Communicating Your Data to Statisticians

BERDC Special Topics Talk 4

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UNIVERSITY OF NORTH DAKOTA

Introduction

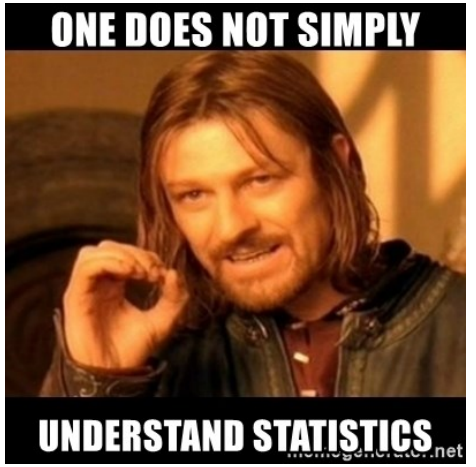
Goal: Improve communication between researchers and statisticians

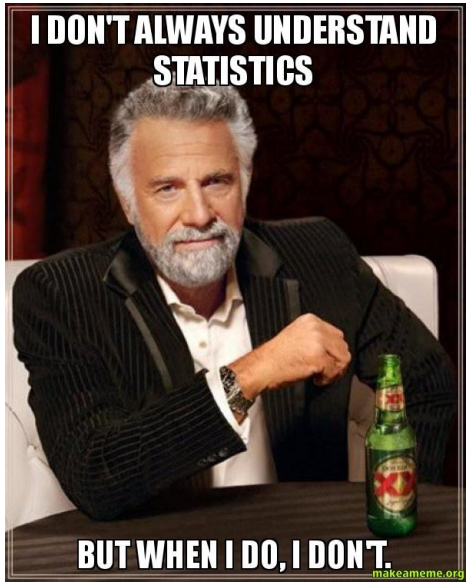
A pragmatic approach to statistics:

- Do I understand what the statistical method is doing?
- Is it interpretable?
- Is it comparable to other studies?
- Can I justify and defend it?

Bridging the gap between experiment and analysis:

- A major obstacle is the move from the specific (researcher) to the general (statistician)
- The target outcome is to train researchers to generalize their data so that statisticians can grasp it quickly and be able to understand the specifics of it
- The approach I'll be using is what I call "The 4 Big Topics"





DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

The 4 Big Topics

UNIVERSITY OF NORTH DAKOTA

T1: Big Picture Goal

- Sum up your experiment in a few, clear sentences
- Sum up your experiment in terms of Y and X variables

T2: General Test Type

- Style of test (t-test, ANOVA regression, etc.)
- Discrete results (tables/graphs) expected

T3: Variables Information

- Variable name and type
- Variable measurement info

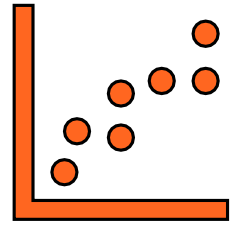
T4: Sample Data

- Provide preliminary or mock data
- Show report of similar experiments

"Explanation"

Y-variable=X-variables

Test-style



Var 1	Var 2	Var 3
Num.	Num.	Cat.
Cont.	Discr.	3 Grps
50-100	1-5	A,B,C

Var 1	*Var 2*	*Var 3*
50.2	*1*	*A*
65.0	*2*	*B*
75.7	*5*	*C*

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

T1: Big Picture Goal

UNIVERSITY OF NORTH DAKOTA

1. **Sum up your experiment in a few, clear sentences**
   - Translate details into the big picture
   - Abstracts work well here
   - May be several sub-experiments to explain
2. **Sum up your experiment in terms of Y and X variables**
   - Examples:
     - *We wanted to model **Y-variable** as a function of **X-variable(s)**.*
     - *We wanted to see if **X-variable(s)** predicted **Y-variable.***
     - *We wanted to see if **Y-variable** was affected by **X-variable(s).***
   - Can also add more information:
     - … while correcting for **confounding-variable(s).**
     - … across time.
     - … etc.

Examples

We wanted to model nematode growth rate (Y) as a function of gene copy number (X).

We wanted to see if the presence of nuclear power reactors in a state (X) predicted brain cancer incidence rates (Y).

We wanted to see if COVID-19 county death rates (Y) were affected by county rural-urban status (X)

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

T2: General Test Type

UND UNIVERSITY OF NORTH DAKOTA.

