Introduction to Within-Person Analysis and RM ANOVA

- Today's Class:
  - From between-person to within-person
  - ANOVAs for longitudinal data
  - Model comparisons under ML (and now REML!)

The Two Sides of a (BP) Model

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + e_i$$

Our focus today

- **Model for the Means (Predicted Values):**
  - Each person's expected (predicted) outcome is a weighted linear function of his/her values on X and Z (and here, their interaction), each measured once per person (i.e., this is a between-person model)
  - Estimated parameters are called fixed effects (here, β_0, β_1, β_2, and β_3)
  - The number of fixed effects will show up in formulas as k (so k = 4 here)

- **Model for the Variance ("Piles" of Variance):**
  - $e_i \sim N(0, \sigma_e^2)$ → ONE residual (unexplained) deviation
  - e_i has a mean of 0 with some estimated constant variance σ_e^2, is normally distributed, is unrelated to X and Z, and is unrelated across people (across all observations, just people here)
  - **Contains residual variance only in above BP model**

Review: Variances and Covariances

Variance:
Dispersion of y

$$\text{Variance}(y_t) = \frac{\sum_{i=1}^{N}(y_{ti} - \hat{y}_{ti})^2}{N - k}$$

Covariance:
How y's go together, unstandardized

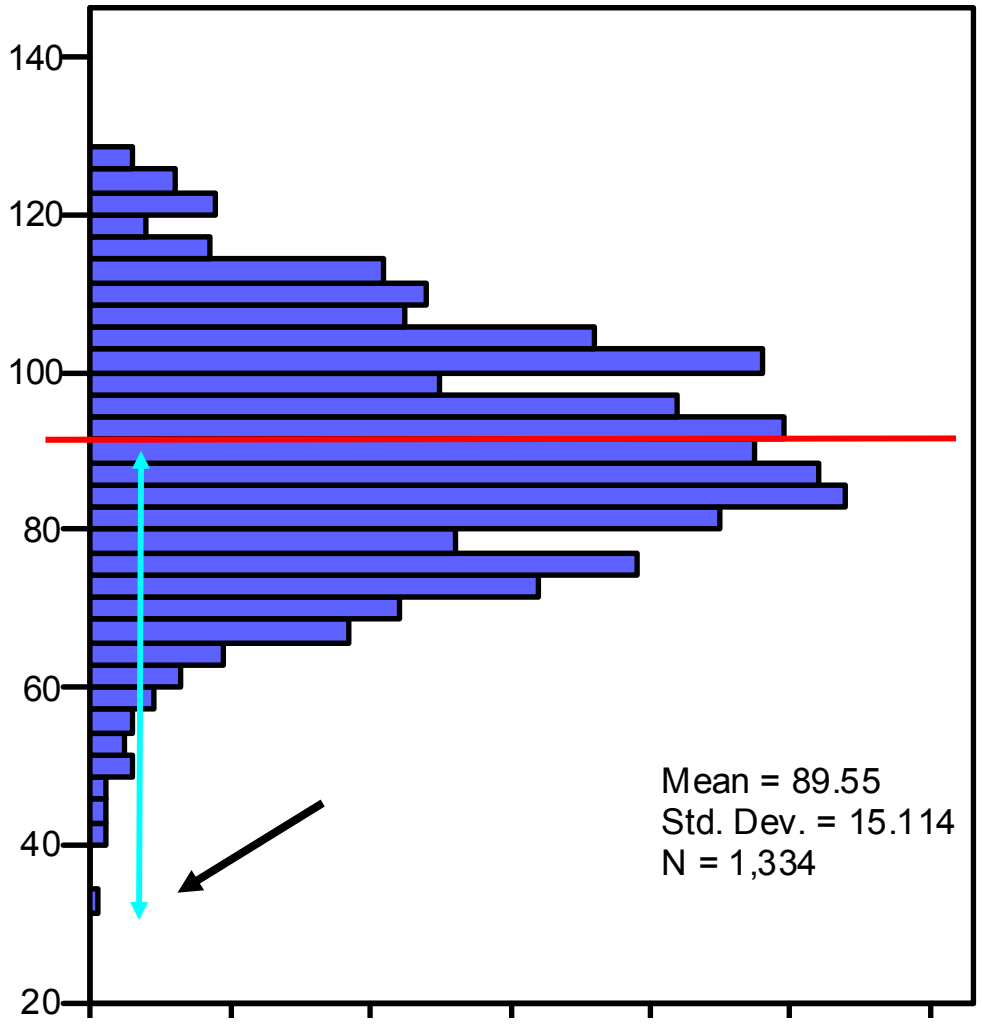
$$\text{Covariance}(y_1, y_2) = \frac{\sum_{i=1}^{N}(y_{1i} - \hat{y}_{1i})(y_{2i} - \hat{y}_{2i})}{N - k}$$

Correlation:
How y's go together, standardized (−1 to 1)

$$\text{Correlation}(y_1, y_2) = \frac{\text{Covariance}(y_1, y_2)}{\sqrt{\text{Variance}(y_1)} * \sqrt{\text{Variance}(y_2)}}$$

N = # people, t = time, i = person
k = # fixed effects, $\hat{y}_{ti}$ = y predicted from fixed effects

An Empty Between-Person Model (i.e., Single-Level)



$$y_i = \beta_0 + e_i$$

Filling in values:

