Statistical Science.

Part III. The General Linear Model.

Chapter 11 Review of GLM, Single Explanatory Variable

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10)
11.1 Model Based versus Ritual Statistics
11.2 List of Terms
11.3 Commentary on Generic Recipe
11.4 Review Questions

Table 11.1 at end
+ In class worksheet

on chalk board

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops,
which combined models (what is the relation of scallop density to substrate?)
with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

ReCap (Ch 9, 10) The General Linear Model is more useful and flexible than a collection of special cases.
Regression is a special case of the GLM. We saw examples where the explanatory variable X was fixed and where the explanatory was measured with error.
ANOVA is another special case of the general linear model.
The relation of the response to explanatory variable is expressed as set of means.
Factor consists of fixed effects or random effects. For the fixed effects, interest is in the source of the differences. For random effects, interest is in whether there is variance among groups, above and beyond variance within groups.

Today: Review of the GLM, Single Explanatory Variable.

Wrap-up.

The model based approach in this course fosters sound diagnosis (e.g. residual versus fit plots). It points to improved therapeutics (e.g., model revision). It avoids ill-founded 'cures' including those that do more harm than good.
Commonly used analyses in biology are special cases of the GLM (See Table 11.1).

11.1 Ritual Statistics

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an American Statistical Association discussion forum:

Q:Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q:Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on yes/no decision at $p<0.05$:
"We teach it because it's what we do; we do it because it's what we teach."

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133,

$p < 0.05$ is not the only example of rituals learned and then taught.
Here are several widely held beliefs and rituals that persist because they are taught.

Exhibit A. 'Check assumptions before undertaking statistical analysis'	X
Exhibit B. 'Use hypothesis testing to check assumptions.	X
Exhibit C. 'Reject the null ($p<0.05$) and accept the alternative hypothesis	X
Exhibit D. 'Accept the null hypothesis ($p > 0.05$)	X
Exhibit E. 'Compute the power of a test if the null hypothesis is accepted'	X
Exhibit F 'Your analysis is pseudoreplicated.'	X

Practices learned by reviewers and examiners several decades ago persist today.
Here are some examples BDC (Before Digital Computers)

Exhibit G. 'Use a non-parametric test if your data do not meet the assumptions for ANOVA or regression'	BDC
Exhibit H 'Use non-parametric test if your data do not meet assumptions'	BDC
Exhibit I 'Use an arcsin transform with percentages'	BDC
Exhibit J 'Use a square root transform for count data'	BDC
Exhibit K 'A two way ANOVA requires equal sample size in each cell'	BDC

Practices marked X are misunderstandings or incorrect extensions of logically correct statements. Practices marked BDC were sound advice in the days before digital computers, but no longer. Just as computational power has improved medical diagnostics and therapeutics in the latter decades of the 20th century, so has best practice in statistical analysis and concomitant development of statistical theory been vastly improved by computational capacity. BDC practices are the equivalent of a physician using 19th century rules of thumb instead of science based diagnostics and therapeutics developed in the later half of the 20th century.

11.1 Statistical malpractice

Bad practices do harm they become malpractice. They violate a principle that should guide data analysis as much as the practice of medicine: *primum non nocere* - above all do no harm (to the patient or the data). Practices that do harm include those that reduce the amount of information (reducing data to ranks, making sample sizes equal by throwing away data), that restrict the interpretive scope (eliminating interaction terms when adopting rank based tests), that create uninterpretable models (as with arcsin transforms), or that discard interpretable parameters (means, slopes, odds) in favor of less informative statistics.

11.1 A Tour of myths, misunderstandings, and malpractice.

Exhibit A. 'Check assumptions before undertaking statistical analysis'

This is a common and persistent myth. It is myth because the assumptions for computing p-values depend on the distribution of the residuals, not the distribution of the response variable. Residuals can meet the assumptions even when 'the data' do not. Checking assumptions is good practice, but only if the correct assumption is checked.

Because the myth is common don't be surprised to hear it. What do you do if someone insists that you check your assumptions before computing residuals?

First, state the assumptions correctly.

> "Type I error calculated from an F or t-distribution of course assumes that the residuals from my model were homogeneous and normal. I examined the residuals, found they did not strongly violate assumptions, and so I used the p-values calculated by the statistical package."

