An Overview of the Statistical Challenges to Understanding the Ecology and Management of Regulated Rivers

Joe Thorley

May 7th, 2015

The challenges are numerous

The statistical challenges associated with understanding the environment, ecology and management of regulated rivers are *numerous*.

The challenges can be divided into **data** and **analytic** challenges.

This presentation briefly defines and illustrates each of the challenges and provides recommendations.

Data Challenges

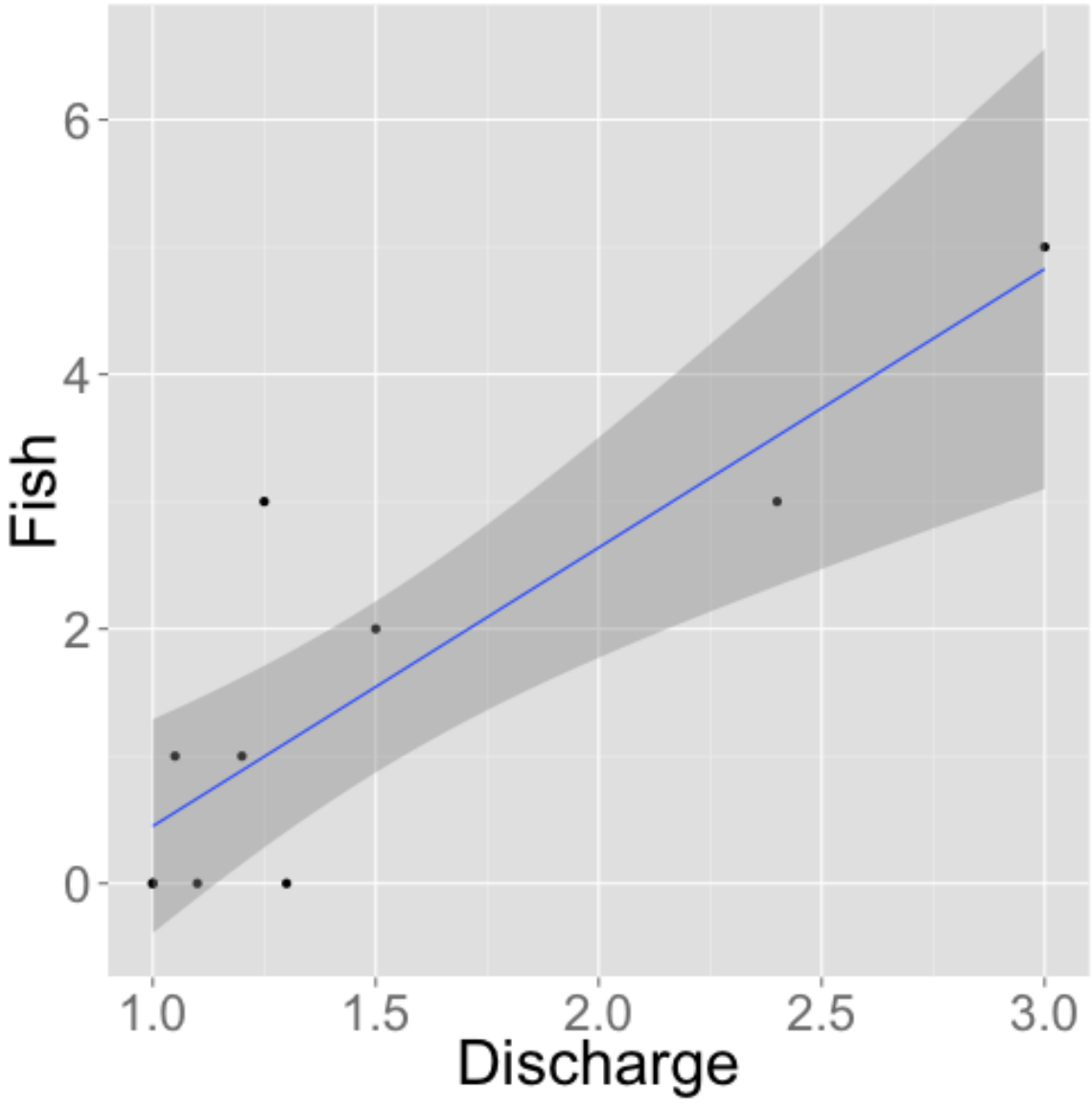
There at least seven types of data challenge

