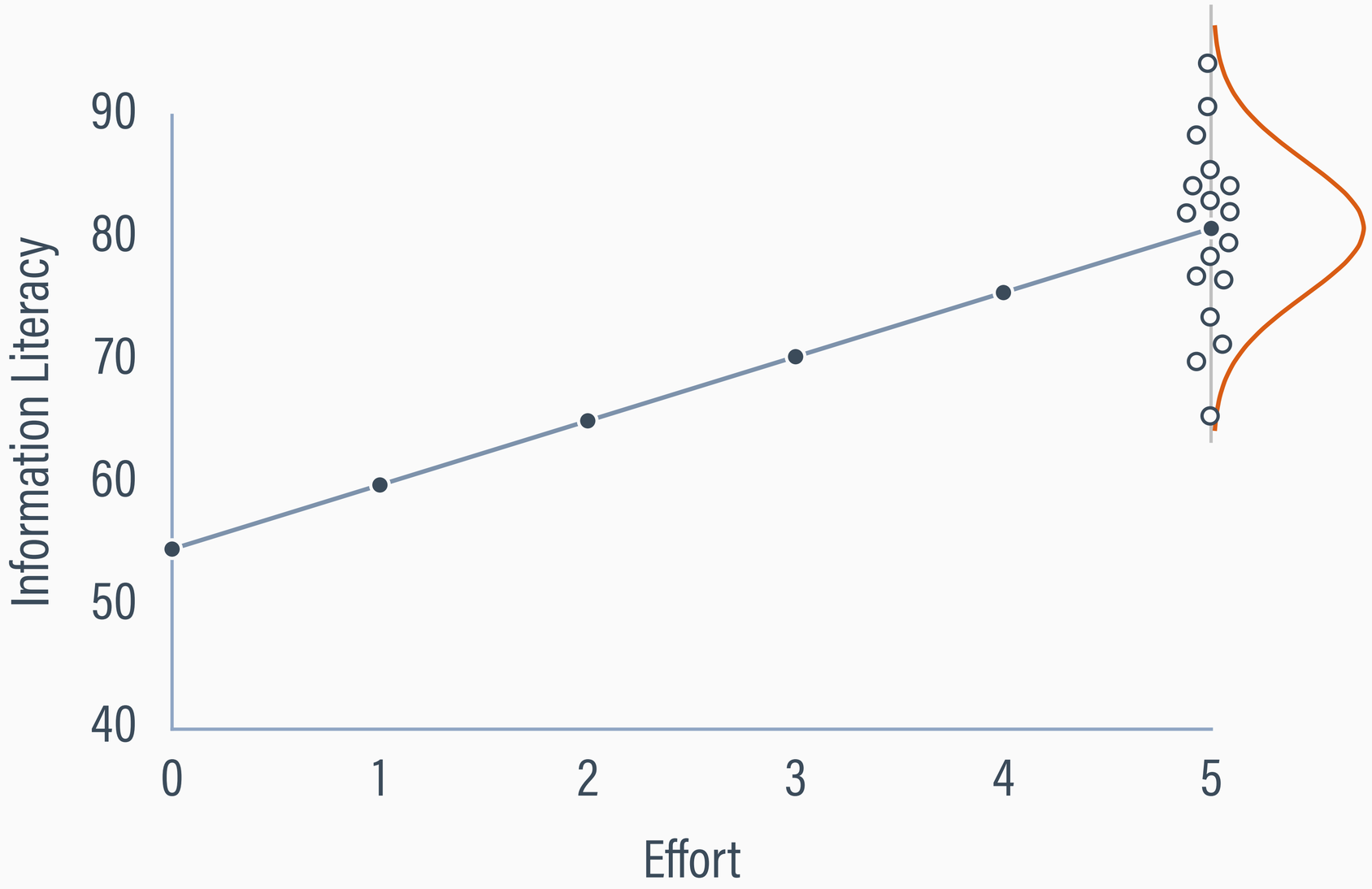
● = randomly selected imputation



IMPUTATION = PREDICTION + NOISE

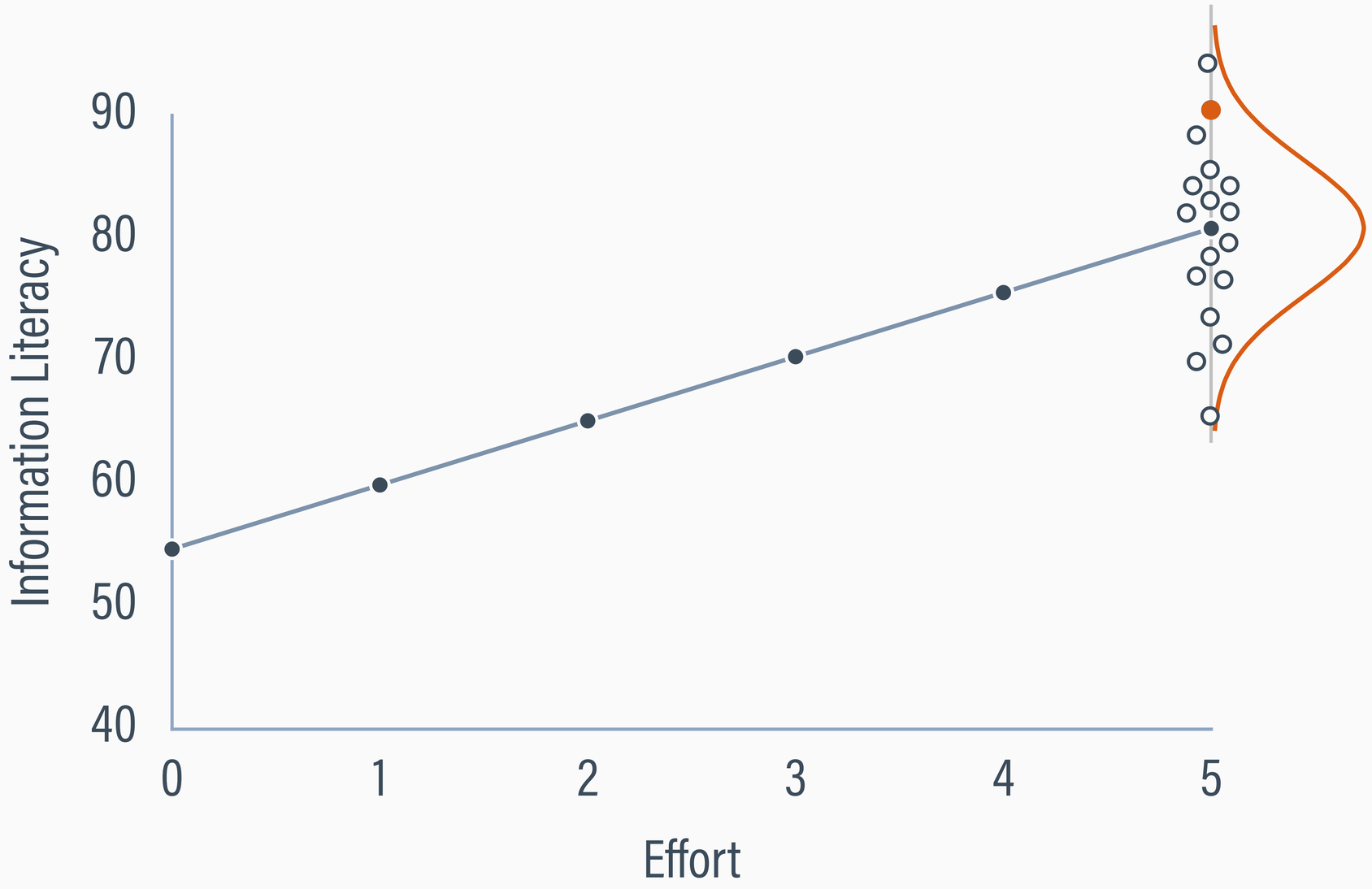


IMPUTATIONS FOR HIGH EFFORT

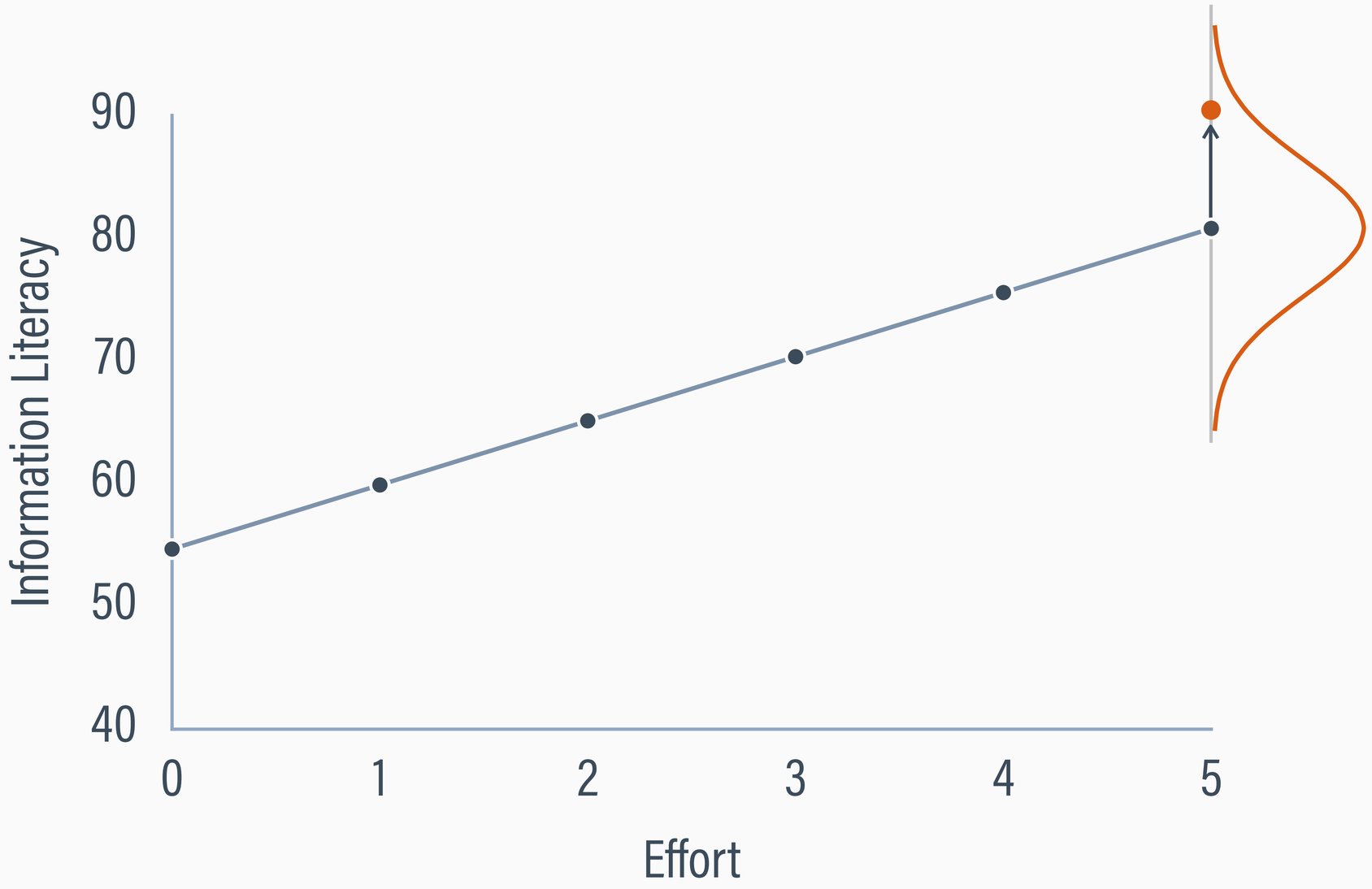


DRAW IMPUTATION AT RANDOM

● = randomly selected imputation



IMPUTATION = PREDICTION + NOISE

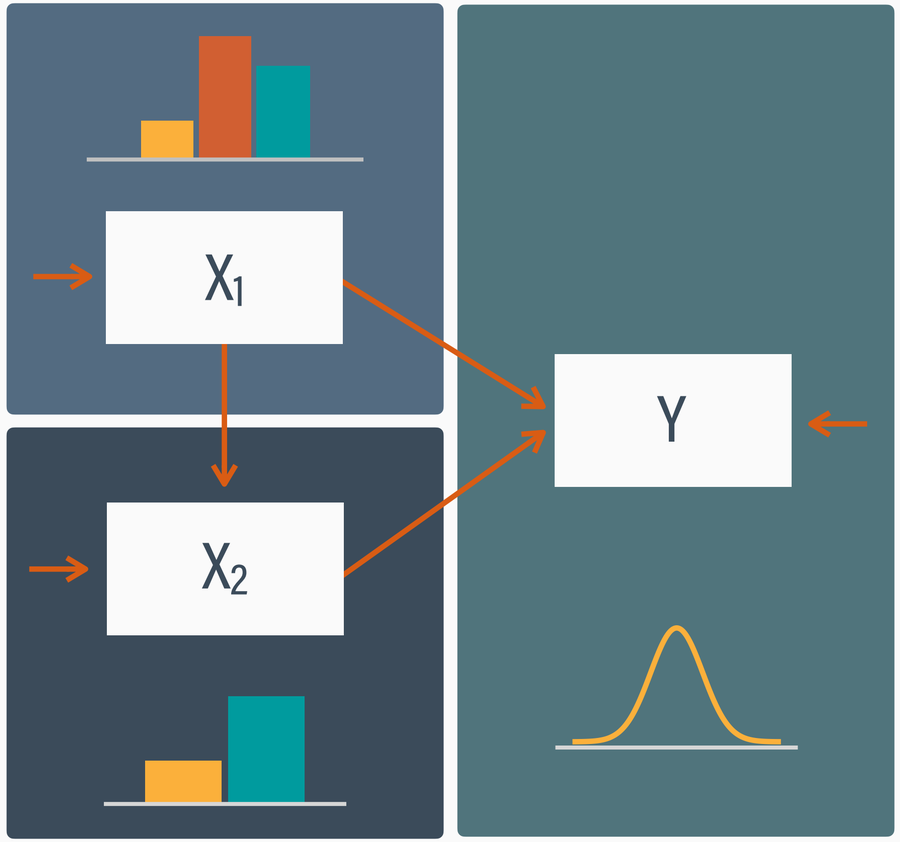


INCOMPLETE PREDICTORS

- Incomplete predictors require their own model and distributional assumptions
- Multivariate normal methods can mis-specify the data distributions in a way that introduces bias
- Factored regression uses a modular specification where a sequence of submodels replaces a multivariate model

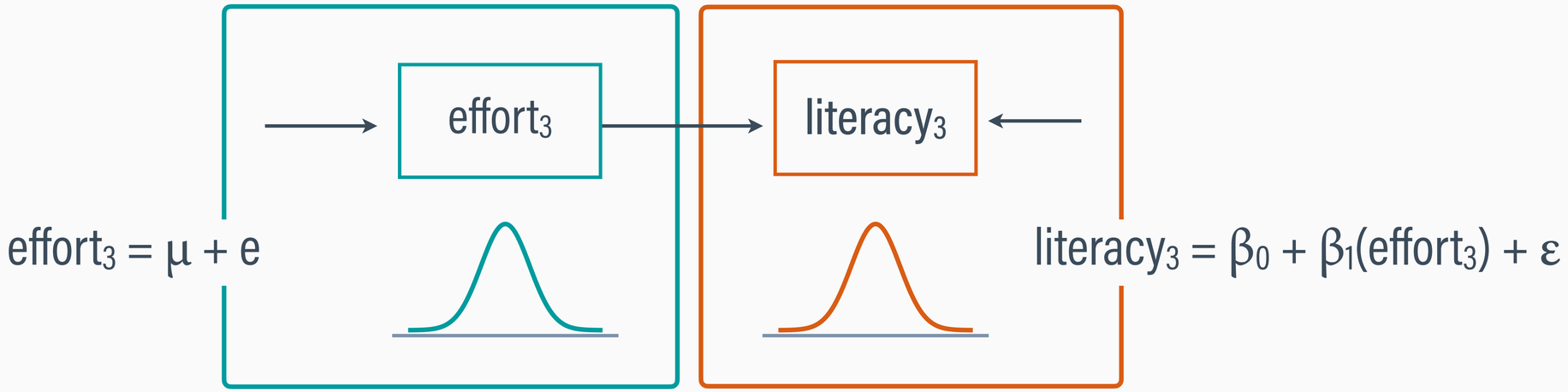
FACTORED REGRESSION SPECIFICATIONS

- MCMC uses a factored regression specification that invokes a unique distribution for each variable
- The analysis consists of a collection of univariate regression models
- Each model can include terms that are at odds with multivariate normality



