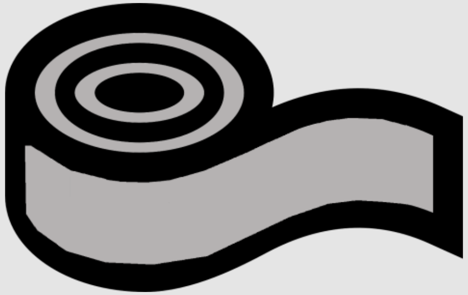
Statistical Rules to Tape to Your Forehead

BERDC Special Topics Talk 6



DaCCoTA

DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

Opening

<u>**Goal**</u>: present a clear list of rules for running a statistical analysis that are ***useful*** and ***memorable***

- Sources: personal experience and the expertise of others
- Not exhaustive, inviolable, or strictly linear
- Example using shark data

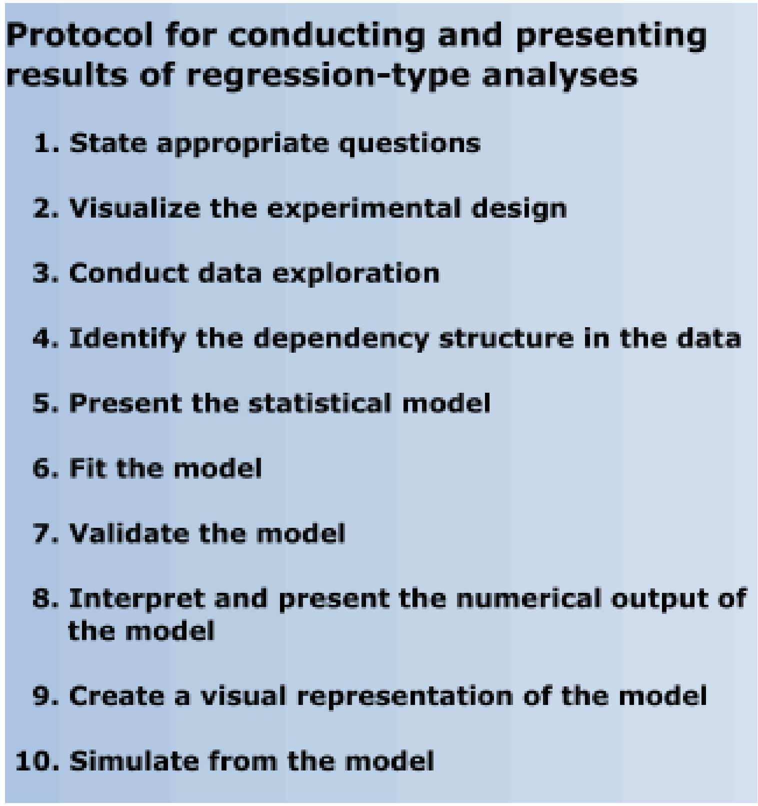
Before Moving On:

Pre-test: https://und.qualtrics.com/jfe/form/SV_ebTk64vnD3yrrPo

Sources

Zuur & Ieno, 2016

A protocol for conducting and presenting results of regression-type analyses



Protocol for conducting and presenting results of regression-type analyses

1. State appropriate questions
2. Visualize the experimental design
3. Conduct data exploration
4. Identify the dependency structure in the data
5. Present the statistical model
6. Fit the model
7. Validate the model
8. Interpret and present the numerical output of the model
9. Create a visual representation of the model
10. Simulate from the model

Kass et al, 2016

Ten Simple Rules for Effective Statistical Practice

Rule 1	**Statistical Methods Should Enable Data to Answer Scientific Questions**
Rule 2	**Signals Always Come with Noise**
Rule 3	**Plan Ahead, Really Ahead**
Rule 4	**Worry about Data Quality**
Rule 5	**Statistical Analysis Is More Than a Set of Computations**
Rule 6	**Keep it Simple**
Rule 7	**Provide Assessments of Variability**
Rule 8	**Check Your Assumptions**
Rule 9	**When Possible, Replicate!**
Rule 10	**Make Your Analysis Reproducible**

Goodman et al, 2014

Ten Simple Rules for the Care and Feeding of Scientific Data

ASA, 2018

Ethical Guidelines for Statistical Practice

The Rules

1) Annunciation
2) Exploration
3) Assumptions
4) Building
5) Testing
6) Justifying
7) Presentation
8) Interpretation

Rule 1: Annunciation

Clearly state your research questions.