1. insufficient data
2. missing data
3. biased data
4. erroneous data
5. messy data
6. undocumented data
7. lost data

Example Data

Date	Discharge	Fish
2010-01-01	1.10	0
2010-01-02	1.50	2
2010-01-03	3.00	5
2010-01-04	1.25	3
2010-01-05	1.05	1
2010-01-06	1.00	0
2010-01-07	1.30	0
2010-01-08	2.40	3
2010-01-09	1.20	1
2010-01-10	1.00	0

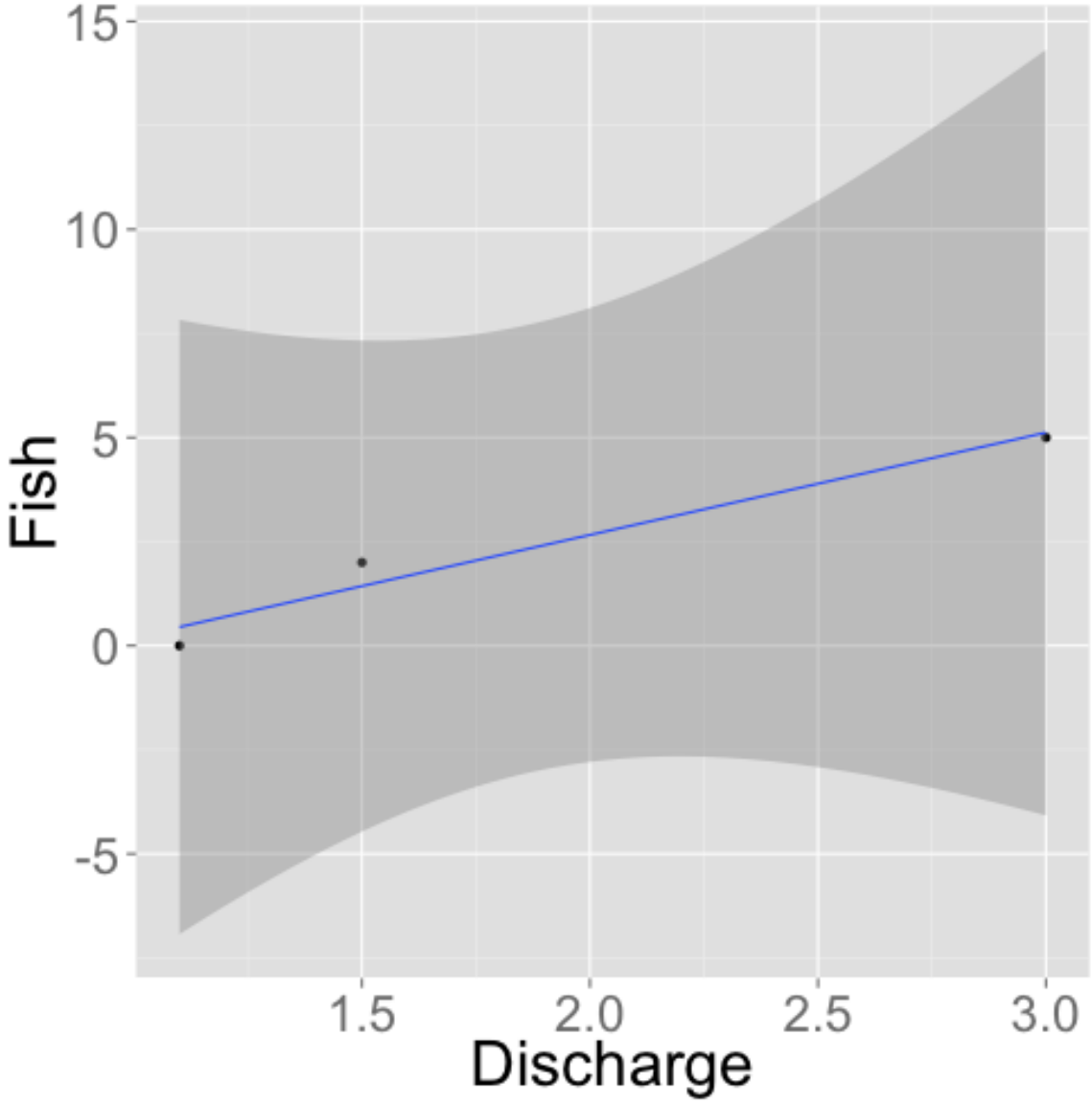
Example Analysis



Insufficient Data

Date	Discharge	Fish
2010-01-01	1.1	0
2010-01-02	1.5	2
2010-01-03	3.0	5

Analysis of Insufficient Data



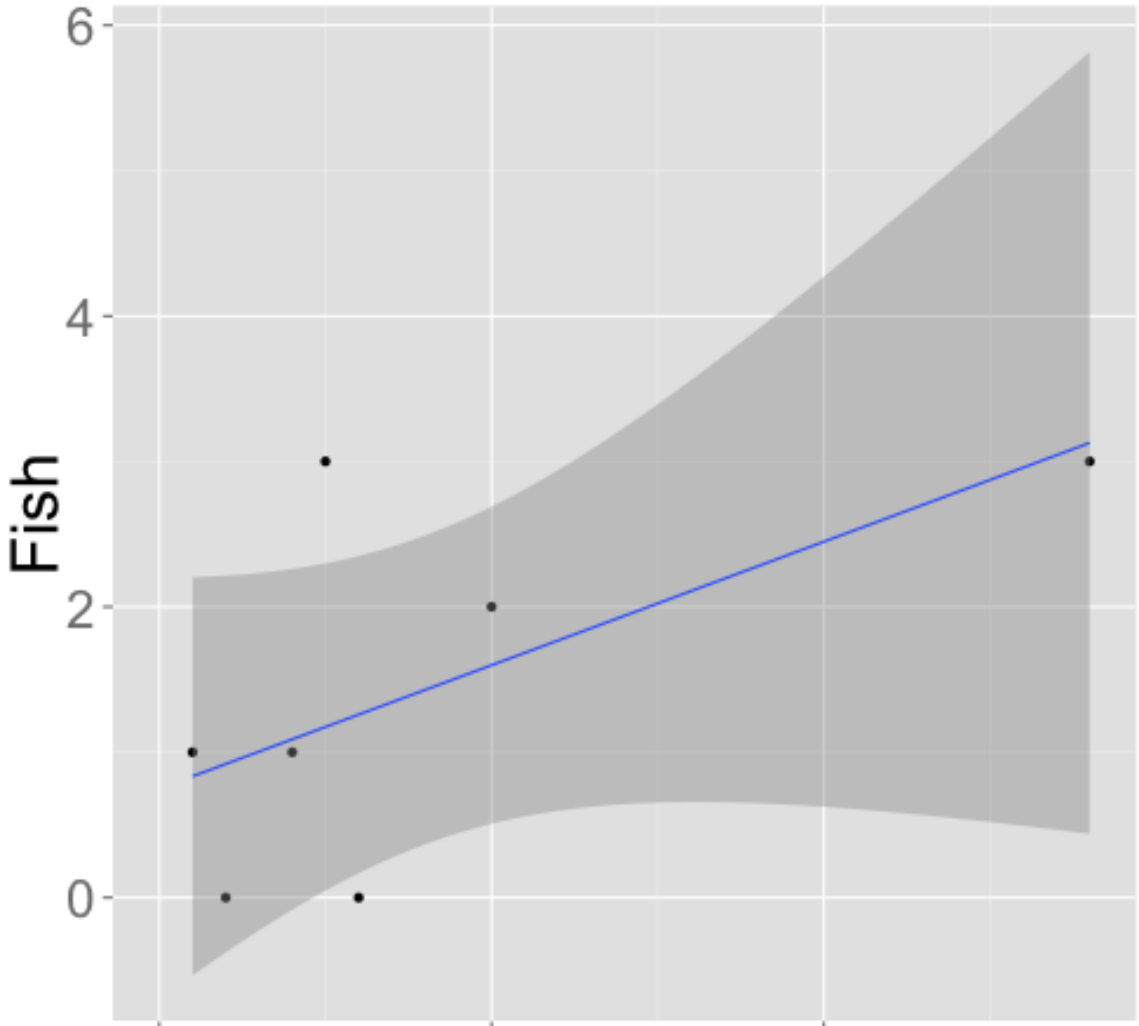
Missing Data

Date	Discharge	Fish
2010-01-01	1.10	0
2010-01-02	1.50	2
2010-01-03	NA	NA