$$32 = \underbrace{90}_{\text{Y pred}} + (-58)$$

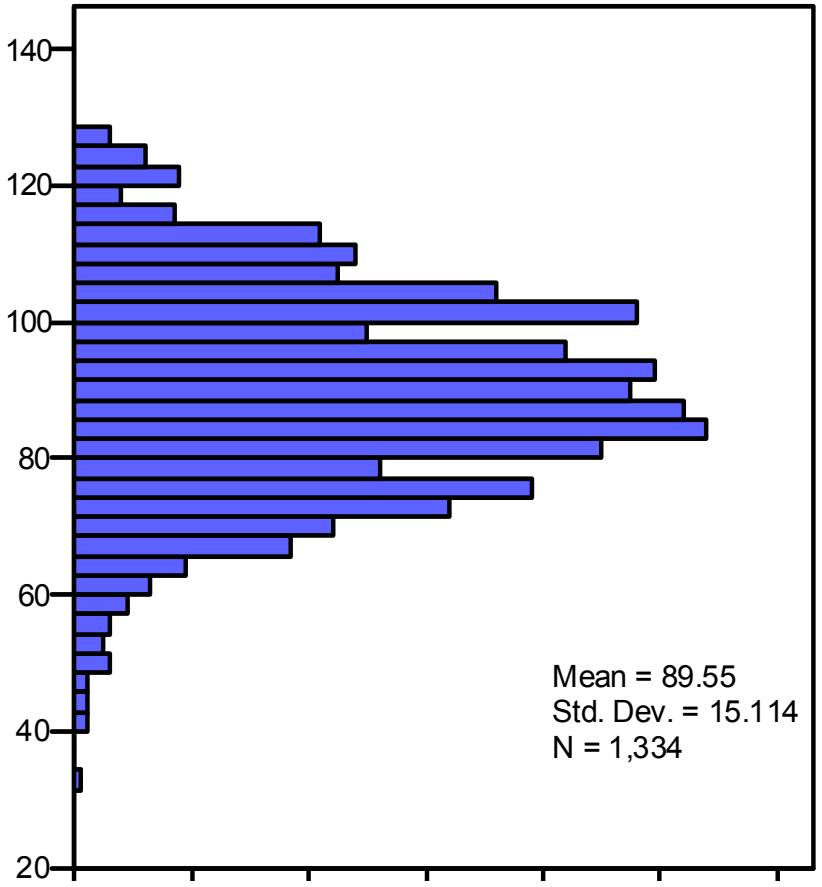
Model for the Means

Y Error Variance:

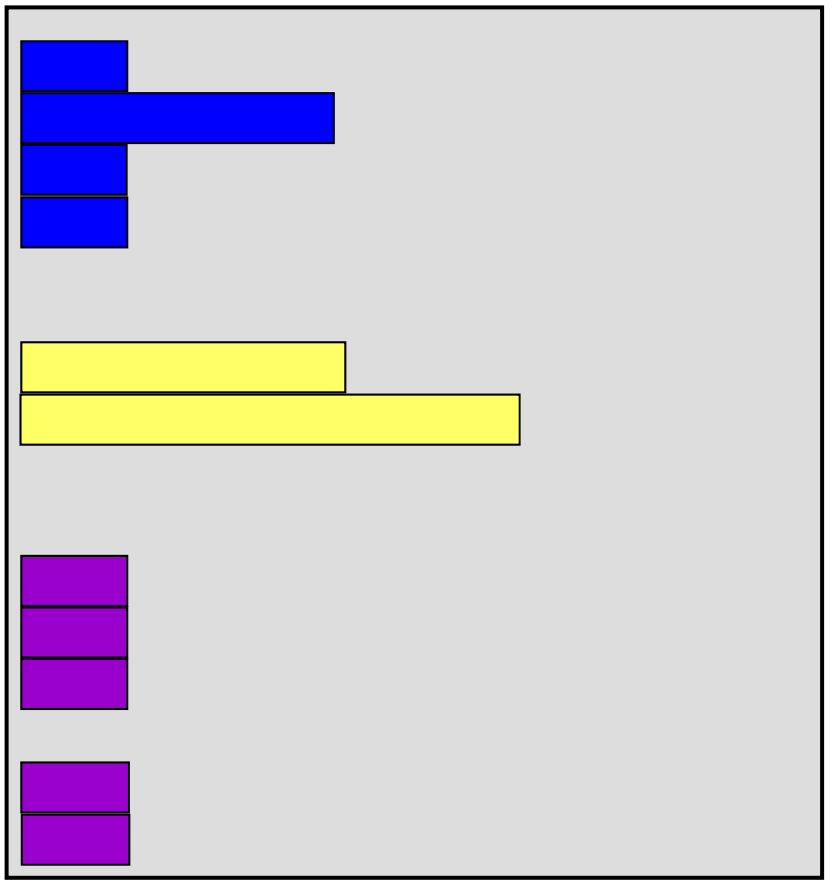
$$\frac{\Sigma (y - y_{pred})^2}{N - 1}$$

Adding Within-Person Information… (i.e., to become a Multilevel Model)

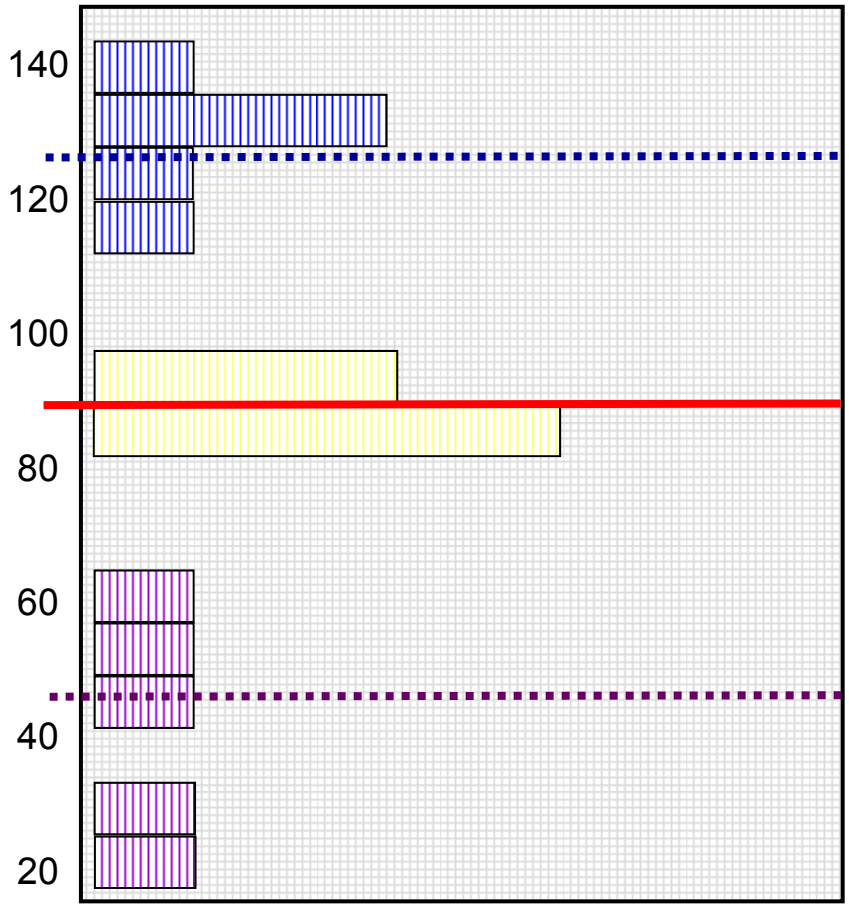
Full Sample Distribution



3 People, 5 Occasions each



Empty +Within-Person Model



Start off with Mean of Y as "best guess" for any value:

