# Statistically Speaking Lecture Series

Sponsored by the Biostatistics Collaboration Center

## *Longitudinal Data Analysis: Challenges and Opportunities*

**Nathan Gill, PhD**

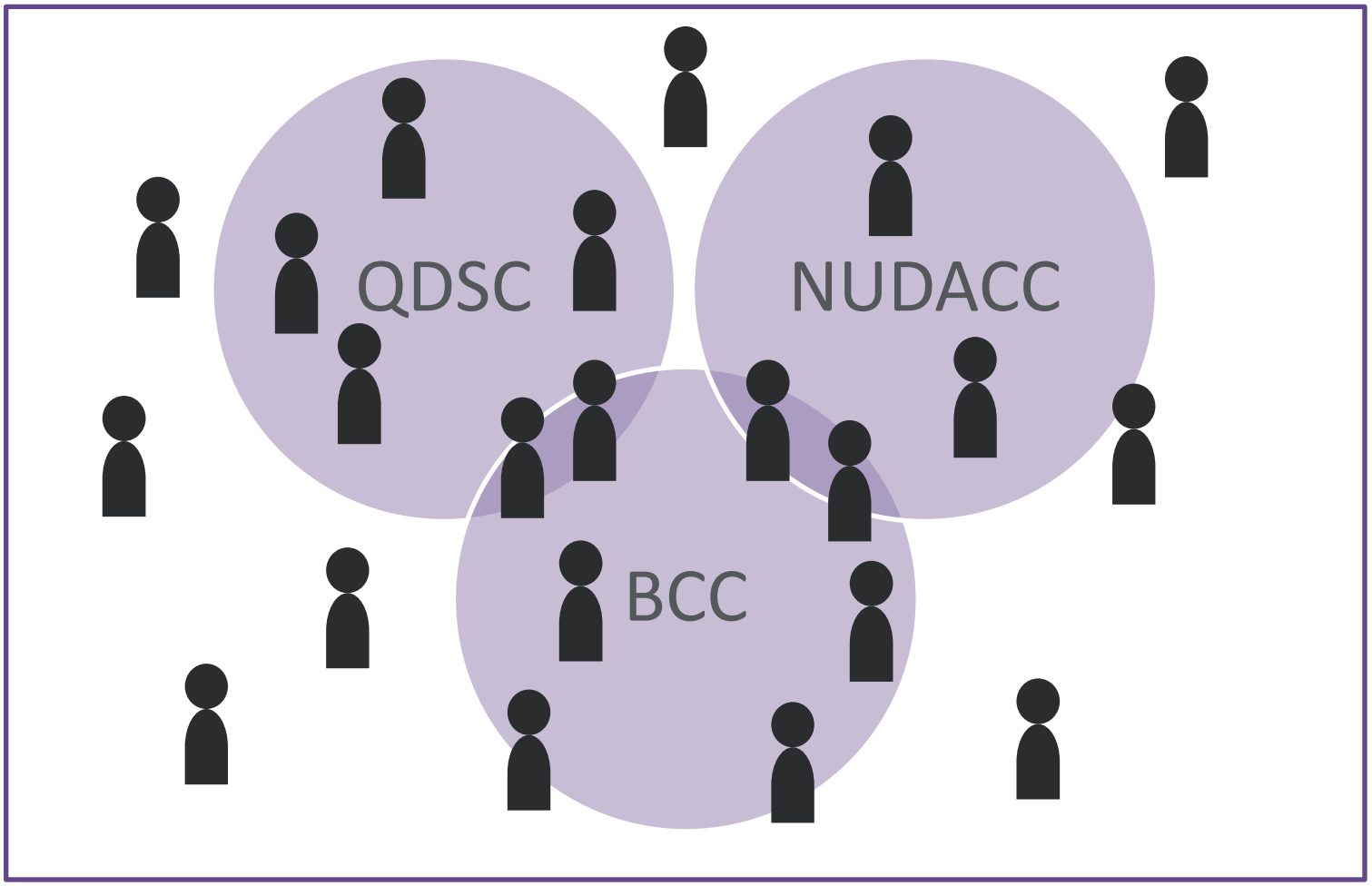
Assistant Professor

Department of Preventive Medicine – Biostatistics

# Biostatistics at NU

## Overview

Division of Biostatistics (Chief: Denise Scholtens),
Department of Preventive Medicine (Chair: Donald Lloyd-Jones)



# Biostatistics Centers and Cores

## Overview

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute | Robert H. Lurie Comprehensive Cancer Center of Northwestern University | Stanley Manne Children's Research Institute | Ann & Robert H. Lurie Children's Hospital of Chicago | Shirley Ryan AbilityLab

**Biostatistics Collaboration Center (BCC)**

- Supports **non-cancer** research at NU
- Initial 1-2 hour consultation subsidized by FSM Research Office
- Grant, Hourly
- https://www.feinberg.northwestern.edu/sites/bcc/

**Quantitative Data Sciences Core (QDSC)**

- Supports **cancer-related** research at NU
- Free to Lurie Cancer Center (LCC) members
- Grant
- https://www.cancer.northwestern.edu/research/shared-resources/quantitative-data-sciences.html

**Northwestern University Data Analysis and Coordinating Center (NUDACC)**

- Prospective, large **multicenter research**
- Comprehensive support (e.g., clinical monitoring, data analysis, project management)
- Grant
- https://www.feinberg.northwestern.edu/sites/nudacc/

# Upcoming *Statistically Speaking* Lectures

Tuesday, May 16
12-1pm

**Unseen Worlds: How Missing Data Impact Statistical Analyses**

Lucia C Petito, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine

*Location: McGaw - Daniel Hale Williams Auditorium & Zoom**

# Outline

1. What is longitudinal data?

2. Why is it difficult to analyze? Examples.

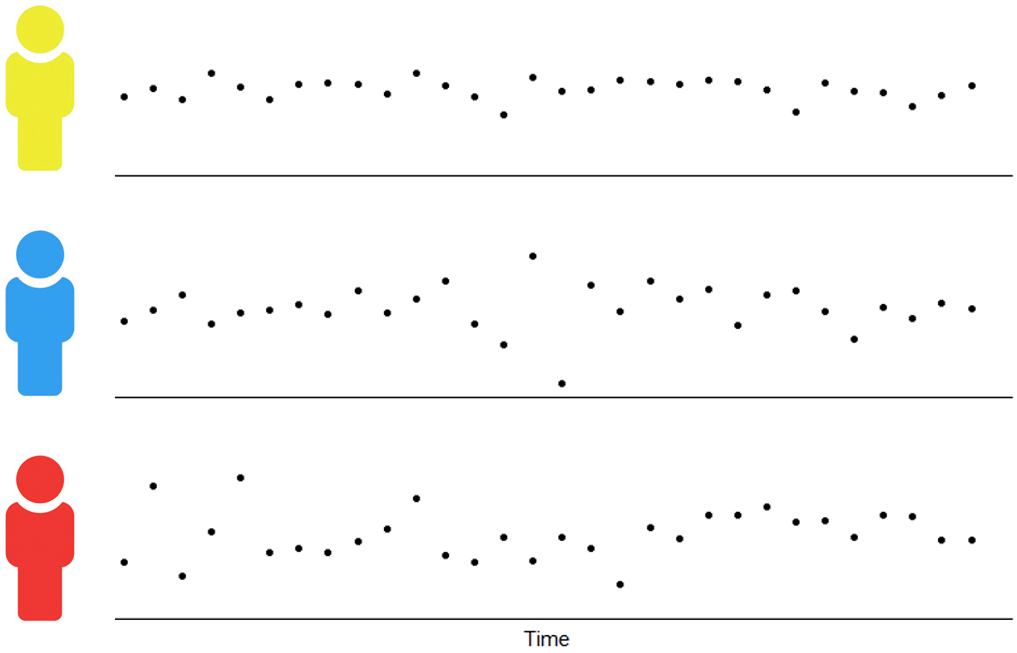
3. What opportunities does it present, and which modeling approaches take advantage of these?

