


Design of Prospective Studies

Kyoungmi Kim, Ph.D.

June 14, 2017

This seminar is jointly supported by the following NIH-funded centers:

UCDAVIS CLINICAL AND TRANSLATIONAL SCIENCE CENTER

UCDAVIS MIND INSTITUTE

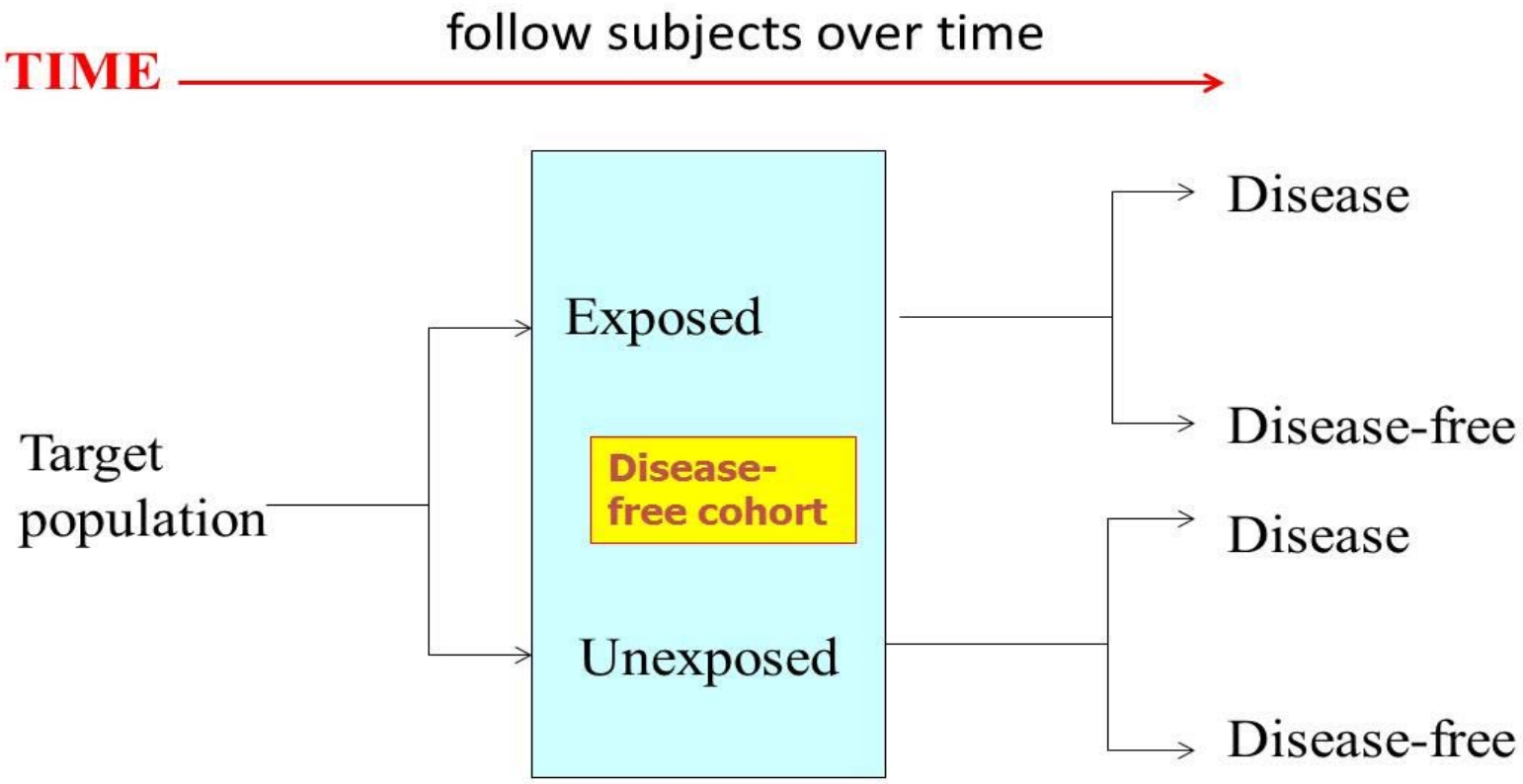
UCDAVIS COMPREHENSIVE CANCER CENTER

