Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 1:

Course Structure;

Overview of Scientific Approach

Course Structure

Course Structure

- Instructor: Timothy A. Thornton, Ph.D.
- TAs:
  - Austin Schumacher
  - Subodh Selukar
  - Edward Zhao
- Time and Place
  - Lectures: MWF 9:30-10:20 A.M., HSB T435
  - Discussion Sections
    - 517 AA: Mondays 8:30-9:20 A.M., HSB T639
    - 517 AC: Tuesdays 8:30-9:20 A.M., HSB T359
    - 517 AB: Wednesdays 11:30-12:20 P.M., HSB T639
    - 514 AA and 514 BA: Fridays 8:30-9:20 A.M., HSB T498

Office Hours

- Times
  - Monday 1:00 – 2:00 PM; Tim
  - Tuesday 1:30-3:00 PM; Subodh
  - Wednesday 10:30-11:30 AM; Tim
  - Wednesday 4:00-5:30 PM; Edward
  - Thursday 10:30 – 12:00 PM; Austin
  - Friday 1:00 – 2:30 PM; Austin

- Location:
  - Tim's office hours held in HSB F658
  - TAs office hours held in the Health Sciences Library

Course Web Page

- UW Canvas Course Web Site:
  - Go to http://canvas.uw.edu, login with your UW netid and select BIOST 514 or 517 from the Courses pulldown menu
- Content will include:
  - Syllabus: Please Read! You are responsible for knowing this information
  - Lecture handouts
  - Recordings of lectures
  - Homework assignments and keys
  - Datasets
  - Supplemental materials not discussed in class
    - Handouts
  - Discussion Board: announcements from the instructional team and questions from the students.

Overview of Course

- This course provides an introduction to applied statistics, with an emphasis on medical and epidemiological data.
- The course is designed for graduate students in public health and medical fields.
- The major topics covered are data summary, introduction to statistical inference including “simple” (i.e. one-covariate) regressions, and statistical testing. The role of these in scientific applications is stressed throughout.

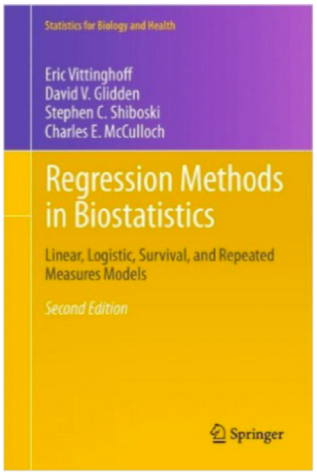
Recording of Lectures

- For students' convenience, recordings of lectures will be posted on the Canvas Course Web Page
  - Posted approximately 24 hours after class

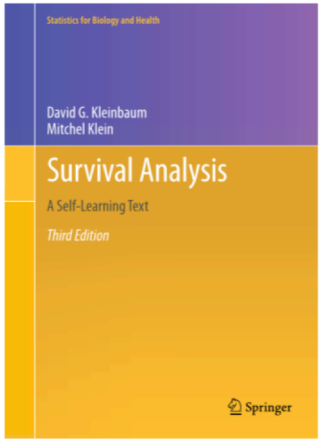
- No guarantees: "Mistakes happen"
  - This is not a distance learning class, and students are responsible for all material in lectures regardless of whether a recording is made available.

Textbooks

- There is no single ‘course book’. But material from these may be helpful;

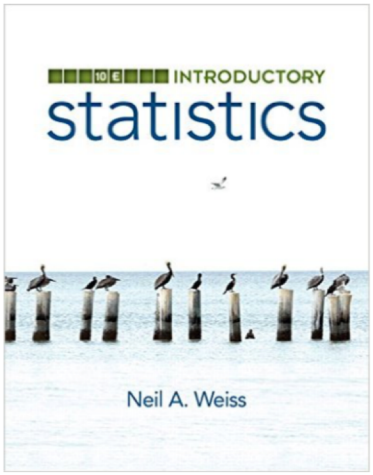


Vittinghoff *et al*
Chapters 2 & 3 cover much of our material – briefly. Also useful for 515/518, next quarter.