    - Random/mixed effects models

    - Generalized estimating equations (GEE)

    - Other methods

# What is Longitudinal Data?




Any data that has multiple observations from the *same subjects* over time

Not all data with observations from multiple time points is longitudinal

Very easy to get misleading results if you ignore longitudinal structure!

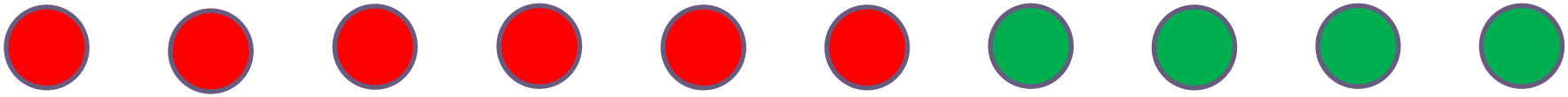
Example:

Patient satisfaction survey

# Are patients happy with their care?

= No = Yes

Responses:

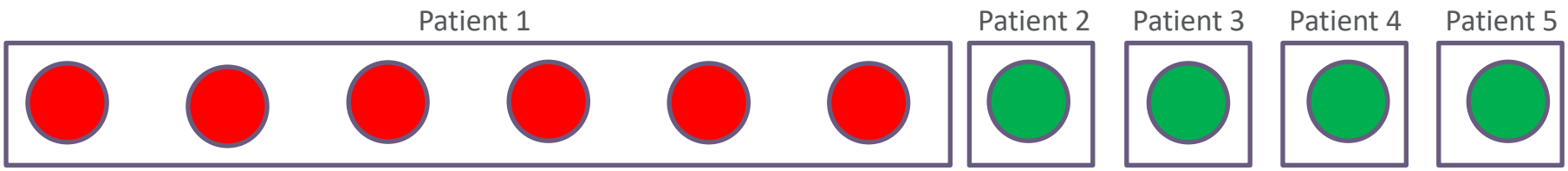


40% are happy: 4 yes out of 10 total

# Are patients happy with their care?

= No = Yes

Responses:

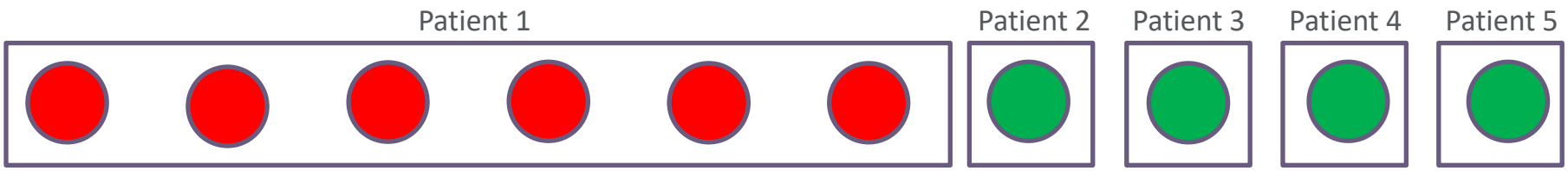



40% are happy: 4 yes out of 10 total

# Are patients happy with their care?

= No = Yes

Responses:



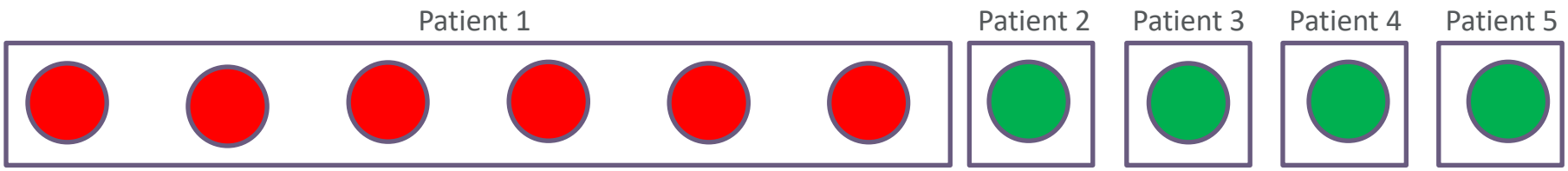

40% are happy: 4 yes out of 10 total

What about 80%?

# Are patients happy with their care?

= No = Yes

Responses:

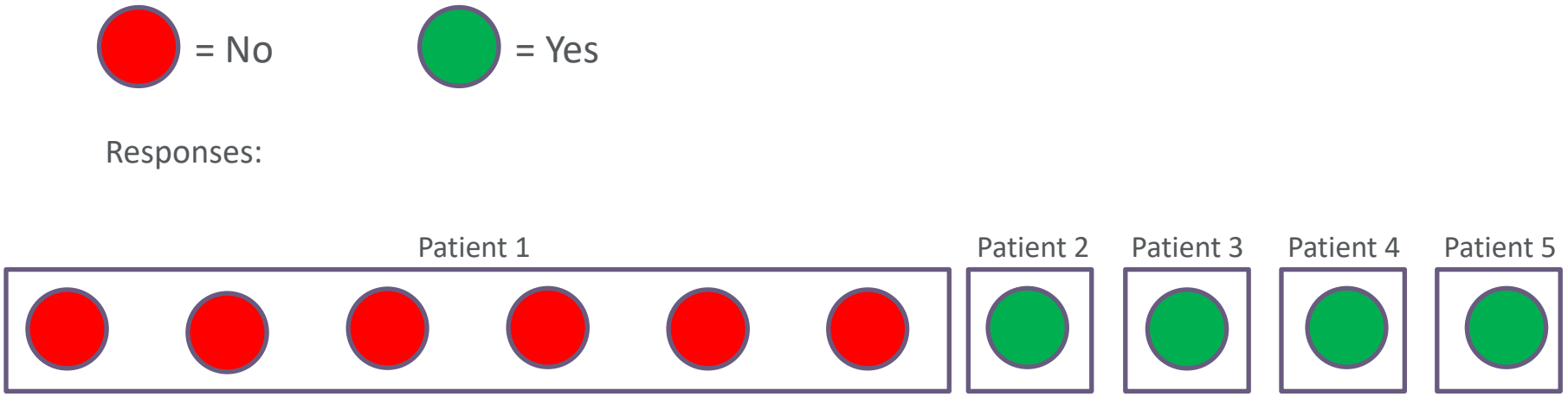



40% are happy: 4 yes out of 10 total

What about 80%?