2010-01-04	1.25	3
2010-01-05	1.05	1
2010-01-06	1.00	NA
2010-01-07	1.30	0
2010-01-08	2.40	3
2010-01-09	1.20	1
2010-01-10	NA	0

Analysis of Missing Data

```
## Warning: Removed 3 rows containing missing values (stat_
```

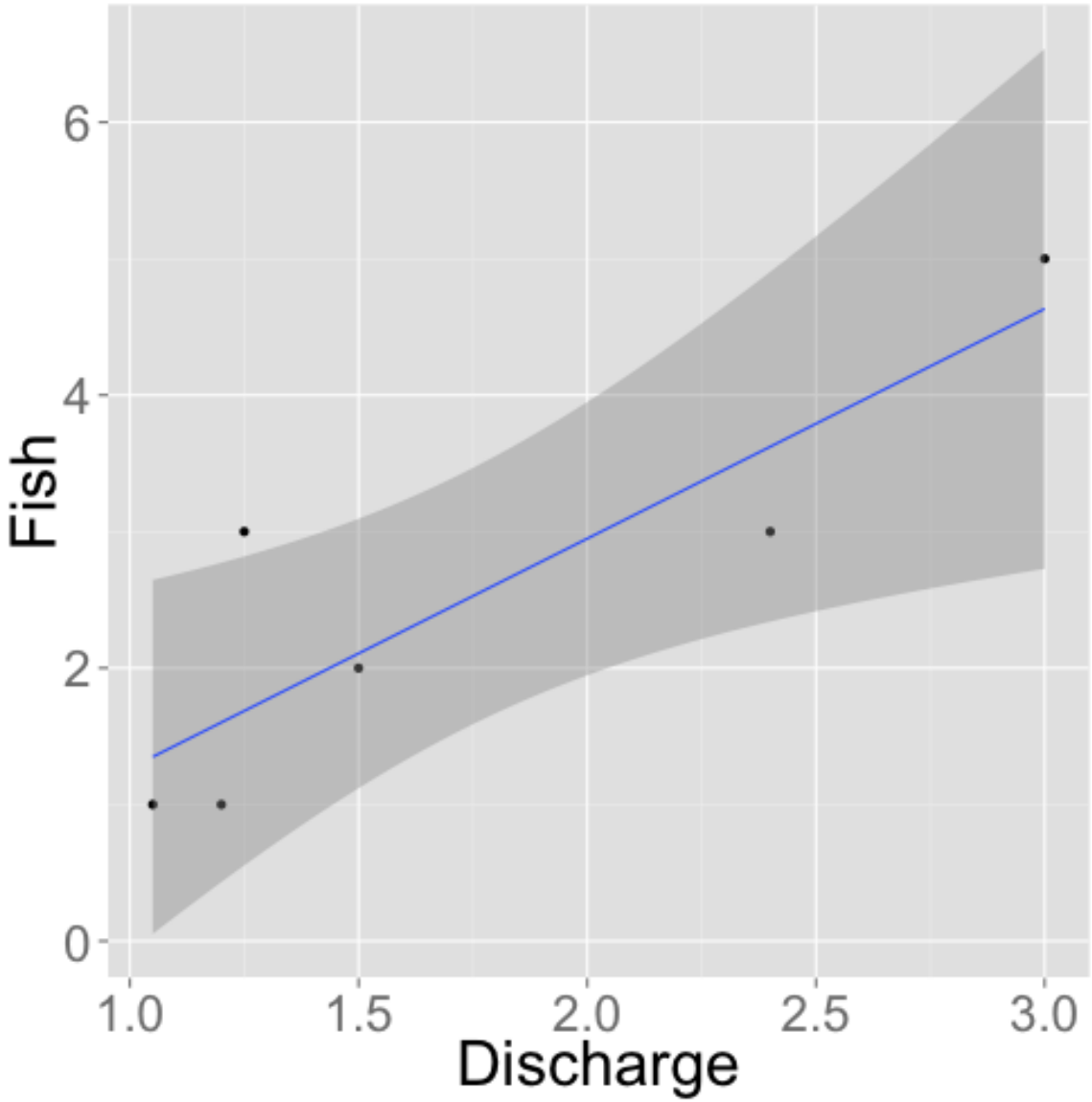
```
## Warning: Removed 3 rows containing missing values (geom_
```