= Grand Mean

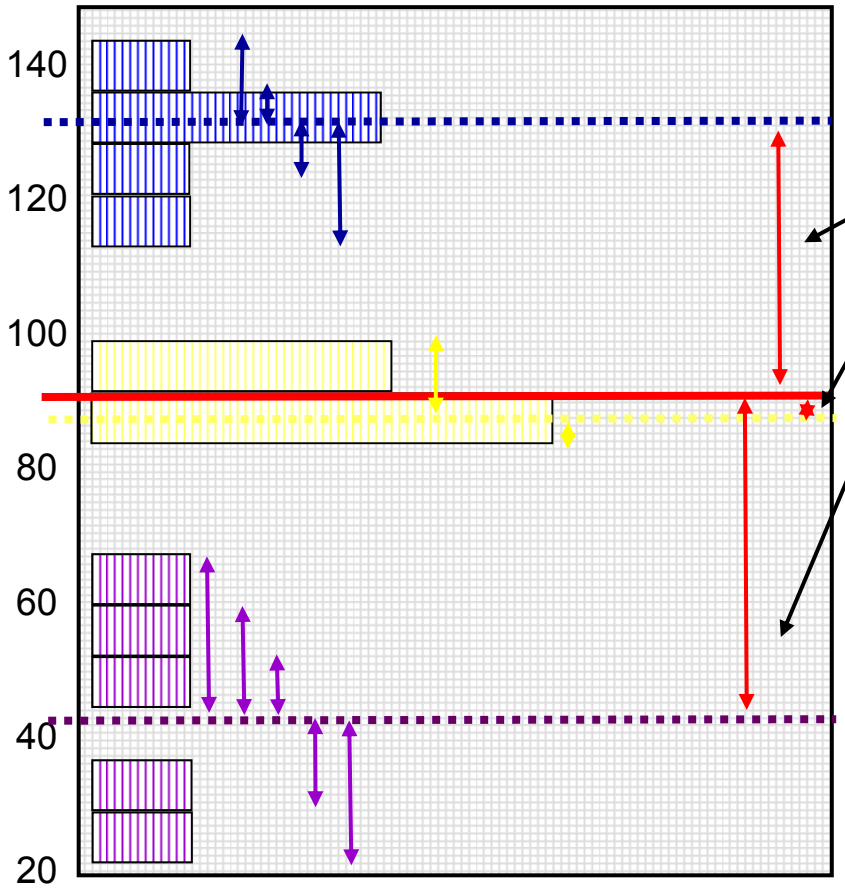
= Fixed Intercept

Can make better guess by taking advantage of repeated observations:

= Person Mean

→ Random Intercept

Empty +Within-Person Model



Variance of Y → 2 sources:

Between-Person (BP) Variance:

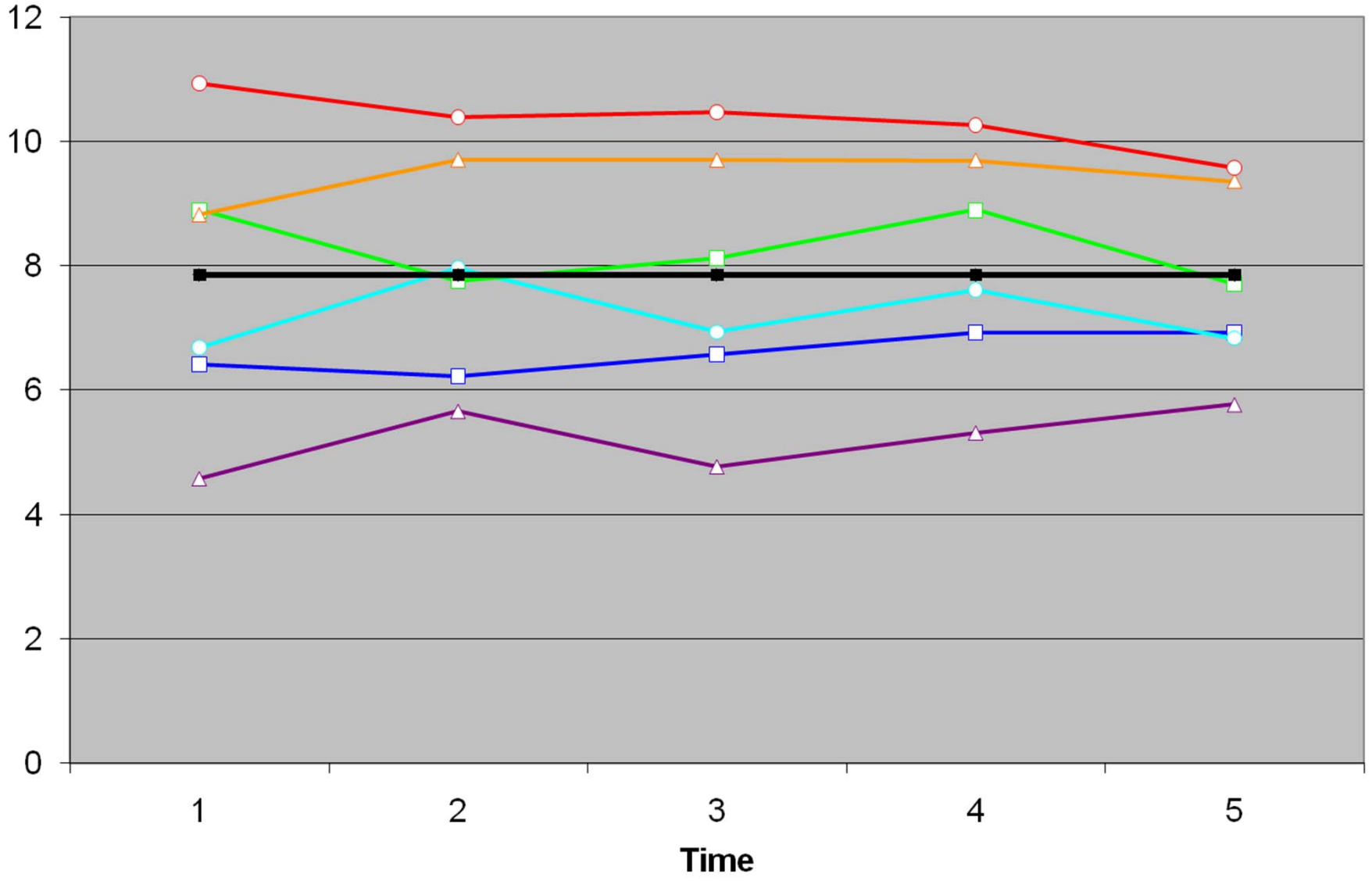
- → Differences from **GRAND** mean
- → **INTER**-Individual Differences

Within-Person (WP) Variance:

