Welcome to 140.655!

This week in the course:

- You should look over the course syllabus and watch the short “Introduction” panopto, which highlights the structure of the class, lab sessions and key due dates
- Lecture 1
- Prior to next Monday (1/27), you should complete Lab 0.
In this Introduction, we will

- Introduce longitudinal data along with key definitions and study designs
- Understand why longitudinal data requires special methods (in-class exercise)
- Understand factors that contribute to correlation in longitudinal data

Introduction

- This course describes statistical methods for the analysis of longitudinal data, with a strong emphasis on applications in the biological and health sciences
Definitions

- **Univariate data**: each subject gives rise to a single measurement
- **Multivariate data**: each subject gives rise to a vector of measurements
- **Longitudinal data**: each subject gives rise to a vector of measurements, but these represent the same response measured at a sequence of observation times
  - Repeated responses over time on independent units (persons)

- **Longitudinal study**: people are measured repeatedly over time
- **Cross-sectional study**: a single outcome is measured for each individual
- Sometimes, cross-sectional and longitudinal studies can answer the same question
- One clear advantage of a longitudinal design is the ability to separate changes within each subject ("aging" or "longitudinal" or "within-subject") effects from differences among subjects ("cohort" or "among-subject") effects
  - We can’t separate within and among-subject effects in cross-sectional studies
Illustrative Example

- Observational study
- Primary exposure: age
- Outcome: reading ability

Reading ability appears to be poorer among older children.

Let’s assume that the reading ability of each child has been measured twice.

In this case, we observe a negative cohort effect; while the aging effect is positive.
• Let’s assume that the reading ability of each child has been measured twice
• Another possible pattern!

Here we observe both a negative cohort and a negative aging effect

- LDA can distinguish between changes over time within individual & differences among people in their baseline levels
  - changes over time within subjects (aging effects)
  - differences among subjects in their baseline levels (baseline or cohort effects)
  - This partitioning of the baseline and aging effects would apply to any exposure that varies within a person over time (e.g. mental health, physical function, quality of life).
  - Can also think of baseline effects = among-subject effects, aging effects = within-subject effects.
• Depending on the design and the primary objective, we may not always have information on cohort effects
• What are advantages of LDA in this setting?
  – We can measure change within an individual over time
  – Ability to describe dependent of change on predictor variables
  – Measuring data from the same person over time controls for time-invariant characteristics of the person
  – In exposure studies (baseline and then follow-up measures on same subjects), the subjects serve as their own control, reducing effect of confounding by time-invariant variables.

Examples of Longitudinal Study Designs
• Pre-post designs
  – Specific event of interest; e.g. policy change, diagnosis of disease, hospitalization
  – Single measure before and after the event of interest
  – Use repeated measures on each unit to control for time-invariant unit characteristics
  – Focus is to estimate the aging or longitudinal effect or within-unit (i.e. change from pre to post)
**Examples of Longitudinal Study Designs**

- **Cohort studies**
  - May recruit based on population subset of interest (physicians health study) OR clinical event of interest (e.g. clinical cohort of newly diagnosed HIV patients)
  - Typically contain information on both cohort and ageing effects
  - Primary focus: within-subject change and how within-subject change varies between subjects with different exposures of interest

**Examples of Longitudinal Study Designs**

- **Randomized trial** – important special case of a cohort study
  - May contain information on among-subject effects depending on design (e.g. inclusion/exclusion criteria.)
    - This is NOT typically of interest
  - Primary focus: within-subject change and how within-subject change varies among the treatment groups
Analytic strategy

- In pre-post design, analysis can focus on change (post – pre measurement)
- In designs with > 2 observations, how do we quantify changes in the outcome of interest over time?

Analytic strategy: Example

- Randomized trial
- Primary exposures: time and treatment
- Time = 0 clearly defined as initiation of treatment

Values in the figure are mean change from baseline (i.e. average of Day j score – Baseline score, j=0,1,…,7) with corresponding 95% CI

Example (contd.)

• Possible analytic approaches:
  – Compute a summary measure for each subject, compare the mean summary measure across treatment groups

Possible summary measures:

1. Change from baseline to last follow-up (day 7 vs baseline)
2. "Lowest" score on treatment
3. Cumulative change (area under the curve)


Example (contd.)

• Possible analytic approaches:
  – Compare the mean response profile across treatment groups
  – Construct a model with time, treatment and interaction

Possible models:

1. 2-way ANCOVA, treating time as a factor
2. Model the mean change in delirium rating scale vs. continuous time (quadratic, spline)

Test for interaction determines if mean response profiles are the same or different

Longitudinal Data Requires Special Methods, WHY??