Biased Data

	Date	Discharge	Fish
2	2010-01-02	1.50	2
3	2010-01-03	3.00	5
4	2010-01-04	1.25	3
5	2010-01-05	1.05	1
8	2010-01-08	2.40	3
9	2010-01-09	1.20	1

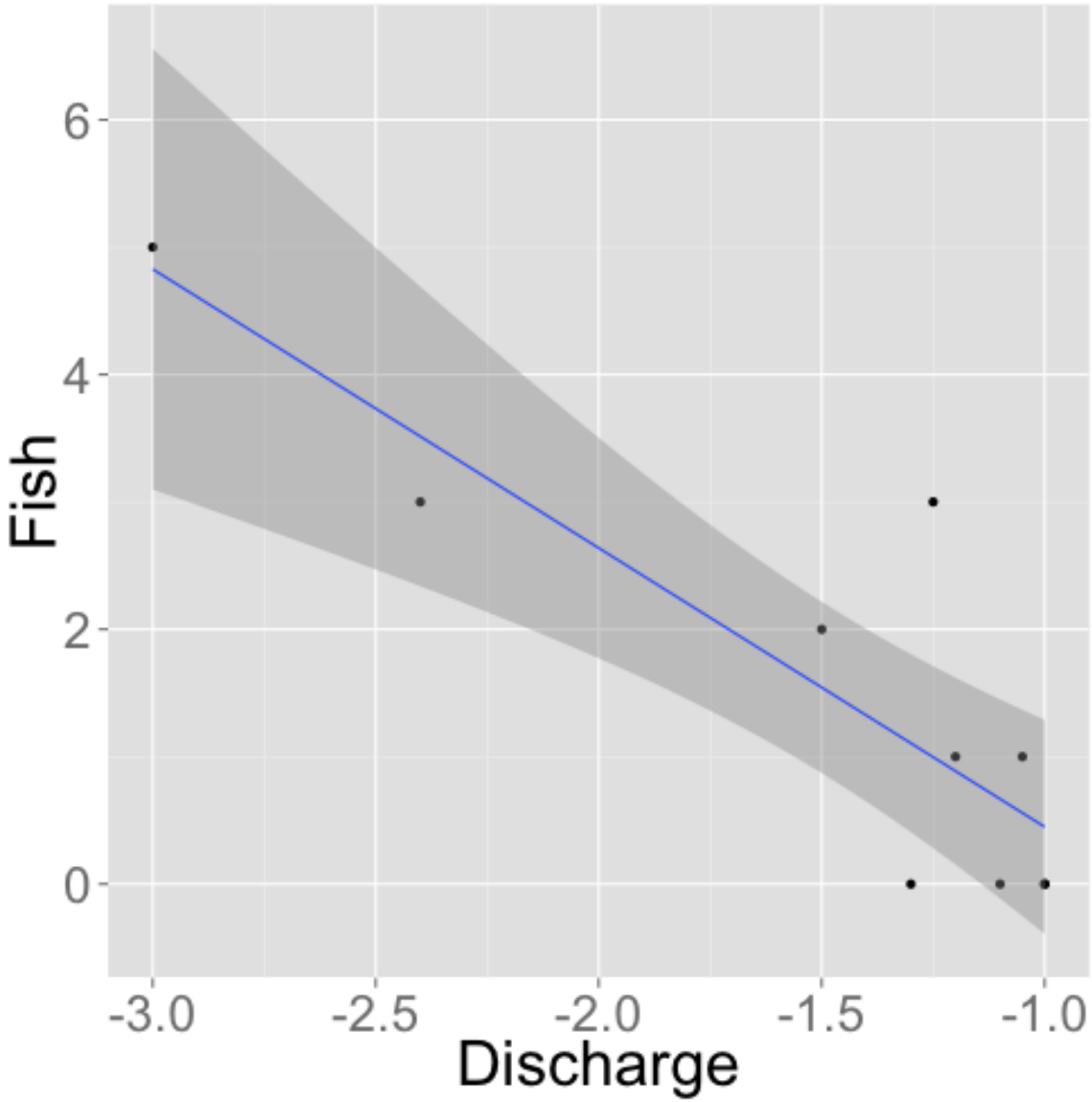
Analysis of Biased Data



Erroneous Data

Date	Discharge	Fish
2010-01-01	-1.10	0
2010-01-02	-1.50	2
2010-01-03	-3.00	5
2010-01-04	-1.25	3
2010-01-05	-1.05	1
2010-01-06	-1.00	0
2010-01-07	-1.30	0
2010-01-08	-2.40	3
2010-01-09	-1.20	1
