Systematic Review Hypotheses

Willson Gaul

10 August 2018

*Note* - Wikipedia says Chi-square test assumes categories are mutually exclusive. Jon thinks it's actually that they need to be independent (counts in cells need to be independent), so mutually exclusive categories isn't actually required as long as the categories aren't correlated.

**About 26 logistic or Poisson regressions planned**

**Expect 300 studies (n = 300)**

Number of tests could be reduced if I can find something like a chi-squared test that just does an overall test for independence between two variables.

## Authors and Data Providing Institutions

**Who choses to use biological records data?**

: The first author is more likely to use data held by their own institution than by other institutions.

: The number of studies with a lead author from the institution that provided the data is the same as expected by chance.

*Proposed Test*: Permutation test of lead author institutions.

## Temporal Extent

: The mean temporal extent of studies using only what, where, when data is longer than the mean temporal extent of studies using richer data types.

: Data type does not affect the mean temporal extent of studies. // The mean temporal extent of studies using what, where, when-only data is the same as the mean temporal extent of studies using other data types.

*Proposed Test*: Linear regression with natural-log transformed time (in years) as the outcome and data types as predictors.

## Study Question Paradigm

There are two questions about this topic.

**What is the most popular study question paradigm?**

: Some broad study question paradigms are more common than others in studies using biological records.

: The number of studies pursuing each of the broad study question paradigms is the same.

*Proposed Test* One Poisson regression with study question paradigm as predictor.

**What data types are used for each study question paradigm?**

Drop compiled individual from coding. For this analysis, use only community and individual species question types. Keep coding "Methodology" but don't include it here, because I think in most cases the methodolgy development will be developing methods for analysing either individual species or community questions. So the community/individual sp paradigm dichotomy still works.

: Different broad study question paradigms use different data types.

: Data type does not affect study question paradigm.

*Proposed Test* Analyze as a table (some sort of chi-squared) (table will be 2 x 7 or so).

Two separate Poisson regressions with data types as predictors and counts of number of studies using each study question paradigm as responses.

After meeting with Jon (10 Aug), could do something like:

count ~ dt\_abundance + dt\_samplingEffort + sq\_community + sq\_individualSp + dt\_abundance:sq\_community + dt\_abundance:sq\_individualSp + ...

OR, could do something like this:

|  |  |  |
| --- | --- | --- |
| Module | Var2 | Count |
| Maths | Irish + Male | 8 |
| Maths | Irish + Femal |  |
| Maths | Not Irish + Male |  |
| Maths | Not Irish + Female |  |
| History | Irish + Male |  |
| ... | ... | ... |

Where Var2 has every combination of the variables along the top row of a pseudo-contingency table (in my case, all the data type variables). Could then use Poisson regression to fit:

count ~ Module + Var2 + Module:Var2

and

count ~ Module + Var2

to test for than interaction (whether there is a dependency between Module and Var2, or in my case between data type and study question paradigm). This would test the null hypothesis I've written above, but isn't actually very interesting, b/c it we would immediately then want to know *which* data types affect *which* study question paradigms.

*Notes* This question in particular has a problem of what should be the predictor and what should be the outcome in a model-based test. Using data type as predictor (the data type exists out there always and everywhere, like the temperature) and number of studies as response suggests Poisson regression (unbounded count of studies in an interval of time). However, if the causation goes the other way, then there are a fixed number of studies for each question paradigm, and there is a probability that each of those studies uses only what, where, when data (which would suggest logistic regression).

## Data Type and Analysis Approach

The question here is whether the data type affects the analysis approach.

: Different types of data are analysed with different broad analysis approaches.

: Different types of data are analysed with the same broad analysis approaches.

: What, where, when data are analysed with inference and/or prediction less often than richer data are.

: What, where, when data are analysed with the same broad analysis approach as other data.

*Proposed Test* Logistic Regression with probability of statistical / model-based results as the outcome and data types as predictors.

$logit(P\_model) ~ \_0 + $

## Specific Study Focus

: Studies with different specific focuses use different data types. // Data type affects the rate at which the data are used to to answer questions with each of the main focuses.

: The data type used is independent of the specific main focus of the study. // The rates at which studies with each main focus occur is not affected by data type. //

*Proposed Test*: Poisson regression or logistic regression depending on outlook.

*Notes*: There are different ways this process might work.