Somewhere between 40% and 80%?

# Are patients happy with their care?




40% are happy: 4 yes out of 10 total

What about 80%?

Somewhere between 40% and 80%?

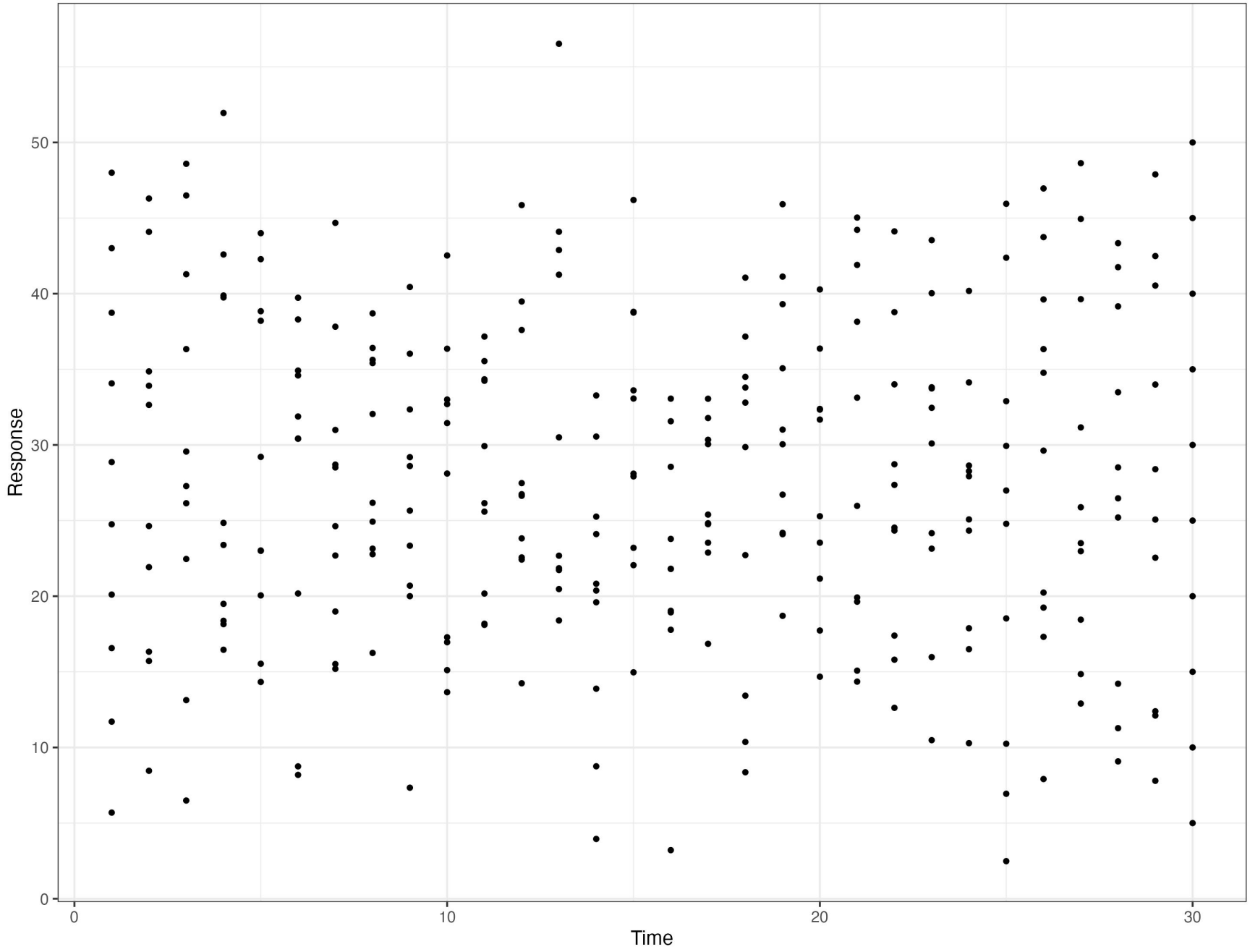
Uncertainty? Variance of the estimate is given by

