What this course won't teach you – count outcomes, survival analysis, correlated data, missing data, and more

REVIEW

Types of Data

- Continuous
- Categorical (focus on binary)
- Count
- Time-to-event
- Qualitative

DISCLAIMER: This course will *not* prepare you to analyze every type of data you encounter.

Count Data

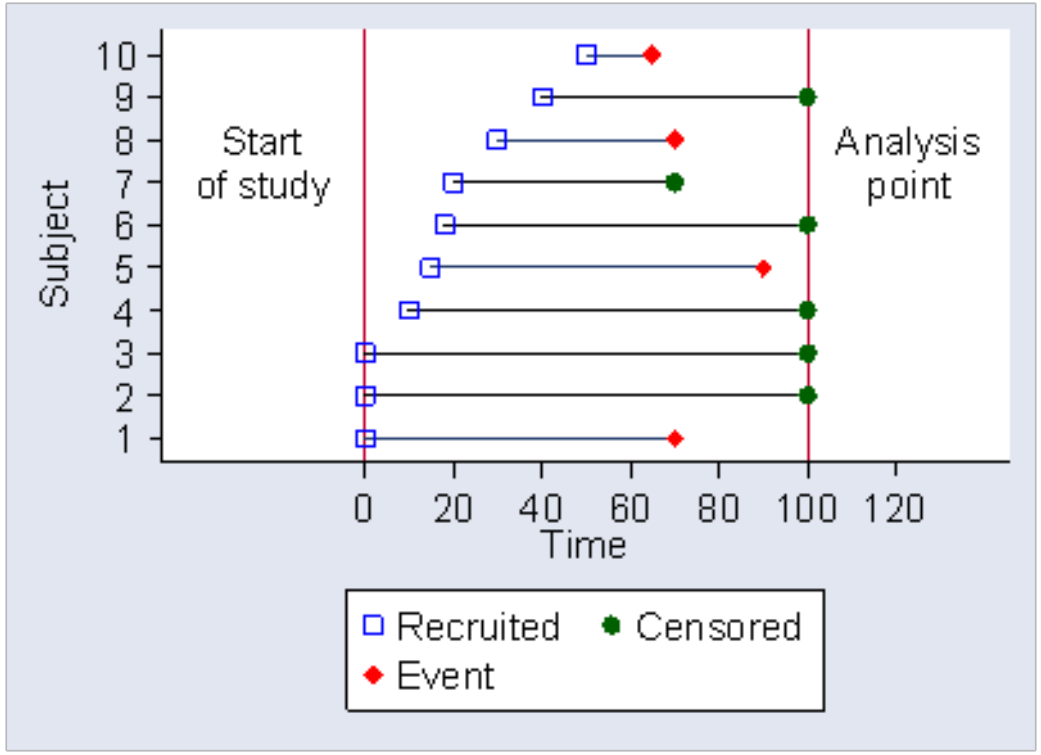
- Count outcomes can be treated like continuous outcomes if they're not close to 0 and there are a large number of possible values
  - Example: number of children vaccinated last year in each state
- This isn't always the case, though
  - Example: number of children each woman has
  - Challenge: linear regression can give us negative predicted values and the assumptions most likely won't be satisfied