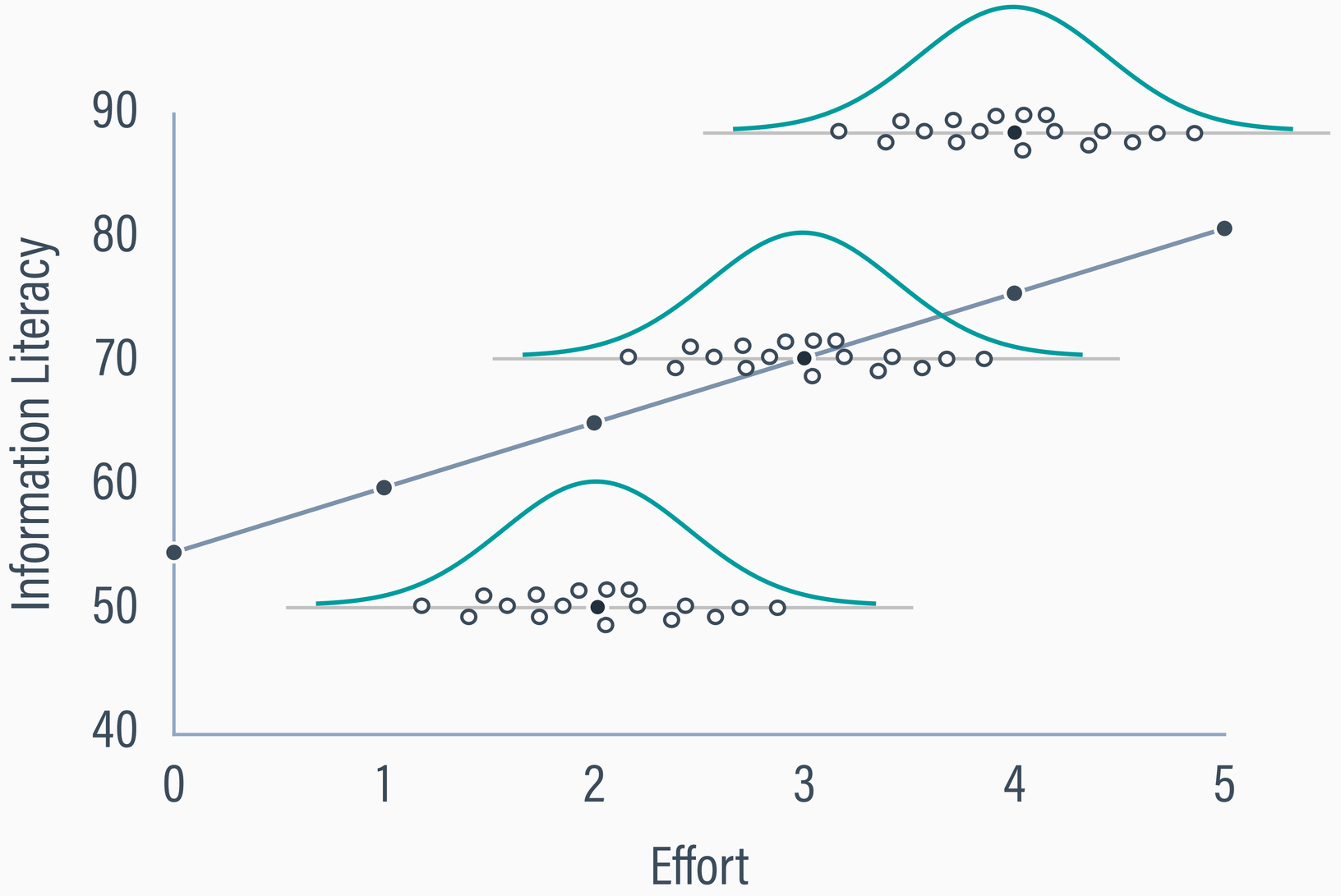
INCOMPLETE PREDICTORS

- Effort is the regressor in the focal model and an outcome in its own empty model



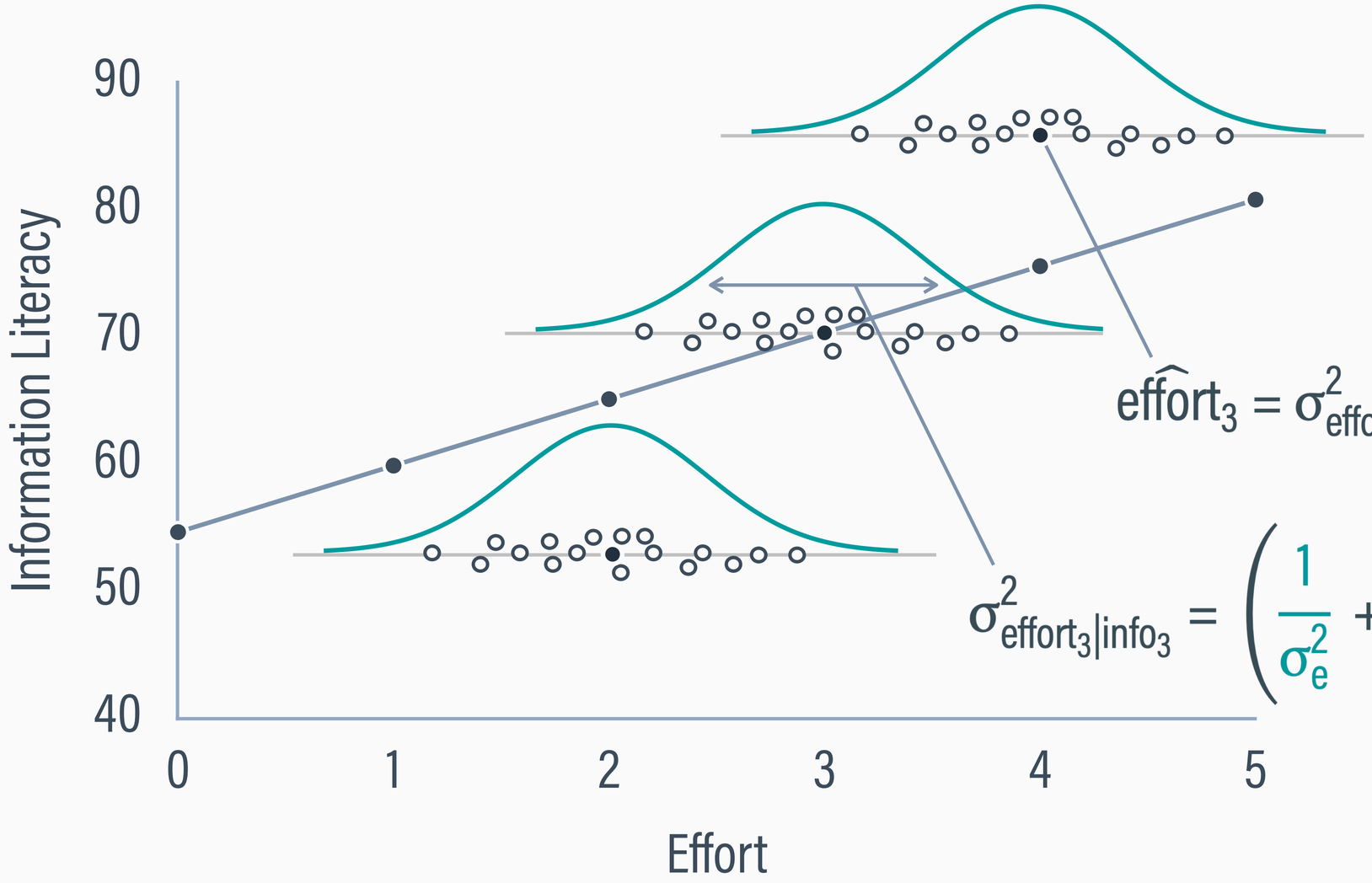
- Both models inform the distribution of predictor imputations

DISTRIBUTIONS OF IMPUTATIONS



PREDICTED VALUES AND VARIATION

Multiple sets of model parameters define the mean and spread of the imputations



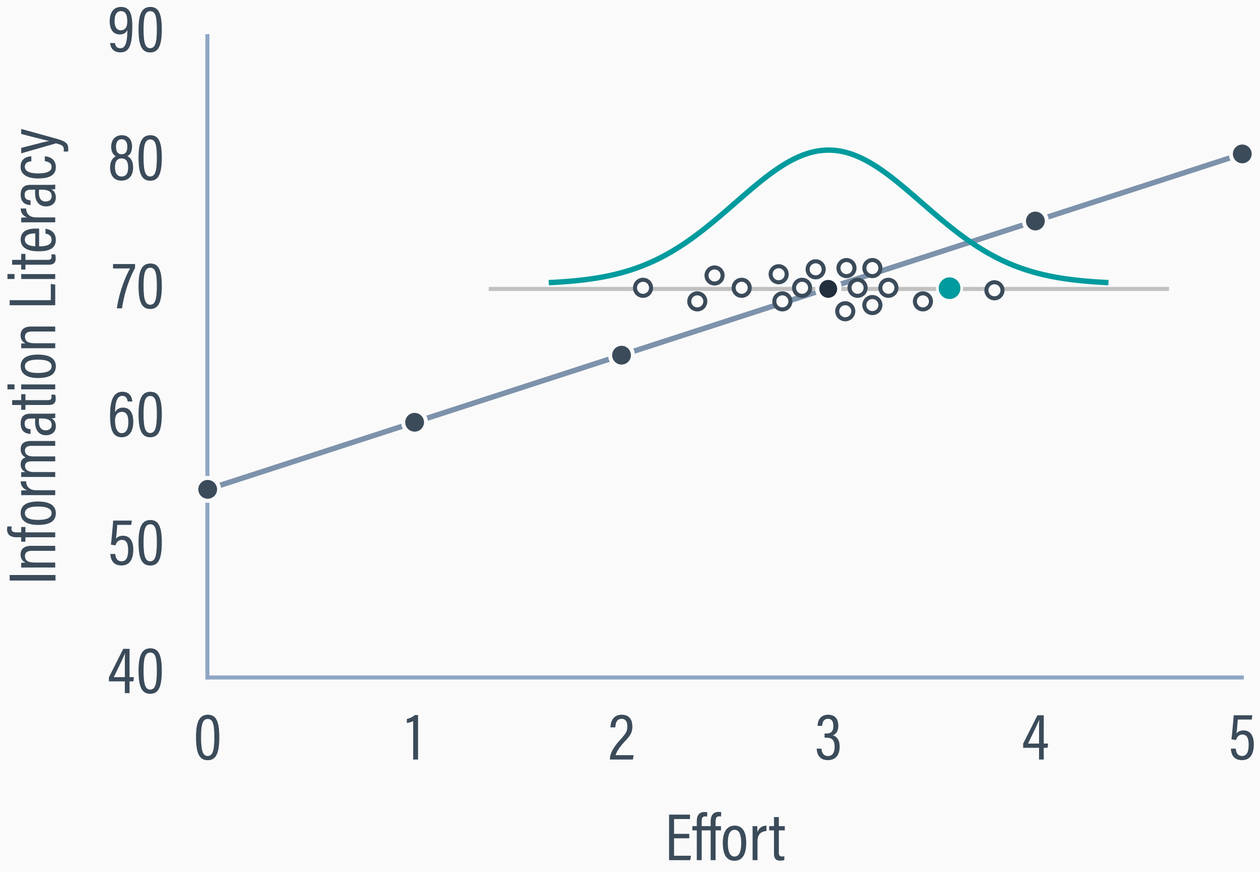
$$\widehat{\text{effort}}_3 = \sigma^2_{\text{effort}_3|\text{info}_3} \times \left(\frac{\mu}{\sigma^2_e} + \frac{\beta_1(\text{info}_3 - \beta_0)}{\sigma^2_\varepsilon} \right)$$

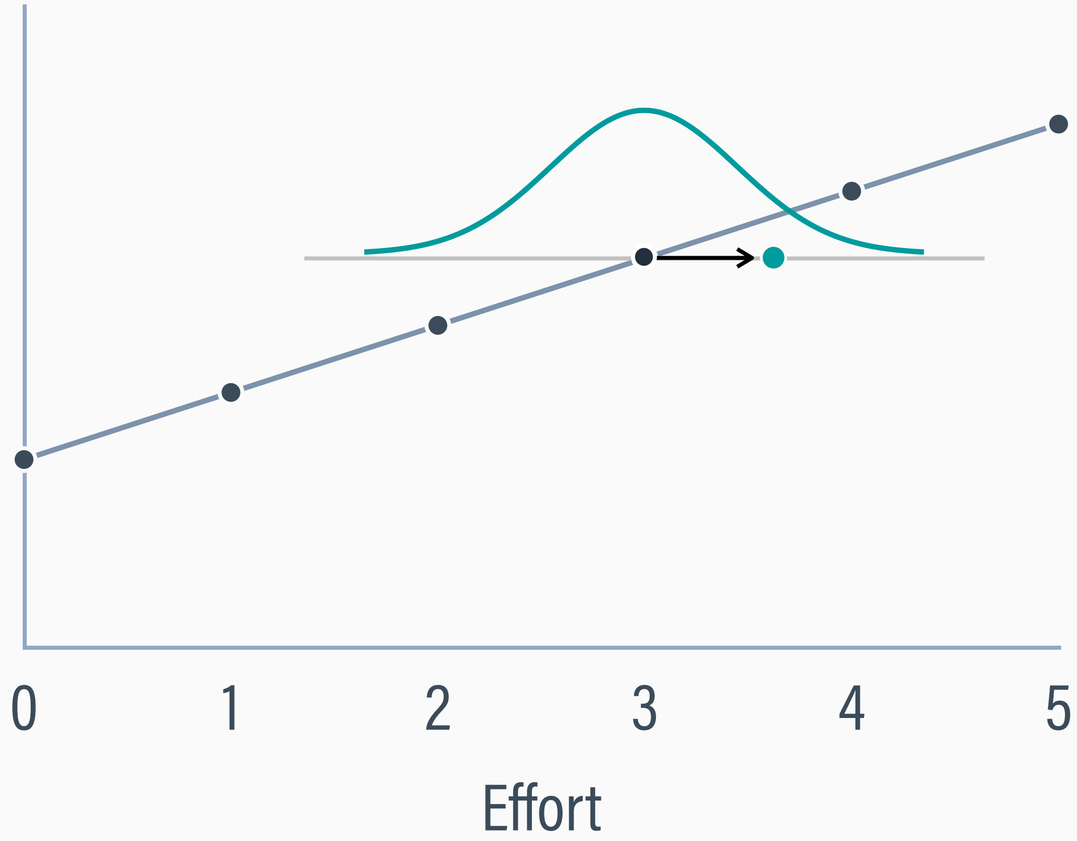
$$\sigma^2_{\text{effort}_3|\text{info}_3} = \left(\frac{1}{\sigma^2_e} + \frac{\beta_1^2}{\sigma^2_\varepsilon} \right)^{-1}$$