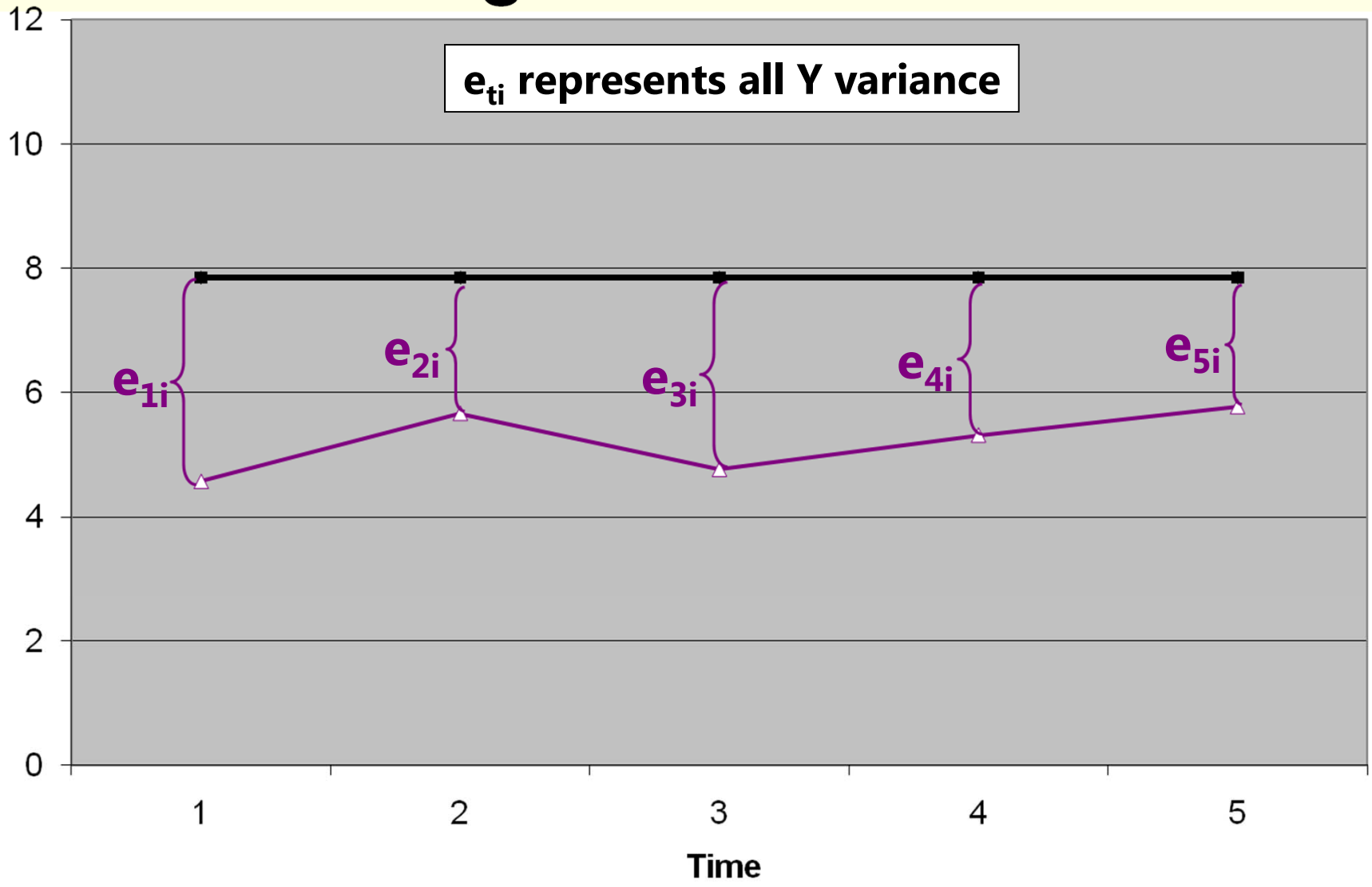
- → Differences from **OWN** mean
- → **INTRA**-Individual Differences
- → This part is only observable through longitudinal data.

Now we have 2 piles of variance in Y to predict.