- **Write one or two sentences**
- **Be painfully explicit**
- **Keep identity of variables in mind**

Abstract Examples:

Here, we use genome-wide CRISPR–Cas9 screening to establish that a T cell receptor (TCR) recognized and killed most human cancer types via the monomorphic MHC class I-related protein, MR1, while remaining inert to noncancerous cells (Crowther et al. 2020)

Using a combination of adrenalectomy, denervation, chemogenetics, cell ablation and knockout of the adrenergic receptor specifically in melanocyte stem cells, we find that the stress-induced loss of melanocyte stem cells is independent of immune attack or adrenal stress hormones (Zhang et al. 2020)

Here we sequenced whole genomes of 632 colonies derived from single bronchial epithelial cells across 16 subjects. Tobacco smoking was the major influence on mutational burden, typically adding from 1,000 to 10,000 mutations per cell; massively increasing the variance both within and between subjects; and generating several distinct mutational signatures of substitutions and of insertions and deletions (Yoshida et al. 2020)

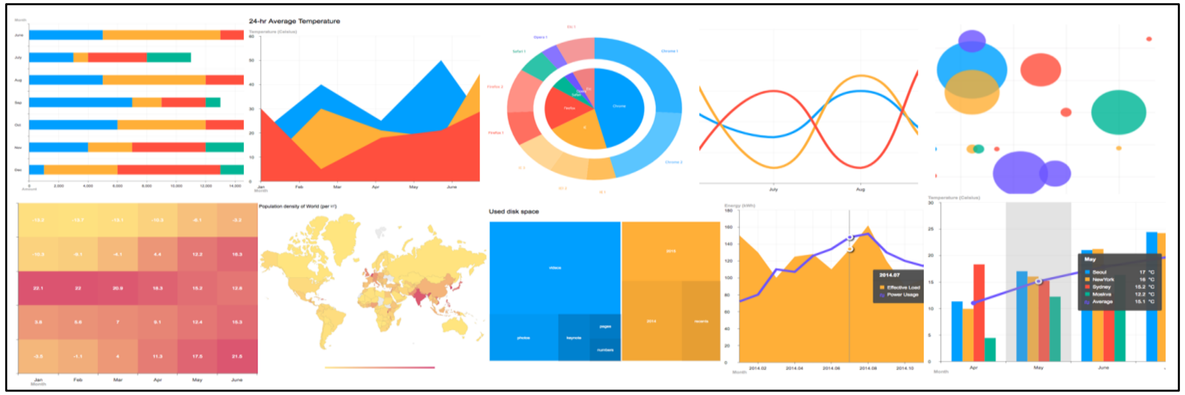
DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