Count Data

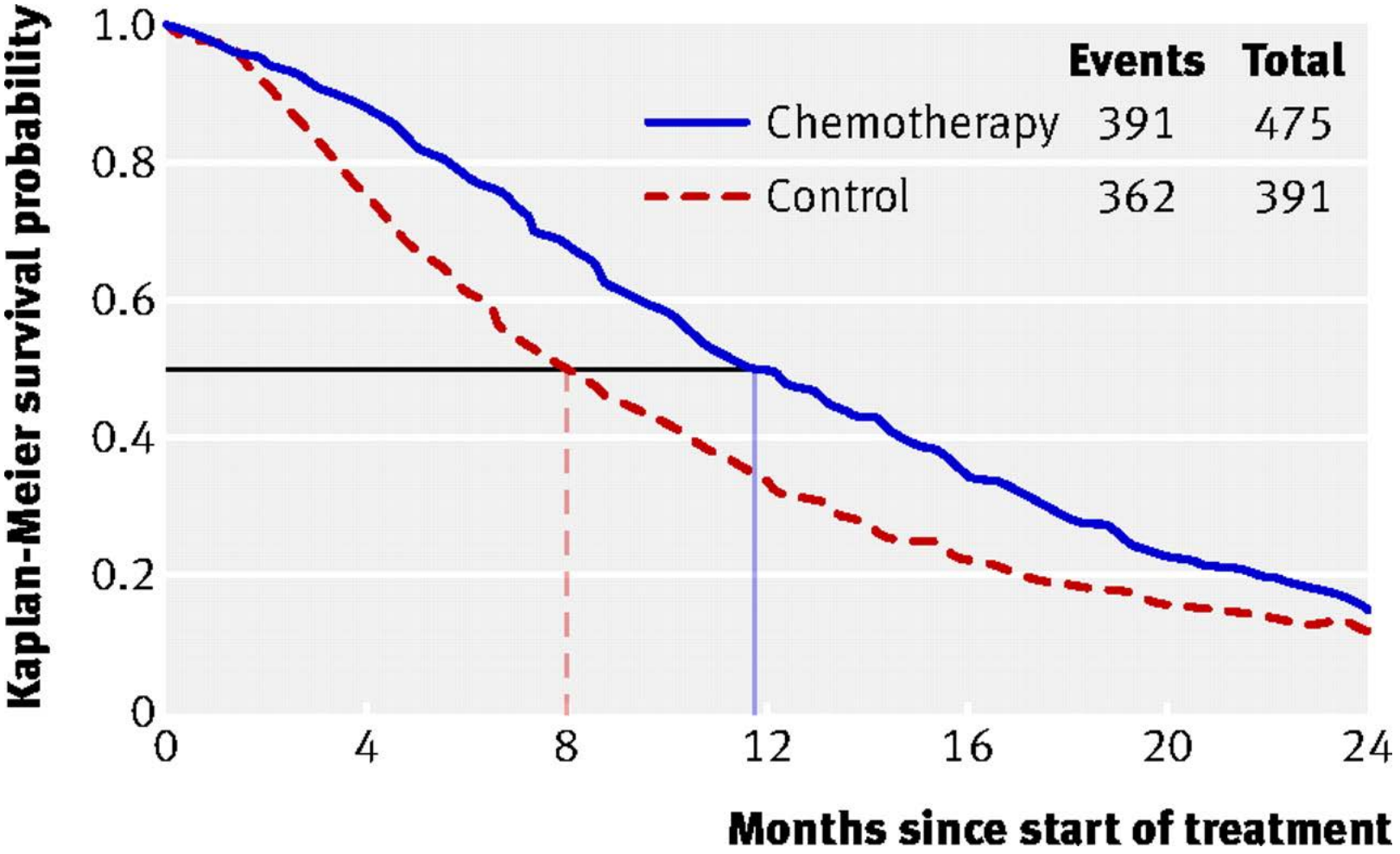
- Generalized linear models (GLMs)
  - Linear regression models and logistic regression models are both special cases of the generalized linear model
- GLMs can accommodate continuous data, binary data, categorical data with >2 categories (nominal or ordinal), count data, etc.
- Requires stronger foundation in probability and distributional theory
  - Look for courses in advanced regression or generalized linear models

Time-to-Event Data

- Interest is not just in whether or not the event (outcome) occurred, but in *when* the event occurred
- Censoring = leaving the study before the event occurs or the event not occurring before the end of the study
  - The presence of censoring makes GLMs inadequate for analyzing time-to-event data



Time-to-Event Data: Survival Analysis



Link to article:
"Survival analysis part I: Basic concepts and first analyses", Clark et al. 2003

Correlated Outcomes

- All of the methods we've discussed in this course assume that the outcomes for each subject are independent
  - Except paired data, but we took care of that by using only the difference between observations for each pair, and we assume that the differences are independent of each other
- Sometimes subjects' outcomes are not independent
  - Does knowing one subject's outcome give you any information at all about what another subject's outcome might be?
  - Are two observations more related to each other than to the other observations?