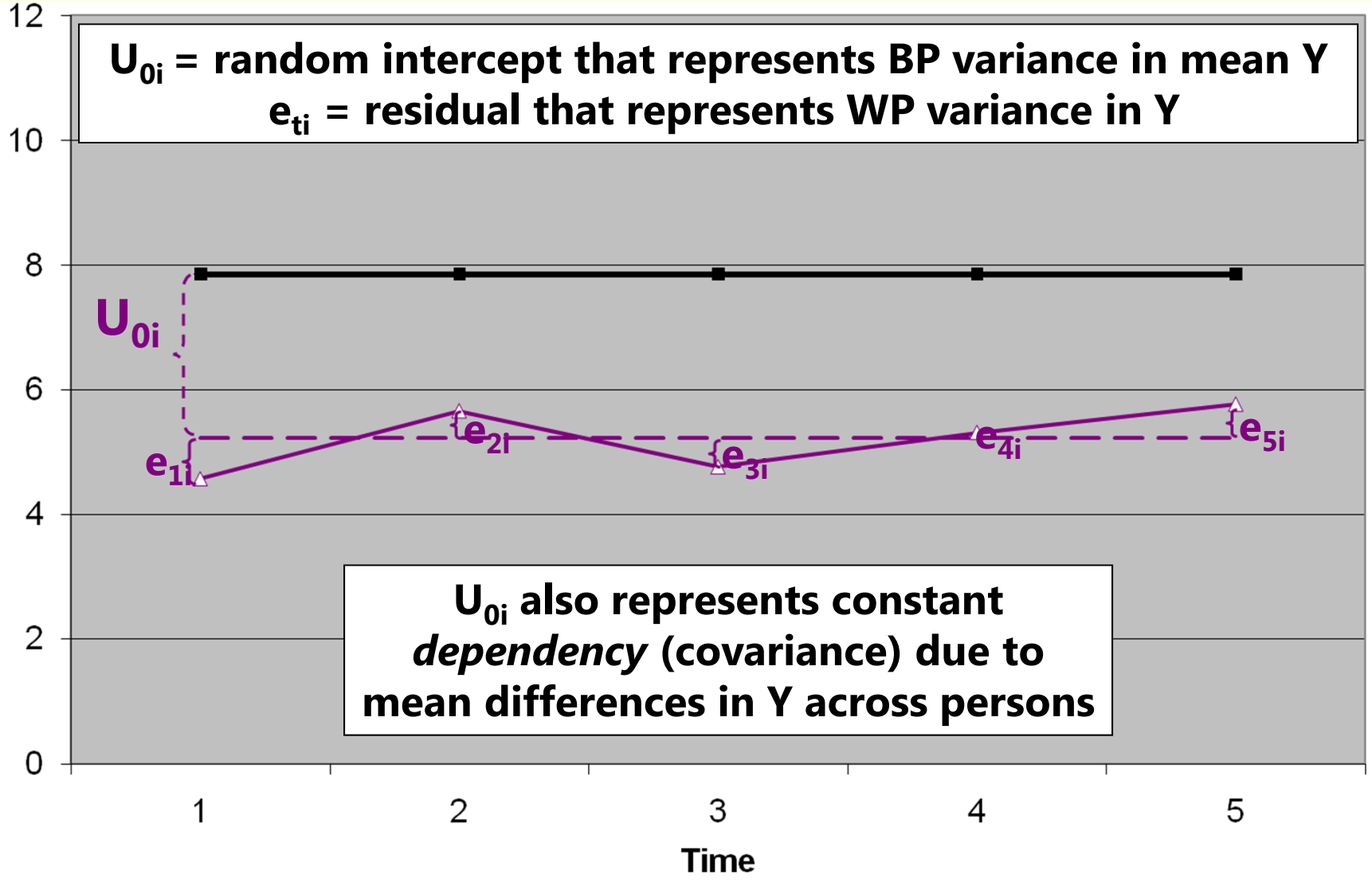
Hypothetical Longitudinal Data



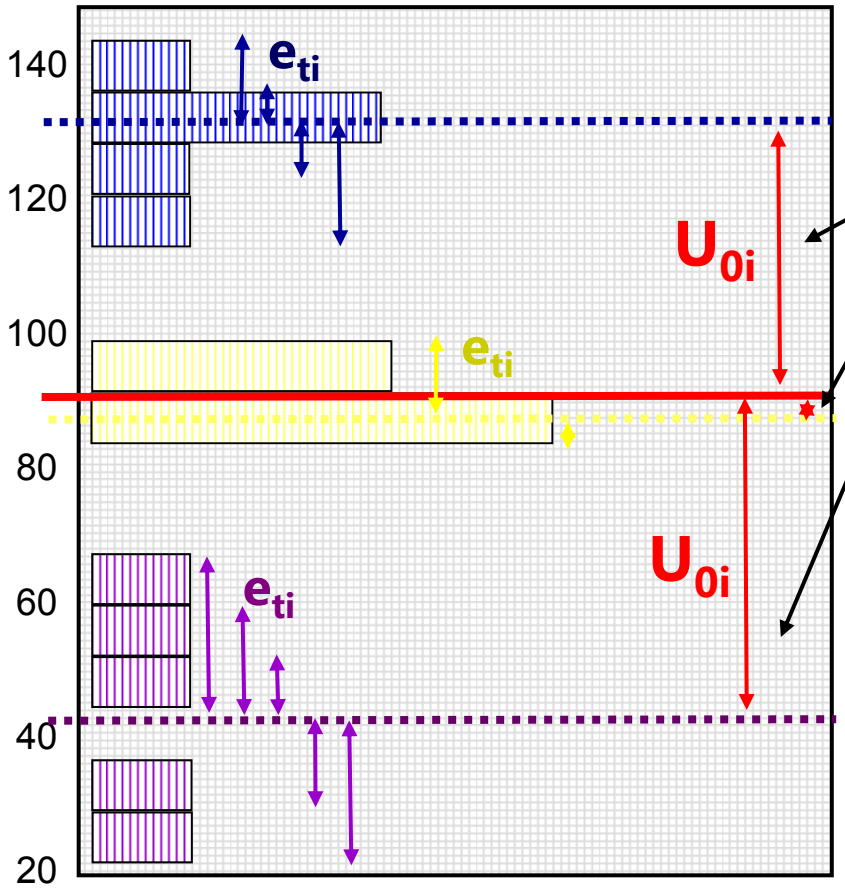
"Error" in a BP Model for the Variance: Single-Level Model



"Error" in a +WP Model for the Variance: Multilevel Model



Empty +Within-Person Model



Variance of Y → 2 sources:

Level 2 Random Intercept Variance (of U_{0i}, as $\tau^2_{U_0}$):

→ **Between**-Person Variance

→ Differences from **GRAND** mean

→ **INTER**-Individual Differences

Level 1 Residual Variance (of e_{ti}, as σ^2_e):

→ **Within**-Person Variance

→ Differences from **OWN** mean

→ **INTRA**-Individual Differences

BP vs. +WP Empty Models

- Empty **Between-Person** Model (used for 1 occasion):

$$y_i = \beta_0 + e_i$$

  - β_0 = fixed intercept = grand mean
  - e_i = residual deviation from GRAND mean

- Empty **+Within-Person** Model (>1 occasions):

$$y_{ti} = \beta_0 + U_{0i} + e_{ti}$$

  - β_0 = fixed intercept = grand mean
  - U_{0i} = random intercept = individual deviation from GRAND mean
  - e_{ti} = time-specific residual deviation from OWN mean

Intraclass Correlation (ICC)

Intraclass Correlation (ICC):

$$\text{ICC} = \frac{\text{BP}}{\text{BP} + \text{WP}} = \frac{\text{Intercept Variance}}{\text{Intercept Variance} + \text{Residual Variance}}$$

$$= \frac{\boldsymbol{\tau^2_{U_0}}}{\boldsymbol{\tau^2_{U_0} + \sigma^2_e}}$$

R matrix

$$\begin{bmatrix} \sigma_e^2+\tau_{u_0}^2 & \tau_{u_0}^2 & \tau_{u_0}^2 \\ \tau_{u_0}^2 & \sigma_e^2+\tau_{u_0}^2 & \tau_{u_0}^2 \\ \tau_{u_0}^2 & \tau_{u_0}^2 & \sigma_e^2+\tau_{u_0}^2 \end{bmatrix}$$

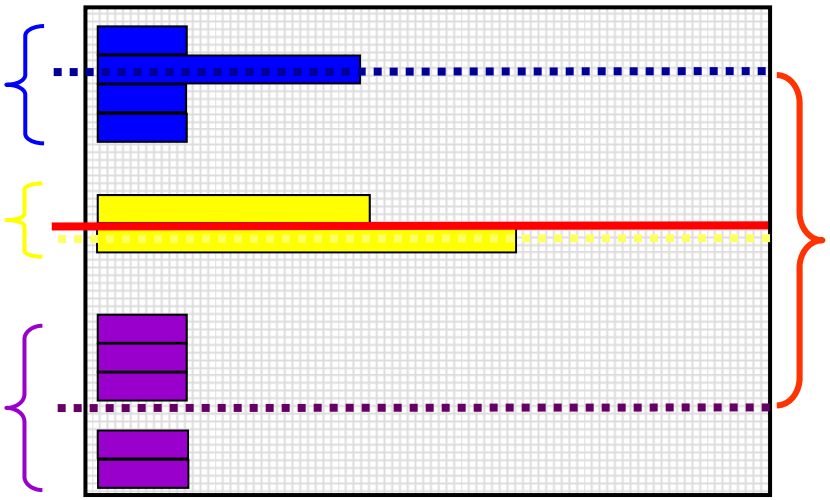
R CORR Matrix

$$\begin{bmatrix} 1 & \text{ICC} & \text{ICC} \\ \text{ICC} & 1 & \text{ICC} \\ \text{ICC} & \text{ICC} & 1 \end{bmatrix}$$