Some people might still argue. If they do, cite a text written by a statistician or well informed biometrician. Here's the phrasing to use, with citations.

> "According to Neter et al (1983), Seber (1966), and Sokal and Rohlf (2012) the assumptions for computing p-values from F, t, and Chisquare distributions are that the residuals are normal, homogeneous, independent, and sum to zero."
>
> Neter, J., W. Wasserman, M.H. Kutner (1983). Applied Linear Regression Models. Homewood Illinois, Richard D. Irwin, Inc. (page 31, 32, 49)
>
> Seber, G.A.F. 1966. The Linear Hypothesis: A General Theory. London, Griffin.
>
> Sokal, R.R., F.J. Rohlf. 2012. Biometry. 4th edition. Freeman.

Not all texts get it right. Some texts fail to state assumptions clearly. Many texts fail to state that assumptions are checked by examining the residuals.

Exhibit B. 'Use hypothesis testing to check assumptions.

This is a common misunderstanding that results in malpractice—doing harm. The statistical literature warns against statistical tests to evaluate assumptions and advocates graphical tools (Montgomery & Peck 1992; Draper & Smith 1998, Quinn & Keough 2002). Läärä (2009) gives several reasons for not applying preliminary tests for normality. These include: most statistical techniques based on normal errors are robust against violation; for larger data sets the central limit theory implies approximate normality; for small samples the power of the tests is low; and for larger data sets the tests are sensitive to small deviations (contradicting the central limit theory).

Tests of normality reliably lead us to conclude that data are not normal when sample sizes are large, which is when violations are less likely to distort estimates of p-values or confidence limits. Conversely, tests of normality will be insensitive to violations at small sample sizes, when violations can have a large effect on p-values and confidence limits. The practice of "checking the residuals first" reliably wastes time and provokes unnecessary remedies.

Statistical tests of assumptions are illogical. If the assumptions for hypothesis testing are in doubt for an analysis, why would we then use hypothesis testing with the same assumptions to test the normal error assumptions?

Draper and Smith

Läärä, E. 2009. Statistics: reasoning on uncertainty, and the insignificance of testing null. — Ann. Zool. Fennici 46: 138–157.

Montgomery and Peck

Neter, JW, MH Wasserman, MH Kutner.1983. Applied linear regression models. Homewood Illinois, Richard D. Irwin, Inc.

Quinn, G and MJ Keough. 2002. Experimental Design and Data Analysis for Biologists. p 110, 280

Exhibit C. 'Reject the null ($p<0.05$) and accept the alternative hypothesis.

This seems logical but alas, it is a fallacy. It is called denying the antecedent.

1. If H_o then not H_A
2. Not H_o (at $p < 5\%$)
3. Therefore H_A

There are several reasons that H_o can be true. The effect we observe, at a level beyond just chance, may have resulted from something other than the experimental intervention. A classic example is cannibalistic learning (McConnell 1962, Hartry et al 1964.) Did planarians "learn" a maze by ingesting a trained worm? Or is it that the investigators did not clean the apparatus of slime trails by prior runs of worms?

Exhibit D. 'Accept the null hypothesis ($p > 0.05$)

Another denial of the antecedent.

1. If $p < 5\%$ then reject H_o
2. p not $< 5\%$
3. Therefore accept the H_o

Exhibit E. 'Compute the power of a test if the null hypothesis is accepted'

We can correct the phrasing here to remove the fallacy.

'Compute the power of a test if the null hypothesis can't be rejected.'

Once we have the results the sample size and variance are fixed, not variable. Consequently the power goes down as the p-value rises. So post-hoc calculation does nothing more than recast the p-value as Type II error (Hoenig and Heisey 2001). These authors further demonstrate that calculation of minimum detectable effect size or of sample size to detect the observed effect do nothing to strengthen the analysis or modify the conclusion of no significant effect. Calculations of detectable effect and sample size are of course valuable in planning the next study.

John M. Hoenig and Dennis M. Heisey. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. The American Statistician 55(1); 19-24.

Exhibit F "Your analysis is pseudoreplicated."

This term (Hurlbert 1984) conflates two problems--spatially autocorrelated measurements (labeled "non-independent") with malformed F-ratios that fail to isolate a fixed effect from mixed effects, also labeled as "non-independent." Incorrectly nested F-ratios apply only to random factors. If a reviewer uses the term, find someone who can write out the correctly nested F-ratios. If the F-ratios are correct, report it to the journal editor and ask for a new reviewer.

