Statistical Science Workshop 6-7 March 2025 University of Waikato – Tauranga
10-11 March University of Auckland – Leigh Marine Lab

Instructor: David Schneider. Memorial University, St. John's, Canada

> *..statistics must be relevant to making inferences in science and technology. The subject should be renamed statistical science and be focused on the experimental cycle, design-execute-analyse-predict.*
> *John Nelder 1999*

What is statistical science?

Statistical science is defined as the application of inferential statistics to the analysis and interpretation of scientific measurements.

Statistical science is not a collection of statistical methods.
Statistical science is not the search for the "best" statistical test.
Statistical science is not the pursuit of $p < 0.05$

Statistical science is founded on writing a model appropriate to the data generated by a research question.
It uses likelihood ratios to compare statistical models (Fisher 1925).
It uses likelihood ratios to replace the erroneous use of p-values as evidence (Goodman 1993).
It requires a model checking loop (Nelder 1999).
It entails distinguishing three modes of inference, all based on likelihood ratios.

Frequentist	Inference from sample to a population via the law of large numbers (Laplace)
Priorist ("Bayesian")	Inference from prior to posterior probability (Laplace 1812, Keynes 1921).
Evidentialist	Inference from data to model parameters (Royall 1997, Nelder 1999).

In this workshop you will learn to

- Translate a research question into a statistical model
- Execute the model and apply the model-checking loop
- Calculate a measure of evidence for the research hypothesis (the likelihood ratio)
- Calculate a measure of uncertainty on the likelihood ratio (p-value / confidence limit))
- Report effect sizes with a measure of uncertainty
- Interpret parameter estimates in light of the research question

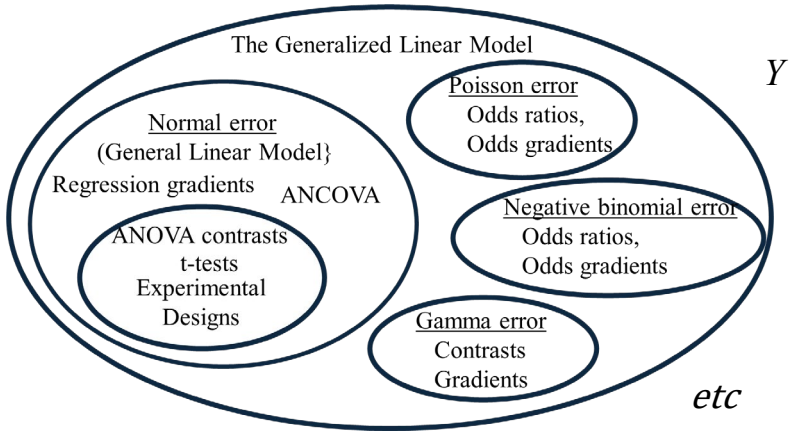
Goal of the third session	Writing the statistical model for the Generalized Linear Model GzLM
Goal of the fourth session	Executing a GzLM in a statistical package
	Using the model checking loop
	Interpreting computer output
	Interpreting the parameter estimates

Statistical Science Workshop Session 3 – Writing the model for a Generalized Linear Model.

Definitions.

Quantities -- Definition
A well-defined quantity has 5 parts:
a name;
a procedural statement that prescribes the conditions for measurement, or calculations from measurements;
a set of scaled numbers generated by the procedural statement;
units on one of several types of measurement scale;
a symbol that stands for the set of scaled numbers.

The Generalized Linear Model uses parameters β_x to relate one or more response variables Y to one or more explanatory variables XA, XB, *etc*



Y ~ Error model (*e.g.* Gamma error)

$\eta = \beta_o + \beta_{XA} XA_1 + \beta_{XB} XB_1$

$Y = e^{\eta}$ (log link for Gamma error)

Regression: β_{XA} = Gradient (change in Y/change in XA)

ANOVA: β_{XA} = Contrasts: $\beta_o - \beta_{XA\_1}$, $\beta_o - \beta_{XA\_2}$,

Statistical Science Workshop Session 3 – Writing the statistical model from GzLM

Second example

Does inversion heterozygosity (HZYG) change with elevation above sea level (*Hsl*) in *Drosophila persimilis*).
Data are from Dobzhansky (1948) as reported in Brussard (1984).
One measurement of HZYG at each of 7 different elevations.

Response variable with symbol _____ _____

Explanatory variable with symbol _____ _____

Source	df
	5

The response variable is bounded at zero and one.
The normal distribution extends beyond these bounds.
The normal distribution performs poorly for parameter estimates near the boundary;
The 20th century solution is the arcsin transform.
This transforms the response variable to value beyond the bounds.
This addressws the problem of non normal error distribution.
It results in uninterpretable estimates of parameters.
The 21st century solution is a GzLM with betabinomial error.
This error distribution talks the logit link.
Write the model in three part form.

Distribution _____

Structural model _____

df _____

Link function _____

Emilie A. Geissinger, Celyn L. L. Khoo, Isabella C. Richmond, Sally J. M. Faulkner, David C. Schneider 2022
A case for beta regression in the natural sciences.
https://esajournals.onlinelibrary.wiley.com/doi/10.1002/ecs2.3940

Statistical Science Workshop Session 3 – Writing the statistical model.

Fourth example (ANCOVA)

Does change in inversion heterozygosity (HZYG) with elevation above sea level (Hsl) in *Drosophila pseudoobscura*) differ from that of *D. persimilis* at the same locations?
Data are from Dobzhansky (1948) as reported in Brussard (1984).
One measurement of HZYG at each of 7 different elevations in two species.

Source	df
	10

Response variable with symbol _____ _____

Explanatory variable V1 with symbol _____ _____

Second explanatory variable V2 with symbol _____ _____

Interactive effect (compares slopes) V1 x V2 _____ _____

Model _____

df _____

Statistical Science Workshop Session 1 – Writing the statistical model.

Review of Session 3

The learning goal was to write a generalized linear model to address a research question.

This replaces the use of transformations, which often result in uninterpretable parameters

Once learned, we can write a model for which we do not know the name.
For example, students can execute a beta-binomial analysis, even though they do not know the name of the test.

Along the way, we learned several important concepts:
Explanatory vs response variables
Parameters relate response to explanatory variables.
Categorical (ANOVA) vs ratio scale (regression) variables.
Use of contrasts to compare means of a categorical variable.
Partitioning the degrees of freedom in an ANOVA table

We set up the model for four examples – two regressions, an ANOVA, and an ANCOVA.

In Session 3 we will learn to use a generic recipe for statistical analysis, based on writing the model.
The recipe will be demonstrated for Gamma error, where the normal error produced unacceptably heterogeneous residuals.
As time permits, another example will be demonstrated while you carry it out in the statistical package.