Weill Cornell Medicine

Communicating with a Statistician

Anjile An, MPH, Kat Hoffman, MS

Division of Biostatistics, Department of Population Health Sciences

Slides adapted from Debra D’Angelo, MS

Today's Topics

1) The statistical consulting process
    - 10 tips for a successful project, start to finish
    - Providing clean data

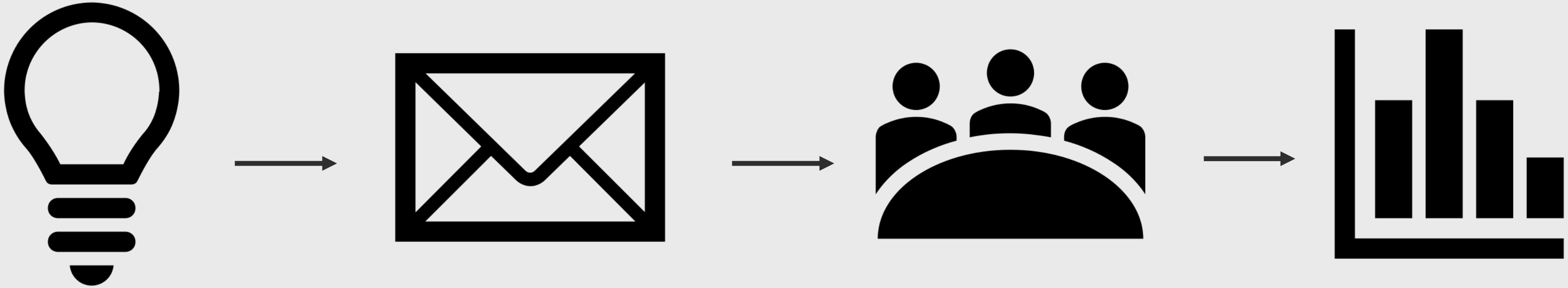
2) Common statistical tests
    - Continuous outcomes
    - Ordinal outcomes
    - Categorical outcomes
    - Multivariable analysis