IMPUTATION EXAMPLE

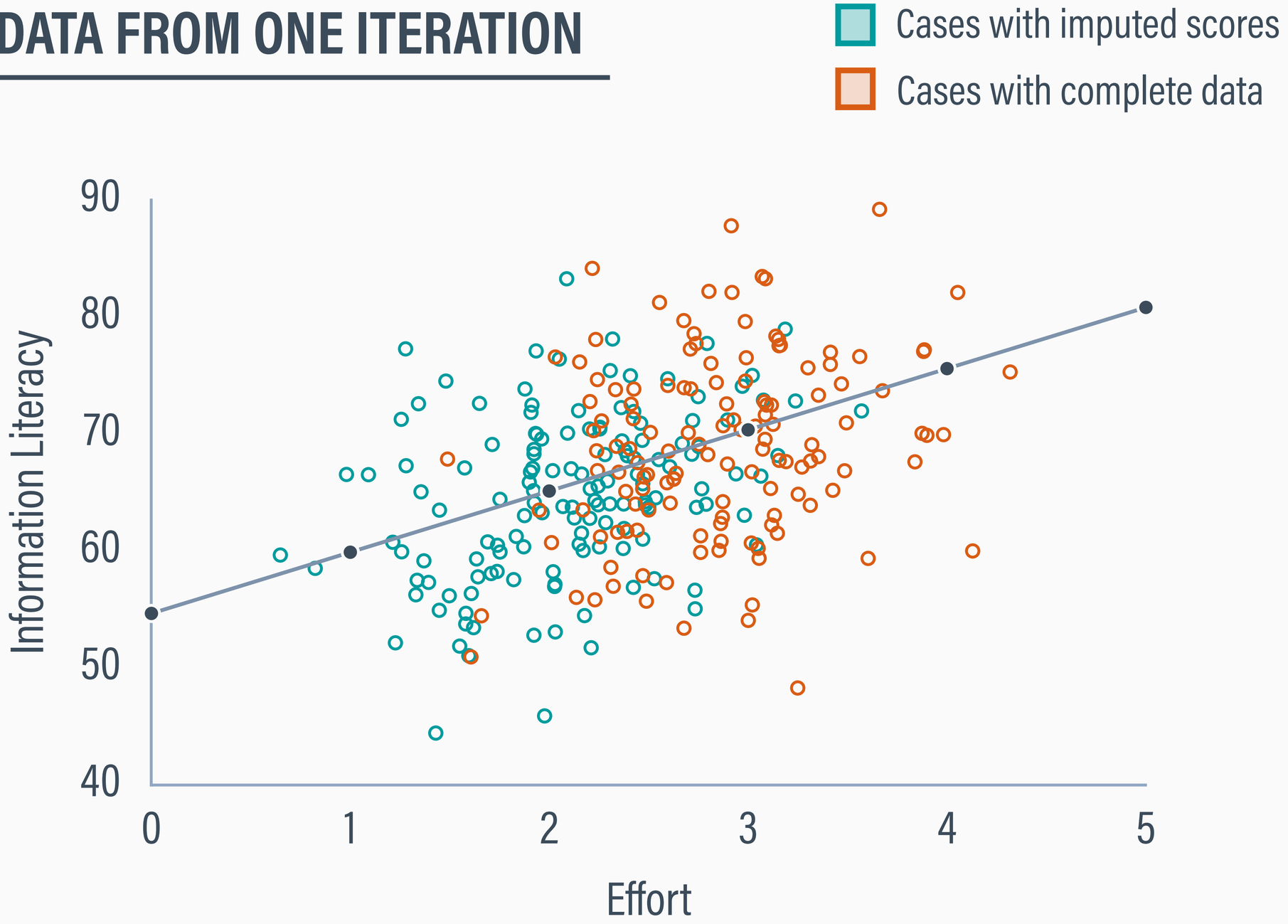
● = randomly selected imputation

Imputation = predicted value + random normal noise



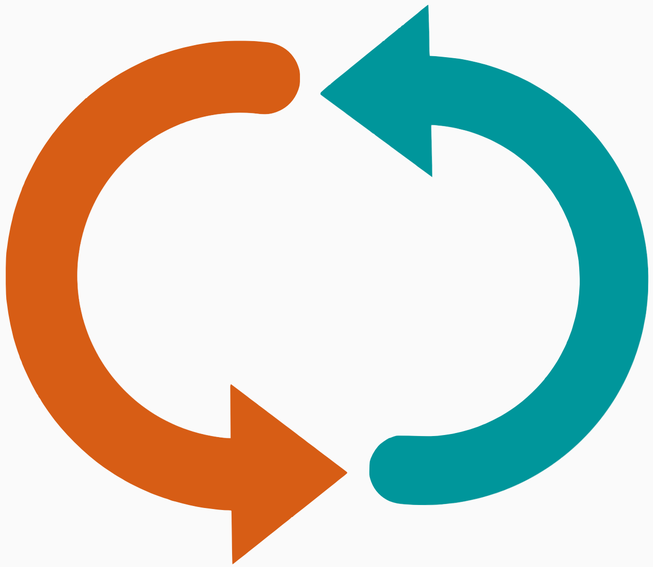


FILLED-IN DATA FROM ONE ITERATION



MCMC ESTIMATION

Estimate model parameters



Impute missing values

Do for t = 1 to 10,000 iterations

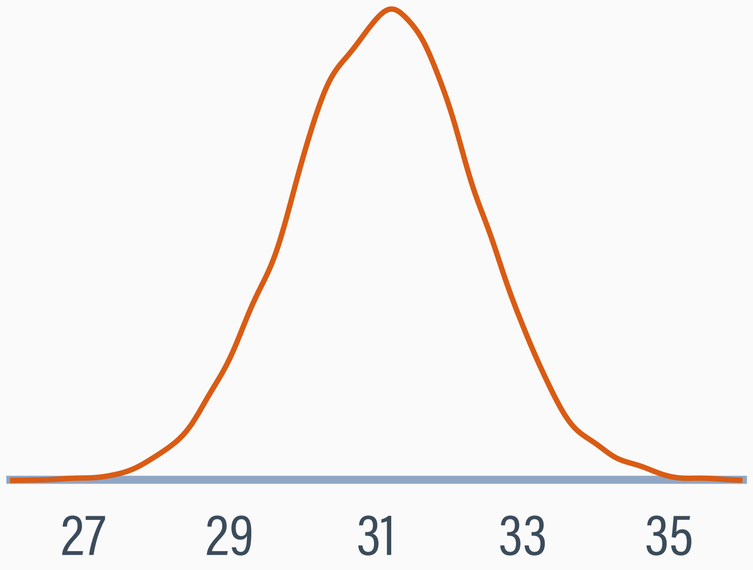
- Estimate model parameters, conditional on the filled-in data
- Impute missing values, conditional on the model parameters

Repeat

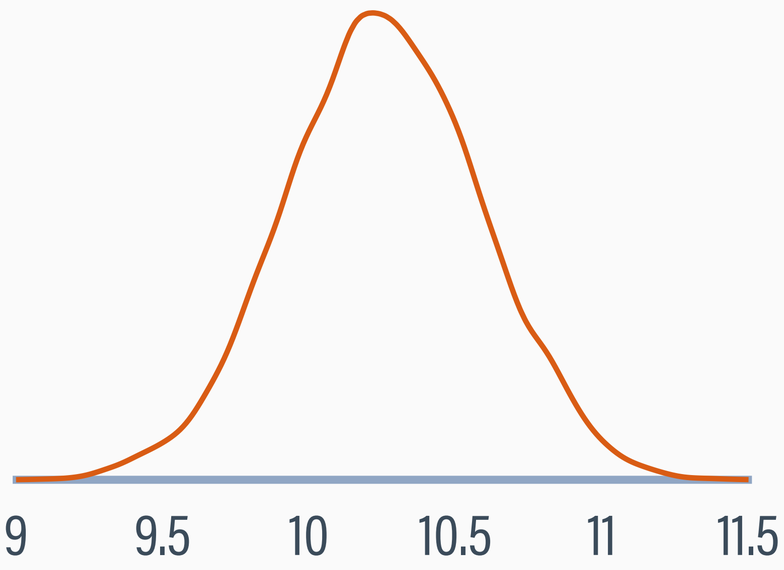
Summarize model parameters

PARAMETER (POSTERIOR) DISTRIBUTIONS

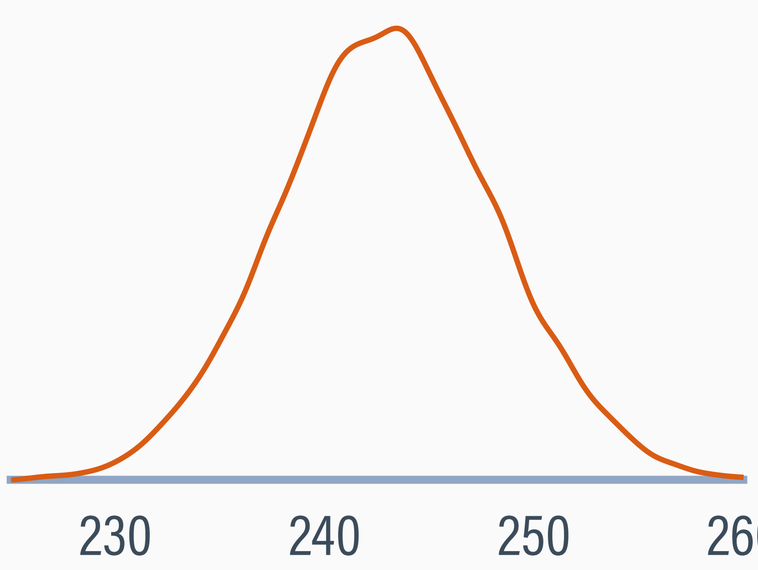
Median = 31.12
Std. Dev. = 1.26
95% CI = (28.65, 33.58)



Median = 10.25
Std. Dev. = 0.34
95% CI = (9.60, 10.91)



Median = 242.98
Std. Dev. = 5.23
95% CI = (232.45, 252.97)



ESTIMATOR COMPARISON

The two estimators are numerically equivalent!!!

	MCMC			FIML		
Parameter	Median	SD	95% CI	Est.	SE	95% CI
Intercept	31.12	1.26	(28.65, 33.58)	31.12	1.26	(28.64, 33.58)
Effort	10.25	0.34	(9.60, 10.91)	10.25	0.34	(9.60, 10.91)
Residual variance	242.98	5.23	(232.85, 253.46)	242.71	5.23	(232.45, 252.97)
R^2	.18	.01	(.16, .20)	.18	—	—

MCMC AS COMPUTATIONAL FREQUENTISM

- Researchers adopting a computational frequentism view can use MCMC results as surrogates for frequentist inference (Levy & McNeish, 2021)
- In this scenario, MCMC is a flexible way to estimate frequentist quantities when FIML solutions are unavailable (e.g., missing data)

Bayesian Inference

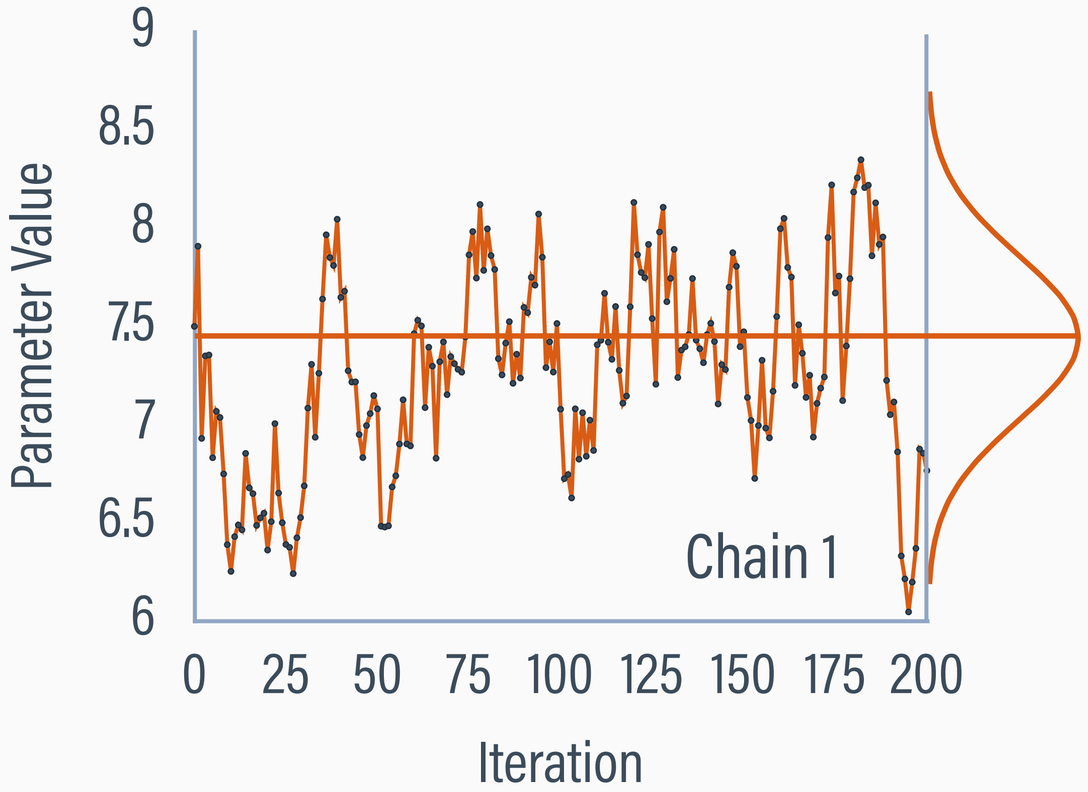
Computational Frequentism

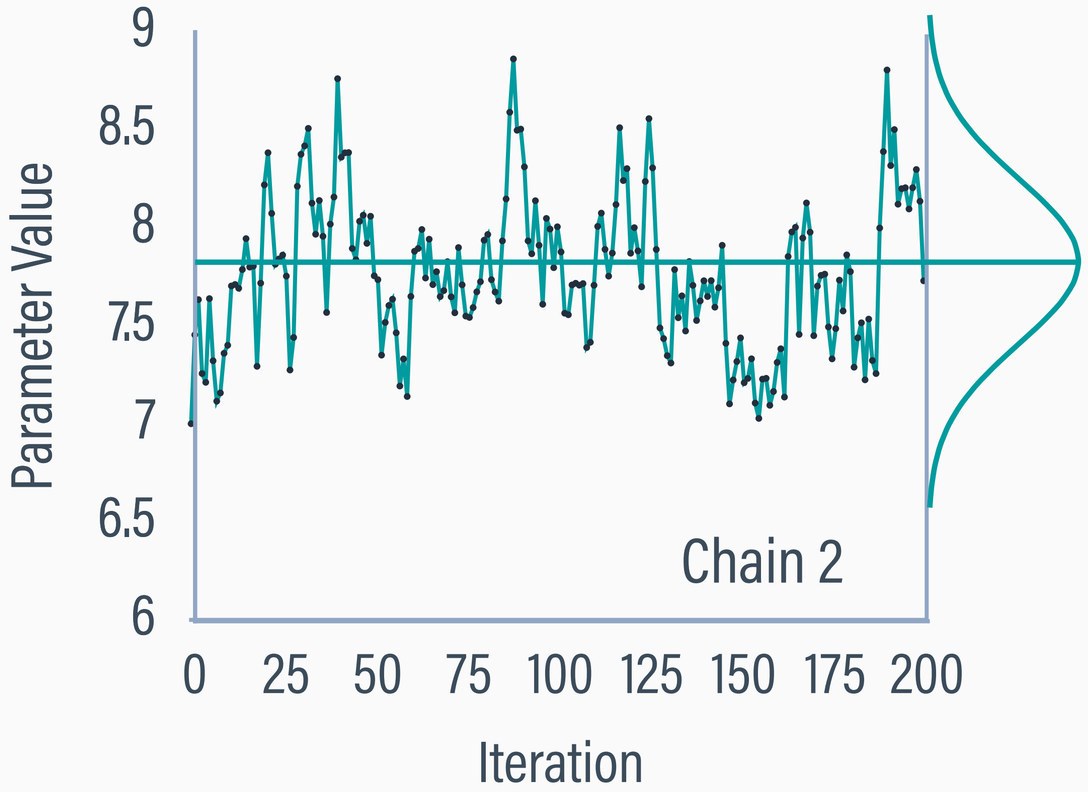
MCMC CONVERGENCE

- With missing data, it is especially important to evaluate whether MCMC is converging and producing reasonable results
- MCMC converges when parameter estimates oscillate around a stable mean, and variation doesn't change with more iterations
- The potential scale reduction factor (PSRF) compares the similarity of parameters generated from two MCMC processes

POTENTIAL SCALE REDUCTION FACTOR

$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains + within-chain variation}}{\text{within-chain variation}}}$$