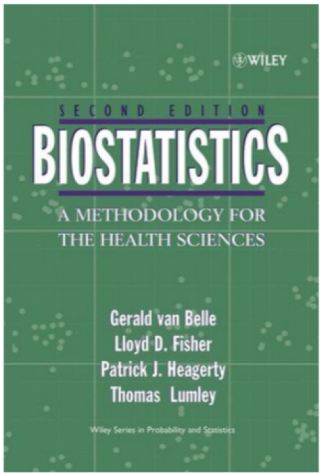
Kleinbaum & Klein
Chapters 1 & 2 introduce analysis of time-to-event outcomes, i.e. *survival* data

Textbooks



Weiss

Good descriptions and many examples of methods we'll discuss, but no time-to-event material



Van Belle *et al*

Extensive coverage (~900 pages) of almost everything in this course, but focuses on 2004-era computing & data resources

Electronic versions for these books are available from the UW Library

R Computer Software

- The R statistical software package will be used for data analysis in this course
- This course does not assumes familiarity with R.
  - The first few discussion sections will focus on learning the basics of R
- R is the ultimate in flexible statistical languages
  - Interactive
  - Many user-supplied functions and extension packages
- Graphical functions generally very good
- Open source and free
- Weekly homework assignments will involve statistical analyses conducted using R.

R Computer Software

- The R statistical software package will be used for data analysis in this course
- This course does not assumes familiarity with R.
  - The first few discussion sections will focus on learning the basics of R
- R is the ultimate in flexible statistical languages
  - Interactive
  - Many user-supplied functions and extension packages
- Graphical functions generally very good
- Open source and free
- Weekly homework assignments will involve statistical analyses conducted using R.

Downloading R and Rstudio

- For this course you will need to download to your computer both R and Rstudio, which is a more user-friendly front-end for R
- R can be downloaded from the following website:

https://www.r-project.org/ (Links to an external site.)

- R studio can be downloaded from the following website:

https://www.rstudio.com (Links to an external site.)

- Note: Both software packages need to be downloaded, as Rstudio requires R to run.

Statistical Software: Comments

- Provides tools for statistical analysis
- Designed for people who know statistics, but do not want to write basic functions
- Tries to be all things to all people
  - Much output that you will not want
- Important: You are responsible for understanding how the tools should be used.

Guiding Principles

- This is a course in biostatistics, not a course in the statistical software package R.
- Will cover how you can get statistics of interest taught in this course from R
- Note that there are often multiple ways to obtain/extract the same results with statistical software.
  - Will typically show one of them
- BIOST 509 “Introduction to R for Data Analysis in the Health Sciences” for more in depth coverage of R

Written Homeworks

- Homeworks approximately every week: analysis of real data
  - Questions directed toward specific analyses
    - But questions will still be stated in as scientific terms (as opposed to statistical) terms as possible
  - Work handed in is expected to be organized scientifically
    - I expect nicely formatted tables, figures
    - Unedited R code is totally unacceptable
- Homeworks must be submitted:
  - Electronically on the course's canvas website
  - On-time (exceptions only in the most dire circumstances)

Errors to Avoid

- Unedited R output is TOTALLY unacceptable
- Any assignments that are handed in should be your work
- Electronically submitted homeworks should be a **pdf** or **Microsoft Word compatible file** (e.g., .doc or .docx)
- Name file appropriately: e.g., if you are submitting HW3, include "HW3" in the file name

Discussion Section

- Discussion section will be used for multiple purposes:
  - learning the basics of R for statistical analysis of data
  - discussing and expanding on course material
  - discussion additional relevant topics
  - actively applying methods to datasets.
- The primary focus of discussion section is applying concepts from class to specific scientific problems and datasets.
- You are expected to attend one of the four discussion sections each week.
- Let me know if you are plan to regularly attend a different discussion section this quarter than the one you are currently registered to attend.

Course Grade Contributions

- 30% Homeworks (approx 8-10)
- 30% One Midterm (in class, closed book)
- 40% Final Exam (in class, closed book)

Grading: Homework

- Homework is an important part of the learning process and should be taken very seriously
- Late homework is not accepted, even for good reasons.
  - This policy is only fair if I don't make exceptions.
- Because of this strict policy, I drop the lowest homework when computing the final grade
- You can miss 1 homework assignment without penalty.
  - Save this "free pass" for a true emergency.