UCDAVIS ENVIRONMENTAL HEALTH SCIENCES CENTER

Seminar Objectives

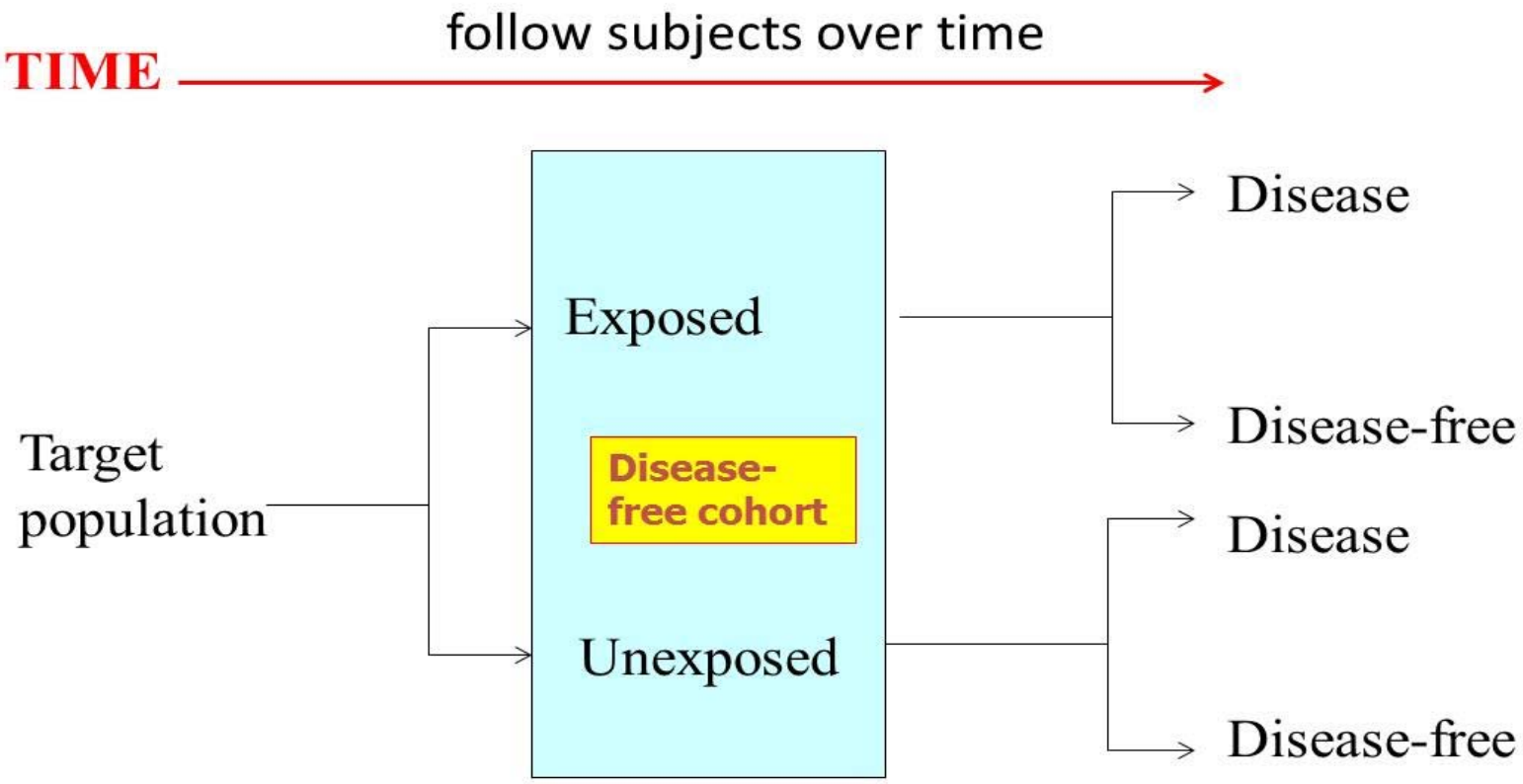
- **Discuss about design of prospective longitudinal studies**
- **Understand options for overcoming shortcomings of prospective studies**
- **Learn to determine how many subjects to recruit for a follow-up study**