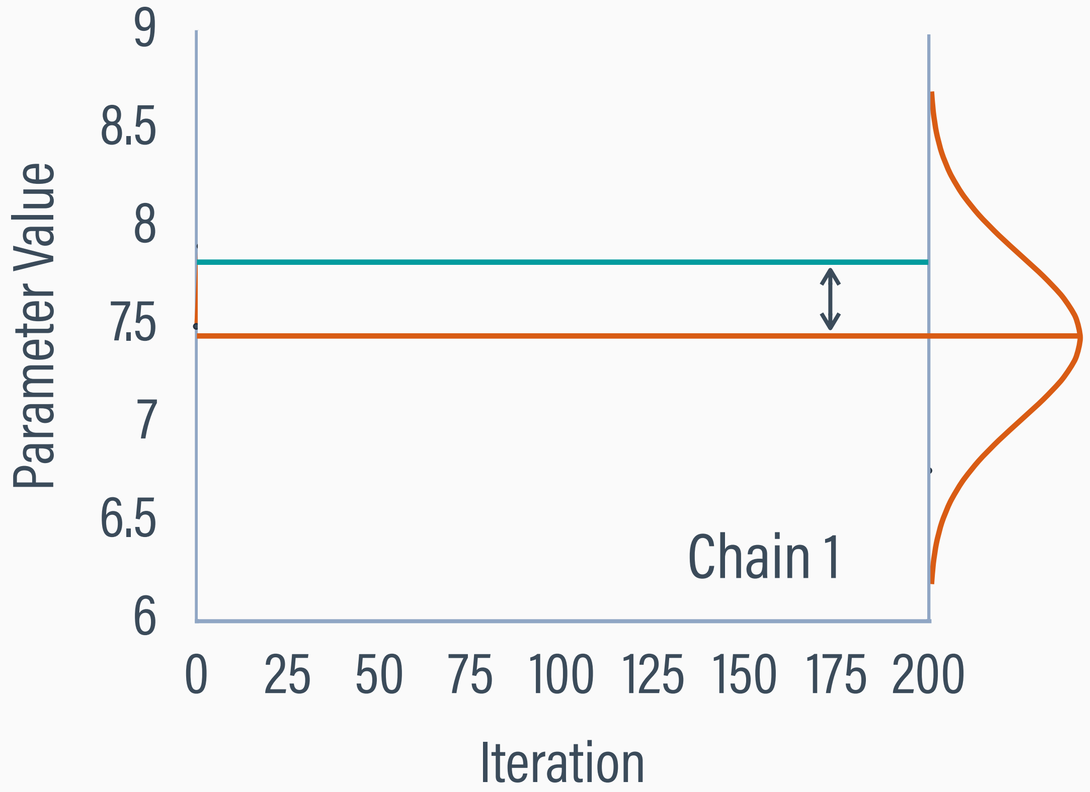


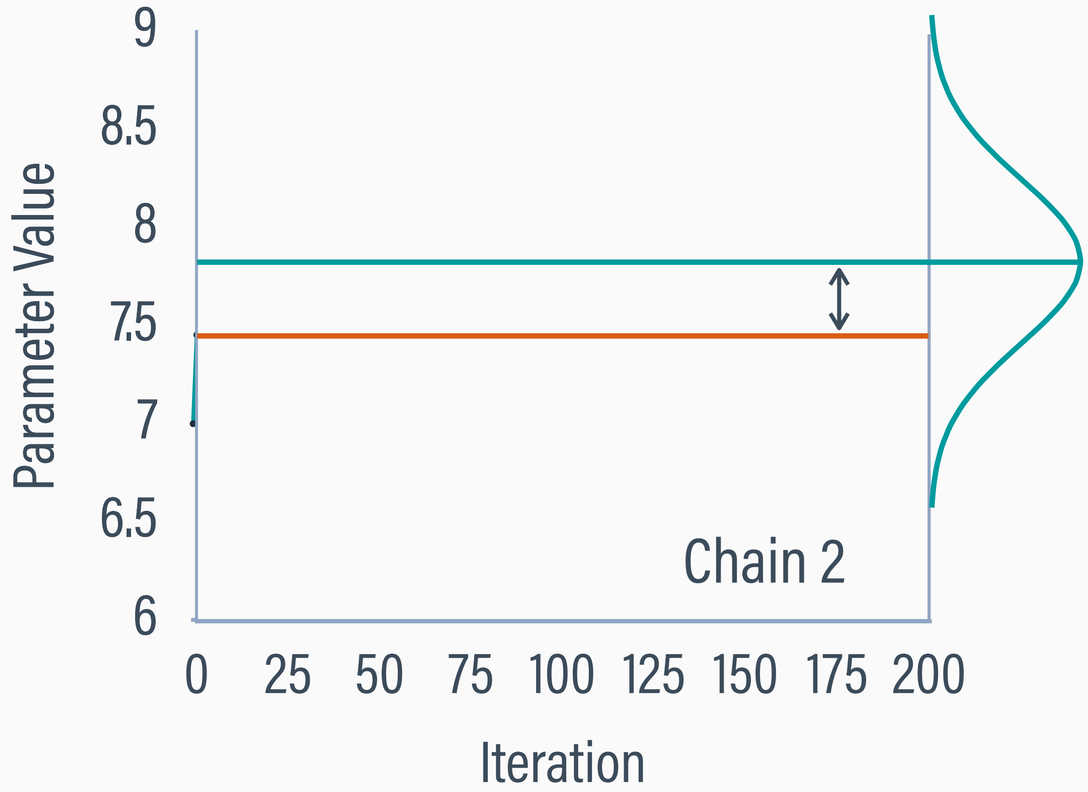


BETWEEN-CHAIN MEAN DIFFERENCE

MCMC converges when mean estimates from two chains are the same (PSRF ≈ 1)

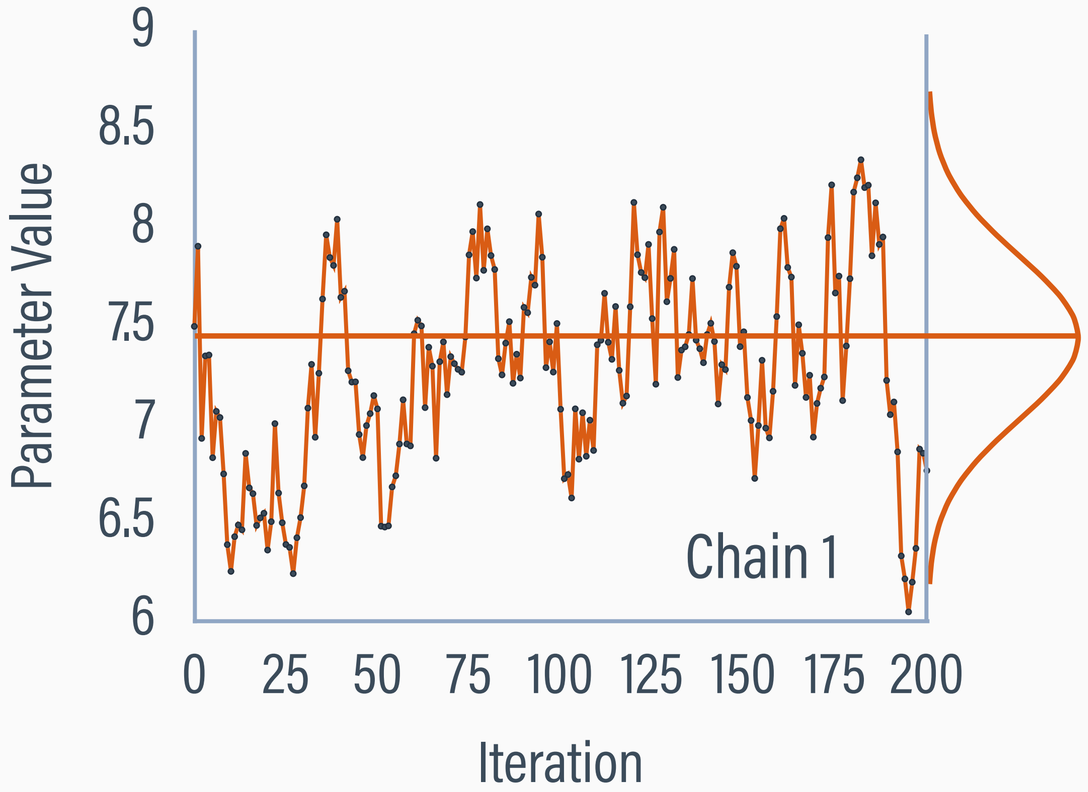
$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains} + \text{within-chain variation}}{\text{within-chain variation}}}$$

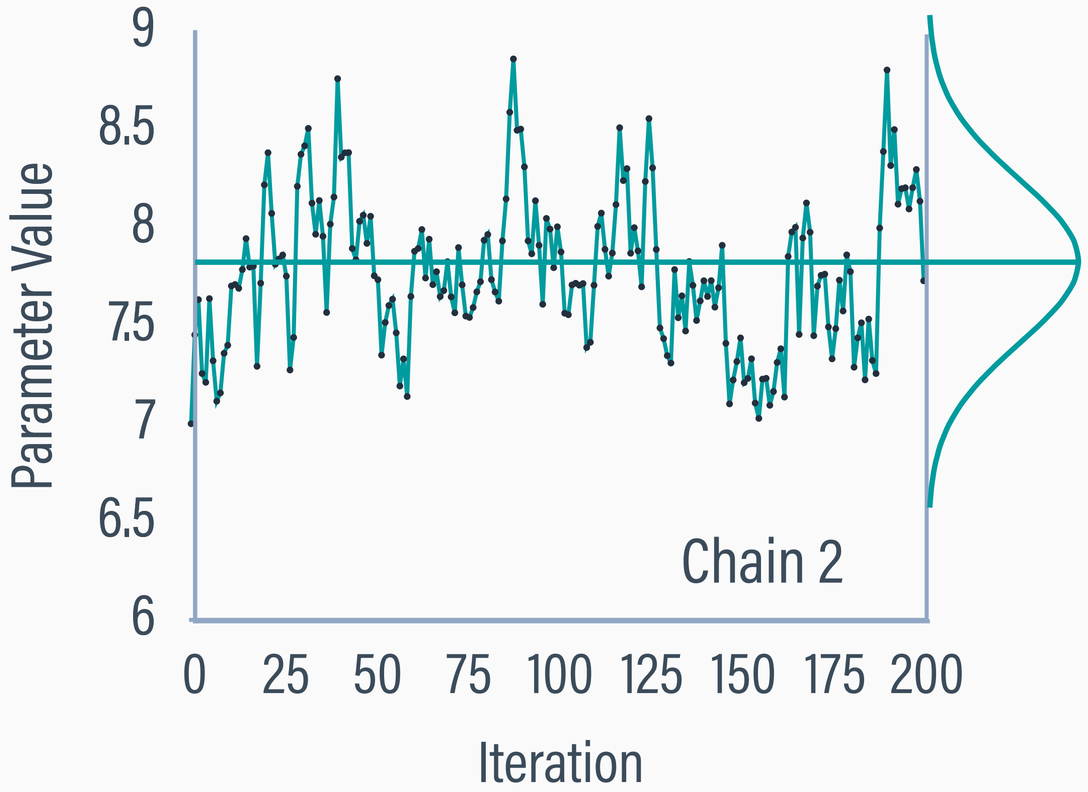




WITHIN-CHAIN VARIATION

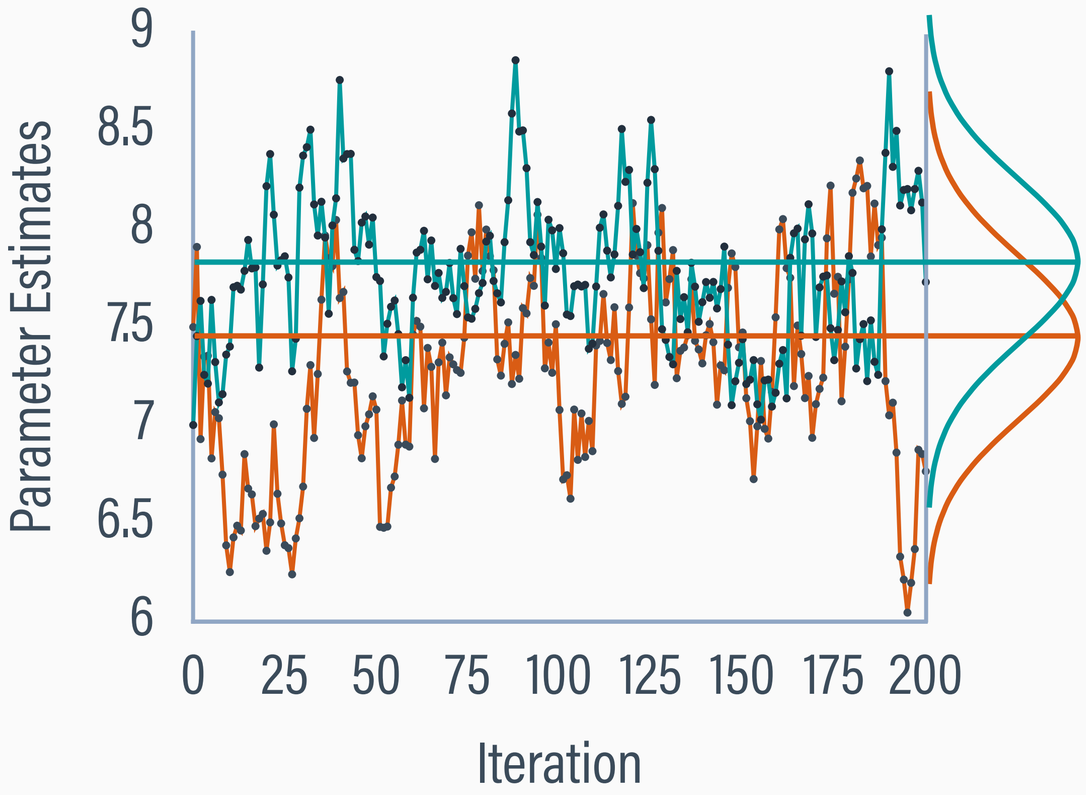
$$\text{PSRF} = \sqrt{\frac{\text{mean difference between chains + within-chain variation}}{\text{within-chain variation}}}$$



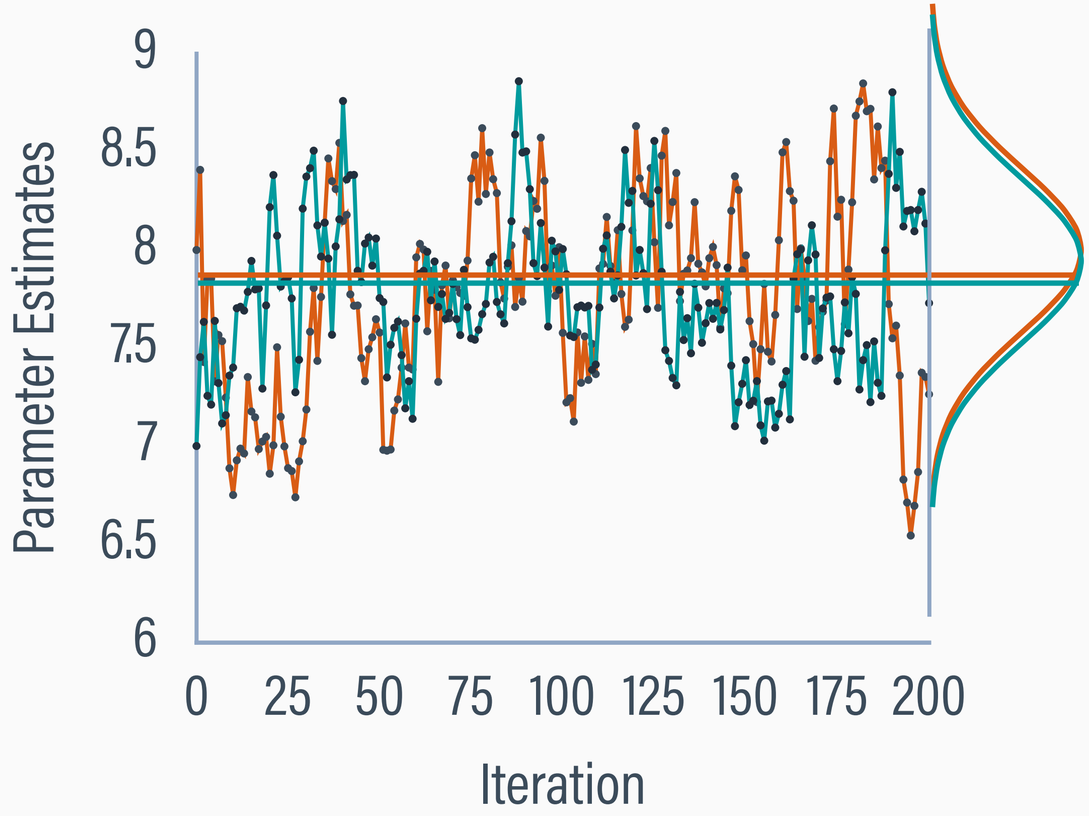


CONVERGENCE

MCMC has not converged because between-chain mean difference is large (PSR > 1.05)



MCMC has converged because between-chain mean difference is very small (PSR < 1.05)

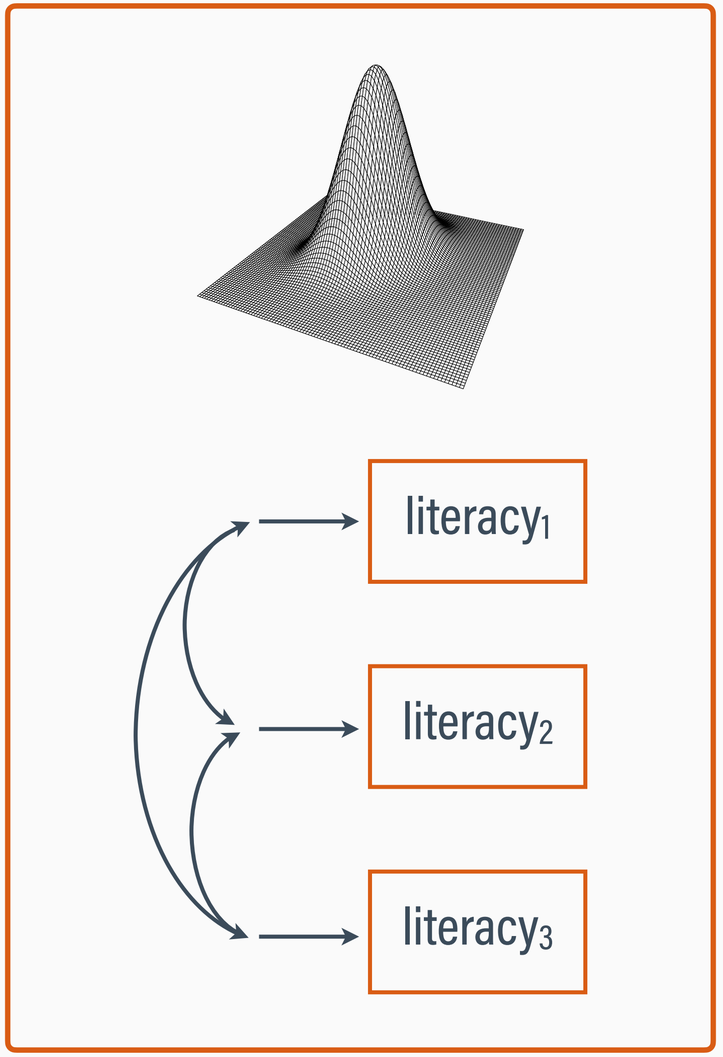


AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

CARS ANALYSIS EXAMPLE

- CARS wants to assess whether information literacy changes over time
- Descriptive statistics are obtained by specifying a saturated model with all possible means, variances, and covariances
- Blimp's MCMC estimator adopts a multivariate model for focal variables and univariate submodes for sequential auxiliary variables