Submitting Homework

- Submit homework online by the deadline day and time
  - Although late assignments are not accepted, there is a brief "grace period"
  - As long as your assignment uploads it will be accepted and eligible for full credit (even if the system marks it "late")

Homework Assignments

- Weekly homework: an occasional "mathematical" exercise, but mostly mini-analyses of real data
  - As much as possible I will state questions in scientific terms
  - Scientific question → data question → statistical analysis
  - Submitted work should be organized scientifically
    - I expect nicely formatted tables, figures constructed with care
      - It takes thought and effort to make good tables and figures
    - E.g., unedited R output is not acceptable
      - Not acceptable = no credit
- Students may consult with each other, the instructor, and the TAs during office hours on homework. However, the work that is handed in should reflect only that student's work.
  - That is, obtaining help from other students in order to learn the METHODS of solution is allowed, but copying another student's answer is NOT.

Homework Keys

- Keys to the homeworks will be posted on Canvas
  - Answers in keys may exceed what I expected you to do
  - You are responsible for any new information in the homework keys, even if that information is not otherwise presented in class
  - Annotated R commands and/or output may appear in keys
    - Even though these are NOT appropriate for what you turn in

Assumed Prior Knowledge

- Students are expected to have completed a course in second year algebra, and to be conversant with graphs, linear equations, e, natural logarithms, and summation notation.
- Students who may find this course's presentation too fast or technical are encouraged to consider the Biostat 511/512/513 sequence instead, or Biost 508
- Prior statistical coursework
  - Not necessary for this course
  - (If you have had prior courses, unlearn
    - Need for Normal data to test means
    - P value as entire summary of analysis
    - Significance testing to detect confounding
    - …)

What this course is and isn't

- This is a course in applied statistics
- This is not a math course
- My view: Statistics is not a sub-field of mathematics. Statistics is a field that uses a lot of math. Similar to other fields, e.g., physics.

Two measures of central tendency: mean and median

Task	Discipline	Difficulty
Compute the mean and median	Mathematics (arithmetic)	Easy
Derive the sampling distributions* of the mean and median	Theoretical Statistics	Harder
Decide whether I should compute the mean or the median for a scientific question of interest	Applied Statistics	Hard

*sampling distribution

- I have not yet defined this concept (we will)
- Deriving sampling distributions is outside the scope of the class (theoretical statistics)
- However, the concept of a sampling distribution is one of the most important concepts to grasp in this course

What can statistics do?

- Association: group comparisons to learn about causal relationships
- Prediction: group summaries to predict or classify unmeasured or future observations
  - To a statistician, “prediction” does not necessarily refer to the future
- Cluster analysis: find or define subgroups
- Dimension reduction: reduce many variables to a smaller number of factors

BIOST 517-518-36-37-40 cover “Association” rather thoroughly, touch on “Prediction,” and not much else.

Course Structure

- Biost 517 / 514
  - One response variable; one grouping variable
    - One-, two-, K-sample description and inference
    - Simple regression
  - Stratified description and inference
    - Adjustment for confounding, precision

- Biost 518 / 515
  - Multivariable regression

Biost 517 / 514: Topics

- Scientific setting
  - Study structures
  - Scientific questions
  - Statistical role

Biost 514 / 517: Topics

- Descriptive statistics
  - Motivation
  - Types of measurements
  - Univariate summary statistics
  - Univariate depictions of distributions
  - Censored data descriptive statistics
  - Bivariate descriptive statistics

- Note: Time spent on descriptive statistics will inform our choices for summary measures (parameters) to answer inferential questions

Biost 514 / 517: Topics

- Inferential statistics for two variables
  - Standard “frequentist” statistics
    - i.e., not Bayesian
  - Comparing means, geometric means, medians, proportions, odds, hazards, …
  - Point and interval estimates
  - Hypothesis tests
    - t, chi squared, Fisher’s exact, logrank, Wilcoxon
    - Simple regression

Biost 514 / 517: Topics

- Introduction to stratified analyses
  - Confounding, precision, effect modification
  - Descriptive statistics
  - Stratified analyses

Overview of Scientific Approach / Introductory Example

Scientific Method

General Philosophy

"Everything should be as simple as possible, but no simpler."

A. Einstein, paraphrased

General Philosophy

"Everything should be as simple as possible, but no simpler."

A. Einstein, paraphrased

"For every complex problem there is an answer that is clear, simple, and wrong."

H.L. Mencken

What is Biostatistics?

- **Biostatistics** is *the science of obtaining, analyzing, and interpreting data using statistical theory and methods to address problems in the biological, medical, and health sciences.*
- *Wikipedia Definition*:

  "Biostatistics (or biometry) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results."