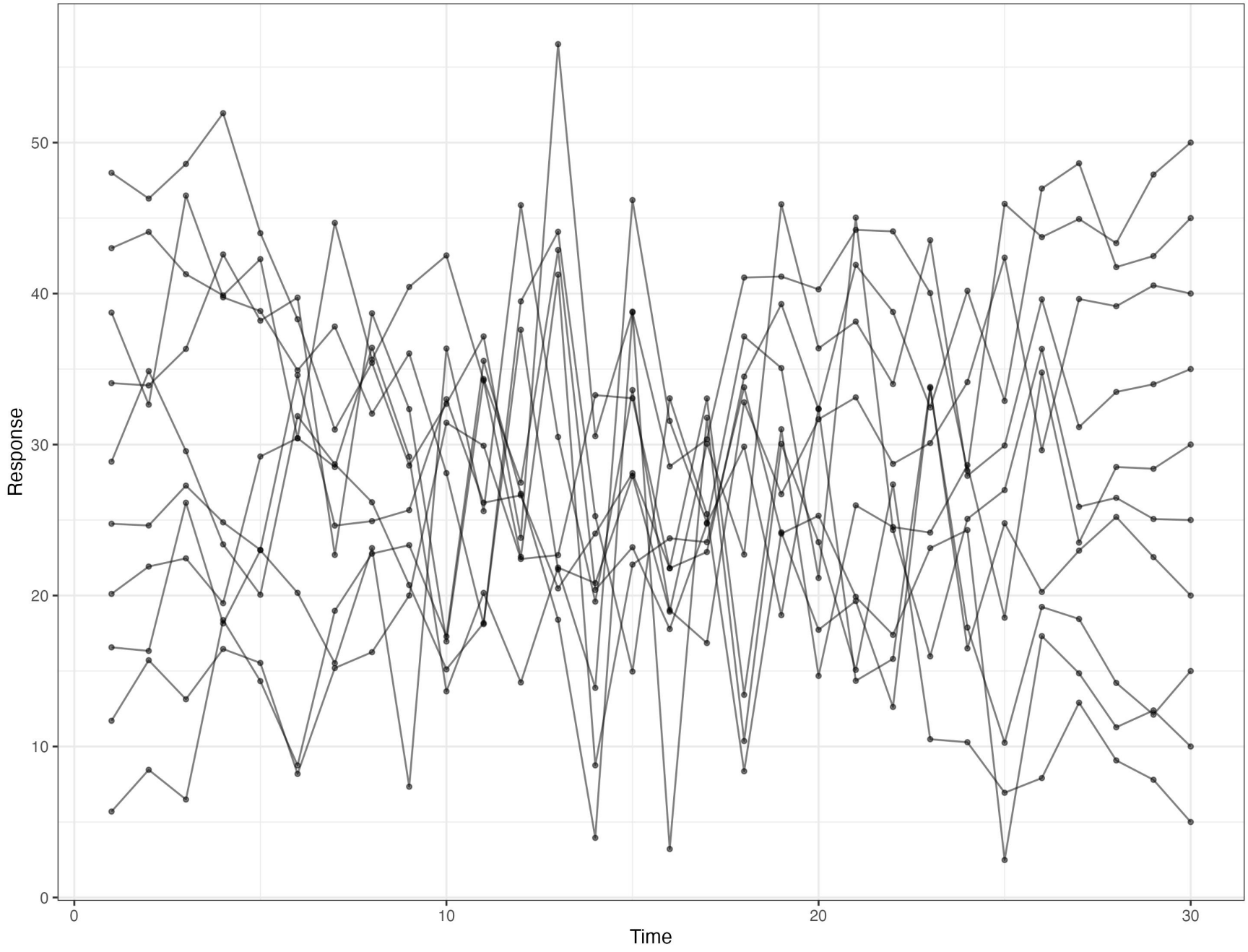
$$\frac{\hat{p}(1-\hat{p})}{n}$$

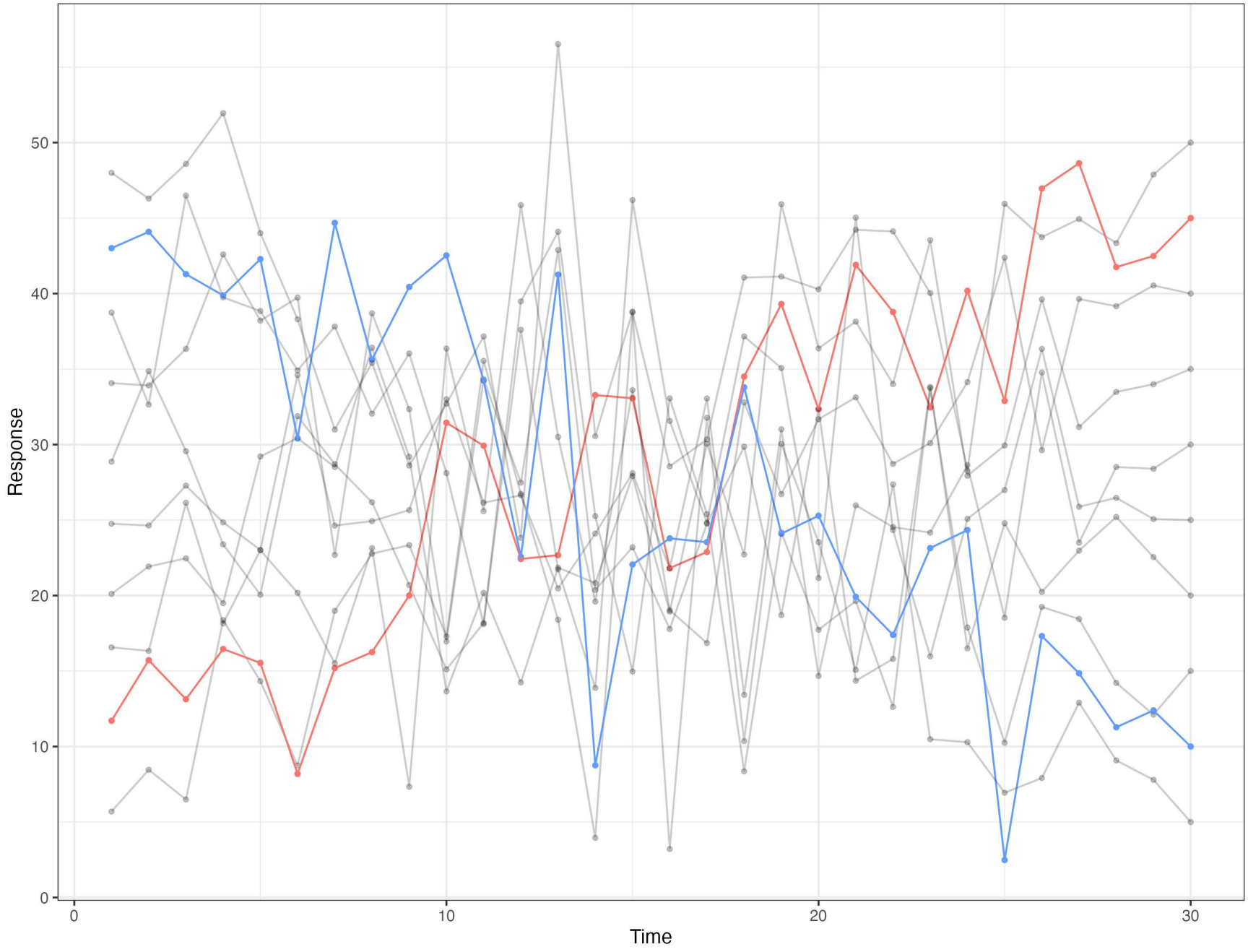
(Standard error is square root of this)

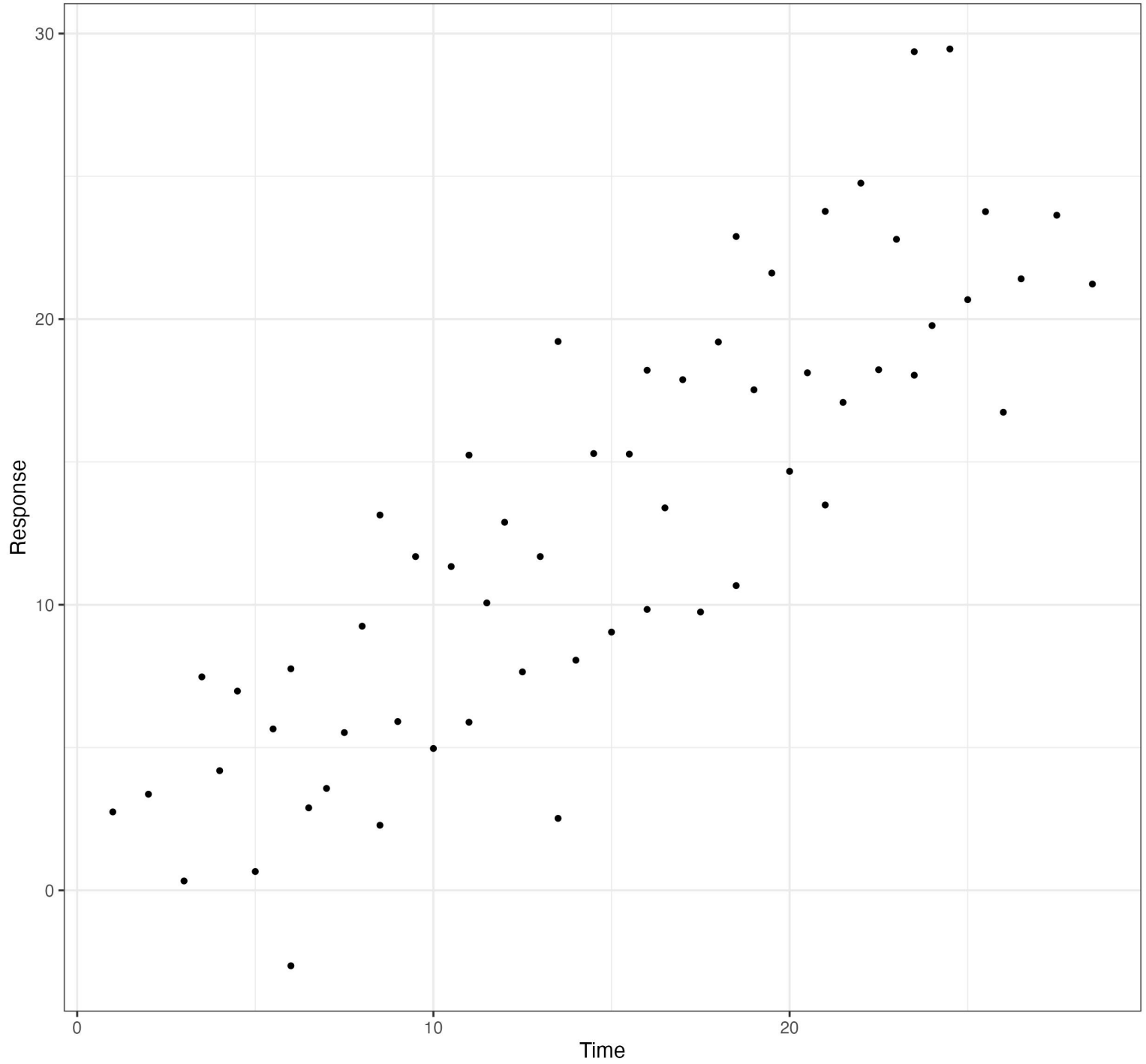
# What goes wrong?

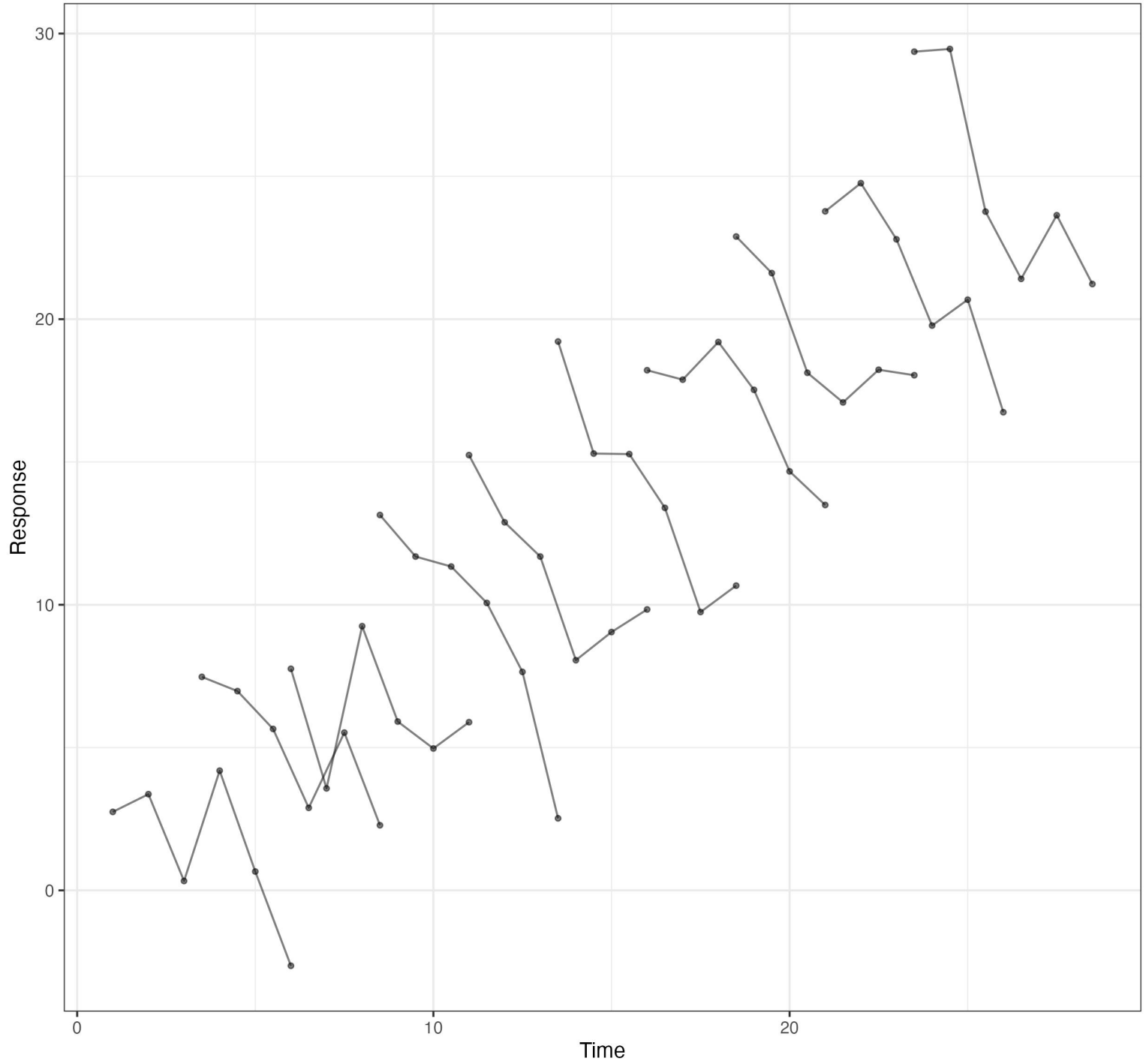
- One of the key assumptions of many familiar statistical methods (like linear regression) is **independence of observations**
- In longitudinal data, this assumption **does not necessarily hold**. Two observations from the same person are likely to be more similar than two observations from different people.
- This can cause two main problems
  - Biased effect estimates
  - Underestimated standard errors
- Standard methods can also miss important features of the data altogether


Response
Time
0
10
20
30
40
50
0
10
20
30


Response
Time
0
10
20
30
40
50
0
10
20
30


Response
Time
0
10
20
30
0
10
20
30
40
50


30
20
10
0
Response
0
10
20
Time


30
20
10
0
Response
0
10
20
Time

# Modeling Approaches

# Mixed (or Random) Effects Model

Extension of the standard linear model to account for repeated observations and modify effect sizes/coefficients accordingly

Versions exist for many common linear models: linear regression, logistic regression, poisson regression, etc.