• Repeated measurements from the same person over time tend to be correlated
  – assumption of independence is violated

• What if we used standard regression methods anyway (ignore correlation)?
  – Correlation may be of scientific focus
  – Incorrect inference (wrong standard error, significance test, confidence interval)

In-class exercise

• Consider a pre-post design:
  
<table>
<thead>
<tr>
<th>Pre-assessment</th>
<th>Intervention</th>
<th>Post-assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Question: does the mean outcome change from pre to post assessments?

• Notation and hypotheses:
  – m = 50 subjects
  – n = 2, the total number of observations per subject
  – $Y_{ij}$ is the value of the outcome for subject $i = 1, \ldots, m$ at time $j = 1, 2$
    • For subject $i$, the data are $Y_{i1}$ and $Y_{i2}$
  – Let $\mu_1 (\sigma_1)$ and $\mu_2 (\sigma_2)$ denote the population mean (standard deviation) for the pre- and post-assessments
  – H0: $\mu_1 = \mu_2$ vs. H1: $\mu_1 \neq \mu_2$
  – H0: $\mu_2 - \mu_1 = 0$ vs. H1: $\mu_2 - \mu_1 \neq 0$
In-class exercise

- Use the “simulation1.csv” and “simulation2.csv” datasets and fill in the following table using the commands provided in InClass1.do or InClass1.R

<table>
<thead>
<tr>
<th></th>
<th>Simulation1</th>
<th>Simulation2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean (SD) for Y1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample mean (SD) for Y2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Two-sample t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Paired t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- After you fill in the table, answer the following:

1. The goal is to estimate the mean change, \( \mu_2 - \mu_1 \). Is your estimate of the mean change the same or different when you apply the two-sample t-test or the paired t-test?
2. The p-value from the paired t-test is **greater than**, **equal to**, **less than** the p-value from the two-sample t-test.
3. The width of the 95% confidence interval for \( \mu_2 - \mu_1 \) from the paired t-test is **larger than**, **equal to**, **less than** the p-value from the two-sample t-test.
### In-class exercise

<table>
<thead>
<tr>
<th></th>
<th>Simulation1</th>
<th>Simulation2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean (SD) for Y1</td>
<td>90.2 (11.6)</td>
<td>90.1 (10.7)</td>
</tr>
<tr>
<td>Sample mean (SD) for Y2</td>
<td>94.7 (13.8)</td>
<td>94.8 (14.0)</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.34</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Two-sample t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td>4.5 (-0.6, 9.5)</td>
<td>4.7 (-0.2, 9.7)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.083</td>
<td>0.060</td>
</tr>
<tr>
<td><strong>Paired t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td>4.5 (0.3, 8.6)</td>
<td>4.7 (2.4, 7.1)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.036</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

1. The goal is to estimate the mean change, $\mu_2 - \mu_1$. Is your estimate of the mean change the same or different when you apply the two-sample t-test or the paired t-test?

2. The p-value from the paired t-test is **greater than, equal to, less than** the p-value from the two-sample t-test.
In-class exercise

<table>
<thead>
<tr>
<th></th>
<th>Simulation1</th>
<th>Simulation2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean (SD) for Y1</td>
<td>90.2 (11.6)</td>
<td>90.1 (10.7)</td>
</tr>
<tr>
<td>Sample mean (SD) for Y2</td>
<td>94.7 (13.8)</td>
<td>94.8 (14.0)</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.34</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Two-sample t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td>4.5 (-0.6, 9.5)</td>
<td>4.7 (-0.2, 9.7)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.083</td>
<td>0.060</td>
</tr>
<tr>
<td><strong>Paired t-test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated mean difference (95% CI)</td>
<td>4.5 (0.3, 8.6)</td>
<td>4.7 (2.4, 7.1)</td>
</tr>
<tr>
<td>P-value</td>
<td>0.036</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

3. The width of the 95% confidence interval for $\mu_2 - \mu_1$ from the paired t-test is larger than, equal to, less than the p-value from the two-sample t-test.

In-class exercise

- Estimated mean difference is the same, regardless of which method we used
- The calculated 95% CI for the mean difference and p-value depend on statistical method applied; both can not valid!
- When estimating the mean change over time, if we account for the correlation using the paired t-test:
  - Smaller width of the valid 95% CI for the mean difference
  - Smaller valid p-value
  - When the correlation is closer to 1 (i.e. comparing simulation1 to simulation2), the width of the 95% confidence interval is smaller and the p-value is smaller when applying the paired t-test.