Role of Statistics

Statistics plays a role at multiple stages in conducting a scientific study

1. Refine the scientific question
2. Study design
3. Descriptive statistics
4. Inferential statistics
   - Computing estimates of population parameters
   - Quantifying strength of evidence in data

First Stage of Scientific Investigation

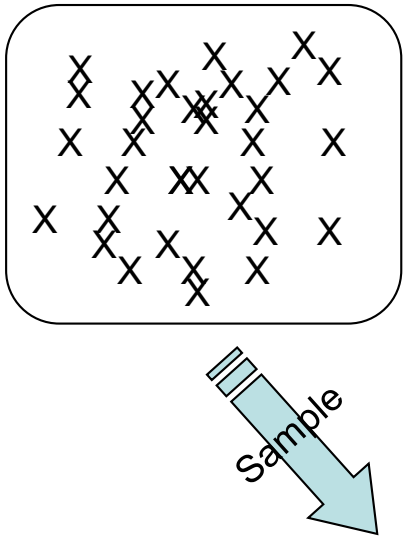
- Hypothesis generation
- Observation
- Measurement of existing populations
- **Population**: the complete set of individuals, objects or scores of interest.
  - Often too large to sample in its entirety
  - It may be real or hypothetical (e.g. the results from an experiment repeated infinitely many times)

Population(s) of Interest

- Population should be clearly defined in a scientific investigation
- For example, what if we are interested in the distribution of BMI for women aged 50 to 100.
  - Is this BMI distribution for all women aged 50-100: present time (e.g., the last decade), or the past (e.g., 1950's)
  - BMI of all possible women aged 50-100: in the U.S., in Washington State
  - BMI of all possible women aged 50-100 of African (European, Asian) descent

Population Parameters

- **Parameters**: Quantities that describe a population characteristic. Usually unknown and goal is often to obtain *statistical inferences* about parameters from a population sample
- In general a sample is drawn from a population and inference about the population is obtained based on measures from the sample

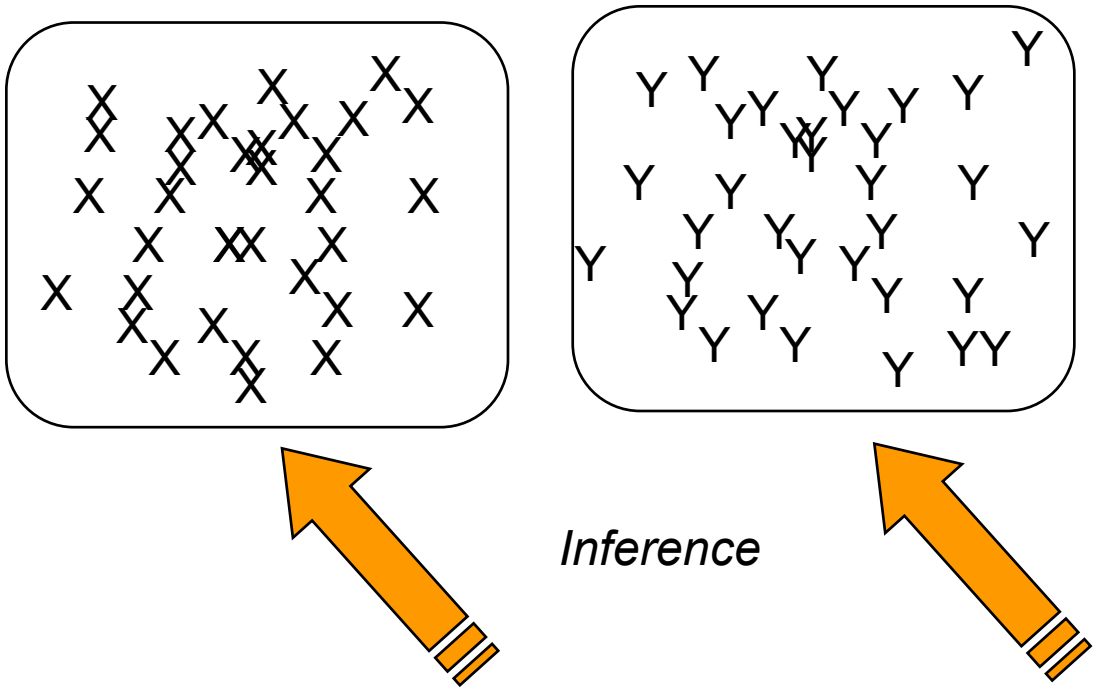


observed:

$$X_1, X_2, X_3, \ldots, X_m.$$

Inference Comparing Two Populations

- Statistics calculated on the samples are used for inference about population parameters.
- What are some common parameters that might be of interest when comparing samples from two populations?