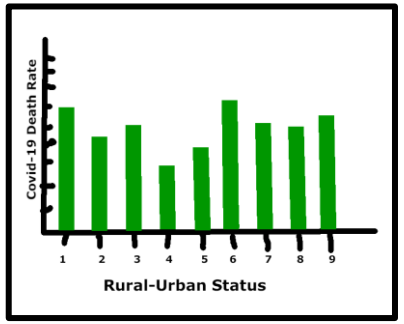
1. **Determine the general style of test**
   - What type of test do you anticipate your experiment(s) to be?
     - Comparing means across 2 groups: T-test
     - Comparing means across 3+ groups: ANOVA
     - Comparing 2 or more numerical values: regression
     - Comparing frequencies across categories: Chi-Square
   - More complicated or specific?
     - Generalized or mixed model?
     - Logistic, Poisson, or other regression?
     - Survival analysis?
     - Etc
2. **Sketch out the expected resulting graph or table**
   - Helps visualize what the outcome will be and therefore what will need to go into it
   - Can be a simple as sketching out hand-drawn graphs and labeling the axes

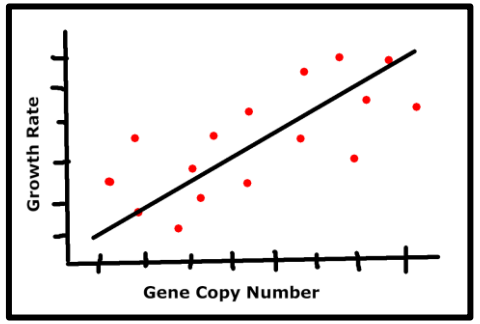
Examples

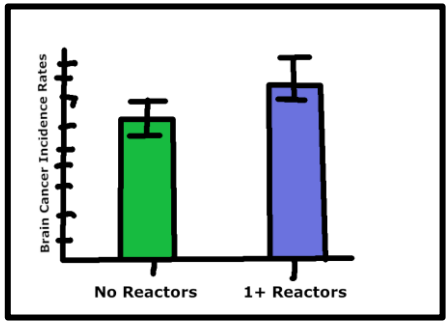
Modeling nematode growth rate as a function of gene copy number: **Regression**

State nuclear power reactor status as a predictor for brain cancer incidence rates: **T-test**

COVID-19 county death rates across rural-urban county type (9 groups): **ANOVA**







DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

T3: Variable Information

UND UNIVERSITY OF NORTH DAKOTA.

1. **Write down the variable names and types**
   - Provide full name rather than abbreviation
     - **LBXWBCSI** [*unhelpful*] - > **White blood cell count** [*helpful*]
     - Include units if possible-> **White blood cell count (1000 cells/uL)**
   - Numerical or categorical:
     - Numerical: continuous or discrete?
     - Categorical: ordinal or nominal?
   - Other considerations:
     - Controlling for confounding variables or random effects?
2. **Write down the variable measurement info**
   - Numerical: central tendency, spread, and distribution
   - Categorical: number of groups, group size

Important point: need to translate from method to measurement

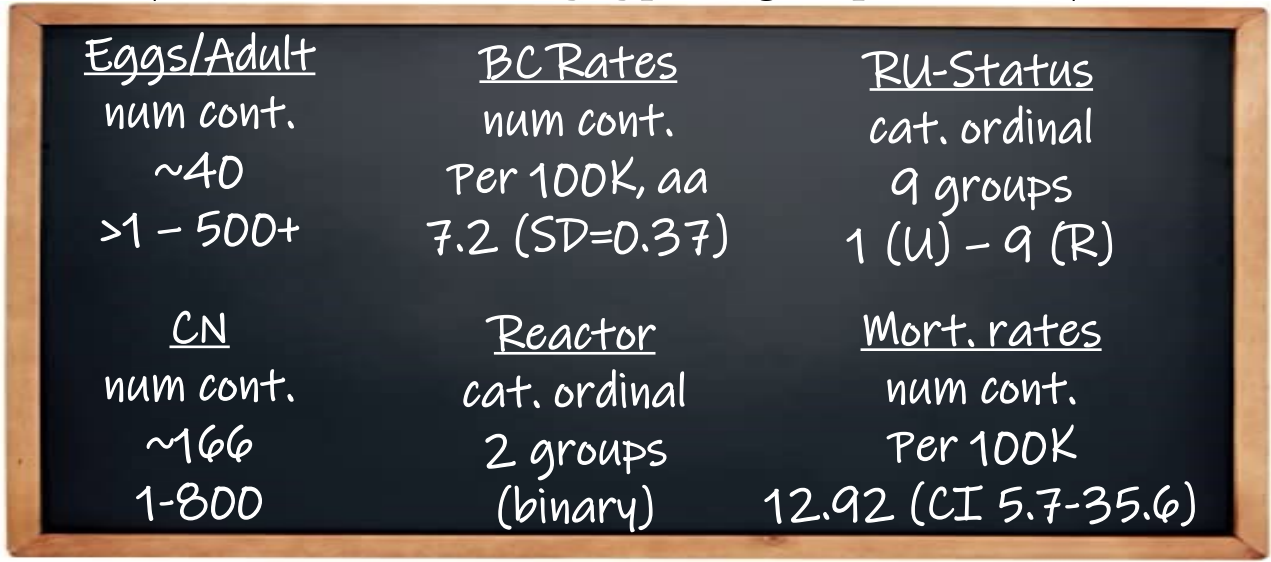
- Ex. 'running a Western Blot' -> translates to generating protein concentrations in (µL)