2010-01-10	-1.00	0

Analysis of Erroneous Data



Messy Data

Variable	2010-01-01	2010-01-02	2010-01-03	2010-01-04	2010-0
Discharge	1.1	1.5	3	1.25	
Fish	0.0	2.0	5	3.00	

Tidy Data

In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

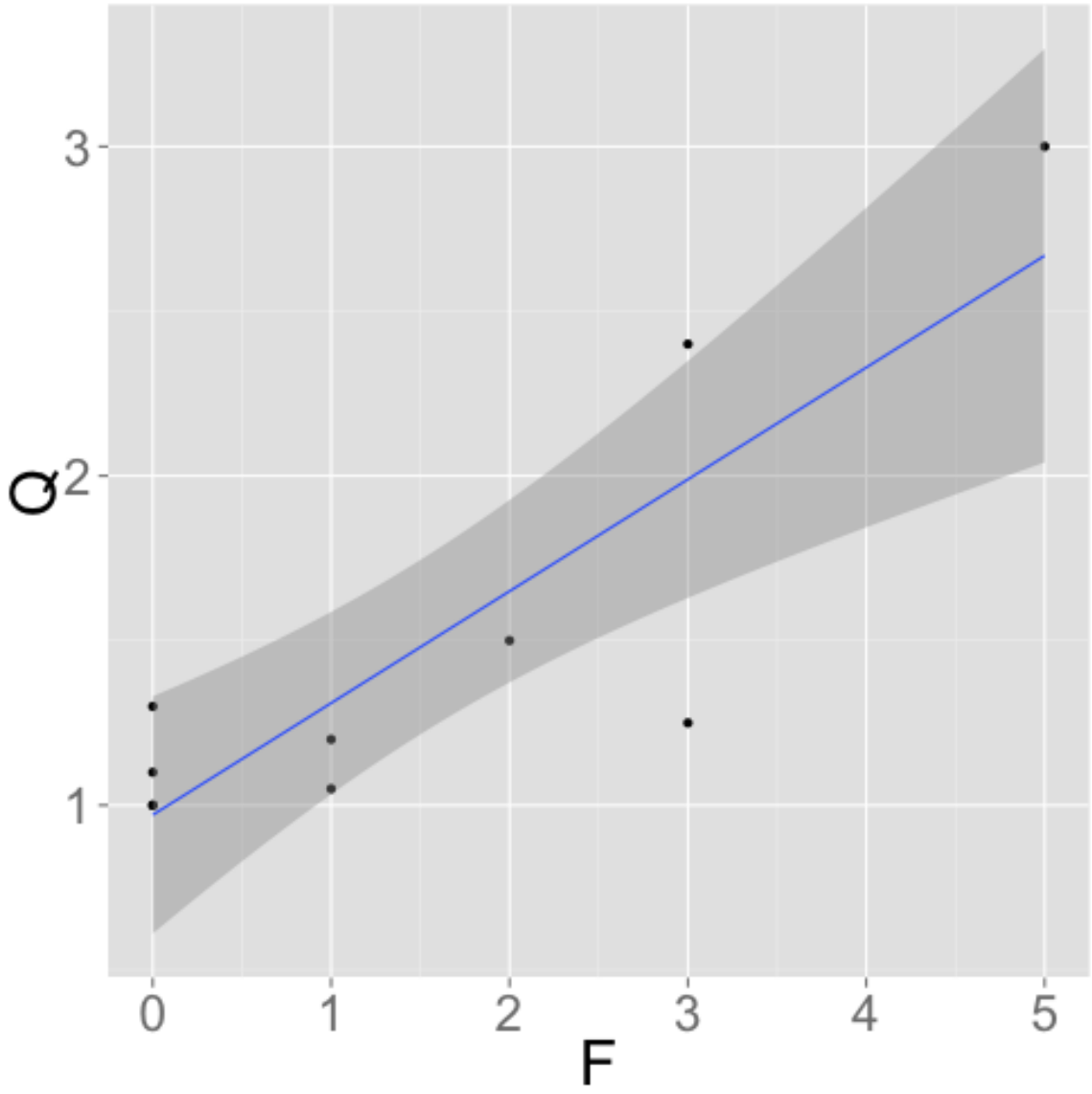
Messy data is any other arrangement of the data.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–22. `http://www.jstatsoft.org/v59/i10/paper`.