$X_1, X_2, X_3, \ldots, X_m.$ $Y_1, Y_2, Y_3, \ldots, Y_n.$

Defining Parameters

- A parameter is (formally) a mathematical operation on the entire population.
- The parameter value is the result of this operation.
- "Inference" means making one or more conclusions about the parameter value
- These could be estimates, intervals, or binary (Yes/No) decisions
- "Statistical inference" means drawing conclusions without the full population's data, i.e. in the face of uncertainty.
- In this class, we will take a Frequentist view: Parameter values themselves are fixed unknowns; they are not "uncertain" or "random" in any stochastic sense.

Defining Parameters

- For any ‘sane’ parameter, it’s reasonable that its value is non-stochastic – because we define it using an entire population.
- However, in some situations, one cannot identify the parameter, even with an infinite sample (e.g. mean BMI of women, when you only have data on men)
- Formal inference then becomes impossible. Instead, you could;
  - Switch target parameters
  - Extrapolate cautiously
  - Not do inference, but ‘hypothesis-generation’
  - Give up
- It is important to determine if parameters are well-defined, i.e. ‘sane’. A parameter is (formally) a mathematical operation on the entire population.

Defining Parameters

"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data"

--John Tukey

Will statistics be helpful?

- Statistics, as a discipline, models the variability from person to person due to unmeasured variables or imperfections in the measurement process as randomness.
- If the scientific question can not be answered when the parameter of interest and/or measurements are available on the entire population, **then statistics will not be of any help**

Further Stages of Scientific Investigation

- Refinement and confirmation of hypotheses
- Experiment: Intervention
- Elements of experiment
  - Overall goal
  - Specific aims (hypotheses)
  - Materials and methods
  - Collection of data
  - Analysis
  - Interpretation; Refinement of hypotheses

Scientific Questions

- Scientific questions are often concerned with investigating cause and effect
  - E.g., in biomedical settings:
    - What are the causes of disease?
    - What are the effects of interventions?

Scientific Questions

- It is often the case that scientific studies make answering such questions difficult even when study results are deterministic (no variation in response)
  - Difficulties in isolating specific causes
    - E.g., isolating REM sleep from total sleep
    - E.g., interactions between genetics and environment
  - Difficulties in measuring potential effects
    - E.g., measuring time to survival
      - length of study
      - competing risks

Scientific Questions

- There is inevitably variation in response across repetitions of an experiment
  - Variation can be due to
    - Unmeasured (hidden) variables
      - E.g., mix of etiologies, duration of disease, comorbid conditions, genetics when studying new cancer therapies
    - Inherent randomness

Scientific Questions

- Scientific questions thus have to be phrased in a manner that acknowledges such variation in response
  - Deterministic:
    - Does statin meditation decrease LDL cholesterol?
  - Probabilistic:
    - Do individuals who take statin meditation tend to have lower LDL cholesterol than individuals who do not.
- When considering a data analysis to answer a scientific question, ask yourself:
  - If my data were completely deterministic, would I know the answer to my question
  - If the answer is “no,” statistics will not help you.
-

Scientific Questions

- Of course, the probabilistic approach only makes sense if we can find a suitable definition for the phrase “tends to”
  - Many possibilities exist for detecting a decrease:
    - A lower average value (arithmetic mean)
    - A lower geometric mean
    - A lower median: Mdn (Trt) – Mdn (Ctrl) < 0.0
    - Median (Treated – Control) < 0.0 , e.g. for a paired study design
    - A lower proportion exceeding some threshold
    - A lower odds of exceeding some threshold
    - Pr (Treated > Control) < 0.5
    - Time average of hazard ratio < 1.0