What is a prospective study?



Study subjects of disease-free at enrollment are followed for a period of time and periodically checked for progress to see **who/when** gets the outcome in question -thus be able to establish a temporal relationship between exposure & outcome.

What is a prospective study?



During the follow-up, data is collected on the factors of interest, including:

- When the subject develops the condition
- When they drop out of the study or become “lost”
- When they exposure status changes
- When they die

Example: Framingham Study

- An original cohort of 5,209 subjects from Framingham, MA between the ages of 30 and 62 years of age was recruited and followed up for 20 years.

- A number of hypotheses were generated and described by Dawber et al. in 1980 listing various presupposed risk factors such as increasing age, increased weight, tobacco smoking, elevated blood pressure and cholesterol and decreased physical activity.

- It is largely quoted as a successful longitudinal study owing to the fact that a large proportion of the exposures chosen for analysis were indeed found to correlate closely with the development of cardiovascular disease.

Example: Framingham Study

- **Biases exist:**
  - It was a study carried out in a single population in a single town, brining into question the generalizability and applicability of this data to different groups. However, Framingham was sufficiently diverse both in ethnicity and socio-economic status to mitigate this bias to a degree.
  - Despite the initial intent of random selection, they needed the addition of over 800 volunteers to reach the pre-defined target of 5,000 subjects- thus reducing the randomization.
  - They also found that their cohort of patients was uncharacteristically healthy.

What is a Longitudinal Study?

- employ continuous or repeated measures to follow particular individuals over prolonged periods of time- often years or decades.

- generally observational in nature, with data being collected on any combination of exposures and outcomes, without any external influenced being applied.

- particularly useful for evaluating the relationship between risk factors and the development of disease, and the outcomes of treatments over different lengths of time (e.g., short-term vs. long-term effect).

What kinds of research questions require longitudinal studies?

Questions about **systematic change over time**

- Espy et al. (2000) studied infant neurofunction
- 40 infants observed daily for 2 weeks; 20 had been exposed to cocaine, 20 had not.
- Infants exposed to cocaine had lower rates of change in neurodevelopment.

1. **Within-person descriptive:** How does an infant's neurofunction change over time?
2 **Within-person summary:** What is each child's rate of development?
3 **Between-person comparison**: How do these rates vary by child characteristics?

Individual Growth Model

Questions about **whether and when events occur**

- South (2001) studied marriage duration.
- 3,523 couples followed for 23 years, until divorce or until the study ended.
- Couples in which the wife was employed tended to divorce earlier.

1. **Within-person descriptive:** Does each married couple eventually divorce?
2. **Within-person summary:** If so, when are couples most at risk of divorce?
3. **Between-person comparison:** How does this risk vary by couple characteristics?

Time-to-Event Survival Analysis

Longitudinal Study Designs

- **Trend Study**
  - an investigator samples randomly from a population over time, with different individuals – purely observational
- **Cohort Study**
  - an investigator randomly samples from a population **selected on the basis of specific exposure characteristics (risk factors)**
- **Cohort-Sequential Samples Study**
  - an investigator repeated measures a cohort group (e.g., people 60 years of age) over time, adding a new cohort (e.g., new 60-year olds) in each sequence in order to differentiate age effects and cohort effects (differences in people resulting from characteristics of the era or social environment in which they grew up)

Example: Trends Over Time

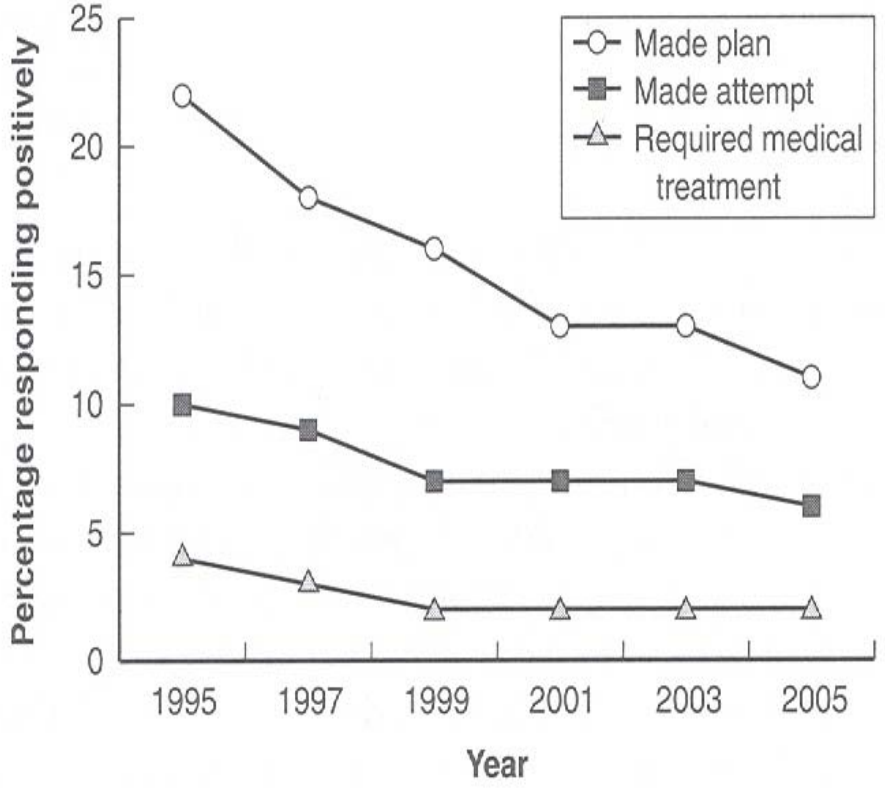
Vermont Trend Study on suicidal ideation

- **Question**
  - How many students have suicide ideation in the State of Vermont?
  - What is the pattern over time?

- **Method**
  - Vermont studied middle- and high-school students regarding their thoughts of suicide since 1995 for 10 years.

Example: Trends Over Time

- **Results**
  - Over 20% of students claimed to have made a plan for suicide in 1995, but the number declined over the next decade to just over 10%.
  - There was also a decline in the number of students who required medical treatment.
- **Conclusion**
  - Although the percentage declined from 1995 to 2005, it would be wise to think about interventions to prevent suicides



Example: Cohort-Sequential Study

Cohort-sequential study on alcohol use in children

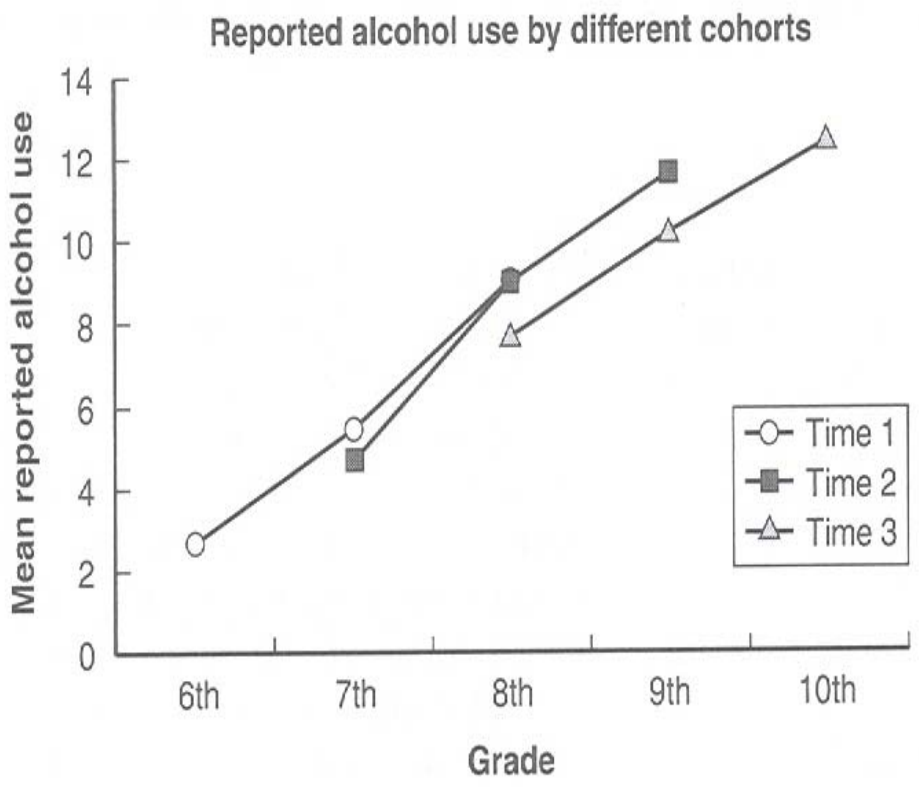
- **Question**
  - When children begin using alcohol, does their consumption increases over time?

- **Method**
  - In a cohort-sequential design, investigate the alcohol intake of different age groups, following a group of 6th grade, 7th grade, and 8th grade children over three years.

Example: Cohort-Sequential Study

- **Result**
  - Among children using alcohol, as they get older, their alcohol use increases in similar ways

- **Conclusion**
  - When children use alcohol early (e.g., in the sixth grade), their consumption increases in consistent ways



Longitudinal Studies: Advantages

- **Able to identify and relate events to particular exposures. Establishing sequence of events**
- **Excluding recall bias in participants, by collecting data prospectively and prior to knowledge of a possible subsequent event occurring**
- **Able to correct for the “cohort effect”- that is allowing for analysis of the individual time components of cohort (range of birth dates), period (current time), and age (at point of measurement).**

Longitudinal Studies: Disadvantages

- **Incomplete and interrupted follow-up of individuals, and attrition with loss to follow-up over time**
- **The potential for inaccuracy in conclusion if statistical analysis fails to account for the intra-individual correlation of repeated measures, incomplete data, and time varying variables.**
- **Generally increased temporal and financial demands associated with this approach.**

Temporal Data Collection

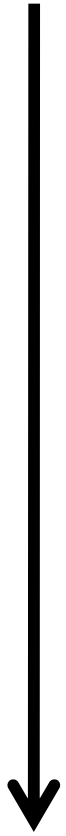
A simple example of a longitudinal study dataset (repeated follow-ups)

ID	*Year*	*Age*	*Gender*	*Employment*	*Marital Status*
1	1991	16	Female	Student	Single
1	1992	17	Female	Student	Single
1	1993	18	Female	Student	Single
1	1994	19	Female	Unemployed	Single
1	1995	20	Female	Employed (ft)	Cohabiting
1	1996	21	Female	Employed (ft)	Cohabiting
1	1997	22	Female	Employed (ft)	Cohabiting
1	1998	23	Female	Maternity Leave	Married
1	1999	24	Female	Family Care	Married
1	2001	25	Female	Employed (pt)	Married

Longitudinal Data Structures

Time invariant variables

ID	Year	Age	Gender	Employment	Marital Status
1	1991	16	Female	Single	
1	1992	17	Female	Single	
1	1993	18	Female	Single	
1	1994	19	Female	Single	
1	1995	20	Female	Cohabiting	
1	1996	21	Female	Cohabiting	
1	1997	22	Female	Cohabiting	
1	1998	23	Female	Married	
1	1999	24	Female	Married	
1	2001	25	Female	Married	



Longitudinal Data Structures

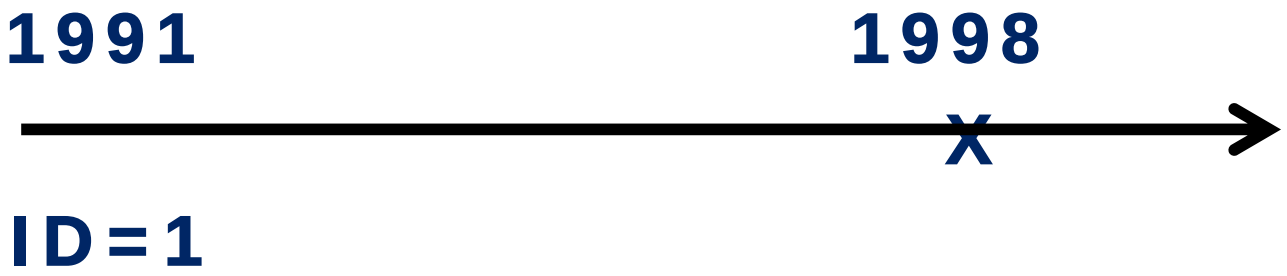
Time varying variables

ID	Year	Age	Gender	Employment	Marital Status
1	1991	16	Female	Student	Single
1	1992	17	Female	Student	Single
1	1993	18	Female	Student	Single
1	1994	19	Female	Unemployed	Single
1	1995	20	Female	Employed (ft)	Cohabiting
1	1996	21	Female	Employed (ft)	Cohabiting
1	1997	22	Female	Employed (ft)	Cohabiting
1	1998	23	Female	Maternity Leave	Married
1	1999	24	Female	Family Care	Married
1	2001	25	Female	Employed (pt)	Married