RBLIMP SCRIPT

```
mymodel <- rblimp(
    data = carsdat,
    ordinal = 'male',  # define binary or ordinal variables
    model = '
      # all possible correlations
      info_t1 info_t2 info_t3 ~~ info_t1 info_t2 info_t3;
      # sequential regression models for auxiliary variables
      extra_t3 cont_t3 om_t3 ag_t3 ne_t3 male ~ info_t1 info_t2 info_t3',
    seed = 90291,  # integer random number seed
    burn = 5000,  # number of warm-up iterations
    iter = 10000) # number of analysis iterations
output(mymodel)   # view output
```

RBLIMP OUTPUT

Outcome Variable: **info_t1**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	226.960	4.864	217.814	236.646	---	---	4878.289
Coefficients:							
Intercept	67.059	0.219	66.627	67.494	93730.685	0.000	3734.402
Proportion Variance Explained							
by Coefficients	0.000	0.000	0.000	0.000	---	---	nan
by Residual Variation	1.000	0.000	1.000	1.000	---	---	nan

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t2**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	369.317	8.807	352.737	387.329	---	---	3445.387
Coefficients:							
Intercept	63.794	0.309	63.179	64.394	42589.707	0.000	2671.121
Proportion Variance Explained							
by Coefficients	0.000	0.000	0.000	0.000	---	---	nan
by Residual Variation	1.000	0.000	1.000	1.000	---	---	nan

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t3**

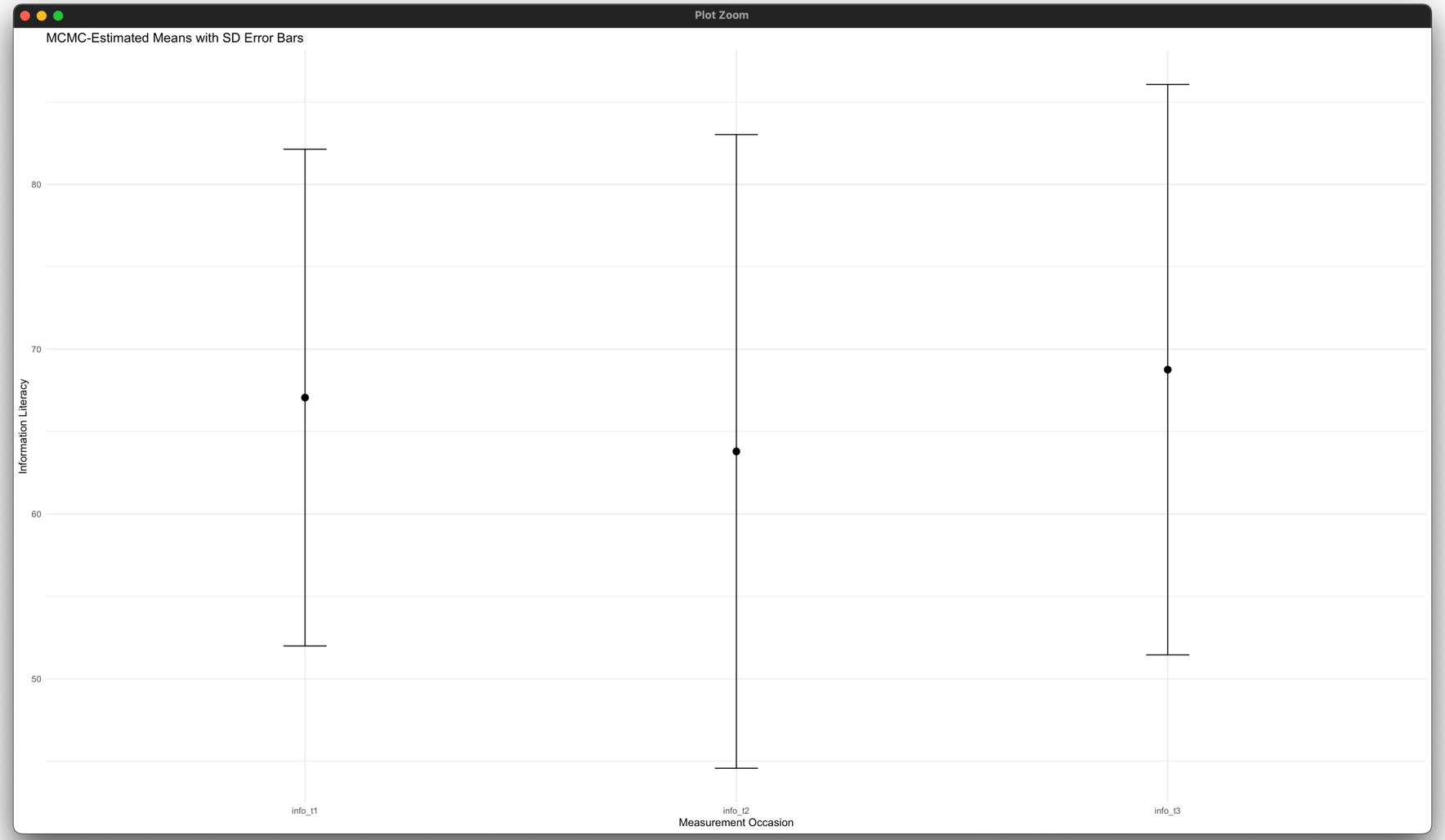
Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	299.273	6.430	287.060	312.294	---	---	4579.102
Coefficients:							
Intercept	68.752	0.253	68.249	69.238	73828.355	0.000	3195.129
Proportion Variance Explained							
by Coefficients	0.000	0.000	0.000	0.000	---	---	nan
by Residual Variation	1.000	0.000	1.000	1.000	---	---	nan

RBLIMP OUTPUT, CONTINUED

Covariance Matrix: info_t1 info_t2 info_t3

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Covariances:							
Cov(info_t1,info_t2)	124.845	5.940	113.184	136.793	441.996	0.000	2115.487
Cov(info_t1,info_t3)	105.674	4.775	96.490	115.207	490.340	0.000	2691.648
Cov(info_t2,info_t3)	160.611	6.542	147.933	173.700	602.934	0.000	2375.172
Correlations:							
Cor(info_t1,info_t2)	0.431	0.017	0.397	0.464	650.657	0.000	1750.850
Cor(info_t1,info_t3)	0.406	0.015	0.375	0.436	708.698	0.000	2417.405
Cor(info_t2,info_t3)	0.483	0.015	0.453	0.513	1016.530	0.000	1929.971

MEANS WITH STD. DEV. ERROR BARS

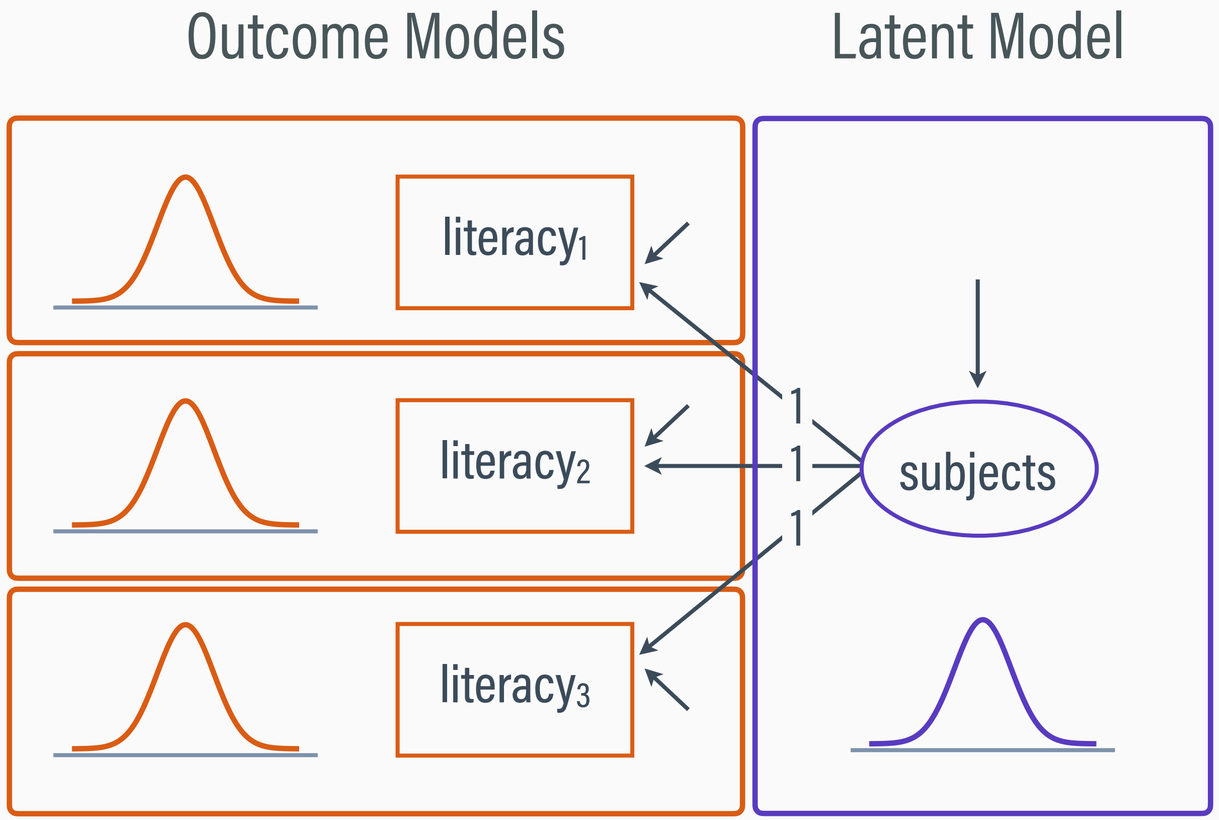


INTERPRETATIONS

- Means exhibit a nonlinear pattern with a decrease at T2
- Standard deviations are unequal with greater variation at T2
- Estimates assume a CMAR process that depends on the observed repeated measures data and auxiliary variables

REPEATED MEASURES ANALYSIS

- MCMC invokes a univariate normal distribution for each variable
- Blimp's MCMC estimator adopts a univariate submodel for every variable
- Distributional assumptions enter on a model-by-model basis



RBLIMP SCRIPT

```
mymodel <- rblimp(
    data = carsdat,
    ordinal = 'male',  # define binary or ordinal variables
    latent = 'subjects',  # define latent variable for subjects factor
    model = '
      subjects ~ intercept@0;  # random subjects factor with mean = 0 and loadings = 1
      info_t1 ~ intercept@mu1 subjects@1;  # @ labels the means and fixes subject factor loadings = 1
      info_t2 ~ intercept@mu2 subjects@1;
      info_t3 ~ intercept@mu3 subjects@1;
      info_t1 ~~ info_t1@res;  # @ sets equal residual variances for compound symmetry assumption
      info_t2 ~~ info_t2@res;
      info_t3 ~~ info_t3@res;
      extra_t3 cont_t3 om_t3 ag_t3 ne_t3 male ~ info_t1 info_t2 info_t3',  # sequential auxiliaries
    waldtest = c( 'mu1 = mu2; mu2 = mu3', 'mu1 = mu2','mu1 = mu3','mu2 = mu3'),  # significance tests
    seed = 90291,   # integer random number seed
    burn = 5000,   # number of warm-up iterations
    iter = 10000)   # number of analysis iterations
output(mymodel)   # view output
```

RBLIMP OUTPUT

Latent Variable: **subjects**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	123.561	4.112	115.359	131.540	---	---	1047.857
Proportion Variance Explained							
by Coefficients	0.000	0.000	0.000	0.000	---	---	nan
by Residual Variation	1.000	0.000	1.000	1.000	---	---	nan

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t1**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	167.366	3.045	161.584	173.545	---	---	1281.032
Coefficients:							
Intercept	67.046	0.249	66.550	67.528	72212.585	0.000	2373.977
subjects	@ 1.000	---	---	---	---	---	---
Standardized Coefficients:							
subjects	0.652	0.008	0.636	0.666	7364.361	0.000	887.714
Proportion Variance Explained							
by Coefficients	0.425	0.010	0.405	0.443	---	---	887.884
by Residual Variation	0.575	0.012	0.554	0.599	---	---	1231.701

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t2**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	167.366	3.045	161.584	173.545	---	---	1281.032
Coefficients:							
Intercept	63.945	0.271	63.400	64.457	55476.123	0.000	1914.174
subjects	@ 1.000	---	---	---	---	---	---
Standardized Coefficients:							
subjects	0.652	0.008	0.636	0.666	7364.361	0.000	887.714
Proportion Variance Explained							
by Coefficients	0.425	0.010	0.405	0.443	---	---	887.884
by Residual Variation	0.575	0.012	0.554	0.599	---	---	1231.701

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t3**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	167.366	3.045	161.584	173.545	---	---	1281.032
Coefficients:							
Intercept	68.752	0.250	68.264	69.243	75631.476	0.000	2257.328
subjects	@ 1.000	---	---	---	---	---	---
Standardized Coefficients:							
subjects	0.652	0.008	0.636	0.666	7364.361	0.000	887.714
Proportion Variance Explained							
by Coefficients	0.425	0.010	0.405	0.443	---	---	887.884
by Residual Variation	0.575	0.012	0.554	0.599	---	---	1231.701

WALD SIGNIFICANCE TEST OUTPUT

```
MODEL FIT:

...

  WALD TESTS (Asparouhov & Muthén, 2021)

  Test #1

    Full:
      [1]  info_t1 ~ Intercept@mu1 info@1
      [2]  info_t2 ~ Intercept@mu2 info@1
      [3]  info_t3 ~ Intercept@mu3 info@1

    Restricted:
      [1]  info_t1 ~ Intercept@mu1 info@1
      [2]  info_t2 ~ Intercept@mu2 info@1
      [3]  info_t3 ~ Intercept@mu3 info@1

    Constraints in Restricted:
      [1]  mu1 = mu2
      [2]  mu2 = mu3

    Wald Statistic (Chi-Square)                228.009
    Number of Parameters Tested (df)                 2
    Probability                                  0.000
```

Omnibus null hypothesis that all means are equal

SIGNIFICANCE TEST OUTPUT, CONT.

```
 Test #2

...
  Constraints in Restricted:
    [1]  mu1 = mu2

    Wald Statistic (Chi-Square)                94.617
    Number of Parameters Tested (df)                1
    Probability                                 0.000

  Test #3

...
  Constraints in Restricted:
    [1]  mu1 = mu3

    Wald Statistic (Chi-Square)                33.071
    Number of Parameters Tested (df)                1
    Probability                                 0.000

   Test #4

...
  Constraints in Restricted:
    [1]  mu2 = mu3

    Wald Statistic (Chi-Square)               225.931
    Number of Parameters Tested (df)                1
    Probability                                 0.000
```

Pairwise comparison

Pairwise comparison

Pairwise comparison

AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

BLIMP VARIABLE TYPES

Exogenous Predictors

Normal (Manifest) | Binary | Ordinal | Nominal

Univariate Outcomes

Normal (Manifest or Latent) | Binary | Ordinal | Nominal | Skewed (Manifest or Latent) | Count | Two-Part (Floor Effects)

Multivariate Outcomes

Normal (Manifest or Latent) | Binary | Ordinal | Skewed (Manifest or Latent)