3) Requesting a biostatistics consult

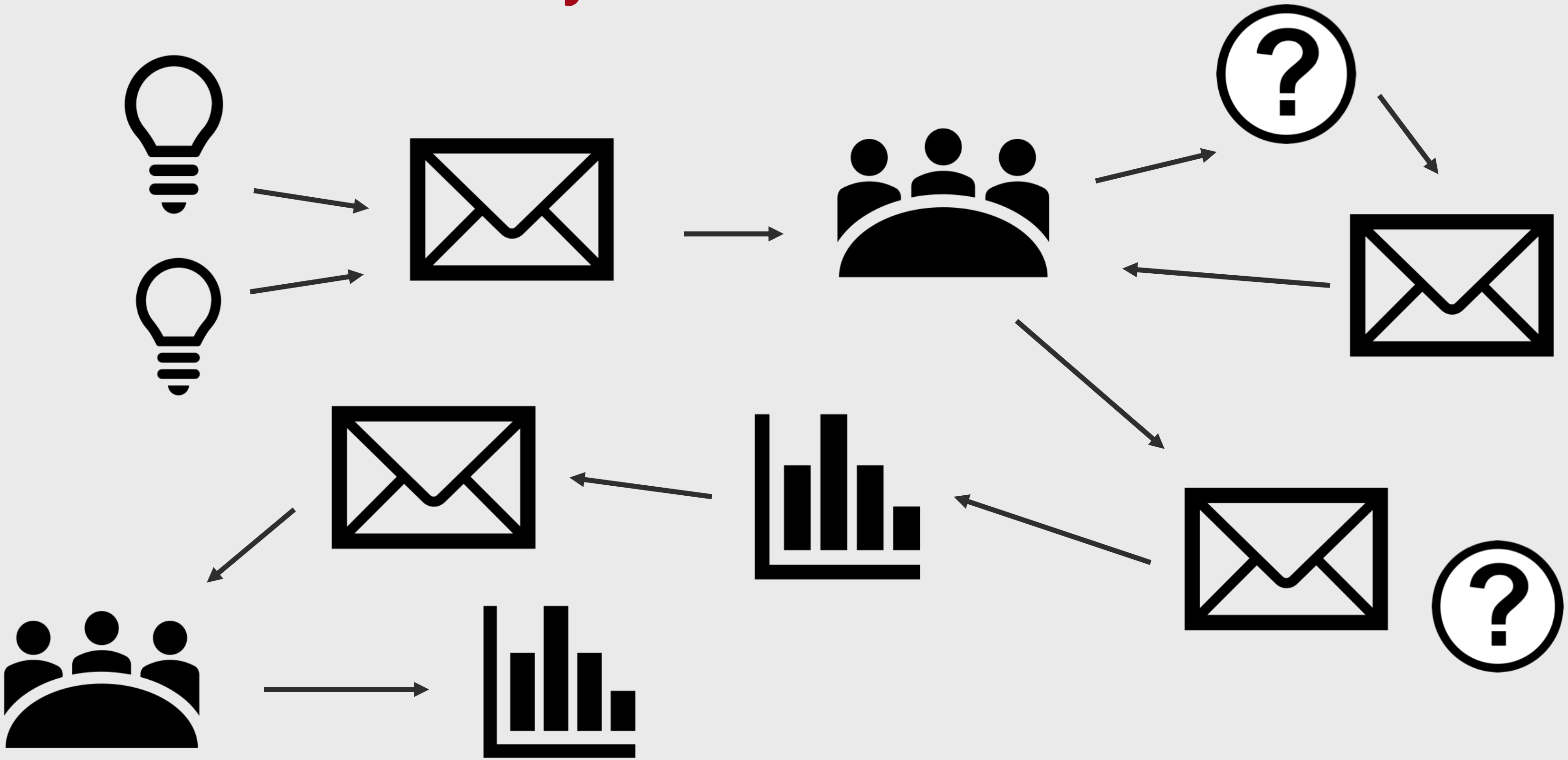
Weill Cornell Medicine

Statistical Consulting Process

What you think it’s going to be like:



What it’s actually like:



Weill Cornell Medicine

Tip 1: Involve statisticians early on

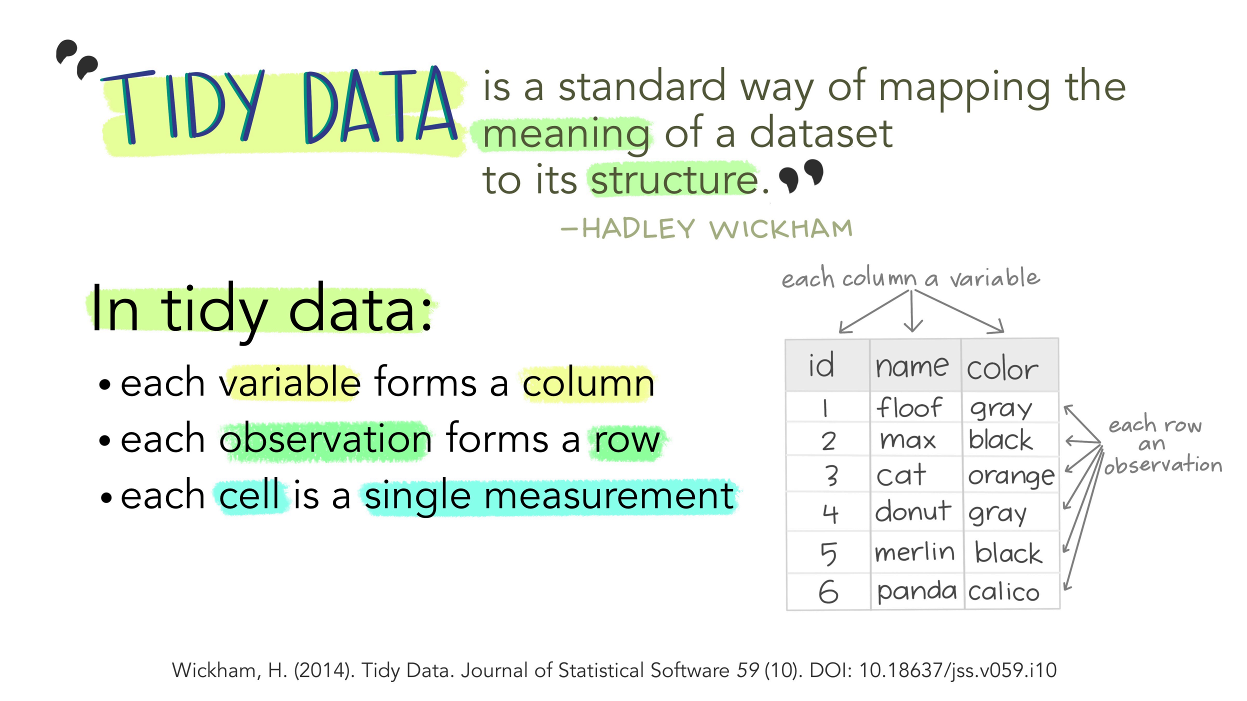
“To call in the statistician after the experiment is done may be no more than asking [them] to perform a post-mortem examination: [s]he may be able to say what the experiment died of.”

R.A. Fisher, 1938

Tip 2: Send relevant materials in advance

- Study protocol/IRB
- Measurement instruments (ie. survey)
- Relevant background literature (papers that are similar in topic, methodology)
- Existing datasets (deidentified!)

On datasets …



"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

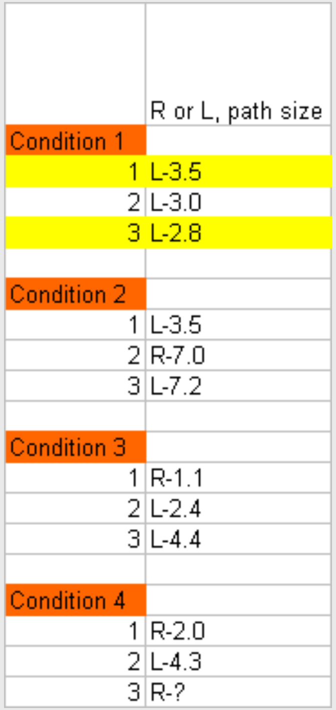
each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

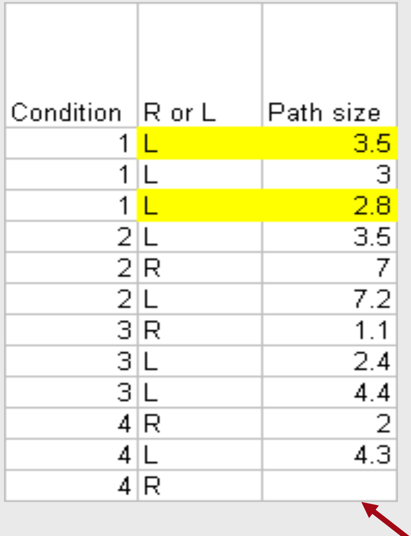
each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Tidying data



	R or L, path size
Condition 1	
1	L-3.5
2	L-3.0
3	L-2.8
Condition 2	
1	L-3.5
2	R-7.0
3	L-7.2
Condition 3	
1	R-1.1
2	L-2.4
3	L-4.4
Condition 4	
1	R-2.0
2	L-4.3
3	R-?



Condition	R or L	Path size
1	L	3.5
1	L	3
1	L	2.8
2	L	3.5
2	R	7
2	L	7.2
3	R	1.1
3	L	2.4
3	L	4.4
4	R	2
4	L	4.3
4	R	

Here we show the correct format to document ‘R or L’ and ‘Path size’

Very important – if the value is unknown, just leave it blank

On HIPAA

Please remove all HIPAA identifiers from datasets!

1. Name
2. Address
3. All elements of dates (except year)
4. Phone Numbers
5. Fax Numbers
6. Email Addresses
7. Social Security Numbers
8. Medical Record Numbers
9. Health Plan Beneficiary Numbers
10. Account Numbers
11. Certificate/License Numbers
12. Vehicle Identifiers
13. Device Serial Numbers
14. Web URLs
15. IP Addresses
16. Finger or Voice Prints
17. Photographic images
18. Other Unique Identifiers

When in doubt, Google: 18 HIPAA Identifiers

Removing HIPAA and tidying data

- Instead of using name/MRN, just add an arbitrary ID number (1, 2, 3, 4, etc) and send the statistician the version with arbitrary ID

- Remove any variables that are not relevant to the study

- Remove any free text variables (ie. Notes) unless relevant

Tip 3: Start with layman’s summary

- Statisticians work on lots of projects, and are likely not experts in your clinical area

- Start discussion with a summary of your study in layman’s terms

- Explain clinical terms that are relevant to the study

Tip 4: State the research question

Should clearly define exposure (X), outcome (Y) and population (P)

“Is X associated with a change in Y in population P?”

Should also be SMART

SMART Research Questions

- Specific?
  - One primary objective?
- Measurable?
  - Quantifiable endpoint?
- Attainable?
  - Do you have the appropriate resources?
- Relevant?
  - Is it novel?
  - Is it clinically meaningful?
- Timely?
  - How long will data collection take?

Tip 5: Define your sample

- What clinical population will you study? What timeframe?
- How will you collect a sample of that population? (EHR, RedCap, registry, surveys etc)
- Are there logistical constrains in collecting the sample? (ie. lab samples)

On sample size

It is not “one size fits all!”

The statistician can help you determine the sample size needed for your study based on some expected parameters, but this will vary based on your research question, the types of variables collected, expected effect sizes, etc.

- Helpful if you have pilot data!

Sometimes the sample size is fixed due to logistics. In this case, we work backwards and calculate power or detectable difference.

Tip 6: Define variables of interest

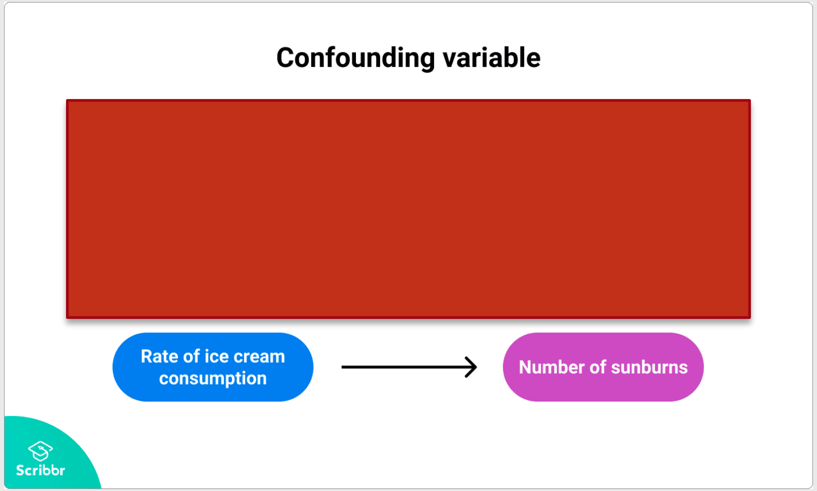
Are the outcomes and exposures of interest measured as:

- Continuous?
- Ordinal?
- Nominal?
- Categorical?