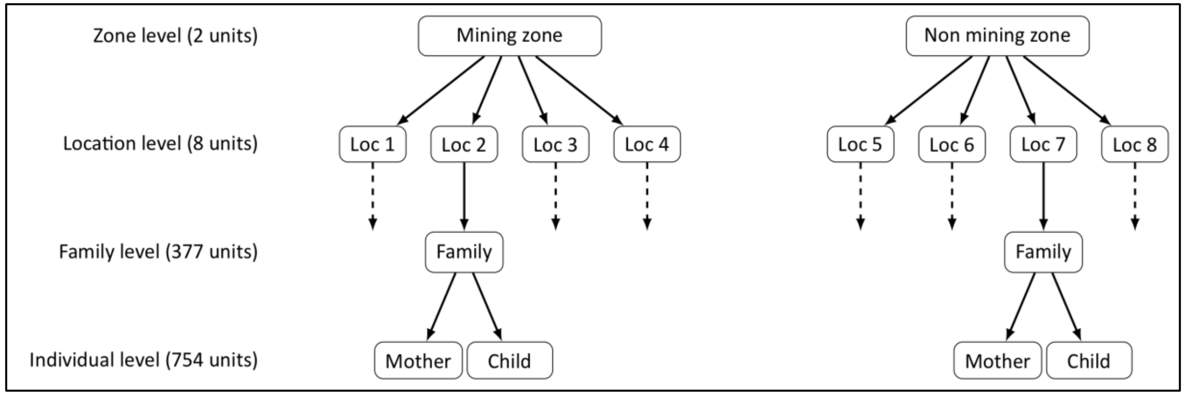
Rule 2: Exploration

UND
UNIVERSITY OF
NORTH DAKOTA

Get to know your data in and out.

- Summary statistics are your friend
- Graphs, graphs, graphs!
- Determine design and dependencies

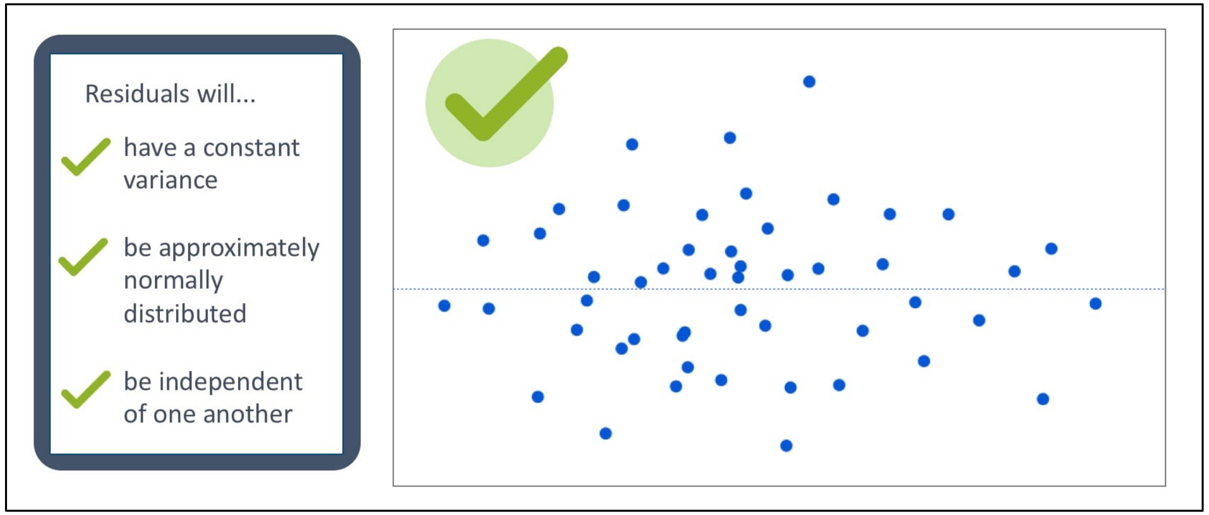



Zone level (2 units)
Location level (8 units)
Family level (377 units)
Individual level (754 units)

Rule 3: Assumptions

Ensure that your data meets test assumptions.

- **Check during design, before analysis, after analysis**
- **Be ready to try different types of tests**
- **Use many piece of information**



Residuals will...