Exhibit G. "Use a non-parametric test when assumptions are violated."

Non-parametric tests are just that—no parameters. No means, no slopes, no odds ratios. Medians are not model parameters. Why would anyone do a non-parametric test when a randomization allows us to report means, slopes, and odds ratios? The answer "Because that was what they were taught."

A variant on this is 'The means differ significantly (Kruskal Wallis test $H = 38$, $p<0.001$. The error here is to take this non-parametric test as an assumption free substitute for the tests on means. The test does not compare means. Strictly speaking it does not even compare medians. It merely computes whether observed rankings ae improbable.

Exhibit H. "Use non-parametric tests if your data do not meet the assumptions for ANOVA or regression."

Non-parametric tests based on reduction of data to ranks were best practice before 1980, when personal computers made high speed computation widely available. Back then, the benefits (ease of computation by hand, assumptions in computing p-values met regardless of data) outweighed the disadvantages. Now that we have enormous computational resources at the click of a mouse or tap on a screen, the benefits are gone and the disadvantages remain.

Disadvantages

-loss of information in reducing ratio scale data to ranks;

-loss of interaction terms, which can be as informative as main effects;

-no estimates of effect size.

Computation by hand no longer matters. With computers it is no longer necessary to reduce data to ranks to obtain tests free of normal error assumptions. We can use randomization to obtain a p-value on any statistic.

Exhibit I: Use an arcsin transform with percentages.

Percentages are bounded at 0 and 1, so in principle residuals will not be homogeneous. In practice, the transform has little effect on the distribution of residuals except near the 0 and 1 boundaries.

Why use a transformation that converts the response variability to an uninterpretable number if it is not necessary? And if it is necessary (non normal residuals for proportions) why not use an appropriate error structure? Be nice to your data. Don't torture it.

See: Warton, David I., and Francis K. C. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92: 3–10. [doi:10.1890/10-0340.1]

Exhibit J: Use a square root transform for count data.

Count data are bounded at zero and so in principle residuals from the analysis of count data will deviate from normality and homogeneity. In practice count data usually do deviate from homogeneity, especially if zero counts are present. The variance will increase along with the mean of groups or fitted values, either in a 1:1 fashion (as with Poisson counts) or as a multiple of the mean (as with counts driven by non-random processes). But while we can expect count data to have a heterogeneous rather than fixed variance, we cannot count on the square root transform to flatten the variance to a constant value. Thus best practice is to examine the residuals, rather than invoking the square root transform whenever count data are analyzed. If residuals are heterogeneous, then a square root transform might be tried if the data show the characteristics of Poisson counts. That is, few counts above 10 or so and the presence of zero counts. For counts with non-Poisson characteristics (zeros and counts above 10 or so) the square root transform will fail to impose homogeneity on the variance. Back transformation of the average of logged data does not produce the mean on an additive scale, as with an arithmetic mean. Back-transformtion produces the geometric mean, which can be interpreted as a measure of central tendency on a multiplicative

scale. Unfortunately, the log transform creates problems of its own (cf Ch9.4). For count data, best practice is to use an appropriate error structure (binomial, poisson, overdispersed poisson, negative binomial) within the computational framework of the generalized linear model. Some simple examples will be shown in Part 5 of this course.

Exhibit K: A two way ANOVA requires equal sample sizes in each cell.
The General Linear model allows correct estimates of unequal sample sizes in multiway designs.

11.2 Review of GLM Concepts

```
Terms in bold on board
Tried in 2003, worked OK
```

Here are terms learned so far in the course.
These terms cover most of the important concepts in this course.

Response (dependent) and **explanatory** (independent) variables.

GLM consists of a **structural** model (explanatory variables) and an **error model**.

Regression (ratio or interval scale, continuous but also counts) variable
vs **Categorical** (nominal scale) variable (factors with levels)

Random vs **fixed** categorical variables.

Arranging data into **Model Format**.

Parameters: **Means, Slopes** μ notation versus β notation

Data Equations

Variability: **Variance, SS, df, MS, Variance ratio (F)**

Analysis of Variance: partition SS

Likelihood ratios calculated from the full and reduced model.

