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 first session	Writing the statistical Model
Goal of the second session	Executing a GLM in a statistical package
	Using the model checking loop
	Interpreting computer output
	Interpreting the parameter estimates

Statistical Science Workshop Session 1 – Writing the statistical model.

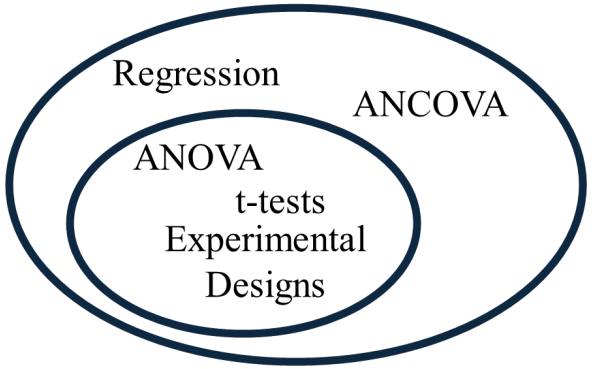
Definitions.

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.

Equations are a mixture of variables and parameters (means, slopes, odds ratios)
Equations have units and dimensions.
Equations must be homogeneous with respect to units and dimensions.

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

The General Linear Model



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

$\beta_o = \Sigma Y/n$ = mean(Y) = $\bar{Y}$

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

ANOVA: β_{XA} = contrasts: $\beta_o - \beta_{XA\_1}$, $\beta_o - \beta_{XA\_2}$, *etc*

Statistical Science Workshop Session 1 – How to write a statistical model

Define Variables

Distinguish response ("dependent") variables from explanatory ("independent") variables

Assign symbols to all variables

Notational conventions

Nominal scale variable ALL UPPER CASE

Ratio scale variables Begin with upper case.

β for fixed effect coefficients (slopes and contrasts)

μ for random effect parameters

Write the model by hand: Response ~ X1 + X2 …. + error

Rewrite the model in ANOVA table format

Calculate the df.

Complete the second column of the ANOVA table

<u>First example</u> (Regression)

Example 9.3.1 from Snedecor and Cochran (1989).
Does the phosphorus content of corn increase when organic soil phosphorus is increased ?
Pcorn and *Psoil* are both ratio scale variables. 9 measurements of *Pcorn*, matched with 9 of *Psoil*.

Quantity of interest is the phosphorus content of corn (*Pcorn* in ppm), in relation to the phosphorus levels in samples of soils with experimentally fixed levels of phosphorus (*Psoil* in ppm).

What is the response variable? _____ Symbol _____

What is the explanatory variable? _____ Symbol _____

How many parameters describe the relation of response to explanatory? _____

One parameter for each regression line

Sketch graph of response vs explanatory

Statistical Science Workshop Session 1 – Writing the statistical model.

First example. Example 9.3.1 from Snedecor and Cochran (1989).

Does the phosphorus content of corn increase when organic soil phosphorus is increased ?
Pcorn and *Psoil* are both ratio scale variables. 9 measurements of *Pcorn*, matched with 9 of *Psoil*.

Using symbols, write the model

Model _____ = _____ + _____
error

df _____ = _____ + _____
n-1 number of parameters error

Source	df
error	7
total	

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt3/Ch9_1.pdf
http://www.mun.ca/biology/schneider/b4605/GLMMworkshop/Data/PCorn.csv

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 _____ _____

Model _____

df _____

Source	df
	5

Statistical Science Workshop Session 1 – Writing the statistical model.

Third example. (ANOVA)

Pea section growth data, from Box 9.4 in Sokal and Rohlf (1995).

Does length depend on treatment (control versus 4 different sugars with auxin present) ?

10 measurements of length of pea section in each treatment group

What is the response variable? _____ Assign a symbol _____

What is the explanatory variable? _____ Assign a symbol _____

Is the explanatory variable on a nominal, ordinal, interval, or ratio scale? _____

How many parameters describe the relation of response to explanatory? _____

Grand mean +

One parameter for each contrast

n = sample size

Sketch graph of response vs explanatory

Using symbols, write the model

Model _____ = _____ + _____
error

df _____ = _____ + _____
n-1 number of parameters error

Source	df
error	7
total	

Statistical Science Workshop Session 1 – Writing the statistical model.

Pea section growth data, from Box 9.4 in Sokal and Rohlf (1995).
Does length depend on treatment (control versus 4 different sugars with auxin present) ?
10 measurements of length of pea section in each treatment group

Length	*Len*	Response variable, ratio scale
Treatment	*TRT*	Categorical explanatory variable

Write the model		Len	$= \beta_o$	$+ \beta_{Trt}\, TRT$	$+ \varepsilon_{Normal}$
Calculate df			(10*5) = 1	+ (5-1)	+ 45

df total = ntot -1 TRT df = number of categories – 1

Fill out first 2 columns of ANOVA table from model

Source	df
TRT	
error	
total	49

http://www.mun.ca/biology/schneider/b4605/LNotes/Pt3/Ch10_3.pdf

Statistical Science Workshop Session 1 – 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 1

The learning goal was to write a statistical modal to address a research question.

This replaces the search for the "right test."

Once learned, we can write a model for which we do not know the name.
For example, students can execute a latin square design, 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 2 we will learn to use a generic recipe for statistical analysis, based on writing the model.
The recipe will be demonstrated for regression, using the first example, phosphorus content in corn.
You will apply this to the second example, fly heterozygosity, another example of regression.
The recipe will then be demonstrated for ANOVA, using the third example, the pea section data.
You will then apply this to a new example, oat yields for treated and untreated plants.
As time permits, the fourth example (ANCOVA) will be demonstrated while you carry it out in the statistical package.

Session 5 – GLMM. The General Linear Mixed Model (Experimental Design)

Fixed versus random factors.

In all four examples in Session 1 the explanatory variables were fixed.

1. Does the phosphorus content of corn increase when organic soil phosphorus is increased ?
   Based on nutrient requirements, we expect a positive relation. We wish to estimate the relation.

2. Does inversion heterozygosity (HZYG) change with elevation above sea level (Hsl) in *Drosophila persimilis*)?
   Hsl is a fixed effect because we expect a decrease in *Hzyg* in harsher environments at higher elevations.
   with elevation (altitude above sea level)

3. Does length depend on treatment (control versus 4 different sugars with auxin present) ?
   Treatment levels are fixed by design. We are inferring only to the levels in the experiment
   TRT is a fixed effect because we are interested in the contrast of 4 means, relative to the control.
   β_{TRT} is a set of unknown fixed effect contrasts.

4. Does change in inversion heterozygosity (HZYG) with elevation above sea level (Hsl) differ in 2 species ?
   The explanatory variable is fixed for several reasons.
      First of all, we expect a decrease.
      Second the study has a repeatable protocol. It can be repeated at the same elevations.
   The categorical variable is fixed because we are only inferring to these two species.
      We are not inferring to all species of fruit fly, based on a sample of only two species.
      Inferring to all species (taking species as random) is too much of a stretch.