✓ have a constant variance

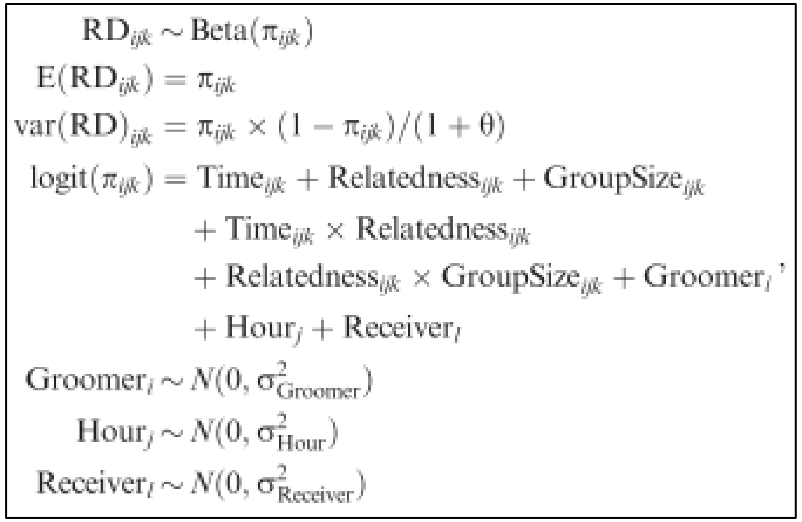
✓ be approximately normally distributed

✓ be independent of one another

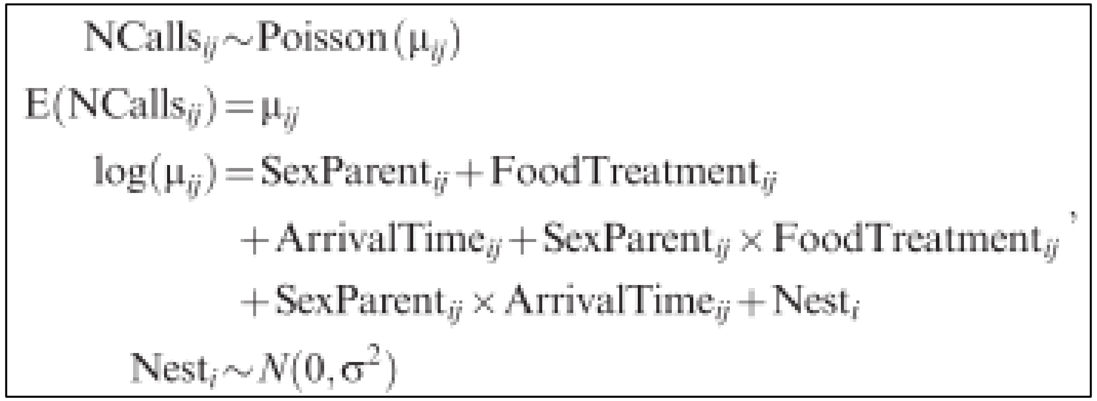
Rule 4: Building

Build one or more statistical models for your data.

- **Model type**
- **Model equation**
- **Model selection**



$$\begin{aligned}
\mathrm{RD}_{ijk} &\sim \mathrm{Beta}(\pi_{ijk}) \\
\mathrm{E}(\mathrm{RD}_{ijk}) &= \pi_{ijk} \\
\mathrm{var}(\mathrm{RD})_{ijk} &= \pi_{ijk} \times (1 - \pi_{ijk})/(1 + \theta) \\
\mathrm{logit}(\pi_{ijk}) &= \mathrm{Time}_{ijk} + \mathrm{Relatedness}_{ijk} + \mathrm{GroupSize}_{ijk} \\
&\quad + \mathrm{Time}_{ijk} \times \mathrm{Relatedness}_{ijk} \\
&\quad + \mathrm{Relatedness}_{ijk} \times \mathrm{GroupSize}_{ijk} + \mathrm{Groomer}_i\,, \\
&\quad + \mathrm{Hour}_j + \mathrm{Receiver}_l \\
\mathrm{Groomer}_i &\sim N(0, \sigma^2_{\mathrm{Groomer}}) \\
\mathrm{Hour}_j &\sim N(0, \sigma^2_{\mathrm{Hour}}) \\
\mathrm{Receiver}_l &\sim N(0, \sigma^2_{\mathrm{Receiver}})
\end{aligned}$$



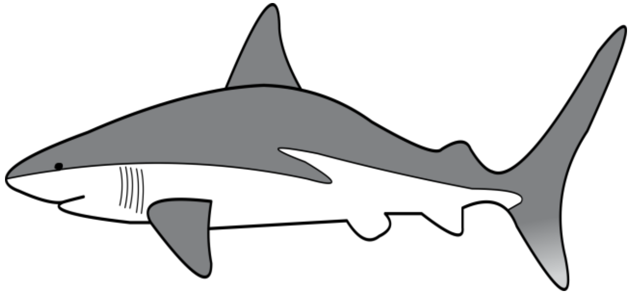
$$\begin{aligned}
\mathrm{NCalls}_{ij} &\sim \mathrm{Poisson}(\mu_{ij}) \\
\mathrm{E}(\mathrm{NCalls}_{ij}) &= \mu_{ij} \\
\log(\mu_{ij}) &= \mathrm{SexParent}_{ij} + \mathrm{FoodTreatment}_{ij} \\
&\quad + \mathrm{ArrivalTime}_{ij} + \mathrm{SexParent}_{ij} \times \mathrm{FoodTreatment}_{ij}\,, \\
&\quad + \mathrm{SexParent}_{ij} \times \mathrm{ArrivalTime}_{ij} + \mathrm{Nest}_i \\
\mathrm{Nest}_i &\sim N(0, \sigma^2)
\end{aligned}$$

Rule 5: Testing

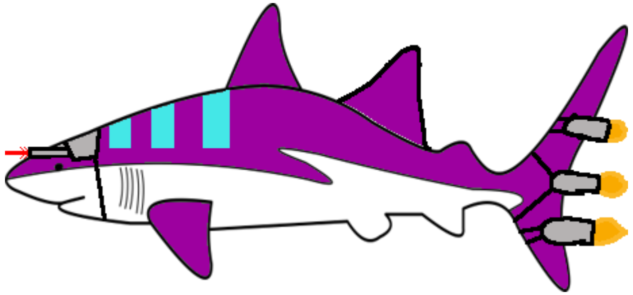
Test your statistical models on your data.

- **Start with the simplest model**
- **Different methods of tweaking**
- **Keep careful notes**

‘Simple Shark Model’



‘Advanced Shark Model’



Rule 6: Justifying

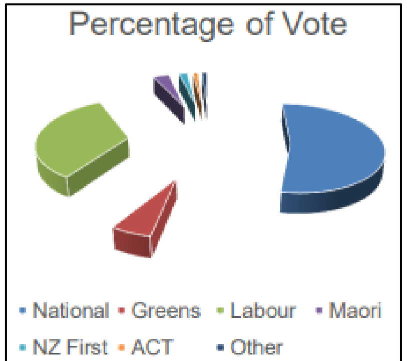
Justify your model results.

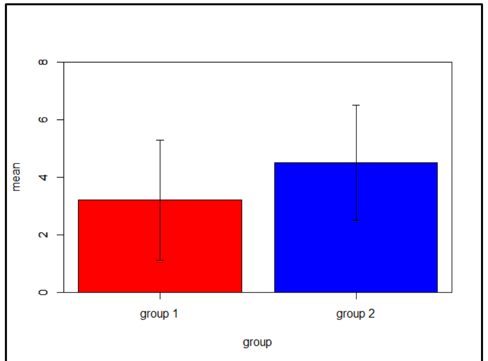
- **Fit statistics**
- **Link to summary stats and research background**
- **Have a 'why' for each piece of the statistical process**

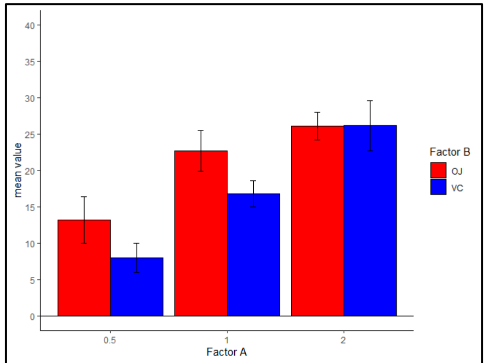
Rule 7: Presentation

Present your results through text, tables, and figures.