Undocumented Data

Date	Q	F
2010-01-01	1.10	0
2010-01-02	1.50	2
2010-01-03	3.00	5
2010-01-04	1.25	3
2010-01-05	1.05	1
2010-01-06	1.00	0
2010-01-07	1.30	0
2010-01-08	2.40	3
2010-01-09	1.20	1
2010-01-10	1.00	0

Analysis of Undocumented Data



Lost Data

Date	Discharge	Fish
NA	NA	NA

Data Solutions

Fortunately most of these data-related challenges can be overcome through recognition of the potential problems and a *study design* that includes where necessary

- a pilot study (all)
- power analyses (sufficient)
- equipment redundancy and forms (complete)
- field protocols (unbiased)
- equipment redundancy and crew training (valid)
- relational database (documented, tidy)
- long-term data curation (archived)

Analytic Solutions

There at least six types of analytic challenge

1. vague questions
2. hidden assumptions
3. derived indices
4. pseudo-replication
5. over-reliance on significance testing
6. researcher degrees of freedom

Vague Questions

Does discharge affect fish abundance?

Which species, life-stage and location (population)?

Which discharge metric? Olden and Poff (2003) identified 171 metrics!

When?

By how much?

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Research and Applications, 19(2), 101–121. http://doi.org/10.1002/rra.700

Hidden Assumptions

A classic example is to collect fish abundance data at *index* sites and then use it to draw river-wide conclusions.

The hidden assumption is that changes in fish density at the index sites are representative of changes at the other sites.

Often this is not the case.

Derived Indices

$$CPUE = \frac{Catch}{Effort}$$

Analysing *CPUE* (as opposed to *Catch* and *Effort*) precludes

1. use of a readily interpretable distribution, i.e., *Poisson* for *Catch*
2. accounting for reduced uncertainty when lower effort
3. testing for catch depletion