The model will tell you:

(1) How your specified variables (e.g. age, sex, time, disease or treatment status) impact the overall mean of the response, just like in a regular linear regression

(2) How much the response varies between different people in the study population beyond the effects of covariates

(3) For continuous responses, how much the response varies within a single person in the study population

# Generalized Estimating Equations (GEE)

- Designed to target incorrect standard errors when standard linear models are applied to longitudinal data

- Basic idea (simplified):
  - First, apply the standard methods (linear regression, logistic regression, etc.) to get an effect estimate
  - Second, **adjust the standard errors** of that estimate to account for the ignored longitudinal structure
    - Usually this means making them bigger, but not always

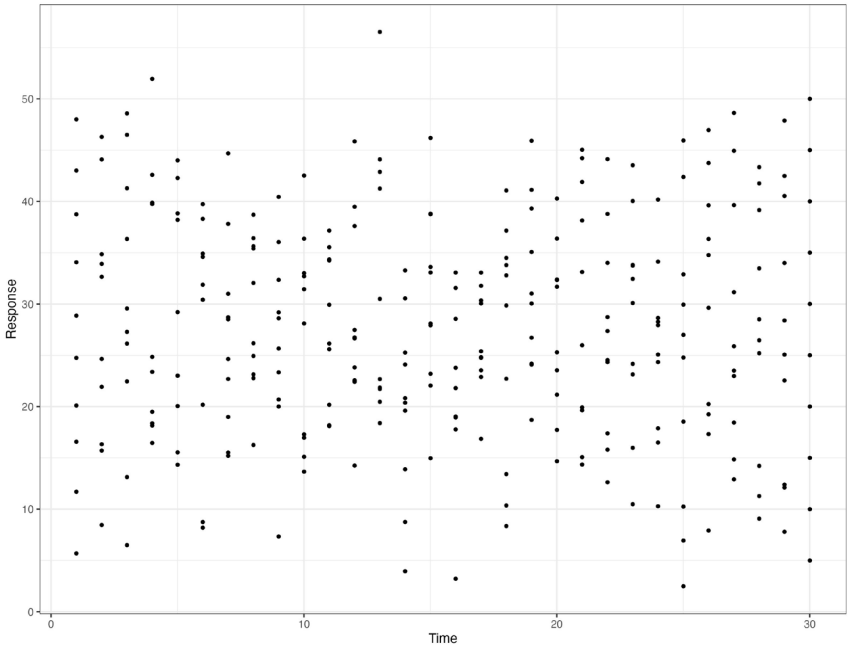
# Mixed Effects Model vs GEE

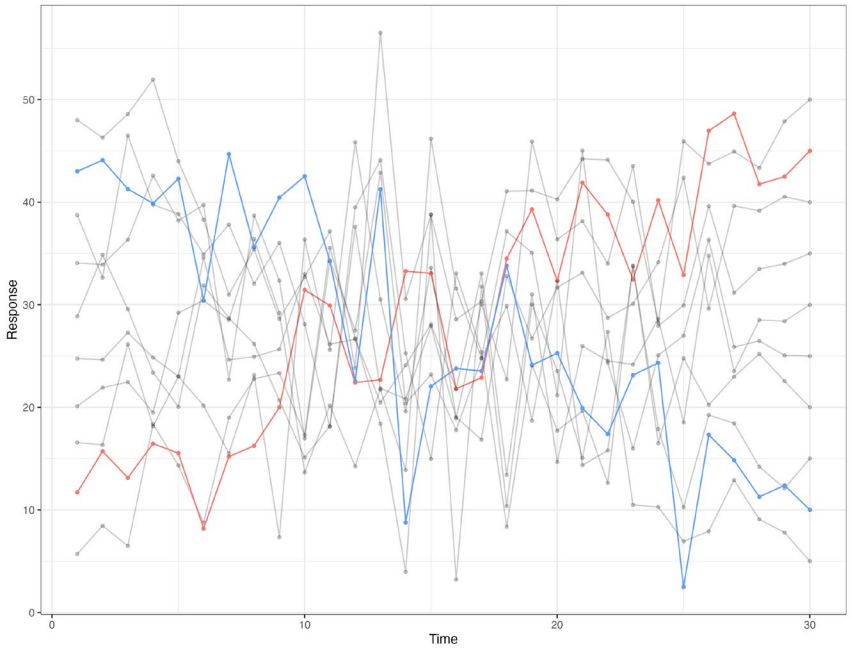
- In most cases, either works!

- Both can treat time as a continuous covariate
  - Don't need to data at fixed time points
  - Good to keep in mind when designing a study

- **GEE**
  - Computationally simpler than mixed models
  - Useful if interested in overall average effects only
  - Point estimates the same as standard methods, standard errors adjusted to account for repeated measures

- **Mixed Effects Model**
  - Useful if you want to know how variability is divided between vs within people, or if there are very unequal sample sizes
  - Require more assumptions than GEEs
  - Computationally more challenging than GEEs




## Linear Regression

Intercept: 28.08 (1.29)

Slope: -0.03 (0.07)

Residual standard error: 10.87

## GEE

Intercept: 28.08 (4.49)

Slope: -0.03 (0.31)

Residual standard error: 10.87

## Mixed effects model

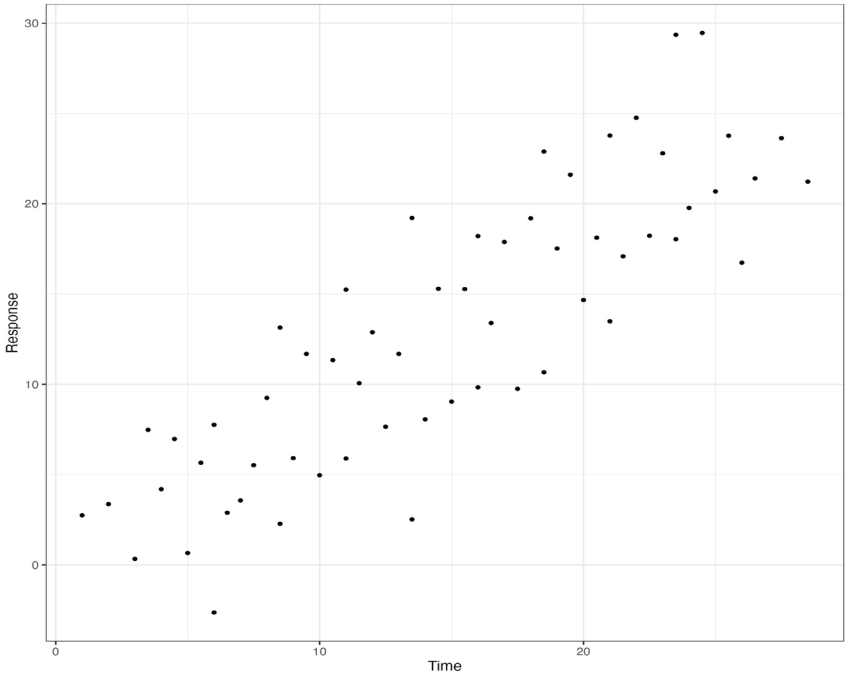
Intercept: 28.08 (4.72)