- **Written results should include estimates, test statistics, and significance values**
- **Tables and figures should add clarity, not banality**
- **Make uncertainty a key feature**







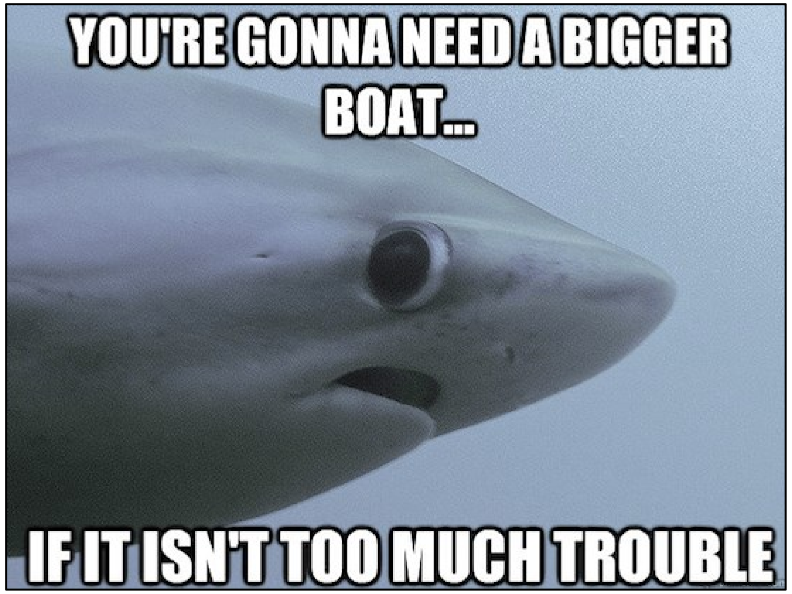
DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF
NORTH DAKOTA

Rule 8: Interpretation

Clearly explain what your results are and what they mean.

- **Follow numerical results with clear wording**
- **Relate back to research question**
- **Project outside the study**


YOU'RE GONNA NEED A BIGGER BOAT...
IF IT ISN'T TOO MUCH TROUBLE

The Rules (again)

1) Annunciation 👅
2) Exploration 👣
3) Assumptions 🦴
4) Building 💪
5) Testing ✋
6) Justifying 🦶
7) Presentation 👁
8) Interpretation 🧠

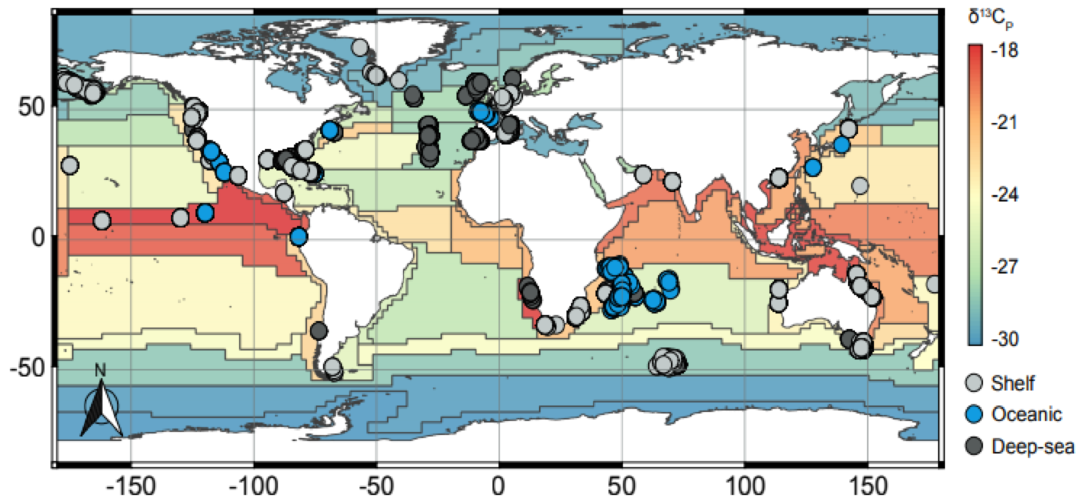
Shark Example

A Global Perspective on the Trophic Geography of Sharks

1) Annunciation
2) Exploration
3) Assumptions
4) Building
5) Testing
6) Justifying
7) Presentation
8) Interpretation

