TOPICS IN STUDY DESIGN AND STATISTICAL ANALYSIS

DR. ANN HESS (SHE/HER)

CSU STAT LAB

STATLAB.COLOSTATE.EDU

OUTLINE

1. Graybill Statistics and Data Science Laboratory (aka “Stat Lab”)
2. Study Design Topics
3. Formatting Data for Analysis
4. Statistical Analysis Topics
5. Learning Resources

1. GRAYBILL STATISTICS AND DATA SCIENCE LABORATORY

- Stat Lab Mission: "Provide statistical expertise for the benefit of advancing the quality of scientific research at Colorado State University and beyond while educating the next generation of applied statisticians."
- Faculty/Staff:
  - Dr. Ann Hess (Associate Director, Professor)
  - Dr. Eric Gilleland (Research Associate)

STAT LAB CONSULTING SUPPORT

Appointments with Faculty Consultants:

- Support relating to research proposals, study design, more complicated statistical analyses and manuscript preparation.
- Dr. Eric Gilleland and Dr. Ann Hess

Drop in with Data:

- Guidance with data formatting, data “wrangling”, data visualization and common statistical analyses.
- Staffed by Stat grad students
- No appointment necessary.

STAT LAB FUNDING

- **There is no charge for general statistical consulting for graduate students and for initial faculty meetings.**
- This service is supported by the Statistic department, Ag Experiment Station, CNS and CVMBS.
- We rely on support from individual faculty users.
- Support requested for grant funded projects.
- Fee-for-service data analysis is available for projects that go beyond general consulting.

WHAT CAN A STATISTICIAN HELP WITH?

BEFORE Data Collection

- Statistical support on grant proposals
- Study design and sample size determination
- Data formatting suggestions

AFTER Data Collection

- Data analysis recommendations
- Interpretation of results
- Software assistance

WHAT TO BRING TO AN INITIAL MEETING?

Design stage:

- Project summary and research aims
- Pilot study data and/or literature review

Analysis stage:

- Project summary and research aims
- Data
- Code and output

EXAMPLE PROJECT 1

- PIs: Paxton Sullivan (PhD student), Dr. Mahesh Nair, Dr. Lily Edwards-Callaway,
- Preslaughter factors affecting mobility, blood parameters, bruising, and muscle pH of finished beef cattle in the United States (2024, Translational Animal Science)
- Observational study with n = 454 cattle. Particular interest in serum cortisol as a measure of stress. Many predictor variables.
- Regression, logistic regression and model selection (AIC) used for analysis.

EXAMPLE PROJECT 2

- PIs: Dr. Sarah Matlock, Dr. Caiti Peters
- The Preliminary Efficacy of an Equine-assisted Social-emotional Learning Program for Youth with Diminished Social-emotional Competencies (2026, Journal of Experiential Education)
- Randomized waitlist-controlled trial with n = 31 students (middle school). DESSA (Devereux Student Strengths Assessment) reported by parents and teachers, pre and post intervention.
- Paired t-tests, Wilcoxon signed rank test and mixed models used for analysis.

2. STUDY DESIGN CONSIDERATIONS

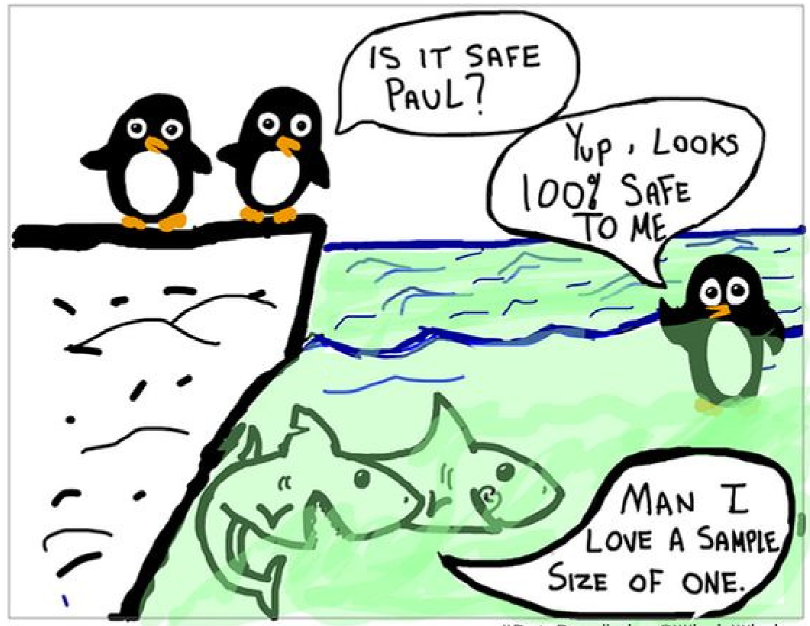
- Study design is everything that happens before data collection.
- Specific aims should be determined in advance.
- **Specific aims drive the design and analysis.**
- Much of the following discussion is taken from "Statistical Principles of Research Design and Analysis" by Kuehl.
- 3 R's of experimental design: randomization, replication, reduce noise!