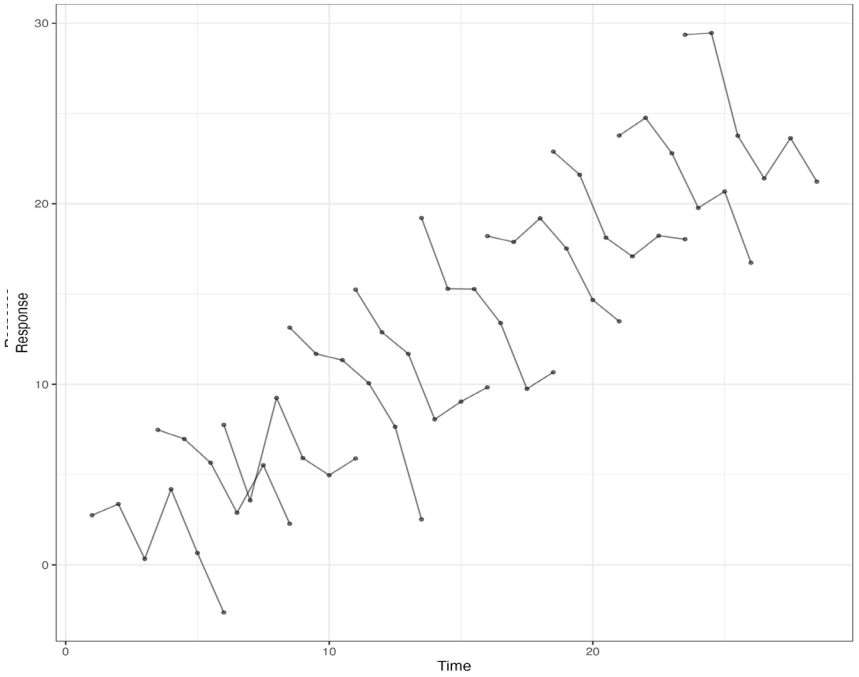
Slope: -0.03 (0.32)

Standard deviation across subjects:

- Intercept: 14.7
- Slope: 1.00

Residual standard error: 6.94




## Linear Regression

Intercept: -0.12 (1.19)

Slope: 0.89 (0.07)

Residual standard error: 4.12

## GEE

Intercept: -0.12 (0.67)

Slope: 0.89 (0.04)

Residual standard error: 4.12

## Mixed effects model

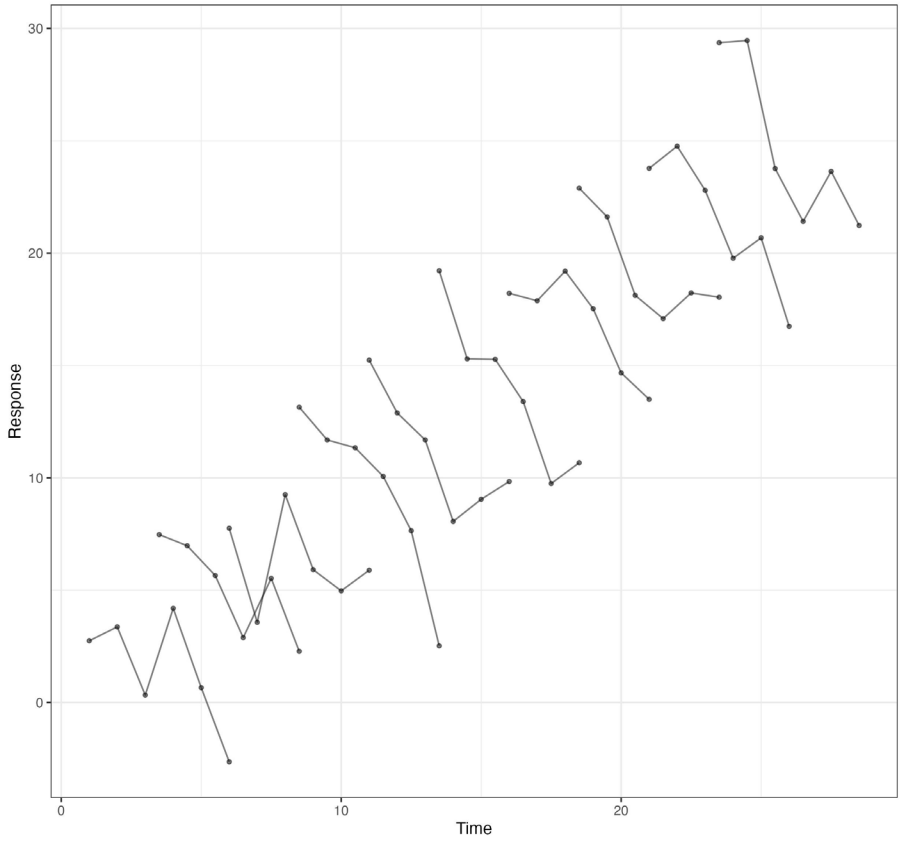
Intercept: 31.30 (6.54)

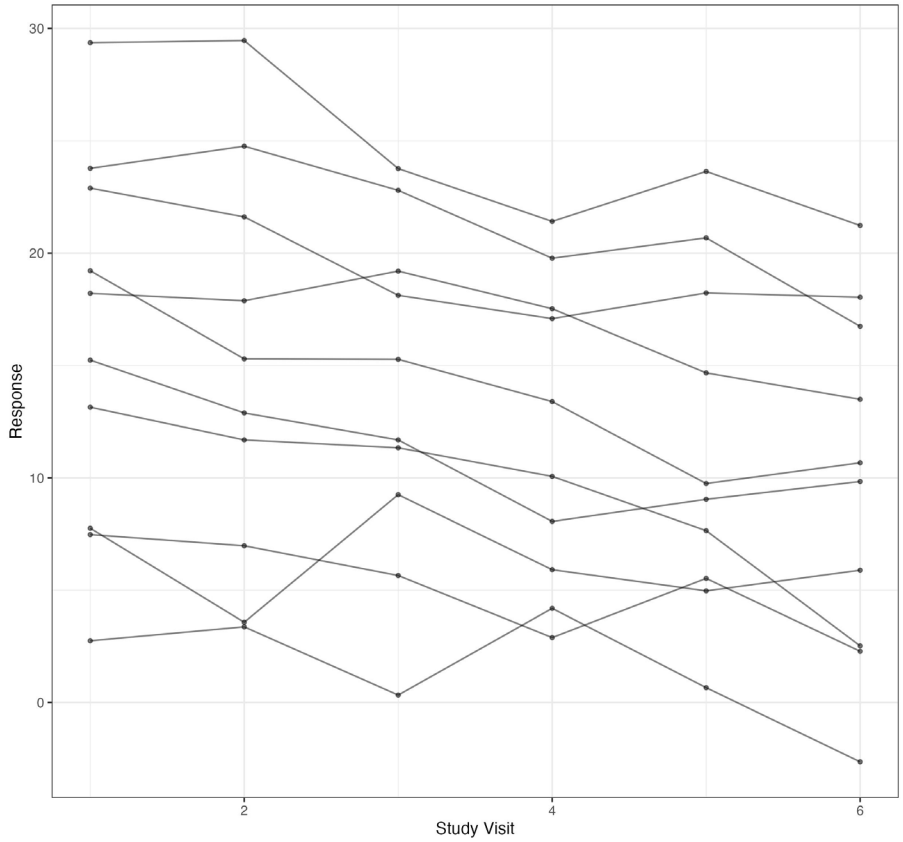
Slope: -1.15 (0.15)