Correlated Outcomes

- Correlated outcomes often occur by design
  - Repeated measures
  - Longitudinal studies
  - Cluster-correlated data
  - Hierarchical data

- Advanced statistical methods exist that account for this correlation structure
  - Failure to use these (assuming outcomes are independent when they're not) can lead to invalid inferences

Most common type of statistical analysis for correlated outcomes - generalized estimating equations...or...generalized linear mixed models

Missing Data

- Missing data occurs very frequently in biomedical data
  - A lab tech drops a sample
  - A doctor makes an error while recording a clinical value
  - A survey respondent skips a question on the survey
  - A study participant doesn’t show up to a follow-up visit
  - A subject withdraws from the study

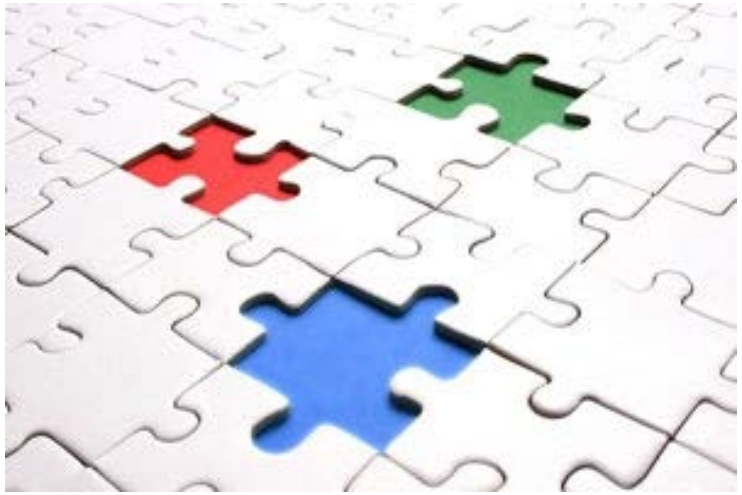
Missing Data

- Complete case analysis = discard subjects who have any missing data
  - Can sometimes be okay if there isn't much missing data
  - Problem: Can lead to throwing out a lot of useful information → loss of power
  - Problem: What if the reasons for missing data are related to what you're trying to study?
    - Example: Study of a drug for treating depression

ID	BMI	Blood glucose	HDL cholesterol	Smoking status
001	28	97	47	NA
002	NA	117	60	Never smoker
003	23	NA	38	Never smoker
004	19	NA	55	Current smoker
005	NA	96	52	Former smoker
006	20	108	NA	Never smoker
007	27	100	49	NA