RANDOMIZATION

- Randomization protects against bias.
- Randomized experiments allow causal conclusions.

REPLICATION

- Replication is critical for valid results!
- Bare minimum n ≥ 2 independent units per treatment.
- Biological vs Technical reps: Biological replicates are essential.



#DataDoodle by @WholeWhale

PSEUDO REPLICATION

- Watch out for "pseudo replication" (defined by Hurlbert):

"The use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated, or replicates are not statistically independent."

DETERMINING SAMPLE SIZE

- Determine sample size to achieve our research goals.
- Justification should match planned analysis and research goals.
- Typically phrased in terms of power.
- Power is the probability of detecting a difference between treatments, given there really is a difference.
- Also consider practical issues (availability, money, time).

CALCULATING POWER

- Most common scenario: Comparing means using two-sample t-test
- Need conjectures for
  - difference between means (μ_1-μ_2)
  - variability (σ)
- Also need significance level of the test (alpha).
- With this info, relatively easy to calculate power for a given sample size using software.

INPUT FOR POWER CALCULATIONS

- My opinion: The hardest part of the power calculation is coming up with reasonable "conjectured" values.
- Use pilot data if available.
- Use summary statistics from published studies.
- Another way to think about the power calculation: power to detect "meaningful difference" between means.

REDUCE NOISE

- Technique
- Selection of experimental units
- Choice of experimental design
- Blocking
- Measurement of covariates

https://behavioralscientist.org/the-benefits-of-statistical-noise/

REDUCE NOISE: TECHNIQUE

- Fundamental tasks (ex: accurate measurement, preparation of media, calibration of instruments)
- Choice of methods or instrument
- Uniform application of treatments

REDUCE NOISE: SELECTION OF UNITS

- Make statistical inference to the population from which we sampled.
- Trade off between range of conditions and uniformity of units.
- Homogeneous/uniform experimental units reduce experimental error.
- Narrow set of conditions restricts the inference base of the study.

REDUCE NOISE: BLOCKING

- Experimental units are grouped such that the variability of units within the "blocks" is less than that among all units prior to blocking.
- Treatments are compared within a block, hence improving accuracy of comparison of treatment means.
- "Block what you can, randomize what you cannot."
- Common blocking criteria: physical characteristics (age or weight), time, management of tasks (lab "batches")