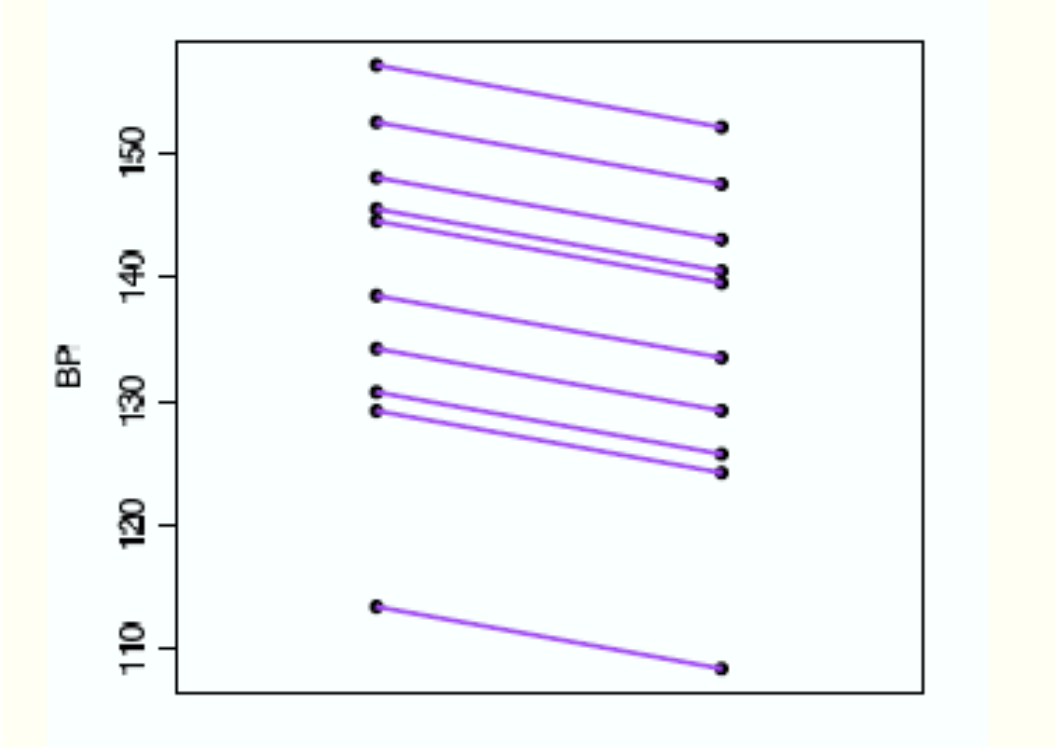
Choice of data summaries

- A big topic of this class is choosing a summary measure for data.
- E.g., we often want to compare two groups to decide which has the higher value of some variable
  - Is survival longer?
  - Is blood pressure lower?
  - Is total medical cost lower?
- We have to decide what we mean by “higher” or “lower”
  - Refine the scientific question to one that can be address statistically

Hypothetical Example

- Consider a measurement of blood pressure on 10 hypertensive people before and after an intervention that has a goal of reducing blood pressure.
  - Each blood pressure falls by 5 mmHg. We can conclude that blood pressure was lower after the intervention.