Type I and II error. **p-value** (Type I error) from **pdf**, from **cdf**

Assumptions for p-values from cdf (4)

Assumptions not met

Hypothesis testing. $\mathbf{H_A}$ / $\mathbf{H_o}$ for parameters,
or for variance due to each term in model

Statistical conclusion. Conventional format: statistic, sample size or df, p-value).

Science conclusion. Report parameters, effect sizes, and interpret statistical results relative to the question that motivated the collection of data.

11.2 Review of Concepts

Two-tailed vs **one-tailed** tests. This question arises frequently.

Two-tailed tests cover both positive and negative outcomes.
The H_A / H_o pair is written in terms of parameters or in terms of variances.

H_A: $\beta_X \neq 0$ $\quad$ H_o: $\beta_X = 0$ $\quad$ for parameters

H_A: $\mathrm{Var}(\beta_X \cdot X) > 0$ $\quad$ H_o: $\mathrm{Var}(\beta_X \cdot X) = 0$ $\quad$ for variances

The cumulative distribution function cdf for the *t* distribution reports the upper tail for positive values of *t*.

```
R> pt(2, 30)
[1] 0.972687
```

```
MTB> cdf 2;
SUBC> t 30.
  2   0.972687
```

Upper tail $p = 1–0.972687 = 0.0273$
Both tails $p = 0.027313 * 2 = 0.0546$

The cdf for the *F*-distribution is used to obtain the right tail for a two-tailed (non-directional) test.

```
R> pf(4, 30)
[1] 0.945375
```

```
MTB> cdf 4;
SUBC> f 1 30.
  2   0.945375
```

The two-tailed value is $p = 1–0.945375 = 0.0546$
The results for the *t* and *F* distributions match in this case because $t^2 = F$

<u>One-tailed</u> tests H_A / H_o pairs are also written in terms of parameters and variances.
For example, we might expect a positive relation between height of sons and their fathers.

H_A: $\beta_{Hfather} > 0$ equivalent to H_o: $\mathrm{Var}(\beta_{Hfather} \cdot Hfather) = 0$

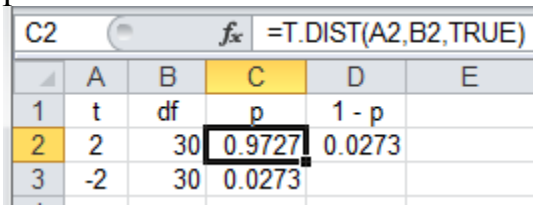
To calculate the p-value on a one-tailed test of a parameter
we use the upper tail of the t distribution: $p = 0.0273$
we cut the *p*-value from the F-distribution) in half: $p = 0.0546/2 = 0.0273$

The cumulative distribution function cdf for the *t* distribution reports the lower tail for negative values of *t*.

```
R> pt(-2, 30)
[1] 0.027313
```

```
MTB> cdf -2;
SUBC> t 30.
  -2  0.027313
```

Statistical tables give the two tailed p-value for the t distribution.
The t.dist function in Excel reports the cdf function: p for negated values of *t*, 1-p for positive values of *t*.



Excel provides other functions for the *t* distribution. These behave in ways idiosyncratic to Excel.

11.3 Commentary on the Generic Recipe for Hypothesis testing with GLM

1. Construct model

Statistics are a way of summarizing pattern in the form of a formal model. The simplest and most familiar is the computation of a mean--a single value that is taken as representative of a set of observations. The General Linear Model is used when we can, on common sense or science grounds, separate the response variable(s) from the explanatory variable(s). The GLM includes such familiar procedures as regression and ANOVA. The GLM relates one (or more!) response variables to one or more explanatory variables.

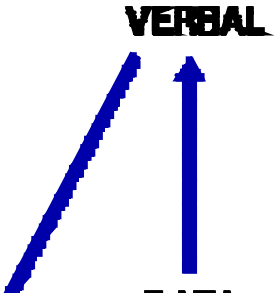
$$Y = \beta_o + \beta_{X1}X_1 + \ldots\ldots + \text{residuals}$$

$$Y = \Sigma\beta_i X_i + \text{residuals}$$

The GLM is flexible. It allows the explanatory variables to be on a nominal (categorical) measurement scale, on a ratio type of scale, or on both.