Missing Data

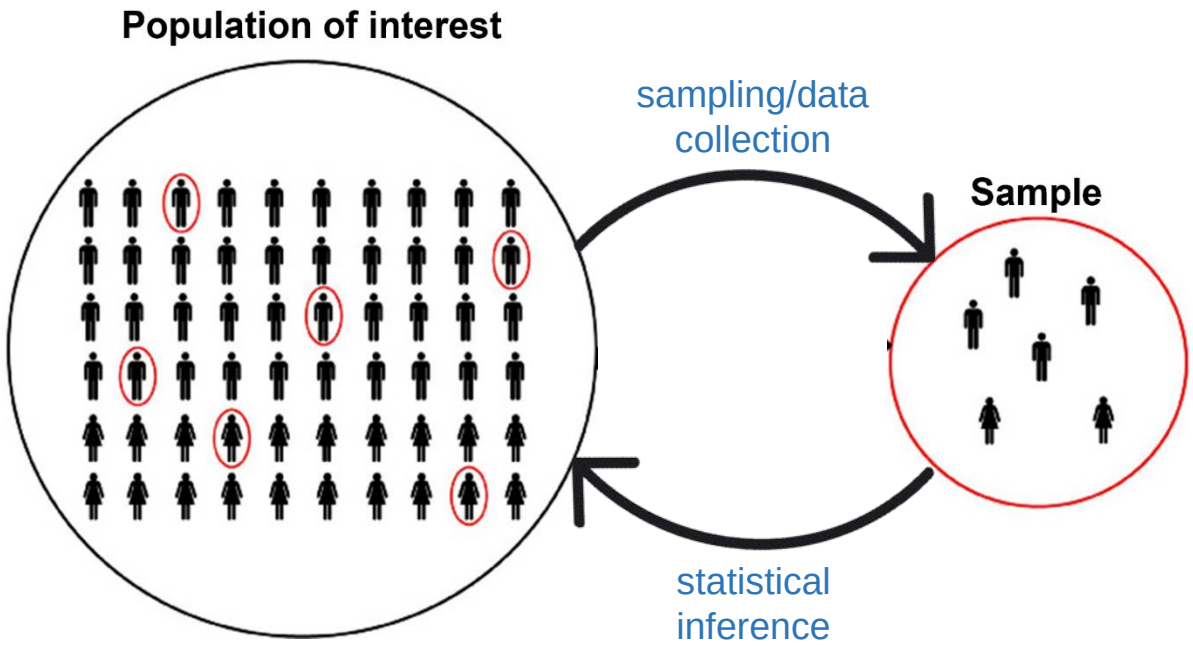
- Advanced statistical methods exist to handle missing data when the reasons for missingness aren't completely random
- Selecting the appropriate method requires knowledge of why the values are missing
- The best way to deal with missing data is to prevent it from happening
  - Think about this when designing studies



Link to article: "The prevention and handling of missing data", Kang 2013

Non-representative Samples

- All of the methods we've discussed in this course assume that the sample you have is representative of the population of interest



Non-representative Samples

- Sometimes the sample isn't selected randomly from the population
  - Certain types of people may be more likely to respond/enroll
  - A random sample may not be logistically possible
- Inferences about the population that are made from a non-representative sample can be misleading or downright incorrect

Non-representative Samples

- **Weighting methods** are used to make the sample representative of the population
  - Example: I know the population is made up of 50% men and 50% women. My sample of 100 people has 25 men and 75 women. If I give each male in my sample 3x the weight that I give to each female, then the sample will be representative of the population.
  - It gets way more complicated than this…