Give feasible ranges/all possible categories for each variable

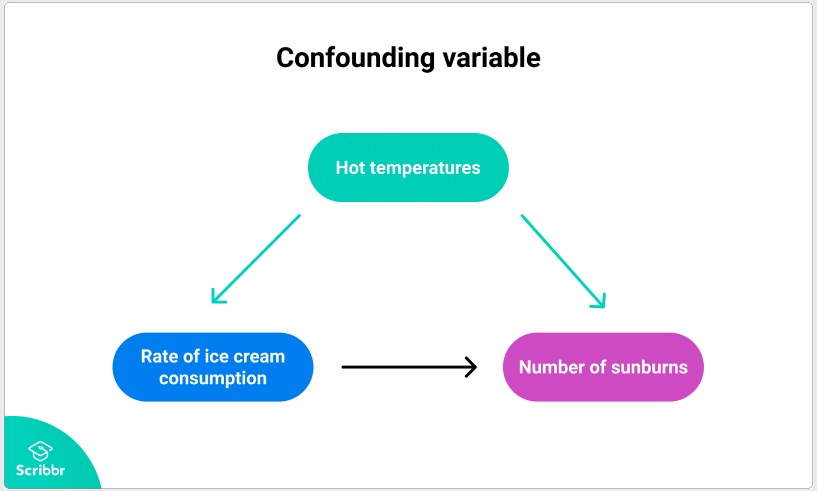
Tip 7: Identify possible confounders

- Sometimes established in literature
- Sometimes based on clinical knowledge



Tip 7: Identify possible confounders

- Sometimes established in literature
- Sometimes based on clinical knowledge



Tip 8: Clearly define deliverables

Statisticians commonly contribute to the following:

- Statistical analysis plans
- Sample size/power calculations
- Statistical analyses (tables, figures, descriptives, regression, etc.)
- Abstract or manuscript writing

Tip 9: Clearly define deadlines

Common ones include:

- IRB submission deadlines
- Abstract deadlines
- Manuscript resubmission deadlines

Please give ample lead time! (Think weeks, not days)

Tip 10: Keep communicating!

Initial meeting is just the beginning; there is often lots of back and forth during collaboration

Don’t hesitate to ask questions until you feel clear that you and the statistician are on the same page!

The statistician will often reach out to you for clarification as well

The 10 Tips

1) Involve statisticians early on
2) Send relevant materials in advance
   - Remove HIPAA and tidy data before sending
3) Start with layman's summary
4) State the (SMART) research question
5) Define your sample
6) Define variables of interest
7) Identify possibly confounders
8) Clearly define deliverables
9) Clearly define deadlines
10) Keep communicating

Weill Cornell Medicine

Common Statistical Tests

Common Statistical Tests

Recall the slide about research questions:

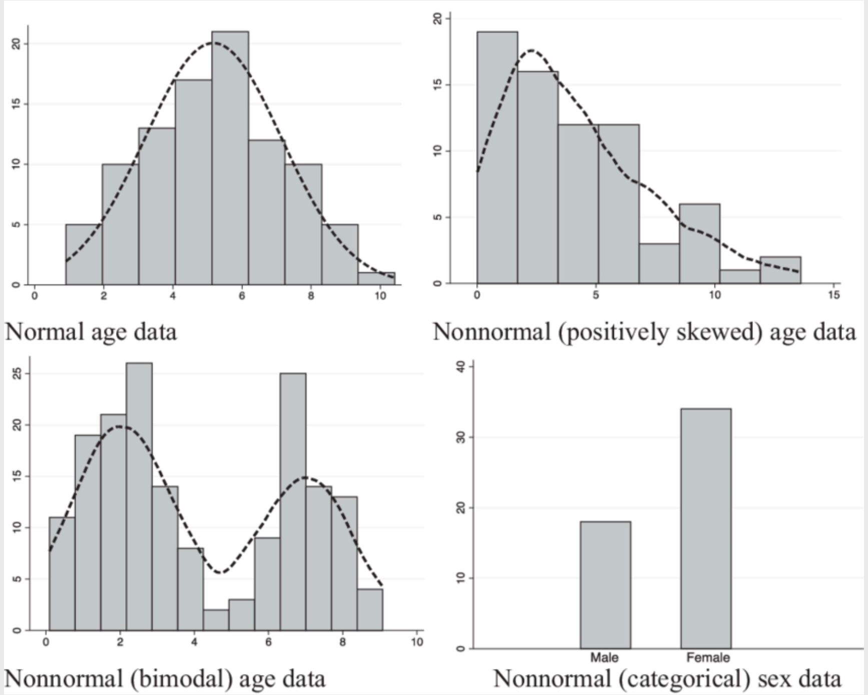
Should clearly define **exposure (X)**, **outcome (Y)** and population (P)