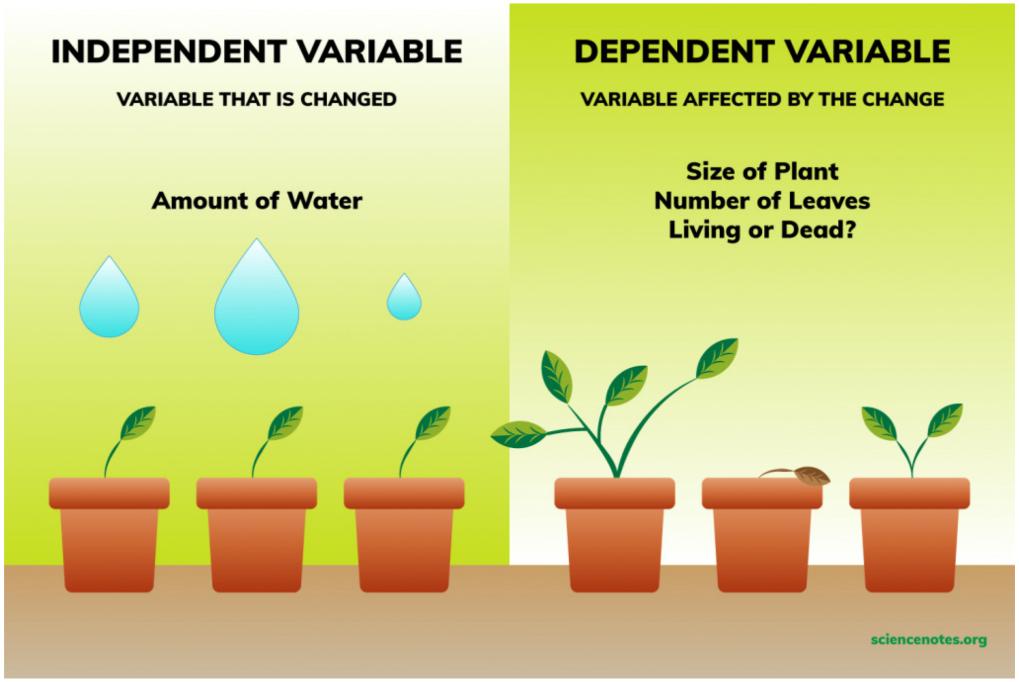
REDUCE NOISE: COVARIATES

- Covariates are "secondary" variables that may be related to the response variable of interest.
- Accounting for covariates, we may be better able to detect differences between the treatments.
- Candidates depend on the particular study. Examples: age, weight or pre-treatment response.
- Covariate should not be affected by the treatment. In most cases, the covariate is measured on the experimental unit **before** the treatment is applied to the unit.

VARIABLES

- What and how many variables?
- If multiple response variables, can we identify a small number of primary responses?
- Meaningful effect size for primary responses

Predictor vs Response Variables



https://sciencenotes.org/independent-and-dependent-variables-examples

3. FORMATTING DATA FOR ANALYSIS

- Common to have multiple versions of the data. Ex: original spreadsheet used for data entry and "clean" analysis-ready data.
- **The original data should never be modified and should be stored in a secure location!**
- Plan ahead to reduce the number of data versions.
- Keep track of the steps you take in your clean up and analysis (just as you would keep track of any other research step).

DATA FORMATTING

- Following suggestions are taken from "Data Organization in Spreadsheets" by Broman and Wu (2017)

DATA FORMATTING: MAKE IT A RECTANGLE

- Single big rectangle with:
  - **rows** corresponding to **units** (ex: animal or human subjects)
  - **columns** corresponding to **variables**.
- First row should contain variable names.
- Do not use more than one row for the variable names.

DATA FORMATTING: BE CONSISTENT

- Consistent codes for categorical variables. Be careful about extra spaces within cells.
- Consistent code (or blank) for any missing values.
- Consistent variable names.
- Consistent data layout in multiple files.

DATA FORMATTING: CELLS

- Think like a computer!
- Put just one thing in a cell.
- Do not merge cells.
- Empty cells are OK for “missing” values.
- Do not use font color or highlighting as data.

"MESSY" DATA EXAMPLE 1

Original Format

	Day1		Day2	
Trt A	Y1	Y2	Y1	Y2
A1	8.4	9.7	9.8	11.2
A2	8.9	8.3	11.1	10.7
	Day1		Day2	
Trt B	Y1	Y2	Y1	Y2
B3	9.2	10.5	8.7	9.8
B4	11.9	10.8	9	10

Tidy Format

ID	Trt	Day	Y1	Y2
A1	A	1	8.4	9.7
A2	A	1	8.9	8.3
A1	A	2	9.8	11.2
A2	A	2	11.1	10.7
B3	B	1	9.2	10.5
B4	B	1	11.9	10.8
B3	B	2	8.7	9.8
B4	B	2	9	10

"MESSY" DATA EXAMPLE 2

Original Format

	Day1		Day2	
Trt A	Y1	Y2	Y1	Y2
A1	8.4	9.7	9.8	11.2
A2	8.9	8.3	11.1	10.7
	Day1		Day2	
Trt B	Y1	Y2	Y1	Y2
B3	9.2	10.5	8.7	9.8
B4	11.9	10.8	9	10

Workable Format

ID	Trt	Day1.Y1	Day1.Y2	Day2.Y1	Day2.Y2
A1	A	8.4	9.7	9.8	11.2
A2	A	8.9	8.3	11.1	10.7
B3	B	9.2	10.5	8.7	9.8
B4	B	11.9	10.8	9	10

Because of the consistent column names, additional formatting can be easily done in R or other programs.