So what is the difference between the two-sample and paired t-test procedures?
The paired t-test accounts for this in the calculation of the statistical variance/standard error

\[ \text{Var}(Y_{i1} - Y_{i2}) \]
\[ = \text{Var}(1 \times Y_{i1} + (-1) \times Y_{i2}) \]
\[ = (1)^2 \text{Var}(Y_{i1}) + (-1)^2 \text{Var}(Y_{i2}) + 2(1)(-1)\text{Cov}(Y_{i1}, Y_{i2}) \]

\[ \text{Cov}(Y_{i1}, Y_{i2}) = \frac{\text{Cov}(Y_{i1}, Y_{i2})}{\sqrt{\text{Var}(Y_{i1})\text{Var}(Y_{i2})}} \]

\[ \text{Cov}(Y_{i1}, Y_{i2}) = \sqrt{\text{Var}(Y_{i1})\text{Var}(Y_{i2})} \times \text{Corr}(Y_{i1}, Y_{i2}) \]

\[ \text{Var}(Y_{i1} - Y_{i2}) = \text{Var}(Y_{i1}) + \text{Var}(Y_{i2}) - 2\sqrt{\text{Var}(Y_{i1})\text{Var}(Y_{i2})} \times \text{Corr}(Y_{i1}, Y_{i2}) \]

In class exercise extended

- Assuming the same setting (i.e. pre-post design)
  - \( m = 50 \) individuals
  - Pre sample mean \( \bar{Y}_{i1} = 90 \), sample sd \( s = 11 \)
  - Post sample mean \( \bar{Y}_{i2} = 95 \), sample sd \( s = 14 \)

**two-sample t-test:**
\[ \frac{95 - 90}{\frac{11^2}{10} + \frac{14^2}{19}} = \frac{5}{2.52} = 1.99, p = 0.0499 \ (df = 98) \]

**paired t-test:**
\[ \frac{95 - 90}{\sqrt{\frac{11^2}{50} + \frac{14^2}{50} - \frac{2(11)(14)}{50} \text{Cov}(Y_{i1}, Y_{i2})}} \]
### In class exercise extended

<table>
<thead>
<tr>
<th>Corr</th>
<th>SE</th>
<th>T</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2.52</td>
<td>1.99</td>
<td>0.050</td>
</tr>
<tr>
<td>0.2</td>
<td>2.26</td>
<td>2.21</td>
<td>0.032</td>
</tr>
<tr>
<td>0.4</td>
<td>1.97</td>
<td>2.54</td>
<td>0.014</td>
</tr>
<tr>
<td>0.6</td>
<td>1.63</td>
<td>3.07</td>
<td>0.003</td>
</tr>
<tr>
<td>0.8</td>
<td>1.19</td>
<td>4.21</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Relative efficiency \( \frac{2.52^2}{2.26^2} = 1.24 \)

If we assume the data are independent, we estimate that the variance of the change in sample means is 24% larger than expected if the correlation is 0.2.

Positive correlation makes estimated “differences” more precise than they would be with independent data. Why?

### Implications

In this case (two-sample test), ignoring the correlation would lead to:

- Overly pessimistic estimate of variance
- Standard errors that are wrong - too large
- Test statistics that are wrong - too small
- Confidence intervals that are wrong - too wide
- P-values that are wrong - too large
Where does the correlation in longitudinal data come from?

- a) Between individual heterogeneity in average responses
- b) Within individual biological variation
- c) Measurement error

Where does the correlation in longitudinal data come from?

Here let:
\[ i = 1, 2 \]
\[ j = 0, 1, 2, 3, ... , 5 \]
\[ Y_{ij} \] is the response for subject i at time j

Consider the correlation between \( Y_{i1} \) and \( Y_{i2} \)

\[
\text{Corr}(Y_{i1}, Y_{i2}) = \frac{\text{Cov}(Y_{i1}, Y_{i2})}{\sqrt{\text{Var}(Y_{i1})\text{Var}(Y_{i2})}}
\]

\[
\text{Cov}(Y_{i1}, Y_{i2}) = \frac{\sum_{i=1}^{m}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{m-1}
\]
Summary of Lecture 1

- Longitudinal data is data collected on repeated occasions for the same unit of analysis (typically a person in our studies)
  - May be randomized trial or observational study
- Primarily interested in the aging (or withinsubject) effect
  - Advantage of the design in the built-in controls for time-invariant confounders
  - Depending on the design, we may also have information about the cohort effect

Summary of Lecture 1

- In analyzing longitudinal data, we must account for correlation of observed data from the same person
  - Statistical inferences can be wrong if we assume data are independent
  - The degree of correlation depends upon the relative size of variation among subjects, within subjects, and measurement error
What is coming up next?

• We will discuss two approaches for constructing regression models for longitudinal data:
  – Marginal model vs. conditional model (random effects model)
• We will discuss exploratory data analysis for the mean and variance
  – Applies to both the marginal and conditional model settings
• We will discuss exploratory data analysis for the correlation structure
  – Applies most to the marginal model setting

• Lab 1 Wednesday Jan 29th, Lab 1 quiz due Friday Jan 31