Verbal and *graphical* models are useful in formulating a general linear model for use in analysis of data. A typical sequence is to start with a verbal model, then go to a graphical model, and finally to a formal model.

In setting up an analysis with the GLM it is important to separate the response from explanatory variables. One of the best ways to help someone who asks for help in "analyzing their data" is to ask them to list their variables and ask them to identify their response variables, separating these from explanatory variables.



The *response* and *explanatory* variables should be defined by a procedural statement, assigned a name and a symbol. Units (for numerical values) and level names (categorical variables) should be reported.

In learning to use the GLM it is important to write out response and explanatory variables, then state the model in words and picture before trying to write it. With practice, writing the model statement comes naturally.

When the GLM is used in inferential statistics, the model is written for the population. By convention parameters of the population appear as greek symbols. Estimates of these parameters appear as roman letters, or as the Greek symbol with a hat, such as $\hat{\beta}$. In GLM notation β is a slope if the explanatory variable is on a ratio type of scale. β is a set of contrasts (dfferences of means) is the explanatory variable is categorical.

2. Execute model

Place data in model format for data not already in model format:

-one column for response variable

-one column for each explanatory variable.

Avoid using numbers for categorical variables.

Code the model statement,

Write a model statement that follows sequence of terms in step 2.

Obtain fits and residuals

GLM routines use the data to make estimates of the model parameters. These estimates are distinguished from the true (and unknown) values of the population parameters by placing a hat over the parameter.

β_o stands for the mean of the entire population.

$\hat{\beta}_o$ stands for the mean estimateed from the sample.

Estimates are made according to two criteria: either minimizing the squared residuals between the model and the data, or maximizing the likelihood of the estimate, given the data. In many cases the latter works out to be the same as the former estimate. The mean, computed according to the familiar formula, is both a maximum likelihood and minimum deviation estimate of the true value. Slopes for linear regression are estimated by minimizing the sum of the squared vertical deviations from the regression line. In simple cases this is accomplished in one step with a formula. In more complex cases, this is accomplished by iterative techniques.

These estimates are used to calculate fitted values and from these the residuals. The residuals are plotted against the fitted values to evaluate the model (next step).

Output from GLM routines.

a. Most routines provide residuals and fitted values as an output option.

b. Most GLM routines provide the parameters for the GLM

These consist of slopes and means, the latter expressed as deviations from the grand mean β_o

Parameter estimates in general linear model format:

$\hat{\beta}_o$	overall mean
$\hat{\beta}_{group}$	deviations from overall mean
$\hat{\beta}_{regression}$	overall slope
$\hat{\beta}_o + \hat{\beta}_{group}$	means in each group

c. Parameters can be estimated outside a GLM routine with functions that estimate slopes and means.

3. Evaluate the model.

A GLM consists of response variables, the structural model (consisting of explanatory variables and parameters), and the error or residual term.

We first evaluate the structural model.
A bowl or arch in the residual plot indicates that the relation of response to explanatory variable is not a straight line. If this proves to be the case the model needs to be revised so that the relation of response to explanatory variables is correctly represented. The straight line assumption does not need to be checked if the GLM consists entirely of categorical (ANOVA) variables.

Another way of looking at this assumption (extra).
Are the residuals associated with the model ? We want to ensure zero covariance between the model and the residuals.

Var(Y) = Var(Model) + Var(Residuals) + Cov(Model, Residuals)

Covariance is detectable as a curved pattern in the plot or residuals against the fitted values, it does not occur as a simple positive or negative association. If there is covariance, then the model is inappropriate and the variance estimates used in hypothesis testing will be in error.
Another model should be used.

Next, we evaluate the error model. For the GLM, the error model is that of homogeneous, normal, and independent errors. If these are violated we have two choices. The first is to revise the model to an error structure that is more appropriate than normal (fixed) error. Examples are logistic and poisson regression, for which the error increases in proportion to the magnitude of the fitted values. These are special cases of the generalized linear model, which allows us to specify any of several error models. The alternative course of action is to continue with the general linear model but use an empirical distribution of outcomes instead of a chisquare (or *t* or *F*) distribution to calculate probabilities and confidence limits. There are many good reasons to adopt the first course of action (McCullagh and Nelder 1987, Myers et al 2002). The first course of action allows inference beyond the data at hand. The second course of action restricts inference to the data at hand.