4. STATISTICAL ANALYSIS CONSIDERATIONS

- Exploratory data analysis
- Repeat observations on the same experimental unit
- Multiple Testing: Parts 1 and 2
- Evaluating assumptions
- Outliers

EXPLORATORY DATA ANALYSIS

- This step is very important!
- Provides first pass information about specific aims.
- Very helpful for data checking.
- Can help inform the analysis.
- Provides a way to spot check the final results.
- Does not have to be complicated: simple graphics (boxplots and scatterplots) and summary statistics (mean and standard deviation).

REPEAT OBSERVATIONS ON THE SAME UNIT

- Independent observations is a standard statistical assumption.
- Repeat observations on the same unit violates this assumption.

Repeated Measures

- Multiple time points or multiple trts observed on the same unit.
- Mixed model analysis

Exp vs Obs Units

- Pens (exp units) assigned to trts
- Measurements taken on individual animals (obs units) within a single pen.
- Average to pen level or use nested analysis

MULTIPLE TESTING 1

- "Classic" multiple testing focuses on pairwise comparisons after ANOVA.
- Suppose we run a two-factor ANOVA model.
- Step 1: ANOVA F-tests for main effects and interactions.
- Step 2: Depending on those results, we might run Tukey adjusted pairwise comparisons for the interaction.

MULTIPLE TESTING 2.1

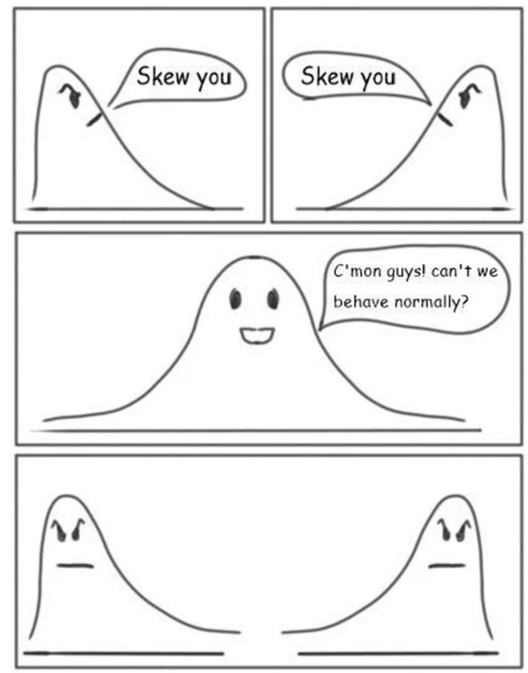
- A different multiple testing situation arises when we fit the same model for multiple response variables.
- Example 1: Study includes 15 response variables focused on animal health.
  - Before starting the study, can we identify a single primary response variable? The other 14 would be considered secondary.
  - We might consider a Bonferroni adjustment for 14 secondary variables.

MULTIPLE TESTING 2.2

- Example 2: RT-PCR study includes 30 genes.
  - Bonferroni or FDR (false discover rate) adjustment
- Example 3: RNAseq study includes 1000 features.
  - FDR adjustment
- Xkcd jelly bean cartoon