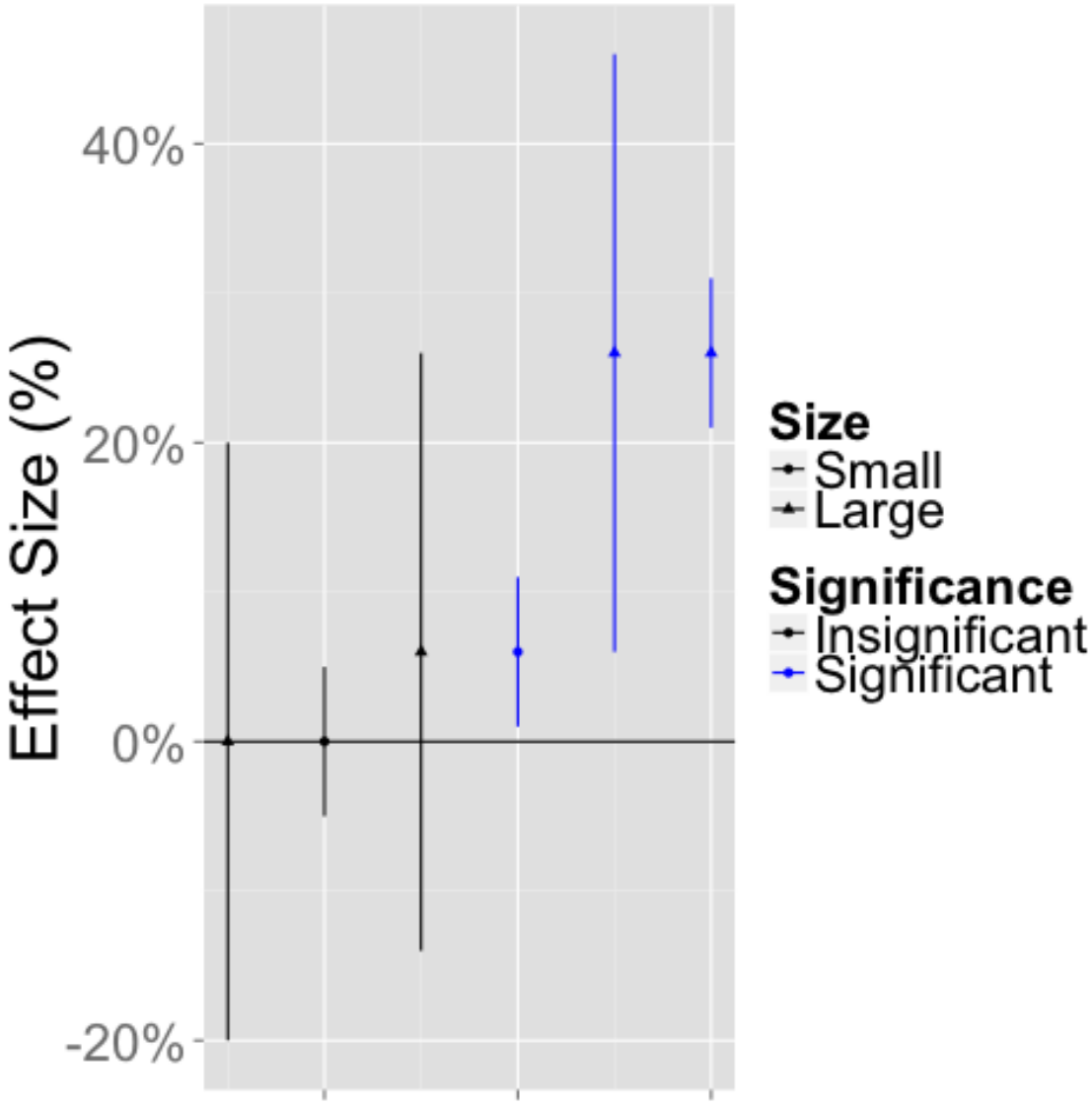
Pseudo-replication

Pseudo-replication occurs when

replicates are not statistically independent

which is almost all the time in ecological studies!

Over-Reliance on Significance Testing



Researcher Degrees of Freedom

Researcher degrees of freedom is an umbrella term for

> *all the data-processing and analytical choices researchers make after seeing the data.*

If researchers decisions are undocumented then the extent to which the probability of getting a false positive has been inflated cannot be assessed.

Analytic Solutions

Most of these analysis-related challenges can be overcome through

hierarchical (account for non-independence), Bayesian scientific (explicitly describe observational and biological processes) models that allow the estimation of secondary parameters (answers to specific questions) with credible intervals (effect sizes) and the release of version-controlled code (researcher decisions).