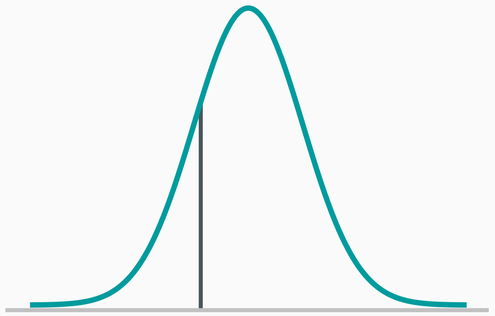
LATENT RESPONSE FRAMEWORK

- MCMC methods for discrete data use an underlying latent response variable that represents a continuous propensity for the construct being measured
- An inverse link function provides a rule for converting the continuous propensities to the discrete metric
- Working with latent responses is essentially a computational trick that allows simpler estimation routines for linear models

LATENT RESPONSE FORMULATION

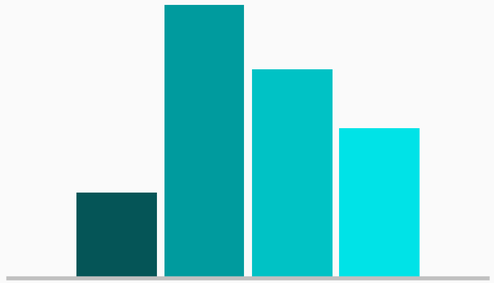
Binary

Discrete Response

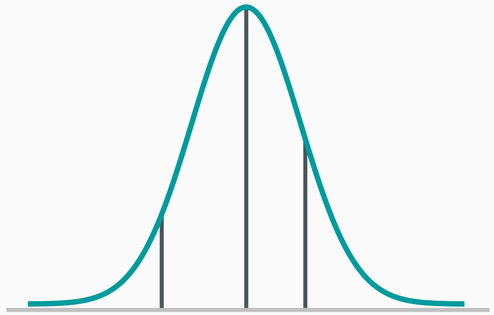


Latent Response

Ordinal

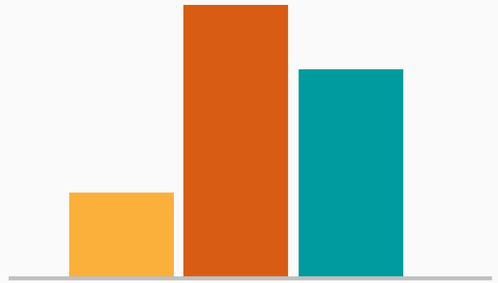


Discrete Response

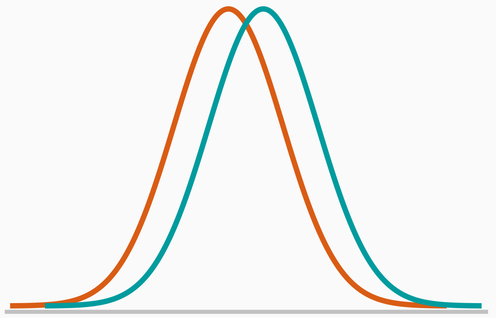


Latent Response

Multicategorical



Discrete Response



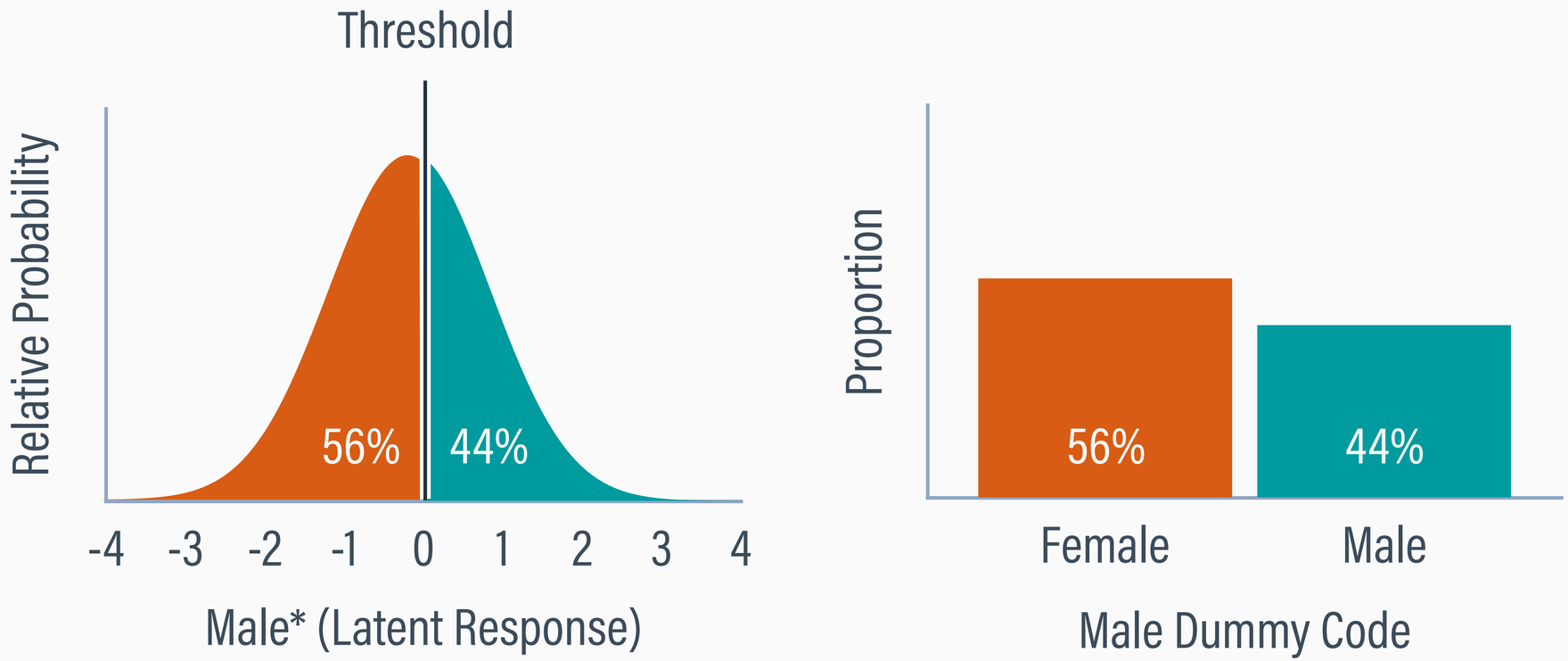
Latent Response

INCOMPLETE GENDER VARIABLE

- Probit regression envisions binary and ordinal variables arising from an underlying normal latent response variable
- Applied to the incomplete gender, the latent variable represents an unobserved, continuous propensity for being male
- A threshold carves the latent distribution into segments that represent the male and female probabilities

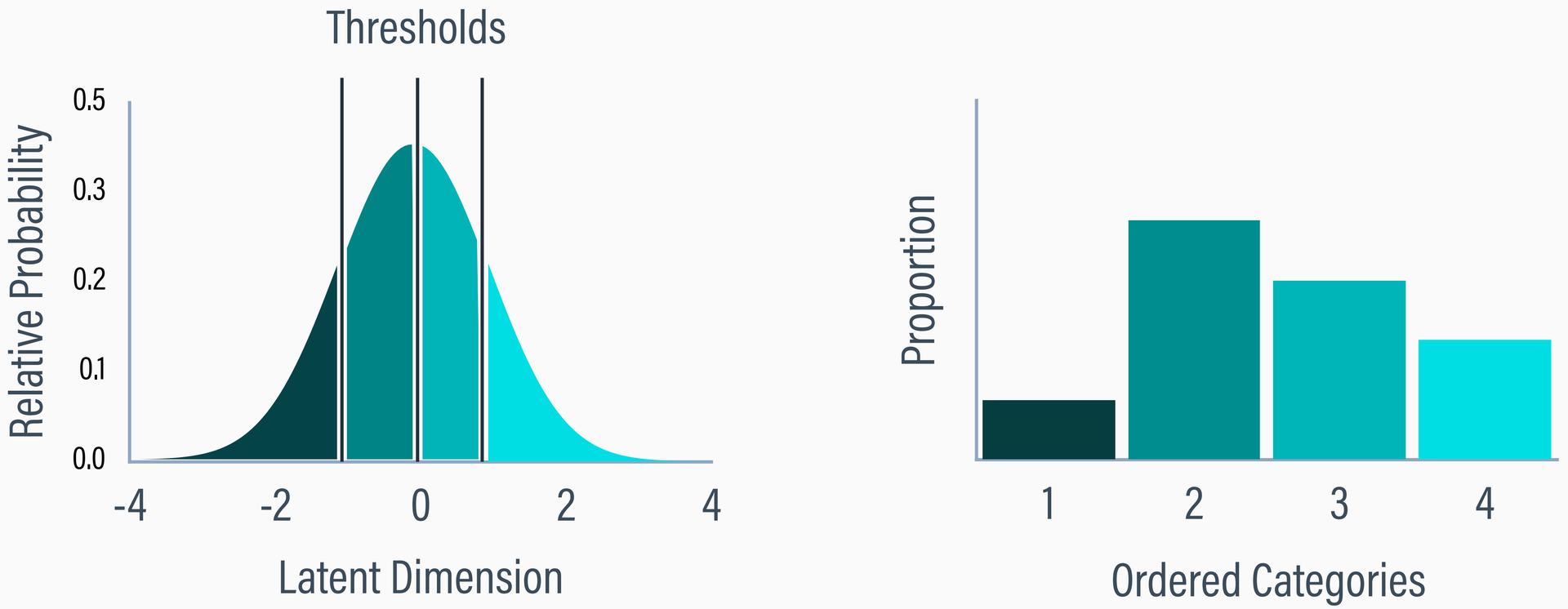
LATENT AND DISCRETE DISTRIBUTIONS

- The threshold parameter divides the latent distribution into segments, with areas under the curve matching the bar plot probabilities



ORDINAL VARIABLES

- Multiple threshold parameters divide the latent distribution into segments, with areas under the curve matching the bar plot

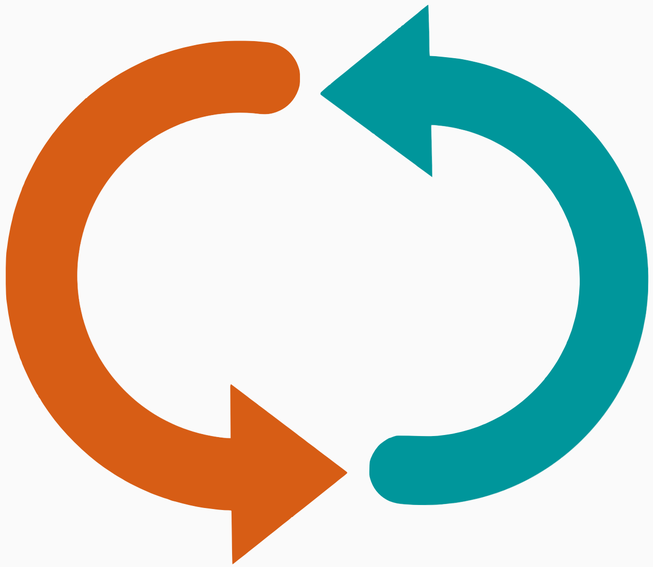


IMPUTING LATENT RESPONSE SCORES

- Latent response scores are 100% missing data that need to be imputed for the entire sample
- MCMC uses computer simulation to “sample” latent response scores from a normal curve, just like any other incomplete variable (imputation = prediction + noise)
- Whether the latent scores are above or below the threshold determines whether the discrete impute equals 0 or 1

MCMC ESTIMATION

Estimate model parameters



Impute missing values

Do for t = 1 to 10,000 iterations

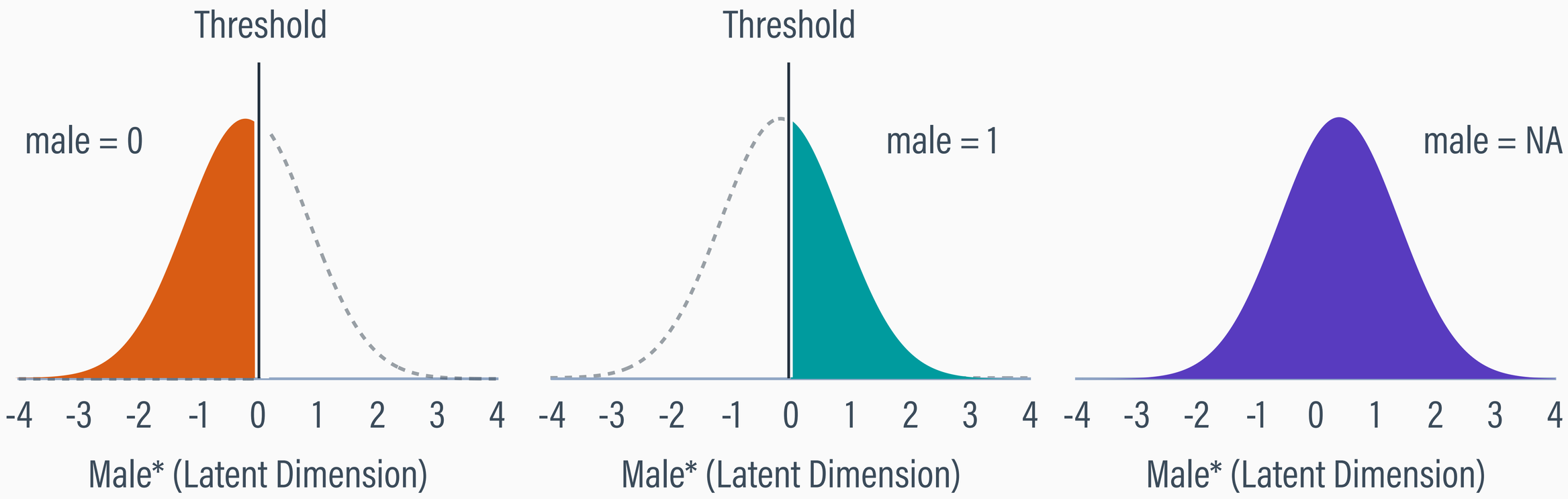
- Estimate model parameters, conditional on the latent and manifest data
- Impute missing values and latent response scores, conditional on the model parameters
- Assign latent imputes to categories

Repeat

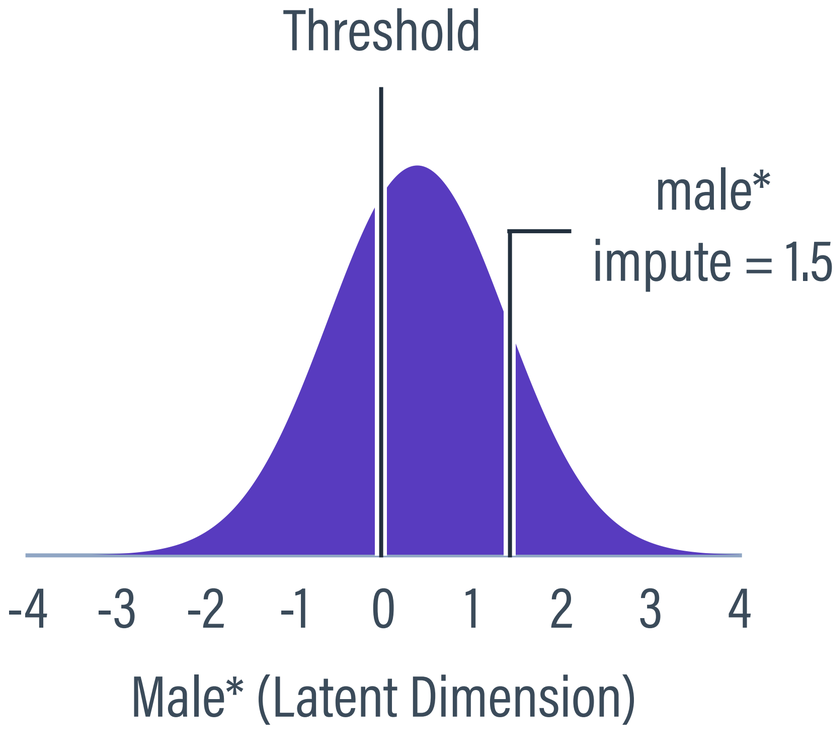
Summarize model parameters

LATENT AND DISCRETE DISTRIBUTIONS

- Latent imputations must fall below or above threshold if the binary variable is observed, and they are unconstrained if missing



Suppose the latent response imputation for a student with a missing gender score was 1.5. What gender group would the probit model assign to this person?