The tests we choose depend on what type of variable X and Y are

Types of outcomes

1) Continuous Outcomes
2) Ordinal Outcomes
3) Categorical Outcomes
4) Multivariable Analysis

Parametric and non-parametric tests



Normal age data

Nonnormal (positively skewed) age data

Nonnormal (bimodal) age data

Nonnormal (categorical) sex data

1) Continuous Outcomes

➢ Outcome of interest: A continuous variable, such as cholesterol level, BMI, a knee rating scale (i.e., 1-100), Hamilton Depression Score, etc.

- Use the two-sample t-test if comparing two independent groups of interest (e.g., men vs. women).
- Use the paired t-test if comparing two non-independent groups of interest (e.g. a set of patients before and after a surgery or intervention).
- Use the ANOVA test if comparing more than two independent groups of interest (e.g., 3 or more types of surgery)

The tests above compare means between groups.

2) Ordinal Outcomes

➢ Outcome of interest: An ordinal variable, such as grade 2 MCL tear, pain score (e.g., rated 1-5), Likert-scale variable, etc.

- Use the Wilcoxon rank-sum test (also known as Mann-Whitney U test) if comparing two groups of interest.
- Use the Wilcoxon signed-rank test if comparing one set of patients before and after a surgery or intervention.
- Use the Kruskal-Wallis test if comparing more than two independent groups (e.g., 3 or more types of surgery).

The tests above compare medians between groups. However, the parametric tests (t-test, paired t-test, ANOVA) tend to work well even with ordinal data (i.e., they are robust to violations of normality).

3) Categorical Outcomes

➢ Outcome of interest: A categorical variable, such as mortality, injury status, redefined pain score (1-3 vs. 4-5 vs. >5), presence of disease (e.g., yes/no), etc.

- Use the Chi-Square test if comparing two or more groups.
- Use the Fisher's Exact test as a substitution for the Chi-Square test if the study sample size is very small (i.e., some cells in the row by column table have few numbers in them).

		Surgery	
		BCS	ME
Age at Diagnosis	<=46	3	2
	>46	2	3

		Surgery	
		BCS	ME
Age at Diagnosis	<=46	33	12
	>46	24	50

4) Multivariable Analysis

➢ Used to determine the independent effect of many variables on a single dependent outcome.

Example: What predicts ACL reconstructions to fail?

- type of surgery
- graft source
- rehabilitation
- age

Four variables would be entered into a multivariable model and each variable would be adjusted for the presence of the other variables (i.e., controlling for other variables). Note - need to have sufficient sample size.

4) Multivariable Analysis

Types of multivariable analysis:

- Multivariable linear regression: for a continuous dependent outcome such as cholesterol level, a knee rating or pain scale (1-100), etc.
- Multivariable logistic regression: for a binary dependent outcome such as mortality, re-operation status (yes/no), injury status (yes/no), etc.
- Cox proportional hazards regression: for a time-dependent outcome such as time to death, time to recurrence, etc. A multivariable extension of Kaplan-Meier survival analysis.

Weill Cornell
Medicine

Requesting a Biostatistics Consult

Requesting a Biostatistics Consult

http://ctsc.med.cornell.edu/BiostatisticsConsult

Weill Cornell
Medicine

Questions?

WEILL CORNELL MEDICAL COLLEGE
CORNELL UNIVERSITY

WEILL CORNELL MEDICAL COLLEGE
CORNELL UNIVERSITY

Weill
Cornell
Medicine