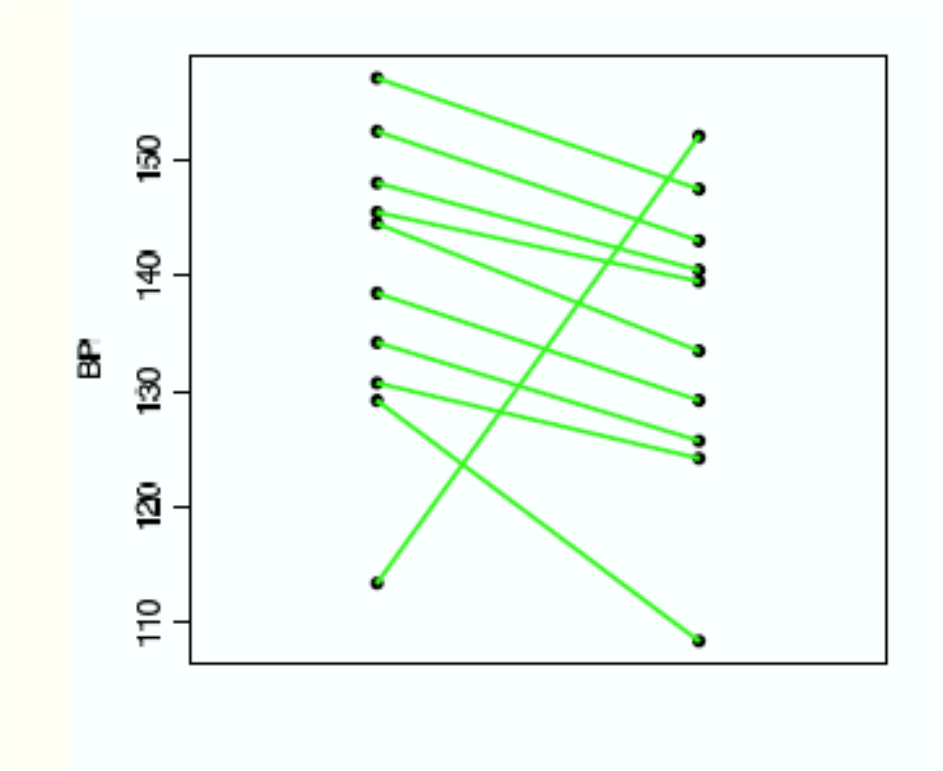
Data A



Hypothetical Example

- Suppose with exactly the same numbers the results actually were:

Data B

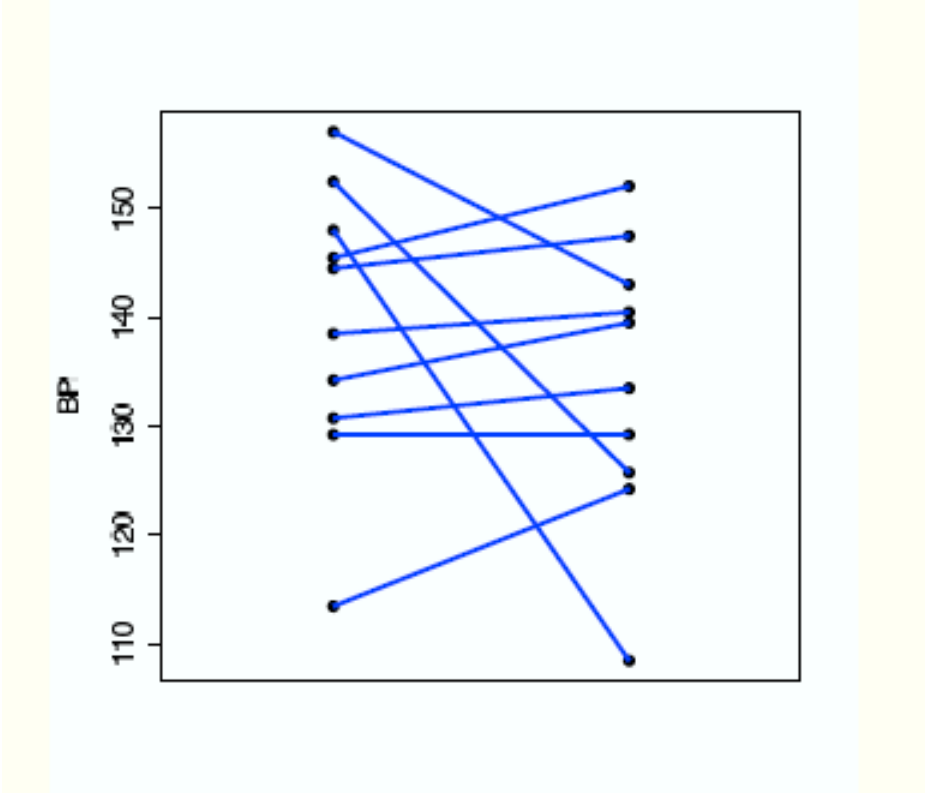


  - Blood pressure goes down for most people, but increases drastically for one person.

Hypothetical Example

- Suppose with exactly the same numbers the results actually were:

Data C



  - 7/10 people had an increase, but blood pressure was greatly reduced for 3/10 patients

Hypothetical Example

- Suppose there is no sampling variability and no measurement error. In other words, suppose any of these pictures exactly represents the population. Would you describe the treatment as successful for lowering blood pressure?
  - This is a scientific question, not a statistical question.

Choice of data summaries

- Biomedical data almost never look like Data A. Data B and Data C are more realistic.
- Thus, if we ask “is blood pressure higher or lower after treatment”, we must refine what we mean by “higher”/”lower”
- We need to summarize data by a single number (a statistic) so that changes in the data that we care about are reflected by changes in that statistic.
  - If the scientific question specifies what summary measure to use, choose that one.
  - Often, competing summary measures differ in how “sensitive” they are to changes. “Sensitivity” can be either desirable or undesirable.
  - If there is no other basis for choosing, recommended to choose a summary that can be estimated precisely

Note: Only the last criterion is statistically based.