- ICC = Proportion of total variance that is between persons
- ICC = Average correlation among occasions (in RCORR)
- ICC is a standardized way of expressing how much we need to worry about *dependency due to person mean differences* **(i.e., ICC is an effect size for *constant* person dependency)**

$$ICC = \frac{Between-Person}{Between-Person + Within-Person}$$

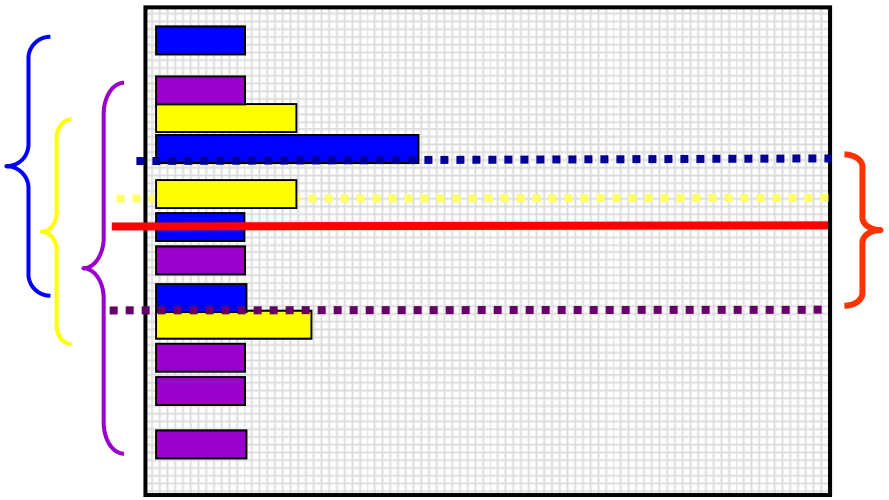
Counter-Intuitive: Between-Person Variance is in the numerator, but the ICC is the correlation over time!



ICC = BTW / BTW + within

→ Large ICC

→ Large correlation over time



ICC = btw / btw + WITHIN

→ Small ICC

→ Small correlation over time

BP and +WP Conditional Models

- Multiple Regression, **Between-Person** ANOVA: **1 PILE**
  - $y_i = (\beta_0 + \beta_1 X_i + \beta_2 Z_i \ldots) + e_i$
  - e_i → ONE residual, assumed uncorrelated with equal variance across observations (here, just persons) → "**BP (all) variation**"

- Repeated Measures, **Within-Person** ANOVA: **2 PILES**
  - $y_{ti} = (\beta_0 + \beta_1 X_i + \beta_2 Z_i \ldots) + U_{0i} + e_{ti}$
  - U_{0i} → A random intercept for differences in person means, assumed uncorrelated with equal variance across persons → "**BP (mean) variation**"= $\tau^2_{U_0}$ is now "leftover" after predictors
  - e_{ti} → A residual that represents remaining time-to-time variation, usually assumed uncorrelated with equal variance across observations (now, persons and time) → "**WP variation**" = σ^2_e is also now "leftover" after predictors

Example Data for BP and WP Models

- 50 kids ages 10 and 11 assigned to control or treatment group
- Hypothesis: Outcome should be higher with age, with a greater age difference in the treatment group

Means *(SE)*	Age 10	Age 11	Marginal
Control	49.08 *(1.14)*	54.90 *(1.13)*	51.99 *(0.89)*
	1.68	**2.04**	**3.72**	**2.70**
Treatment	50.76 *(0.91)*	58.62 *(0.99)*	54.70 *(0.87)*
Marginal	49.92 *(0.73)*	56.76 *(0.79)*	53.34 *(0.64)*

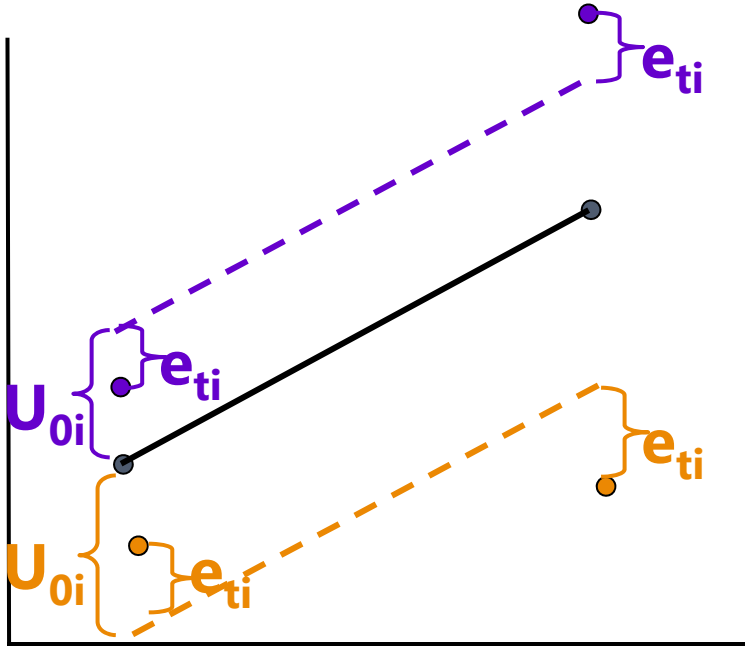
5.82

7.86

6.84

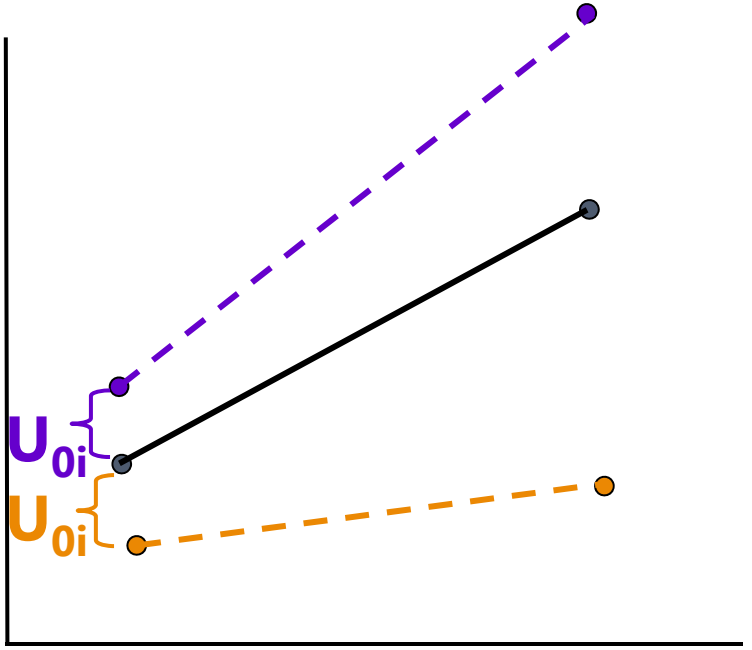
Why error and person*age are the same thing in two-occasion data

Same age slope, so error is leftover



Age (Time)

Different age slope, so no error is leftover



Age (Time)

ANOVA for longitudinal data?