1. "I'm gonna do a study on Bergmann's rule." This would be a trial, and there would be a probability of using what, where, when only data (low probability). This suggests logistic regression, in which each study is a trial, and the specific focus (ecological theory in this case) is a predictor for the probability of using a data type (what where when data). Do a logistic regression for each data type. (7 logistic regressions).
2. "I've got a dataset, I wanna do a study with it, what can I do?" This would also be logistic regression, where it's a given that I'm going to do a study (so a study is a trial). Data type then predicts the probability of chosing a particular specific focus. Do a logistic regression for each specific focus. (8 logistic regressions).
3. There are datasets out there in the world, and they produce studies of different types at different rates. It's not given that I'm doing a study, so this is a Poisson regression of an unbounded count of studies with some specific focus occurring in a time interval. Do a Poisson regression for each specific focus, and in each regression, all data types are binary predictors. (8 Poisson regressions).

## Taxonomic Group

: Different taxa are studied using different data types. // The probability of using each data type depends on the taxonomic group(s) being studied.

: Different taxa are studied using the same data types. // The probability of using a given data type does not depend on the taxonomic group(s) being studied (7 logistic regressions)

*Proposed Test*: Seven logistic regressions with taxon groups as binary predictors and probability of a data type as the responses.

*Notes*: From a process perspective this is probably usually a situation of "I want to study earwigs, what data can I get?", so a study is a trial and taxon groups are predictors in 7 logistic regressions estimating the probability of using each data type.

Alternatively, I suppose I could model it as studies of each taxon being produced at some rate, and that rate depends on the data type (e.g. abundance data being true dramatically reduces the rate at which earwig studies appear). I find this weird from a real-world process standpoint, but maybe it's ok? **Jon?**

### No Statistical Testing? Descriptive Results Only

## Spatial Extent

Question is of interest but I think I might not have the data to answer it b/c I am limiting to UK & Ireland, so distribution of areas will have lots that are UK-sized.

: The mean spatial extent of studies using only what, where, when data is larger than the mean spatial extent of studies using richer data types.

: Data type does not affect the mean spatial extent of studies.

*Proposed Test*: Linear regression with area (km) as the outcome and data types as predictors.

## Spatial Bias Correction

: Spatial bias is corrected for more often in studies using inference than in studies using other analysis strategies.

: Analysis strategy does not affect the probability that a study corrects for spatial bias.

*Proposed Test*: Logistic regression with analysis strategies as predictors and whether the study corrected for spatial bias as the outcome.

*Notes*: I will present descriptive statistics of % of studies that correct for spatial bias. This feels like a weird, weak hypothesis. Maybe this doesn't need a test?

## Role of biological records

No testing planned. I will present descriptive statistics of proportion of studies using biological records as a response, a predictor, or facilitatively.

## Prediction performance measure

: AUC is used more often than other prediction performance measures.

: The number of studies using each prediction performance measure is the same.

*Proposed Test*: prop.test to test if proportions are different.

*Notes*: This feels fringe. I want to code it for my own reference, but perhaps I shouldn't plan to include it in paper. This is a different discussion that I don't need to get into.

## Specific Analysis methods

Present descriptive statistics of relative frequencies of methods (e.g. GLMs, MaxEnt).

: Different analysis methods are used different amounts.

: Analysis methods are all used the same amount. This is a straw man if ever there was one.

*Proposed Test*: Poisson regression? Does it make sense to do a single Poisson regression with analysis methods as predictors? This assumes there's some baseline rate at which studies appear and that rate is either increased or decreased by the use of each analysis method. This isn't actually the process, but is it a valid way to test the hypothesis?

## Authors associated with data provider

**Is any author associated with the data provider?**

**drop?**

: Studies are more likely to have at least one author from the institution that provided the biological records than from the institution(s) that provided other prepared data (e.g. climate data, land use/cover, functional traits, night lighting, NDVI, conservation status listing).

: The probability that at least one author on a study is associated with the biological records provider is the same as the probabilities that at least one author on a study is associated with the provider of other types of prepared data.

## *Proposed Test*: prop.test in R, which seems to report Pearson's chi-squared test statistic. But I think it can estimate the proportion +- 95% CI for each group, and can then compare those proportions, so doesn't require mutually exclusive groups, like a chi-squared test. But I remain unsure about this. Perhaps look at Agresti (2007) *An introduction to categorical data analysis*.

## Variables I might not use

Drop these from coding? Code just in case? How to report that, then?

* data gridded
* grid resolution
* non-detection inferred from other presences
* entire taxonomic group analyzed (DROP)
* results format - hypothesis testing (DROP)
* compiled individual (DROP)

## Variables to add

* last author institutions