EVALUATING ASSUMPTIONS

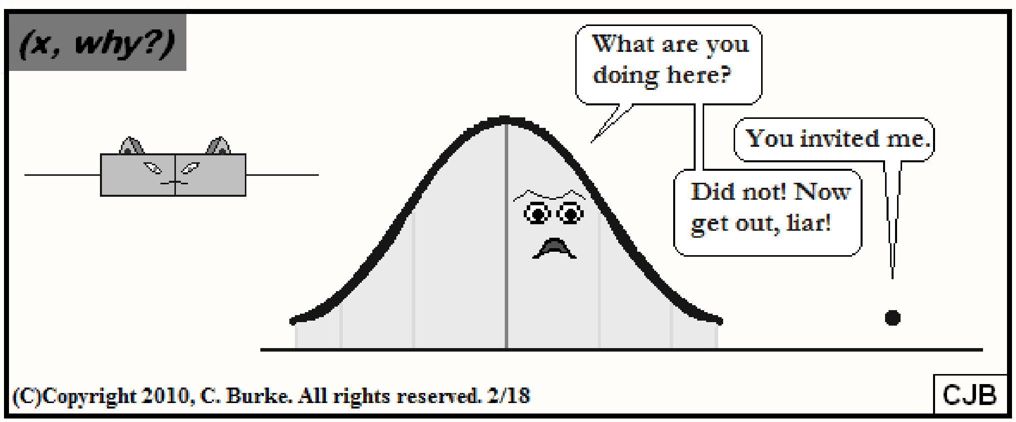
- Standard Assumptions:
  - Normality of residuals
  - Equal Variance of residuals
- Use residual diagnostic plots to check
- Common mistake: Checking normality of the raw response (before modelling) instead of check normality of the residuals (after modelling).



Randy Krum
https://coolinfographics.com/news/2021/10/4/statistic-joke

OUTLIERS I

- **I start every analysis planning to use all the data!**
- Formal test of outliers available but not usually of interest.
- Bigger issue: What to DO about an outlier?
- Assumptions: Small number (1 or 2) outliers and model assumptions satisfied.



OUTLIERS 2

- First: check the data for errors.
- Try analysis with and without the outlier.
- If the conclusions do **not** change, keep the outlier as part of the analysis.
- If conclusions **do** change:
  - Discuss the outlier and how it affected results.
  - May want to discuss both analyses (with and without).
- If data is omitted, this needs to be discussed in the write-up.

5. LEARNING RESOURCES: CODING AND COOKIES

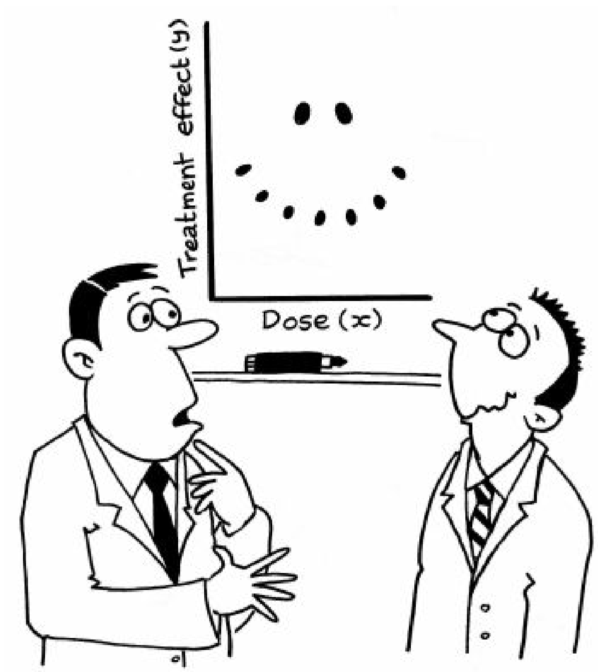
- https://libguides.colostate.edu/coding-cookies
- These free, non-credit 2 hour workshops are sponsored by the Stat Lab and CSU Libraries.
- R topics: R basics, Tidy data with tidyverse, Data visualization with ggplot2
- Python topics: Python basics, Working with Data with pandas, Data Visualization with matplotlib and plotnine
- Version Control with Git/GitHub command line

LEARNING RESOURCES: STAR COURSES

- STAR courses are applied statistics courses for graduate students across CSU (non-Stats students). Note the STAR acronym!
- These courses fill VERY quickly! Early registration is strongly recommended. Use RamWeb waitlist if needed.
- STAR topics courses are offered on a "rotating" basis, but with no set schedule.

LEARNING RESOURCES: STAR COURSES

- STAR 511: Design and Data Analysis for Researchers I (4 credits): Topics include simple linear regression, one-way ANOVA and R software.
- STAR 513: Regression Models (2 credits)
- STAR 531: Generalized Regression Models (2 credits)
- STAR 512: Design and Data Analysis for Researchers II (4 credits): Topics include multiple regression, factorial ANOVA and mixed models.



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

QUESTIONS?

REFERENCES

- “Statistical Principles of Research Design and Analysis” by Robert Kuehl.
- “Data Organization in Spreadsheets” by Broman and Wu (2017)