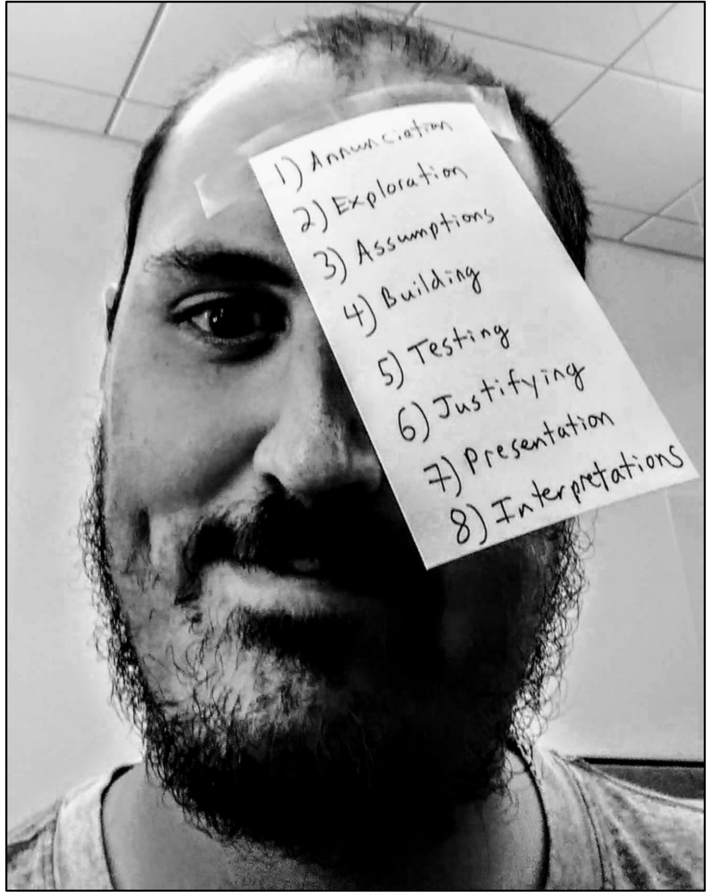
Abstract

Sharks are a diverse group of mobile predators that forage across varied spatial scales and have the potential to influence food web dynamics. The ecological consequences of recent declines in shark biomass may extend across broader geographic ranges if shark taxa display common behavioural traits. By tracking the original site of photosynthetic fixation of carbon atoms that were ultimately assimilated into muscle tissues of 5,394 sharks from 114 species, we identify globally consistent biogeographic traits in trophic interactions between sharks found in different habitats. We show that populations of shelf-dwelling sharks derive a substantial proportion of their carbon from regional pelagic sources, but contain individuals that forage within additional isotopically diverse local food webs, such as those supported by terrestrial plant sources, benthic production and macrophytes. In contrast, oceanic sharks seem to use carbon derived from between 30° and 50° of latitude. Global-scale compilations of stable isotope data combined with biogeochemical modelling generate hypotheses regarding animal behaviours that can be tested with other methodological approaches.



Wrap-up

May these rules 'stick' with you on your statistical journey



Please take the post-test and survey:

Post-test: https://und.qualtrics.com/jfe/form/SV_4Tol5Rj6GcoGkv4
Survey: https://und.qualtrics.com/jfe/form/SV_5iHQvVd2ooBpoSa

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

References

UND
UNIVERSITY OF
NORTH DAKOTA.

Rule Reference Materials:

- https://besjournals.onlinelibrary.wiley.com/doi/epdf/10.1111/2041-210X.12577
- https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1004961
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998871/pdf/pcbi.1003542.pdf
- https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Example Abstracts:

- https://www.nature.com/articles/s41590-019-0578-8
- https://www.nature.com/articles/s41586-020-1935-3
- https://www.nature.com/articles/s41586-020-1961-1

Other:

- https://www.theanalysisfactor.com/three-rules-statistical-analysis-unlearn/
- https://www.jmp.com/en_be/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html
- https://www.iase-web.org/islp/apps/gov_stats_graphing/GoodBad/GoodBadGraphs.pdf
- https://www.thedailybeast.com/democrats-have-already-jumped-the-shark-on-judge-kavanaugh

Shark Data:

- https://nsuworks.nova.edu/cgi/viewcontent.cgi?article=1916&context=occ_facarticles
- https://datadryad.org/stash/dataset/doi:10.5061/dryad.d1f0d

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