Longitudinal Data Structures

Time to an event

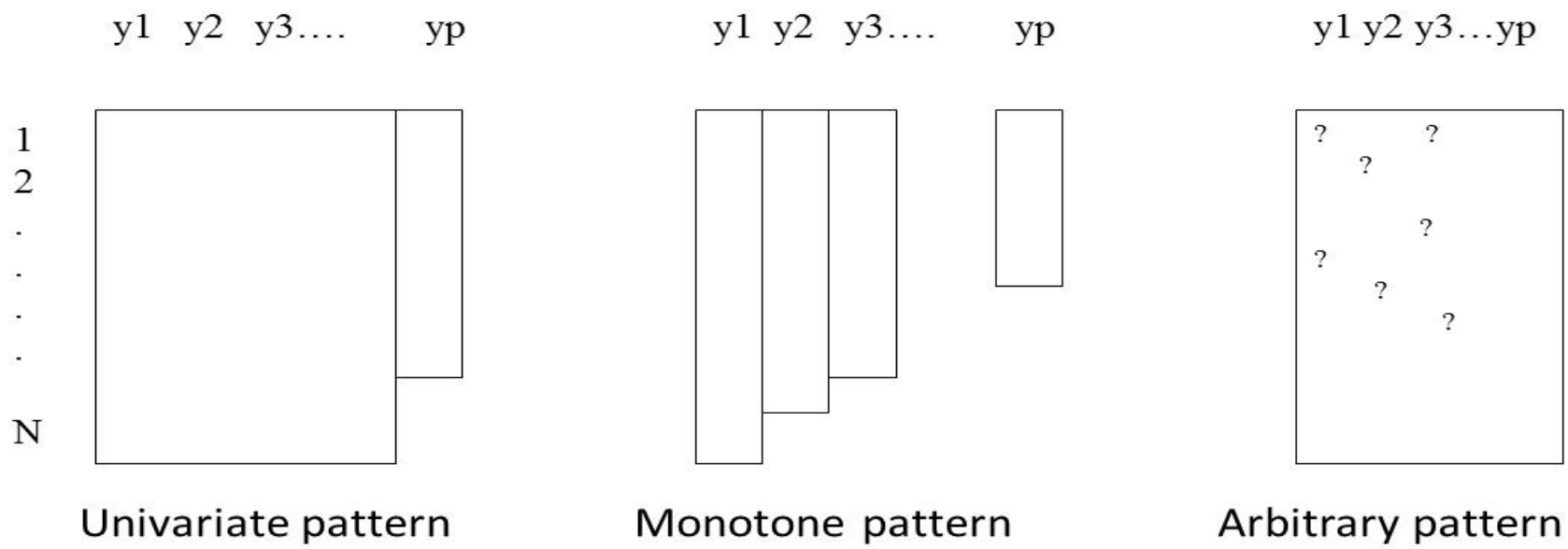
Time to first childbirth



Issue in Longitudinal Research: Nonresponse

- **Attrition**
  - The loss of participants in longitudinal studies due to death, disappearance (incomplete), loss of interest (drop-out), etc
  - Attrition is one of the most serious methodological problems associated with longitudinal studies as data is partially missing.
  - Those who drop out might differ in important ways from those who remain, so conclusions based on studies with significant attrition can be biased.
- **Missing data**
  - Missing is where one or more of the sequences of measurements are incomplete

Different Patterns of Nonresponse



- **Univariate pattern**
  - For some items we have full observations, and for some others we have missing values (no answers). These may be fully or partially missing
- **Monotone pattern**
  - May arise in longitudinal studies with attrition
  - If an item is missing in some period, it continues to be missing in the next periods
- **Arbitrary pattern**
  - Any set of variables may be missing for any unit

Types of Missing

- **Missing completely at random (MCAR)**
  - Whether the data are missing is entirely unrelated statistically to the values that would have been observed.
- **Missing at random (MAR)**
  - Missingness is statistically unrelated to the variable itself.
  - However, it may be related to other variables in the data set.
- **Missing not at random (MNAR)**
  - Nonignorable missing data
  - Missingness conveys probabilistic information about the values that would have been observed.

Example

- **Let us take two variables, education and income.**
- **Education has no missing values while income has.**

- **MCAR**
  - The missing values of income are dependent neither on education nor on income

- **MAR**
  - The missing values of income are dependent on education. That is, education can predict the missing values in income.

- **MNAR**
  - The missingness in income is not independent of the values of the missings, controlling for the prediction of education. That is, for example, high income values are more often missing than low income values.

Methods for Addressing Missing Data in Longitudinal Research

1. Listwise deletion (LD)

- Also known as complete case analysis
- Deleting every case which has any missing value
- How to check?
  - Divide the data observations into groups (e.g., disease vs non-disease) and test whether the groups differ in the proportions of observations with missing data for particular variables.
  - Any such difference indicates the data may be not missing at random.
- Advantage:
  - consistent solution & easy to use
- Disadvantages:
  - not efficient, and causes often a drastic reduction in sample size, especially in studies where multiple variables are involved.

2. Pairwise deletion

- Also called available case analysis
- Calculate each correlation separately
- This method excludes an observation from the calculation when it is missing a value that is needed for the computation of that particular correlation.
- For example,

	1	2	3	4	5	6	7	8	9	10
X	x	x	x	x	x	.	.	.	.	.
Y	x	x	x	x	x	x	x	x	x	x
Z	.	.	.	.	.	x	x	x	x	x

- Advantages:
  - smaller loss of cases than in the LD
- Disadvantages:
  - not efficient, and could create problems in estimation, because the observed correlation matrix may not be positive definite (as correlation between X & Z is invalid).
  - There is no defined N for the sample, since it depends on the computed pair.