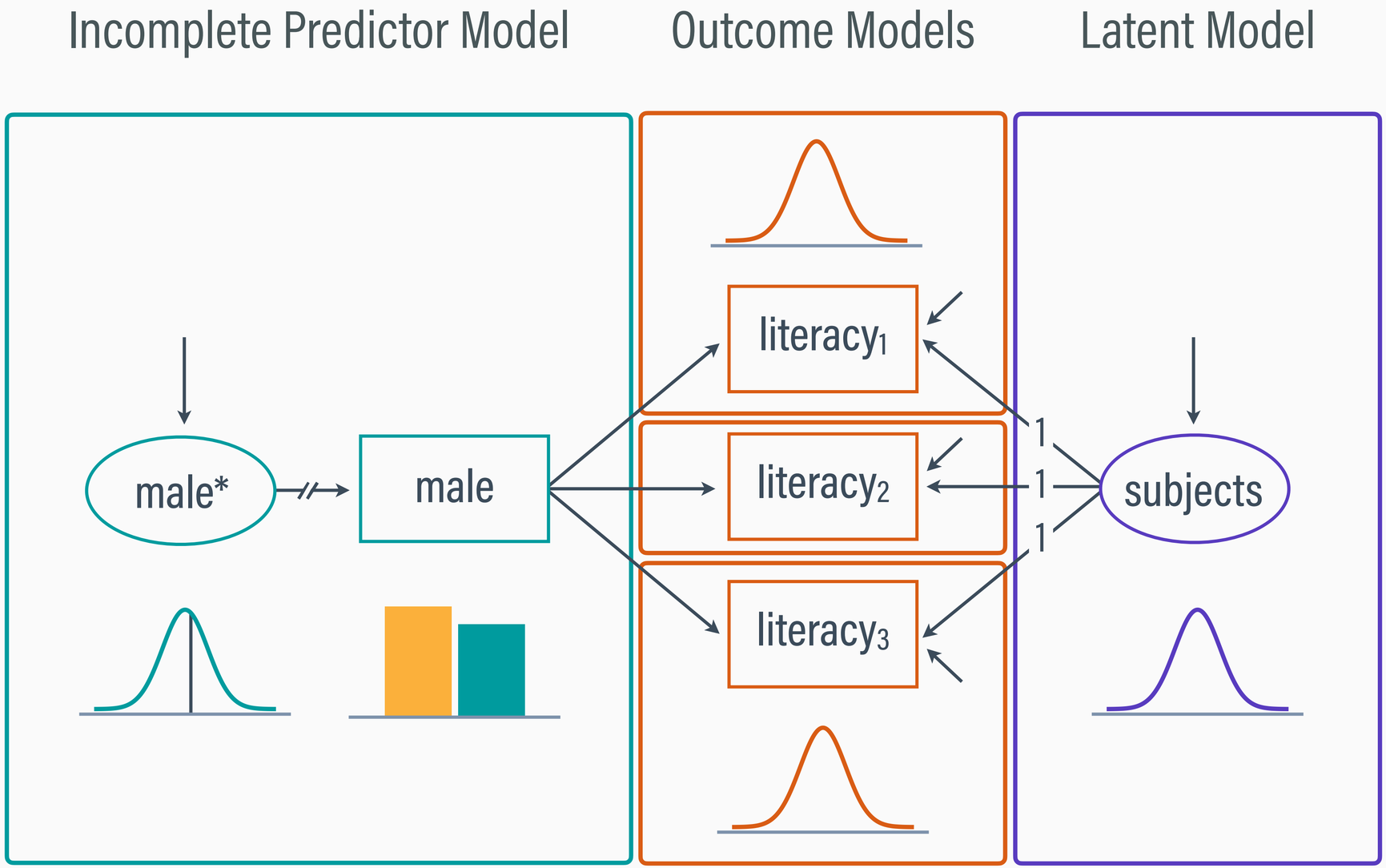
AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

CARS ANALYSIS EXAMPLE

- CARS wants to assess whether information literacy changes over time and whether that changes differs by gender
- Incomplete predictors require distributional assumptions for missing data handling
- MCMC readily accommodates incomplete variables with different metrics

FACTORED REGRESSION SPECIFICATION



RBLIMP REPEATED MEASURES SCRIPT

```
mymodel <- rblimp(
   data = carsdat,
   ordinal = 'male',  # define binary or ordinal variables
   latent = 'subjects',  # define latent variable for subjects factor
   model = '
     subjects ~ intercept@0;  # random subjects factor with mean = 0 and loadings = 1
     info_t1 ~ intercept@mu1 subjects@1 male@dif1; # @ labels means and differences and fixes loadings to 1
     info_t2 ~ intercept@mu2 subjects@1 male@dif2;
     info_t3 ~ intercept@mu3 subjects@1 male@dif3;
     info_t1 ~~ info_t1@res;  # @ sets equal residual variances for compound symmetry assumption
     info_t2 ~~ info_t2@res;
     info_t3 ~~ info_t3@res;
     extra_t3 cont_t3 om_t3 ag_t3 ne_t3 male ~ info_t1 info_t2 info_t3',  # sequential auxiliaries
   waldtest = 'dif1 = dif2; dif2 = dif3',  # wald test that group-by-time interaction = 0
   parameters = ' # define group means
     fem_mu1 = mu1;
     fem_mu2 = mu2;
     fem_mu3 = mu3;
     male_mu1 = mu1 + dif1;
     male_mu2 = mu2 + dif2;
     male_mu3 = mu3 + dif3;',
   seed = 90291,   # integer random number seed
   burn = 5000,   # number of warm-up iterations
   iter = 10000)   # number of analysis iterations
output(mymodel)   # view output
```

RBLIMP OUTPUT

Outcome Variable: **info_t1**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	166.571	3.161	160.439	172.945	---	---	1112.374
Coefficients:							
Intercept	67.954	0.322	67.322	68.598	44445.588	0.000	2157.623
subjects	@ 1.000	---	---	---	---	---	---
male	-2.038	0.499	-3.024	-1.062	16.695	0.000	2062.233
Standardized Coefficients:							
subjects	0.646	0.008	0.630	0.662	6202.366	0.000	778.239
male	-0.060	0.015	-0.088	-0.031	16.762	0.000	2054.457
Proportion Variance Explained							
by Coefficients	0.421	0.011	0.400	0.441	---	---	784.686
by Residual Variation	0.579	0.012	0.555	0.603	---	---	1052.147

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t2**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	166.571	3.161	160.439	172.945	---	---	1112.374
Coefficients:							
Intercept	66.233	0.362	65.513	66.938	33523.741	0.000	1569.466
subjects	@ 1.000	---	---	---	---	---	---
male	-5.303	0.562	-6.415	-4.220	89.203	0.000	1587.404
Standardized Coefficients:							
subjects	0.640	0.008	0.623	0.655	6040.640	0.000	815.201
male	-0.154	0.016	-0.185	-0.123	92.286	0.000	1581.583
Proportion Variance Explained							
by Coefficients	0.433	0.011	0.412	0.453	---	---	777.480
by Residual Variation	0.567	0.012	0.543	0.592	---	---	1030.909

RBLIMP OUTPUT, CONTINUED

Outcome Variable: **info_t3**

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	166.571	3.161	160.439	172.945	---	---	1112.374
Coefficients:							
Intercept	70.894	0.326	70.255	71.530	47302.378	0.000	2123.967
subjects	@ 1.000	---	---	---	---	---	---
male	-4.914	0.502	-5.892	-3.921	95.720	0.000	1979.720
Standardized Coefficients:							
subjects	0.641	0.008	0.624	0.656	6143.042	0.000	795.942
male	-0.142	0.014	-0.170	-0.114	98.396	0.000	1971.855
Proportion Variance Explained							
by Coefficients	0.431	0.011	0.410	0.451	---	---	791.496
by Residual Variation	0.569	0.012	0.546	0.594	---	---	1048.678

GROUP MEAN OUTPUT

GENERATED PARAMETERS:

Summaries based on 10000 iterations using 2 chains.
NOTE: Estimate column based on posterior median.

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
fem_mu1	67.954	0.322	67.322	68.598	44445.588	0.000	2157.623
fem_mu2	66.233	0.362	65.513	66.938	33523.741	0.000	1569.466
fem_mu3	70.894	0.326	70.255	71.530	47302.378	0.000	2123.967
male_mu1	65.918	0.385	65.162	66.672	29283.751	0.000	1986.912
male_mu2	60.924	0.424	60.094	61.763	20657.140	0.000	1547.318
male_mu3	65.984	0.386	65.225	66.747	29225.214	0.000	1752.829

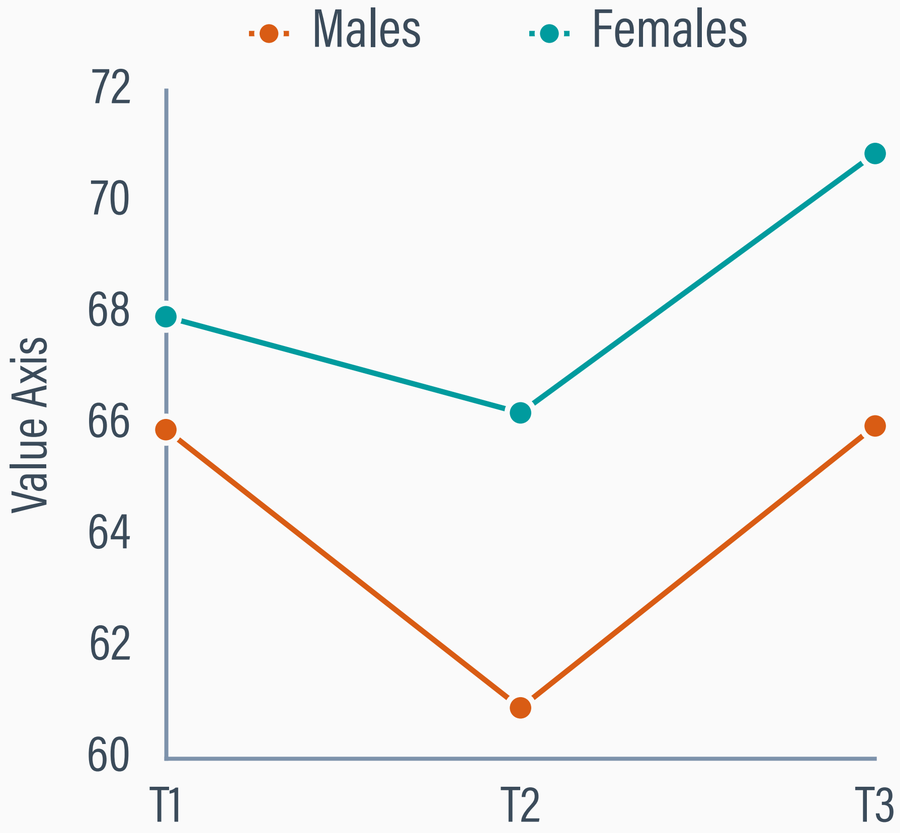
WALD SIGNIFICANCE TEST OUTPUT

```
MODEL FIT:


  INFORMATION CRITERIA

    Conditional Likelihood
      DIC2                    278349.136
      WAIC                    288741.492

  WALD TESTS (Asparouhov & Muthén, 2021)

  Test #1

...

    Constraints in Restricted:
      [1]  dif1 = dif2
      [2]  dif2 = dif3

    Wald Statistic (Chi-Square)                33.123
    Number of Parameters Tested (df)                2
    Probability                                 0.000
```

Null hypothesis that group-by-time interaction = 0 (no time-specific mean differences)

INTERPRETATIONS

- The group-by-time interaction was significant ($\chi^2 = 33.12$, $p < .001$)
- Females decreased at T2 then rebounded to a higher mean at T3
- Males decreased by a larger amount at T2 (about five points versus less than two), and their T3 mean is the same as T1



AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

CARS ANALYSIS EXAMPLE

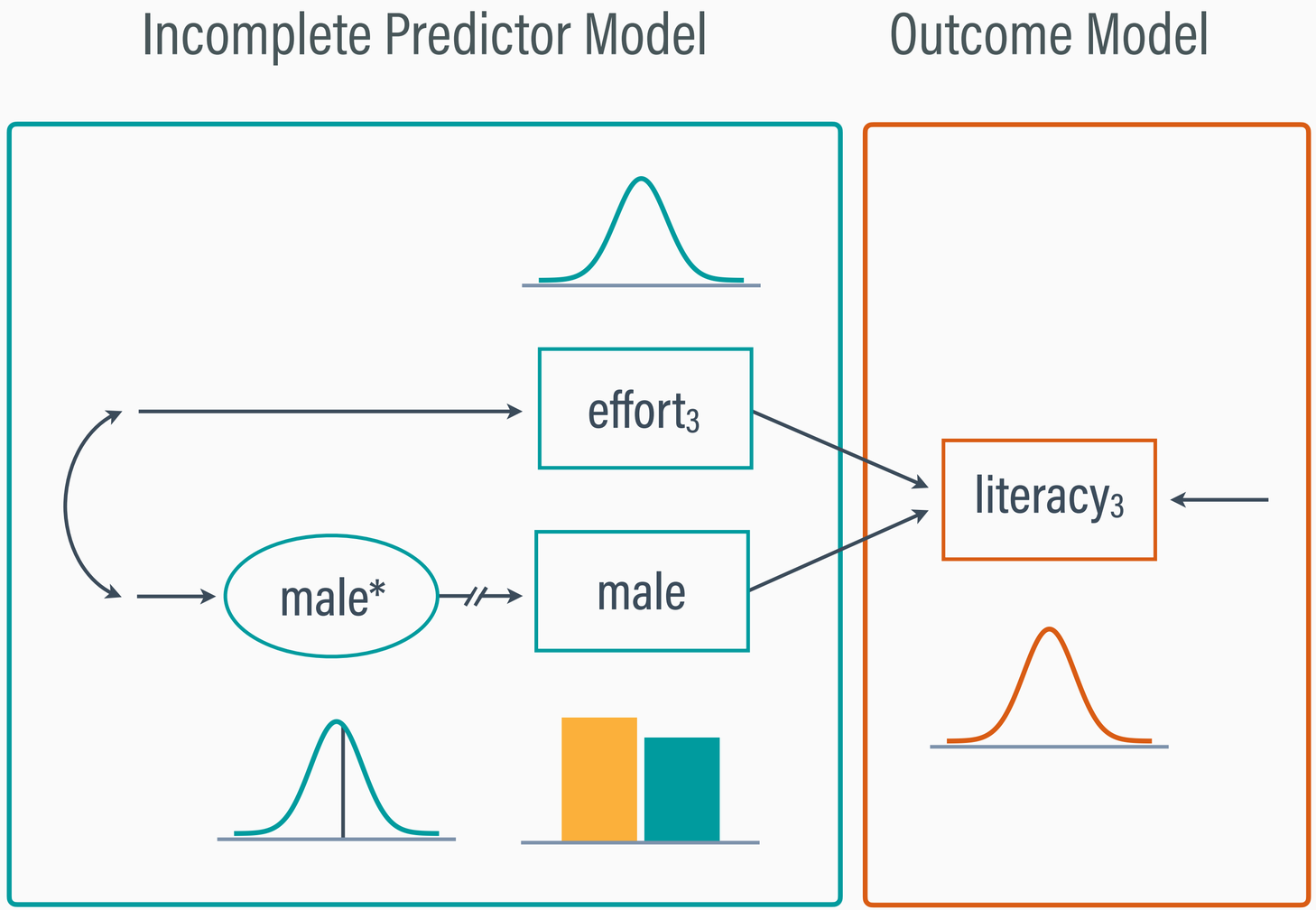
- Are there gender differences in T3 information literacy, controlling for T3 effort?

$$literacy_3 = \beta_0 + \beta_1(effort_3) + \beta_2(male) + \varepsilon$$

- Maximum likelihood estimation in lavaan treats the discrete predictor as normally distributed (conceptually, this “imputes” decimal values instead of 0s and 1s)

- MCMC allows mixtures of categorical and continuous variables

FACTORED REGRESSION SPECIFICATION



RBLIMP SCRIPT

```
mymodel <- rblimp(
    data = carsdat,
    ordinal = 'male admit_type',  # define binary or ordinal variables
    center = 'effort_t3',  # iterative grand mean centering
    model = '
      # focal model
      info_t3 ~ effort_t3 male;
       # sequential regression models for auxiliary variables
      extra_t3 cont_t3 om_t3 ag_t3 ne_t3 admit_type ~ info_t3 effort_t3 male',
    seed = 90291,  # integer random number seed
    burn = 5000,  # number of warm-up iterations
    iter = 10000)  # number of analysis iterations
output(mymodel)  # view output
```

RBLIMP OUTPUT

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	241.488	5.176	231.542	251.849	---	---	4643.665
Coefficients:							
Intercept	70.061	0.330	69.413	70.705	45175.243	0.000	2561.768
effort_t3	10.010	0.338	9.341	10.679	878.118	0.000	4431.186
male	-2.511	0.488	-3.453	-1.541	26.349	0.000	4143.186
Standardized Coefficients:							
effort_t3	0.413	0.012	0.388	0.436	1128.004	0.000	4108.547
male	-0.072	0.014	-0.099	-0.044	26.567	0.000	4155.379
Proportion Variance Explained							
by Coefficients	0.184	0.010	0.164	0.204	---	---	3978.725
by Residual Variation	0.816	0.010	0.796	0.836	---	---	3978.725