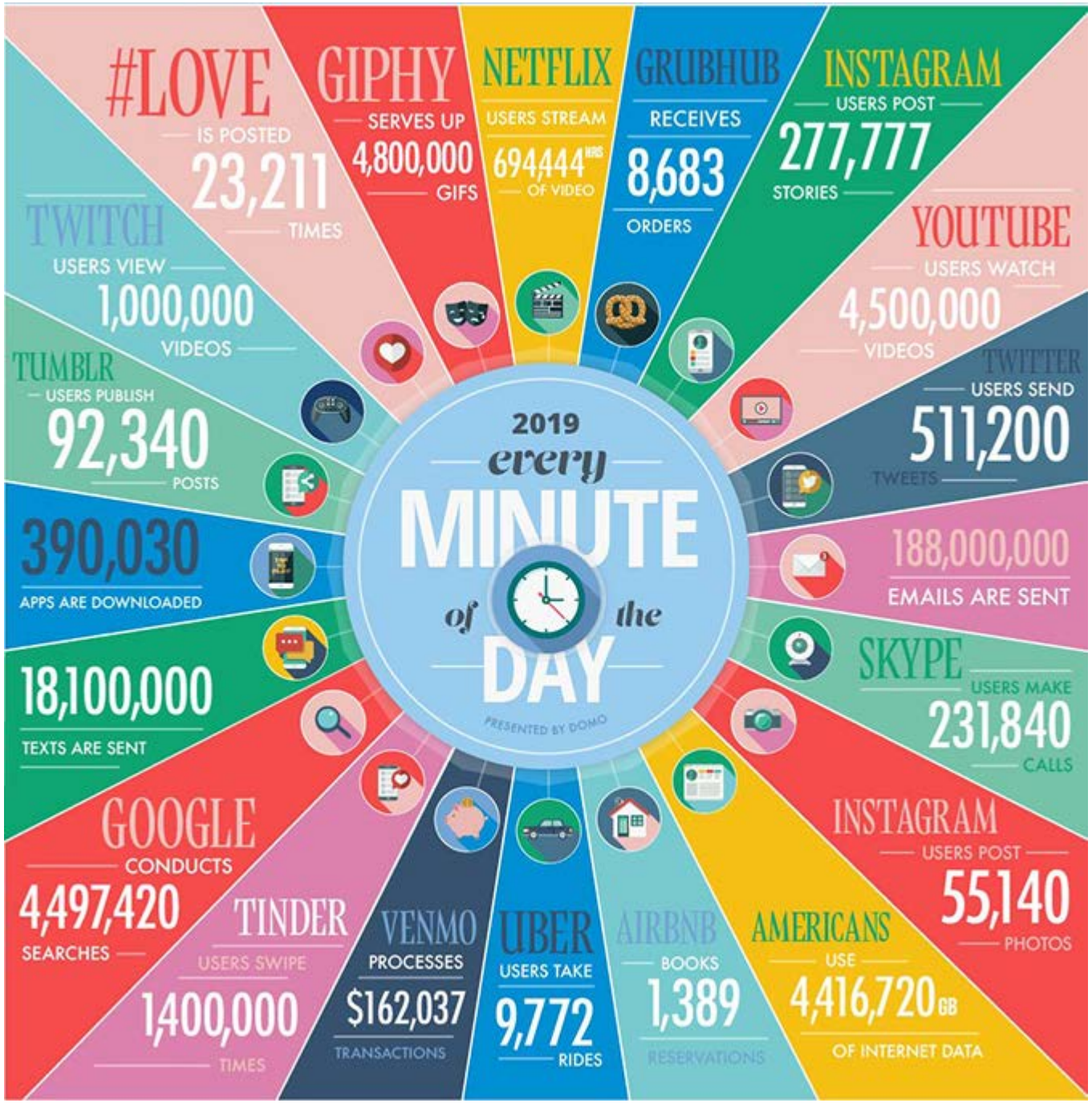
Link to Pew Research Center online article:
"How different weighting methods work",
Mercer, Lau, & Kennedy, 2018

“Big Data”

- Datasets can be “big” if they have a lot of variables or a lot of subjects/observations (or both)
- Many variables
  - Examples: Genetic data, wearable technology data
  - Challenge: The more variables you have in your model, the more subjects you need (for traditional statistical methods)
- Many subjects
  - Examples: Social media data, electronic health record data
  - Challenge: Computing! Storage is not trivial, analyses can take a very long time to run, etc.

Pro tip:
Collaborate with a data scientist and/or bioinformatician

"Over 2.5 quintillion bytes of data are created every single day, and it's only going to grow from there."



“Big Data”

- A relatively new and very popular branch of statistics seeks to extract information from “big” datasets to identify patterns and predict future outcomes and trends
  - Machine learning
  - Artificial intelligence
  - Predictive analytics
  - Data mining

Link to blog post: “Artificial intelligence in medicine: Applications, implications, and limitations”, Greenfield 2019

Important Points

- Know the limits of your statistical knowledge
- Collaborate!
  - You may need to find a biostatistician, data scientist, bioinformatician, or epidemiologist who has knowledge that complements your own
  - Team science is the future (and the present)