Reweighting

- In some non-MCAR situations, it is possible to reduce biases by applying weights.
- After incomplete cases are removed, the remaining complete cases are reweighted so that their distribution more closely resembles that of the full sample or population with respect to auxiliary variables.
- It requires some model for the probabilities of response to calculate the weights.
- Better for the univariate and monotone missing patterns.
- Become complicated to apply if missing is in an arbitrary pattern.

Imputation Methods

- **Imputation means filing in missing values with plausible values, and continuing with the analysis**

- **Advantages:**
  - potentially more efficient than discarding the unit.
  - Prevention of loss of power due to decreasing sample size

- **Disadvantages:**
  - imputation may be difficult to implement well.

- **Imputing unconditional means:**
  - average is preserved, but distributional aspects such as variance are distorted.

- **Hot deck imputation:**
  - filling in data with values from actual respondents randomly. However, selection of ‘matched’ cases may not always be possible- especially in multilevel data
  - It preserves the variable’s distribution, but the method still distorts correlations and other measures of association.

- **Imputing conditional means by regression:**
  - the model is first fit for cases to which y is known. After we have a regression parameter from X to Y, we use it to forecast missing values of Y by known values of X. It is almost optimal with some correlations for standard errors.
  - Not recommended for analyses of covariances or correlations, since it overstates the relation between Y and X.

- **Imputing form a conditional distribution:**
  - distortion of covariances can be eliminated if each missing value of Y is replaced not by a regression prediction but by a random draw from the conditional or predictive distribution of Y given X plus an error term.

Imputation Built on Maximum Likelihood Method

- **Full information maximum likelihood (FIML)**
  - FIML is a direct method in the sense that models parameters and standard errors are estimated directly from the available data.
  - Missing points are not estimated or imputed, and are essentially treated as values that were never intended to be sampled.
  - Advantage:
    - the algorithm uses all the available information, and the method is both consistent and efficient for MAR.
  - Disadvantage:
    - the method is model dependent, as it uses information only from variables in the model (different variables in the model-different results)

Statistical Analyses

- **Commonly applied approaches for longitudinal analysis are:**
  - **Analysis of Variance (ANOVA)**
  - **Multivariate Analysis of Variance (MANOVA)**
  - **Mixed-effect regression model**, focusing specifically on individual change over time while accounting for variation in the timing of repeated measures and missing data
  - **Generalized estimating equation (GEE) models**, relying on the independence of individuals within the population to focus primarily on regression data
  - **Survival Analysis with Cox proportional Hazard models**, focusing on time to event.

- **Details will be discussed next month.**

Power Analysis

- Power computations for longitudinal studies are doable, but depend on parameters that may not be well known.
- Reliability of trend coefficients
- When parameters such as these are known, the computations are straightforward, but there is relatively little information about them that can be used for planning
- To make matters worse, the values of some parameters (such as reliability) depend on the number of measures
- Thus it is often necessary to rely on values of variance components

Power Analysis

- Still some generalizations are possible
  - Power increases with the number of measures
  - Power increases with the length of time over which measures are made
  - Power increases with the precision of each individual measure
- Pilot data (or data from related studies, perhaps non-experimental ones) is more important in planning longitudinal studies.
- Longitudinal studies to look at growth trajectories are attractive, but this is an area at the frontier of practical experience.

Summary

- **Longitudinal methods may provide a more comprehensive approach to research, that allows an understanding of the degree and direction of change over time.**

- **One should carefully consider the cost and time implications, while ensuring completeness in design and process.**

Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30pm in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **MIND IDDRC Biostatistics Office Hours**
  - Monday-Friday at MIND
  - Provide full stat support for the IDDRC projects
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4pm in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC – www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center websites