Residuals and fits are used to evaluate assumptions.
In this course we rely primarily on a normal error structure to estimate parameters and calculate Type I error from likelihood ratios. This entails four assumptions:

3. Evaluate the normal error model.

1. Var(res) = constant. Plot residuals versus fits, check for cones and spindles.
2. E(res) = 0 This will be automatically true for analyses in which parameters are estimated from data, as in most statistical packages, so no need to check.
3. $Cov(res_j\ res_j) = 0$ I.e., residuals are independent.
   This is checked by plotting residuals in some logical order,
   such as order in which data were collected.
   An equivalent check is to plot residuals against neighboring value.
   Create new column of residuals lagged by 1, then plot residuals vs lag(res)
   This check can be extended to multiple lags, not just lag 1.
   This is accomplished with ACF command in a statistical package.
4. Residuals normal. This is checked by
   -looking at histogram of residuals
   -checking the fit to normal distribution with rootogram
   -checking fit with nscore(res) vs residuals.
   Normal residuals adhere to a diagonal line

In this course we will use primarily graphic displays to evaluate the assumptions. The reason for this is that statistical tests of assumptions perform poorly. Statistical tests of assumptions are insensitive to violations at small sample sizes, which is precisely when violations can distort estimates of p-values. Tests of assumptions become increasingly sensitive to minor violations at large sample sizes, which is when violations no longer distort estimates of p-values. Statistical tests of assumptions seem like a good idea, but upon examination turn out to be a bad idea.

4. Partition df and SS according to the model. Calculate LR = weight of evidence

LR < 10.	**Discard model**
10 < LR < 20	**Dubious model**
20 < LR	**-->Step 5**

5. Inferential mode.

We have three choices.

Evidentialist. We report relative evidence as a likelihood ratio. This mode of inference relies on the validity of the error model.

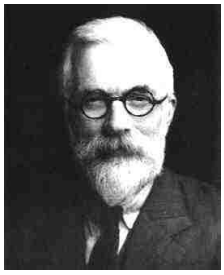
Frequentist. We report the likelihood ratio along with a Type I error rate. This mode of inference relies on the Law of Large Numbers: as sample size increases estimates converge on the true value of a parameter in a population. The population can be finite, as with survey design, where random sampling is from a known frame. The population can be infinite, as with experimental design based on repeated use of the same experimental protocol. The prevailing mode includes a fixed Type I error rate. In the absence of a clear reason to control Type I error rate, we use Fisher sorting to evaluate the evidence.

Priorist. We report a posterior probability based on a generalized likelihood ratio and a prior probability that is overt, public, and cumulative. Sophisticated sampling (Markov Chain Monte Carlo) is no substitute for weak or poorly substantiated prior probabilities.

5. Decide on mode of inference. Is hypothesis testing appropriate?

Hypothesis testing (H_A versus H_o) is usually considered mandatory. This view stems from R.A. Fisher.

> Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis
> --R.A. Fisher 1935 p19



Fisher's approach to null hypotheses was supplanted by the decision-theoretic approach of J. Neyman and R. Pearson. This approach was criticized by Fisher. Throughout the 20th century it has come increasingly into question. By the mid 1990s M.R. Nester (1996) had collected over 125 quotes against the use of hypothesis tests. In 2019 a statement from the American Statistical Association recommended against declaring statistical significance against a fixed Type I error rate. The generic recipe in tnhsi course separates the measure of relative evidence (the likelihood ratio), from the measurement of uncertainty (Type I error_).

6. If hypothesis testing is appropriate , state H_A / H_o pairs..

The alternative hypothesis is that the response variable is related to an explanatory variable. This is usually the basis for undertaking an analysis, so it makes sense to write this first. In the analysis of variance the H_A / H_o pair is expressed about the variance due to a term in the model. The hypothesis pair concerning a term in the model is equivalent to a general hypothesis about the parameters: that the means differ, the slopes are not zero, or that interactive effects are present.

The logic of hypothesis testing with inferential statistics is that all possible outcomes can be divided into <u>two categories</u>: those included under the alternative hypotheses, and those not included. These latter are labelled the null hypothesis. The H_A/H_o pair should include all outcomes. For example, if the H_A were that the expected value for a treated group exceed the expected value for a control group, then the H_o would be that the treated group was less than or equal to the control group.