INTERPRETATIONS

- The female literacy mean was $\beta_0 = 70.06$
- For two students with the same gender, scoring one point higher on the effort measure was associated with a $\beta_1 = 10.01$ increase in information literacy
- For two students with the same effort score, males scored $\beta_2 = -2.51$ points lower than females

AFTERNOON OUTLINE

1. MCMC Estimation
2. Analysis Example 1: Descriptive Statistics and Repeated Measures
3. MCMC With Categorical Variables
4. Analysis Example 2: Repeated Measures With Between-Subjects Predictor
5. Analysis Example 3: Multiple Regression
6. Analysis Example 4: Moderated Regression

MODERATION

- Moderation occurs when a focal predictor's influence on the outcome depends on a third (moderator) variable
- For whom does an effect apply?
- We model moderation effects by including the product of two predictors in the regression model ($Y = X + M + X \times M$)

CARS ANALYSIS EXAMPLE

- Does the influence of effort on literacy vary by gender?

$$literacy_3 = \beta_0 + \beta_1(effort_3) + \beta_2(male) + \beta_3(effort_3)(male) + \varepsilon$$

- β_1 and β_2 are conditional effects: β_1 is the influence of the effort when gender (the moderator) equals zero, and β_2 is the gender difference when effort (the focal predictor) equals zero

- β_3 is the change in β_1 for a one-unit increase in the moderator

INCOMPLETE PRODUCT TERMS

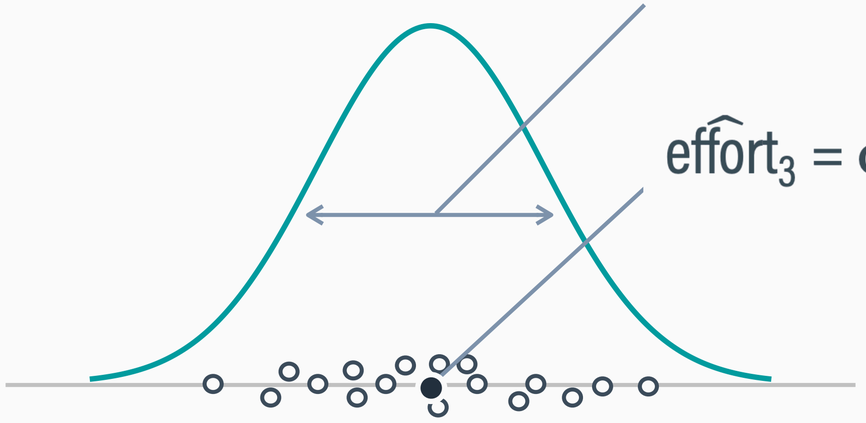
- Incomplete products should not be treated as unique variables (the “just another variable” method), as this causes bias
- Rather, product terms should be viewed as deterministic functions of the incomplete predictors
- When the interaction is non-zero, predictor imputations become non-normal (heteroscedastic) to accommodate the nonlinearity in the focal model

DISTRIBUTION OF IMPUTATIONS

- Multiple sets of model parameters define the mean and spread of the imputations, and variation is heteroscedastic (depends on gender)

$$\sigma^2_{effort_3|info_3,male} = \left(\frac{1}{\sigma^2_e} + \frac{(\beta_1 + \beta_3 male)^2}{\sigma^2_\varepsilon} \right)^{-1}$$

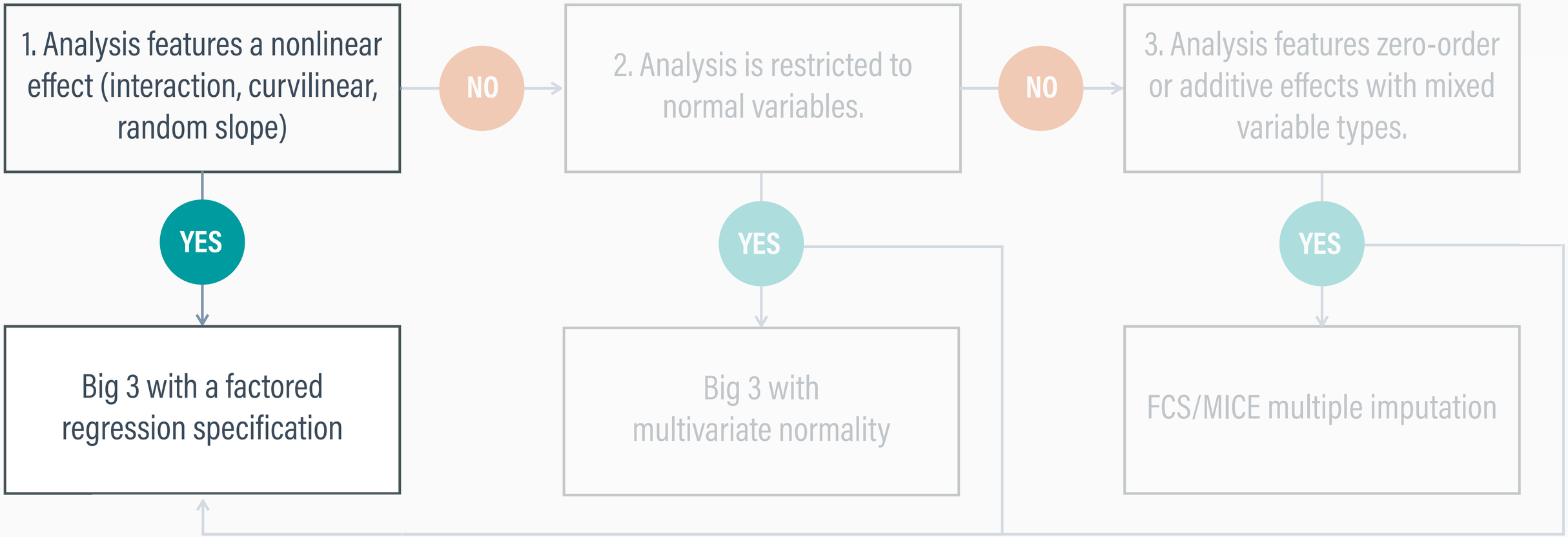
$$\widehat{effort}_3 = \sigma^2_{effort_3|info_3,male} \times \left(\frac{\gamma_0 + \gamma_1 male^*}{\sigma^2_e} + \frac{(\beta_1 + \beta_3 male)(info_3 - \beta_0 - \beta_2 male)}{\sigma^2_\varepsilon} \right)$$



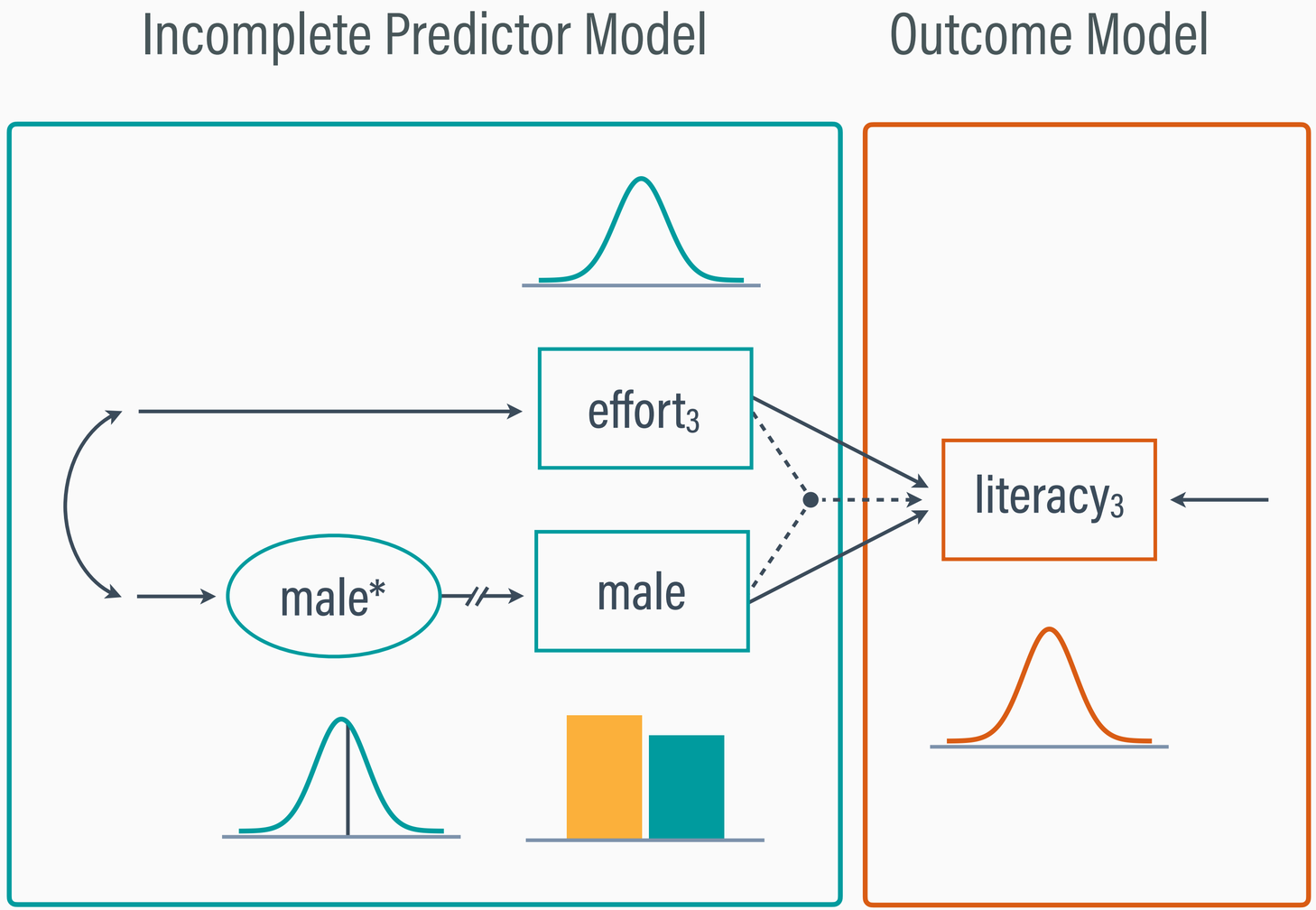
Effort Imputations

MISSING DATA DECISION TREE

Interaction and nonlinear effects require factored specifications with specialized software.



FACTORED REGRESSION SPECIFICATION



RBLIMP SCRIPT

```
mymodel <- rblimp(
    data = carsdat,
    ordinal = 'male admit_type',  # define binary or ordinal variables
    center = 'effort_t3',  # iterative grand mean centering
    model = '
      # focal model
      info_t3 ~ effort_t3 male effort_t3*male,  # the * specifies an interaction
       # sequential regression models for auxiliary variables
      extra_t3 cont_t3 om_t3 ag_t3 ne_t3 admit_type ~ info_t3 effort_t3 male',
    simple = 'effort_t3 | male',  # simple intercepts and slopes
    seed = 90291,  # integer random number seed
    burn = 5000,  # number of warm-up iterations
    iter = 10000)  # number of analysis iterations
output(mymodel)  # view output
simple_plot(info_t3 ~ effort_t3 | male, mymodel)  # plot simple intercepts slopes
```

REGRESSION SUMMARY TABLE

Parameters	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
Variances:							
Residual Var.	240.003	5.223	229.951	250.561	---	---	4670.446
Coefficients:							
Intercept	70.204	0.323	69.563	70.824	47153.662	0.000	2803.239
effort_t3	8.353	0.456	7.450	9.241	335.042	0.000	4451.948
male	-2.436	0.489	-3.393	-1.473	24.857	0.000	4488.756
effort_t3*male	3.694	0.677	2.372	5.042	29.863	0.000	4261.002
...							
Proportion Variance Explained							
by Coefficients	0.191	0.010	0.170	0.211	---	---	3691.631
by Residual Variation	0.809	0.010	0.789	0.830	---	---	3691.631

INTERPRETATIONS

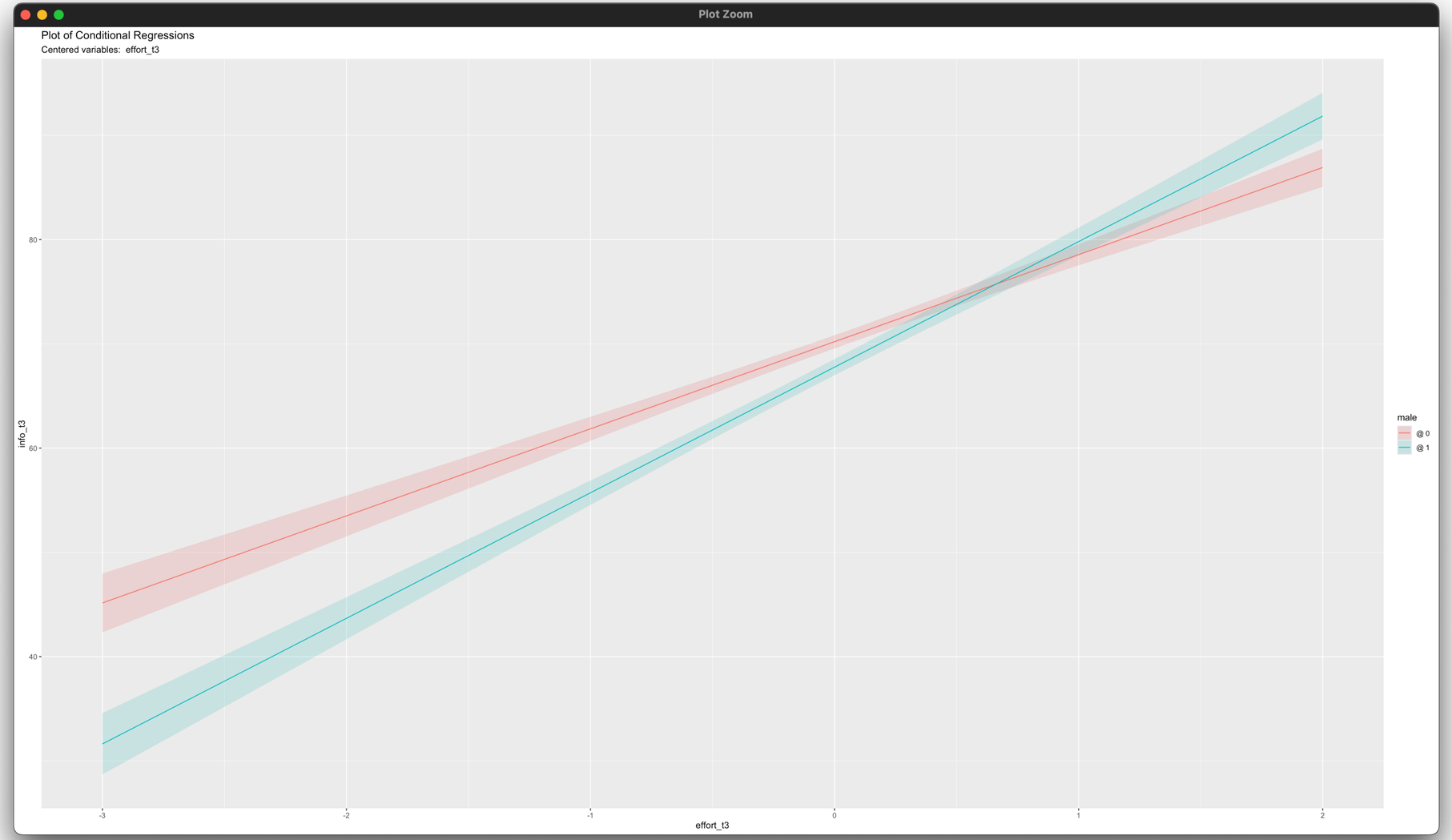
- The mean for a female with average effort was $\beta_0 = 70.20$
- For two females, scoring one point higher on the effort is associated with a $\beta_1 = 8.35$ increase in information literacy
- For two students at the mean of the effort distribution, males scored $\beta_2 = -2.44$ points lower than females
- The male regression slope was $\beta_3 = 3.69$ points higher than the female slope

CONDITIONAL EFFECTS SUMMARY TABLE

Conditional Effects	Estimate	StdDev	2.5%	97.5%	ChiSq	PValue	N_Eff
effort_t3 | male @ 0							
Intercept	70.204	0.323	69.563	70.824	47153.662	0.000	2803.239
Slope	8.353	0.456	7.450	9.241	335.042	0.000	4451.948
effort_t3 | male @ 1							
Intercept	67.765	0.392	66.989	68.526	29887.030	0.000	2383.933
Slope	12.041	0.498	11.073	13.024	584.349	0.000	3993.951

NOTE: Intercepts are computed by setting all predictors
not involved in the conditional effect to zero.

CONDITIONAL EFFECT PLOTS



For more information go to

WWW.APPLIEDMISSINGDATA.COM