- There are 3 possible "kinds" of ANOVAs we could use:
  - Between-Persons/Groups, Univariate RM, and Multivariate RM
- **NONE OF THEM ALLOW:**
  - **Missing occasions** (do listwise deletion due to least squares)
  - **Time-varying predictors** (covariates are BP predictors only)
- Each includes the same model for the means for time: all possible mean differences (so 4 parameters to get to 4 means)
  - **"Saturated means model":** $\beta_0 + \beta_1(T_1) + \beta_2(T_2) + \beta_3(T_3)$
  - **The *Time* variable must be balanced and discrete in ANOVA!**
- These ANOVAs differ by what they predict for the correlation across outcomes from the same person in the model for the variances…
  - i.e., **how they "handle dependency"** due to persons, or what they says the variance and covariance of the y_{ti} residuals should look like…

1. Between-Groups ANOVA

- **Uses e_{ti} only** (total variance = a single variance term of σ_e^2)
- **Assumes no covariance** at all among observations from the same person: *Dependency? What dependency?*
- Will usually be **very, very wrong** for longitudinal data
  - WP effects tested against wrong residual variance (significance tests will often be way too conservative)
  - Will also tend to be wrong for clustered data, but less so *(because the correlation among persons from the same group is not as strong as the correlation among occasions from the same person)*

- Predicts a variance-covariance matrix over time (here, 4 occasions) like this, called "**Variance Components**" (R matrix is TYPE=**VC** on REPEATED):

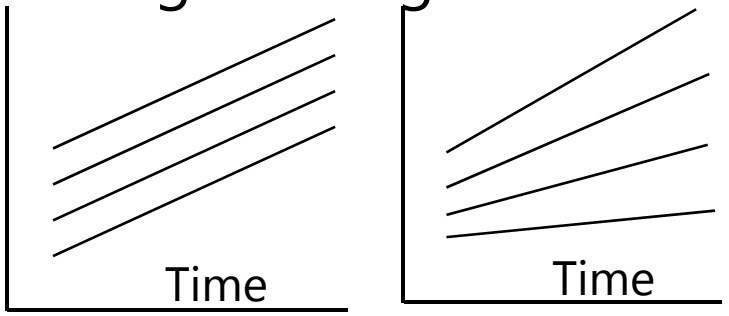
$$\begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

2a. Univariate Repeated Measures

- Separates total variance into two sources:
  - **Between-Person** (mean differences due to U_{0i}, or $\boldsymbol{\tau^2_{U_0}}$)
  - **Within-Person** (remaining variance due to e_{ti}, or $\boldsymbol{\sigma^2_e}$)
- Predicts a variance-covariance matrix over time (here, 4 occasions) like this, called "**Compound Symmetry**" (R matrix is TYPE=**CS** on REPEATED):
  - **Mean differences from U_{0i} are the only reason why occasions are correlated**

$$\begin{bmatrix} \sigma^2_e+\tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} \\ \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} \\ \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} & \tau^2_{u_0} \\ \tau^2_{u_0} & \tau^2_{u_0} & \tau^2_{u_0} & \sigma^2_e+\tau^2_{u_0} \end{bmatrix}$$

- Will usually be at least somewhat wrong for longitudinal data
  - If people change at different rates, the variances and covariances over time have to change, too



The Problem with Univariate RM ANOVA

- Univ. RM ANOVA $\left(\tau^2_{U_0} + \sigma^2_e\right)$ predicts **compound symmetry:**
  - All variances and all covariances are equal across occasions
  - In other words, the amount of error observed should be the same at any occasion, so a single, pooled error variance term makes sense
  - If not, tests of fixed effects may be biased (i.e., sometimes tested against too much or too little error, if error is not really constant over time)
  - **COMPOUND SYMMETRY RARELY FITS FOR LONGITUDINAL DATA**
- But to get the correct tests of the fixed effects, the data must only meet a less restrictive assumption of **sphericity**:
  - In English → **pairwise differences** between adjacent occasions have equal variance and covariance (satisfied by default with only 2 occasions)
  - If compound symmetry is satisfied, so is sphericity (but see above)
  - Significance test provided in ANOVA for where data meet sphericity assumption
  - **Other RM ANOVA approaches are used when sphericity fails…**

The Other Repeated Measures ANOVAs…

- 2b. **Univariate RM ANOVA with sphericity corrections**
  - Based on ε → how far off sphericity (from 0-1, 1=spherical)
  - Applies an overall correction for model df based on estimated ε, but it doesn't really address the problem that data ≠ model

- 3. **Multivariate Repeated Measures ANOVA**
  - All variances and covariances are estimated separately over time (here, 4 occasions), called "Unstructured" (R matrix is TYPE=UN on REPEATED)—it's not a model, it IS the data:

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix}$$

  - Because it can never be wrong, UN can be useful for **complete and balanced longitudinal data** with few occasions (e.g., 2-4)
  - Parameters = $\frac{\#occasions * (\#occasions + 1)}{2}$ so can be hard to estimate
  - Unstructured can also be specified to include random intercept variance $\tau_{U_0}^2$
  - Every other model for the variances is nested within Unstructured (we can do model comparisons to see if all other models are NOT WORSE)

Summary: ANOVA approaches for longitudinal data are "one size fits most"

- **Saturated Model for the Means** (balanced time required)
  - All possible mean differences
  - Unparsimonious, but best-fitting (is a description, not a model)

- **3 kinds of Models for the Variances** (complete data required)
  - BP ANOVA (σ_e^2 only) → assumes independence and constant variance over time
  - Univ. RM ANOVA $\left(\tau_{U_0}^2 + \sigma_e^2\right)$ → assumes constant variance and covariance
  - Multiv. RM ANOVA (whatever) → no assumptions; is a description, not a model

there is no structure that shows up in a scalar equation (i.e., the way $U_{0i} + e_{ti}$ does)

- **MLM will give us more flexibility in both parts of the model:**
  - Fixed effects that *predict* the pattern of means (polynomials, pieces)
  - Random intercepts and slopes and/or alternative covariance structures that *predict* intermediate patterns of variance and covariance over time

3 Decision Points for Model Comparisons

1. Are the models **nested** or **non-nested**?
   - Nested: have to add OR subtract effects to go from one to other
     - Can conduct significance tests for improvement in fit
   - Non-nested: have to add AND subtract effects
     - No significance tests available for these comparisons

2. Differ in model for the **means**, **variances**, or **both**?
   - Means? Can only use ML $-2\Delta LL$ tests (or *p*-value of each fixed effect)
   - Variances? Can use ML (or preferably REML) $-2\Delta LL$ tests, no *p*-values
   - Both sides? Can only use ML $-2\Delta LL$ tests

