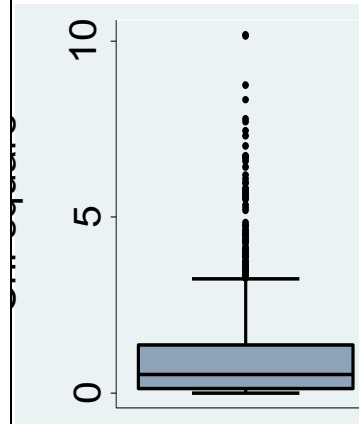


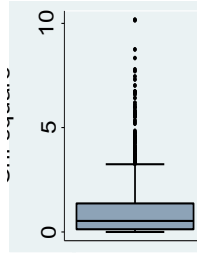
Biostat 202: Accessing Data and Large Datasets

*Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and
Biostatistics*



Biostat 202

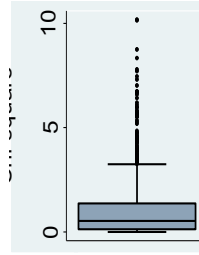
Announcements



- Prerecorded Lecture 3 will be posted to CLE by Tuesday evening
- In class learning poll
- Assignment 1 due Thursday
- Office hours weekly Thursday, 4:00 – 5:00 PM by ZOOM

Outline for today

- File types
- Data formats
- Variable types and roles
- Best practices in data
- Availability of datasets
- UCSF large dataset help
- Research using APeX data

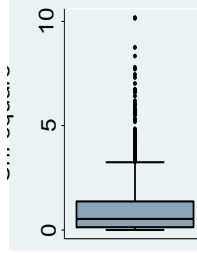


Summary



1. It helps to know the format and layout of a dataset you want to analyze.
2. Paying attention to data formatting issues can save LOTS of time later.
3. Defining variable roles is important in Orange Data Mining.
4. Access to rich data is easy.
5. Finding your way around can be hard.
6. Look for tools (tutorials, browsers),
7. Or get professional help (Library, ARS and CTSI).

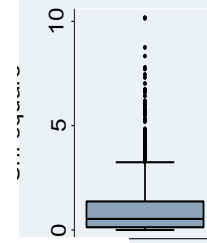
Best data practices: sad but true 1



When looking at the variable titled “patient ID” in the dataset:

- Question: Is the patient with patient ID equal to 2 different from the one with patient ID equal to 2R?
- Answer: No – the patient with patient ID equal to 2 is the left and 2R is the right leg of the same patient.

Best data practices: sad but true 2



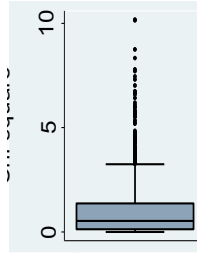
When modeling a response to different dosages of a drug:

- Question: Why didn't you analyze the data using the actual concentrations of the dose?
- Answer: We used the dose variable you sent us in the data which was coded as 0, 1, 2, ..., 11.
- Question: Why? Those weren't the actual doses.
- And this was after scraping the data out of the body of an email.

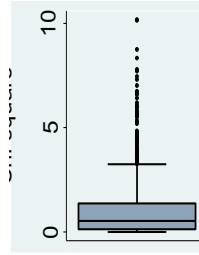
Best data practices: sad but true 3

When asked to compare the treatment and placebo groups in the dataset in which there was no treatment variable:

- Question: How can we tell which patients are in the treatment group?
- Answer: Those are the patient IDs that are in red in Excel.

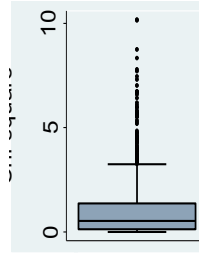


Data check exercise

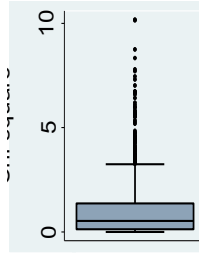


- Split into breakout rooms
- Introduce yourselves, identify a spokesperson
- Identify topic areas of interest to group
- Then spot some bad data habits in the BSI_exercise.xlsx file on CLE under today's lecture
- We'll reconvene after ten minutes and discuss

Data check exercise - mistakes



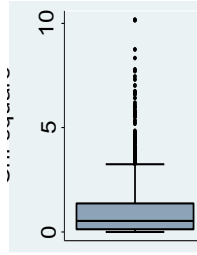
- Variable names not in first row
- Multiple values per column (alcohol/marijuana)
- Highlighting
- Summary statistics inserted
- “?” for missing data, use blanks
- Metadata (e.g. source) inserted within data
- Etc...



Questions?

Use the **chat** to let me know, I will attend in that order.

Discuss project



- Project details can be found at:
<https://courses.ucsf.edu/mod/folder/view.php?id=654109>