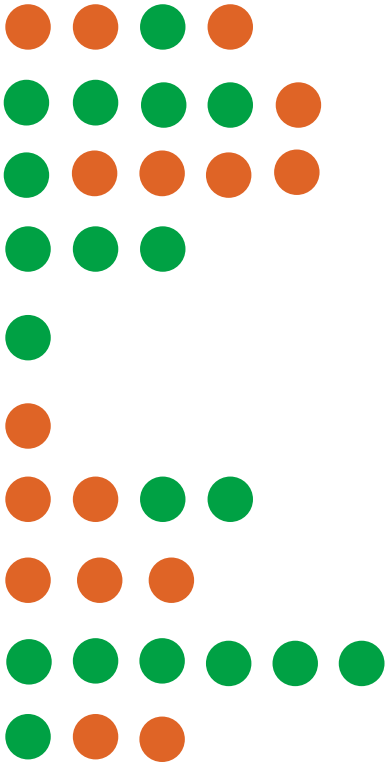
Standard deviation across subjects:

- Intercept: 19.7
- Slope: 0.3

Residual standard error: 1.7


Response
Time
30
20
10
0
0
10
20


Response
Study Visit
30
20
10
0
2
4
6



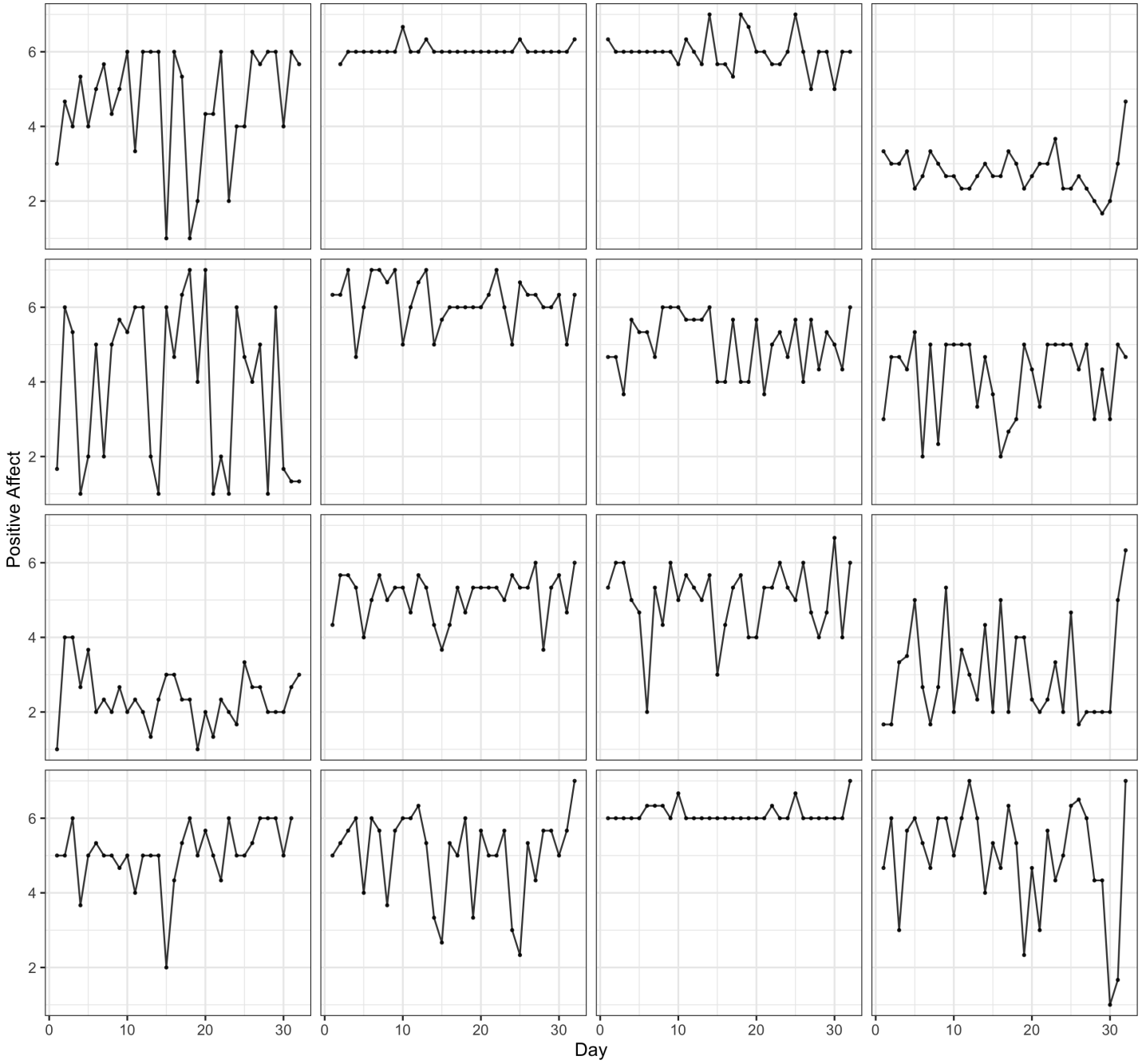
Et cetera…

True proportion: 0.7

Logistic regression:  0.60 (0.55, 0.65)

GEE: 0.60 (0.52, 0.68)

Mixed effects model: 0.73 (0.64, 0.81)


Positive Affect
Day
2
4
6
0
10
20
30

# What about simpler approaches?

- Analyze with a standard linear model (bad)
- Reduce each person's data to a single point (e.g. rate of change), then analyze with a standard linear model
  - Statistically valid, but lose a lot of data
- Reduce each person's data to two points (initial and final), then analyze with a paired t-test
  - Statistically valid, still lose a lot of data
- Compare average responses at each time point (e.g. repeated measures ANOVA)
  - Need data at regular time points

# Other types of data that violate independence assumption

- Multi-site clinical trial data
- Laboratory data with multiple batches, different technicians, different machines
- Education data
- Geographically clustered data
- **Not necessarily longitudinal, but the same difficulties arise, and the same types of models are used**

# Summary

- Ignoring longitudinal structure during analysis produce misleading results
  - Biased effect estimates
  - Underestimated standard errors
- Mixed effects models and GEEs can help solve these problems
  - Each have advantages/disadvantages
  - GEE works well for estimation of average effects with balanced data
  - Mixed models give more subject-level information
- Data reduction methods can be statistically valid, but don't allow you to use all your data
  - Usually better to use a comprehensive modeling approach
- Consult a statistician!

# Thank you!

Biostatistics Collaboration Center

Search web or people

About Us ▾ Research Services and Support Education People

**Biostatistics Collaboration Center**

REQUEST AN APPOINTMENT

## Expertise in biostatistics, statistical programming and data management

Since 2004, the Biostatistics Collaboration Center (BCC) has partnered with Northwestern investigators at every level – from residents and postdoctoral fellows to junior faculty and well-established senior investigators.

### Meet Our Team

Our faculty and staff have collaborated with basic science, clinical, and health services investigators in over 60 Northwestern units across Chicago and Evanston campuses, and with all of NU's clinical partners. Meet our team and learn more about our expertise.

OUR PEOPLE