Examples

Modeling nematode growth rate as a function of gene copy number

State nuclear power reactor status as a predictor for brain cancer incidence rates

COVID-19 county death rates across rural-urban county type (9 groups)



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

T4: Sample Data

UNIVERSITY OF NORTH DAKOTA

1. **Provide preliminary or mock data**
   - Best would be preliminary, pilot, sample data
   - Doesn't need to be full dataset, just enough to get a feel for the analysis
   - Also good for power analysis
   - If not, could also try to create mock data through simulation
   - https://aosmith.rbind.io/2018/08/29/getting-started-simulating-data/
2. **Show report of similar experiments**
   - Previous experimental design and results
   - Published papers (proposal references work well)

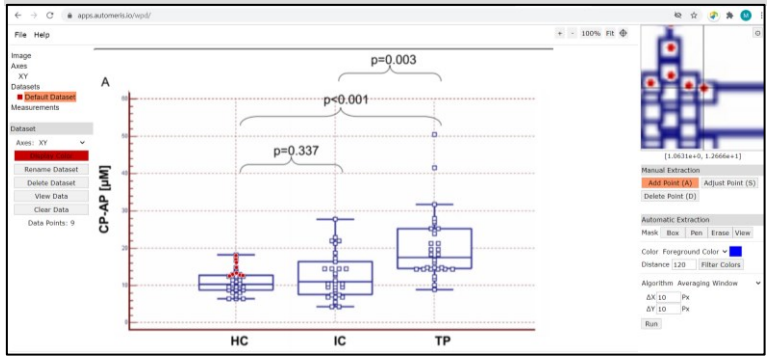
Simulation Example

```
Y <-rnorm(n=100, mean=20, sd=5)
Y <-round(Y,2)
X <-rpois(n=100, lambda=15)
C <-c(rep('A',25), rep('B',25), rep('C',25), rep('D',25))
sim_data <-data.frame(Y,X,C)
View(sim_data)
```

Data Extraction Example

WebPlotDigitizer

Web based tool to extract data from plots, images, and maps



	Y	X	C
1	13.14	18	A
2	19.82	19	A
3	16.81	15	A
4	17.40	11	A
5	22.34	12	A
6	19.61	19	A
7	16.41	19	A
8	27.08	14	A
9	30.76	19	A
10	9.97	21	A

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Other Considerations

UND
UNIVERSITY OF NORTH DAKOTA.

- More involved experiments will have more information to consider
  - Is time involved (ex. repeated measures)?
  - Are there fixed (age, sex, etc.) or random effects (plot, batch, etc.) you want to account for?
  - Is there missing data?
  - Anything else that would make the analysis harder?
- Better to have the researcher start to answer the questions than have the statistician try to answer them with less intimate knowledge of the research
- Don't have to have all the answers
- Useful procedure is to start with a basic setup and add complexity from there

DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

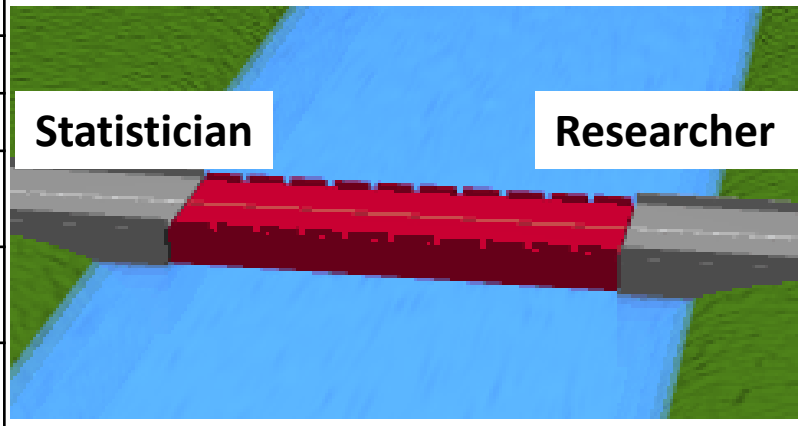
What's Next

UND UNIVERSITY OF NORTH DAKOTA.

Closing points

- 4 Big Topics helps statistician understand research project and helps researcher understand, frame, and communicate their research
- The clearer the design, the better the chance for successful outcome
- Well-defined data plan protects against failure due to faulty planning
- Researchers are encouraged to think about their data

BERDC project support by priority	
1	Career development awards (DaCCoTA CTR)
2	Pilot projects (DaCCoTA CTR)
3	Resident/trainee research
4	Federally funded research
5	General consultation on research design and methods
6	Grant application support for clinical research
7	Industry-funded research and clinical trials
8	Manuscript support
9	Other funded research
10	Privately funded research
11	Unfunded research



Feedback:

- Survey: https://und.qualtrics.com/jfe/form/SV_exoFvxRYpkddzhz
- Questionnaire: https://und.qualtrics.com/jfe/form/SV_dcyUSPLhD4cmP5Q

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF
NORTH DAKOTA.

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY