Different Study Designs Generate Different Types of Data: Implications for Modeling

Brady T. West

Review: Where data come from

When fitting statistical models to data, critically important to understand **how the data were generated**:

Review: Where data come from

When fitting statistical models to data, critically important to understand **how the data were generated**:

- From a **carefully designed probability sample**, featuring cluster sampling?
- From a **convenience sample** / non-probability sample?
- From a **longitudinal study**?
- From a **simple random sample**?
- From a **natural / organic process**?

Why Does It Matter?

- When we fit a model to particular variable in set of data…

Goal = **estimate parameters** that best describe the distribution of that variable

means
variances
correlations

Why Does It Matter?

- When we fit a model to particular variable in set of data…

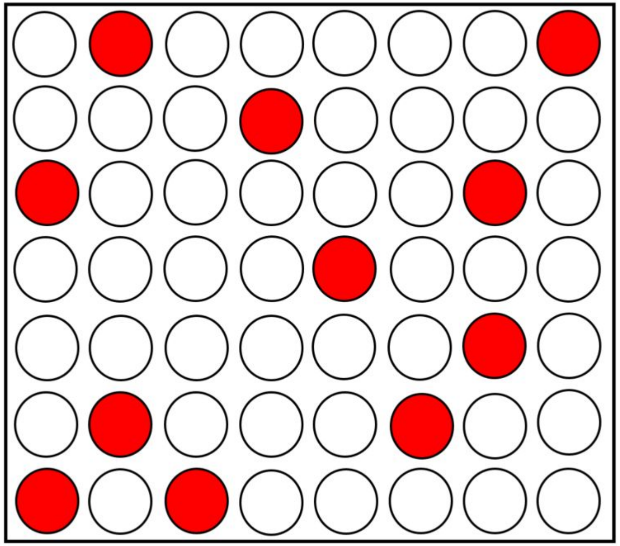
Goal = **estimate parameters** that best describe the distribution of that variable

means
variances
correlations

- If aspects of study design that generated data affect these parameters
  → need to **account for these design aspects** when fitting models!

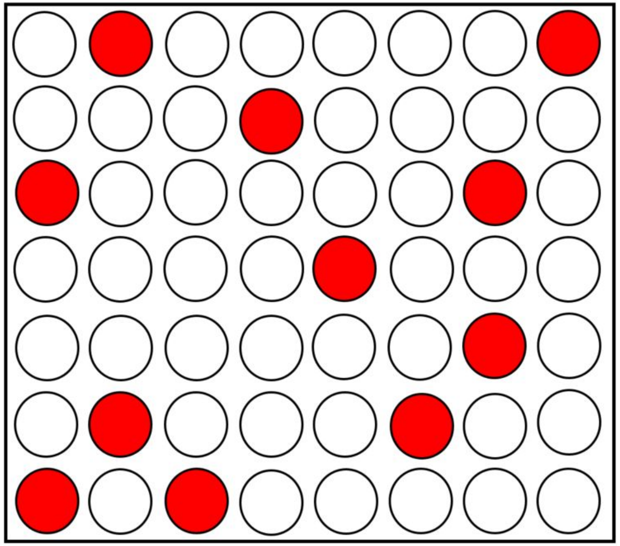
Simple Random Samples

- **Simple random samples (SRS)** from carefully defined populations generally produce observations on variable of interest that are **independent** and **identically distributed (i.i.d.)**



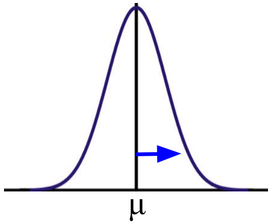
Simple Random Samples

- **Simple random samples (SRS)** from carefully defined populations generally produce observations on variable of interest that are **independent** and **identically distributed (i.i.d.)**
- When fitting models to data from SRS, select distributions for variables with important property that all observations in data are **independent** (unrelated to each other!)



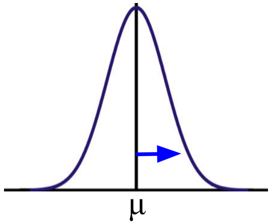
Simple Random Samples Example

Observations on **happiness scale** in SRS come from common normal distribution with some mean and variance, with **zero correlation** between any two randomly selected observations



Simple Random Samples Example

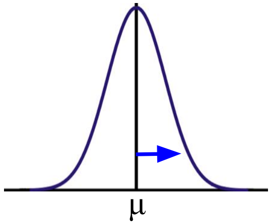
Observations on **happiness scale** in SRS come from common normal distribution with some mean and variance, with **zero correlation** between any two randomly selected observations



Estimate of **standard error** for estimated mean happiness would be computed assuming observations are independent of each other

Simple Random Samples Example

Observations on **happiness scale** in SRS come from common normal distribution with some mean and variance, with **zero correlation** between any two randomly selected observations



Estimate of **standard error** for estimated mean happiness would be computed assuming observations are independent of each other

More unique statistical information
→ smaller SE → more precise estimates!

Simple Random Samples Example

Depending on research question …

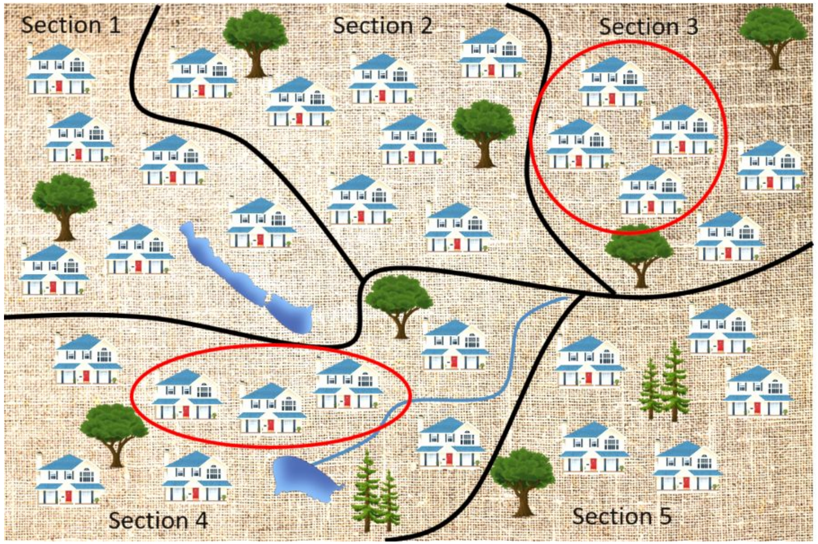
(e.g. *model **difference** in mean happiness between males and females*)

might fit model that **does not** assume observations from same distribution

Example: Mean of normal distribution of happiness scores depends on gender, but once we condition on gender, all observations are independent and have the same variance!

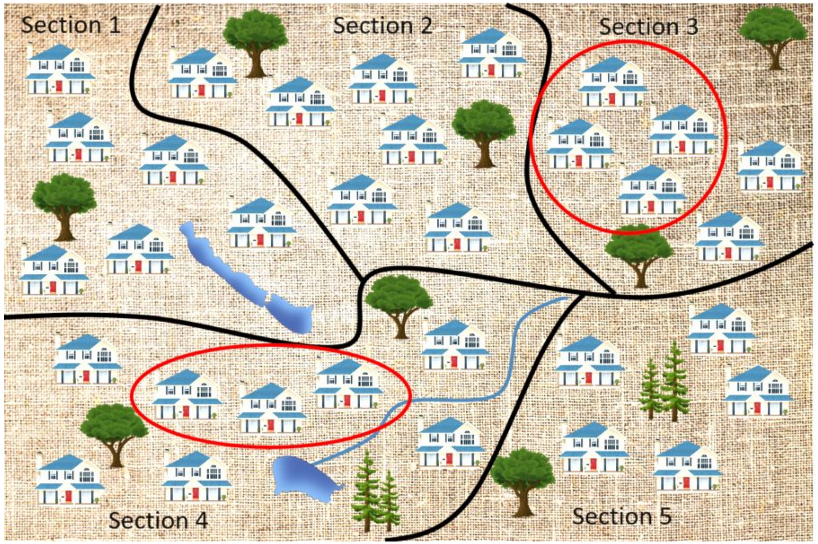
Clustered Samples

- Arise from study designs that generate **clusters of related observations** *(e.g., hospitals, clinics, schools, neighborhoods)*



Clustered Samples

- Arise from study designs that generate **clusters of related observations** *(e.g., hospitals, clinics, schools, neighborhoods)*
- Because observations from same naturally occurring cluster will tend to be similar to each other, **need to account for this correlation** when fitting model to data (unlike models for SRS!)



Clustered Samples Example

- If study design produced several observations of happiness from selected neighborhoods, **observations within neighborhood may well be correlated** with each other
- Model for happiness specified with **additional parameters** capturing this within-neighborhood correlation

Clustered Samples Example

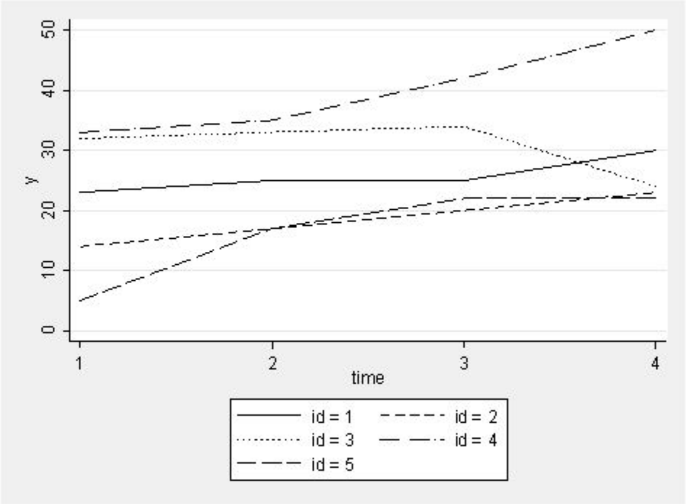
- If study design produced several observations of happiness from selected neighborhoods, **observations within neighborhood may well be correlated** with each other

- Model for happiness specified with **additional parameters** capturing this within-neighborhood correlation
- Standard error of estimated mean would reflect this correlation

Less unique, independent information ⟶ higher SE!

Longitudinal Data

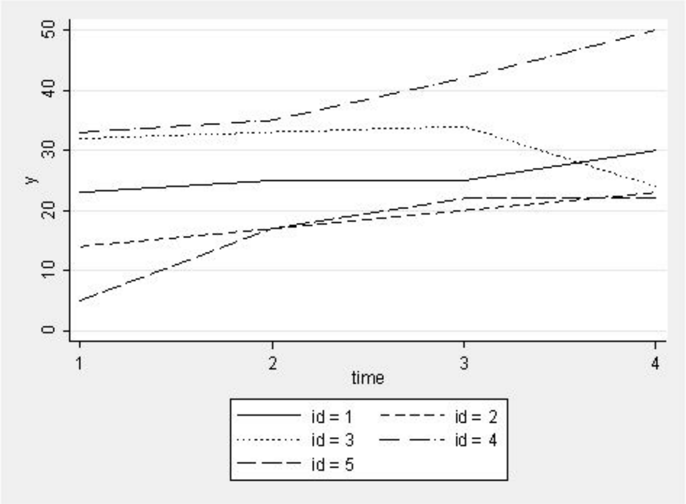
- **Longitudinal data**: repeated measures of same variable, collected from same unit over time → **likely correlated**
- Recorded observations on variable of interest no longer completely independent of each other!



https://stats.idre.ucla.edu/stata/faq/how-can-i-visualize-longitudinal-data-in-stata/

Longitudinal Data

- **Longitudinal data**: repeated measures of same variable, collected from same unit over time → **likely correlated**
- Recorded observations on variable of interest no longer completely independent of each other!
- Models fit to repeatedly-measured variables **need to account for within-unit correlation** (similar to cluster samples!)



https://stats.idre.ucla.edu/stata/faq/how-can-i-visualize-longitudinal-data-in-stata/

Dependent vs. Independent Data

Important distinction between models for:

Dependent data
observations correlated
due to feature of study design
(cluster sampling or
longitudinal measurement)

Dependent vs. Independent Data

Important distinction between models for:

Dependent data	**Independent data**
observations correlated due to feature of study design (cluster sampling or longitudinal measurement)	observations completely independent of each other may/may not arise from common distribution

Dependent vs. Independent Data

Important distinction between models for:

Dependent data	**Independent data**
observations correlated due to feature of study design (cluster sampling or longitudinal measurement)	observations completely independent of each other may/may not arise from common distribution

Want best possible model for a given variable, reflecting important study design features!

What's Next?

- Different **objectives** when fitting statistical models (*inference about relationships* between variables versus *prediction of future outcomes*)

What's Next?

- Different **objectives** when fitting statistical models (*inference about relationships* between variables versus *prediction of future outcomes*)
- Introduce **alternative approaches** to fitting models and making inferences about parameters that define models specified for observed variables:

Frequentist Inference versus Bayesian Inference