3. Models estimated using **ML** or **REML**?
   - ML: All model comparisons are ok
   - REML: Model comparisons are ok for the variance parameters only

Likelihood-Based Model Comparisons

- Relative model fit is indexed by a "**deviance**" statistic → **−2LL**
  - Log of likelihood (LL = total data height!) of observing the data given model parameters, −2*LL so that the differences between model LL values follow ~χ^2
  - **−2LL is a measure of BADNESS of fit, so smaller values = better models**
  - Models are compared using their deviance values (significance tests)
  - Comes in two estimation flavors (labeled as −2 log likelihood on output): Maximum Likelihood (**ML**) or Restricted (Residual) ML (**REML**)

- Fit is also indexed by **Information Criteria** that reflect **−2LL** deviance AND # parameters used and/or sample size
  - **AIC** = Akaike IC = **−2LL** + 2 *(#parameters)
  - **BIC** = Bayesian IC = **−2LL** + log(N)*(#parameters) → penalty for complexity
  - In ML → #parameters = all parameters (means and variances models)
  - In REML → #parameters = variance model parameters only (except in STATA!)
  - No significance tests or critical values, just "smaller is better"

$-2\Delta LL$ (i.e., LRT, Deviance) Tests:
(models must use the same estimator & *N*)

1. Calculate $-2\Delta LL$: $(-2LL_{fewer}) - (-2LL_{more})$
2. Calculate Δdf: $(\#\ Parms_{more}) - (\#\ Parms_{fewer})$
3. Compare $-2\Delta LL$ to χ^2 distribution with df = Δdf
   CHIDIST function in excel will give exact p-values for the difference test

1. & 2. must be positive values!

- Fixed effects $p<.05$: $-2\Delta LL(1)>3.84$, $-2\Delta LL(2)>5.99$, $-2\Delta LL(3)>7.82$
- Some controversy about $-2\Delta LL$ tests when testing random effects variances that cannot be negative (i.e., the "boundary problem")
  - χ^2 is not distributed as usual (mean=df) → is actually a mixture χ^2 with df and df−1, so using the critical χ^2 for actual df results in conservative model comparison test
  - e.g., $-2\Delta LL(df=2)>5.99$, whereas $-2\Delta LL(df=\text{mixture of } 1,2)>5.14$
- Two proposed solutions when testing random effects variances:
  - For random intercepts, can use a 1-tailed test (χ^2 for $p < .10$): $-2\Delta LL(1)>2.71$
  - Use mixture *p*-value = $0.5*prob(\chi^2_{df-1} > -2\Delta LL) + 0.5*prob(\chi^2_{df} > -2\Delta LL)$
  - In practice these assume no relationship among how well variance parameters are estimated, which is suspect → I tend to just use the conservative test and call it good

Critical Values for 50:50 χ^2 Mixtures

	Significance Level				
df (q)	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
0 vs. 1	1.64	2.71	3.84	5.41	6.63
1 vs. 2	3.81	5.14	6.48	8.27	9.63
2 vs. 3	5.53	7.05	8.54	10.50	11.97
3 vs. 4	7.09	8.76	10.38	12.48	14.04
4 vs. 5	8.57	10.37	12.10	14.32	15.97
5 vs. 6	10.00	11.91	13.74	16.07	17.79
6 vs. 7	11.38	13.40	15.32	17.76	19.54
7 vs. 8	12.74	14.85	16.86	19.38	21.23
8 vs. 9	14.07	16.27	18.35	20.97	22.88
9 vs. 10	15.38	17.67	19.82	22.52	24.49
10 vs. 11	16.67	19.04	21.27	24.05	26.07

This may work ok if only one new parameter is bounded … for example:

+ Random Intercept
df=1: 2.71 vs. 3.84

+ Random Linear
df=2: 5.14 vs. 5.99

+ Random Quad
df=3: 7.05 vs. 7.82

Critical values such that the right-hand tail probability =
$0.5 \times \Pr(\chi^2_q > c) + 0.5 \times \Pr(\chi^2_{q+1} > c)$

Source: Appendix C (p. 484) from Fitzmaurice, Laird, & Ware (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley

ML vs. REML (more details to follow)

Remember "population" vs. "sample" formulas for calculating variance?

$$\text{Population: } \sigma_e^2 = \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N} \qquad \text{Sample: } \sigma_e^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$$

All comparisons must have same N!!!	**ML**	**REML**
To select, type…	METHOD=ML (-2 log likelihood)	METHOD=REML *default* (-2 res log likelihood)
In estimating variances, it treats fixed effects as…	**Known** (df for having to also estimate fixed effects is not factored in)	**Unknown** (df for having to estimate fixed effects is factored in)
So, in small samples, L2 variances will be…	**Too small** (less difference after N=30-50 or so)	**Unbiased** (correct)
But because it indexes the fit of the…	**Entire model** (means + variances)	**Variances model only**
You can compare models differing in…	**Fixed and/or random effects** (either/both)	**Random effects only** (same fixed effects)

Rules for Comparing Multilevel Models

All observations must be the same across models!

Compare Models Differing In:

Type of Comparison:	**Means Model (Fixed) Only**	**Variance Model (Random) Only**	**Both Means and Variances Model (Fixed and Random)**
Nested? YES, can do significance tests via…	Fixed effect *p*-values from ML or REML -- *OR* -- ML −2ΔLL only (NO REML −2ΔLL)	NO *p*-values REML −2ΔLL (ML −2ΔLL is ok if big N)	ML −2ΔLL only (NO REML −2ΔLL)
Non-Nested? NO signif. tests, instead see…	ML AIC, BIC (NO REML AIC, BIC)	REML AIC, BIC (ML ok if big N)	ML AIC, BIC only (NO REML AIC, BIC)

Nested = one model is a direct subset of the other

Non-Nested = one model is not a direct subset of the other

Summary: Model Comparisons

- Significance of **fixed effects** can be tested with EITHER their ***p*-values** OR **ML** −2ΔLL (LRT, deviance difference) tests
  - *p*-value → Is EACH of these effects significant? (fine under ML or REML)
  - ML −2ΔLL test → Does this SET of predictors make my model better?
  - *REML −2ΔLL tests are WRONG for comparing models differing in fixed effects*

- Significance of **random effects** can only be tested with **−2ΔLL tests** (preferably using REML; here ML is not wrong, but results in too small variance components and fixed effect SEs in smaller samples)
  - Can get *p*-values as part of output but *shouldn't* use them
  - #parms added (df) should always include the random effect covariances

- My recommended approach to building models:
  - Stay in REML (for best estimates), test new fixed effects with their *p*-values
  - THEN add new random effects, testing −2ΔLL against previous model