$\alpha = 5\%$ This is the conventional fixed criterion for statistics in biology. It is a compromise between Type I and II statistical error. Reducing the Type I error by making α smaller will raise Type II error, the chance of missing a true effect. The criterion used should be stated <u>before</u> undertaking a statistical analysis. A convenient place to do this is in the material and methods section of a scientific report or thesis.

7. ANOVA – Report the model results in ANOVA table.

The ANOVA table displays the model arranged vertically.

Add Figure showing horizontal model to vertical table

The total degrees of freedom (sample size minus 1) are partitioned according to the model statement. This is easily done by hand. Computation of df by hand is a good way to check that the model statement is carrying out the analysis intended.

The sum of the squared deviation of the data from the mean is partitioned according to the GLM written for the analysis. The sum of squares to be partitioned is

$$SS_{total} = \Sigma Y^2 - n^{-1}(\Sigma Y)^2$$

Completing the table.

This is done in the following sequence:

Write out the headings of ANOVA table

Sources df SS MS F

Fill in sources
Fill in df (these are partitioned according to the model)
Add SS_{tot} to bottom of table.
Add partitioned SS to the table from a computer print-out.
Compute MS from SS/df, if not already done by the computer.
Compute F from MS/MS, if not already computed.

It is always a good idea to undertake a quick paritioning of the degrees of freedom by hand. These are checked these against the software results, to make sure the intended model was executed.

Calculate Type I error if appropriate.
It is easier to calculate a p-value from a theoretical distribution than from an observed distribution generated by randomization. p-values from *F, t*, chi-square and other distributions no longer need to be the cut-off values obtained shown in tables. Any good statistical package (e.g. Minitab, SAS) will allow you to calculate exact p-values from F, t, chi-square and other distributions. These are more informative than critical or cut-off values obtained from printed tables.

8. Recompute p-value if necessary

If the assumptions are not met, the remedy is to compute the p-value by randomization. This results in an observed distribution of outcomes from the data, when the H_o has been made true by randomizing the data so as to make the null hypothesis true. Outcomes are tabulated as frequency distribution. From this the p-value of observed statistic (data not randomized) is calculated. Could this statistic, of this magnitude, have arisen by chance ?

Recomputing is not always necessary in a lab setting.
We ask a sequence of questions if assumptions are not met.

df_{err} small ?

If df > 100 little need to recompute p-value, even if residuals terrible.
the p-value won't change by much.
If 30 < df < 100 then may need to recompute if residuals terrible
the p-value may change
If df < 30 then recompute p-value if residuals terrible.

p close to α ?

If p not close (e.g. twice or half α) then recomputation unlikely to change the decision, even though the p-value is incorrect.

Recomputing – reporting results to a wider audience.
In general, p-values do not work well in a public setting. The logic is unfamiliar. In a public setting, the likelihood ratio is more readily grasped. "A relation lung cancer and cigarette smoking is a thousand times more likely than not.
In a public setting measure of evidence is more effective than a measure of uncertainty (the p-value).

9. Report statistical conclusion.

In reporting the results from hypothesis testing in a publication it is important to remember that rejecting the null hypothesis argues against chance as an explanation for a particular outcome. Eliminating chance does not, however, establish causality. The relation of the response variable to the explanatory variable may be due to some factor other than the explanatory variable. An example of this is a regression of food consumption against age. Food consumption appears to rise with age in animals that increase in size with age. Food consumption changes in more complex ways with age, after adjusting for the effects of size.

The conclusion from a statistical analysis with the GLM should contain, at a minimum, the *F*-ratio, df, p-value, and whether the *p*-value was obtained from a theoretical frequency distribution or from frequency distribution generated by randomization.

10. Report science conclusion. Interpret parameters of biological interest.
Report the effect size: How far apart were the means ? How strong is the rate of change estimated by a regression parameter?
Report a measure uncertainty. This can be a confidence limit, a standard error, or a standard deviation. Evaluate the statistical conclusion in light of effect size and uncertainty.

References

Fisher RA. The design of experiments Edinburgh: Oliver and Boyd; 1935 p19

Hartry. AL et al.1964 Planaria: Memory transfer through cannibalism re-examined Science 146: 274-275

Hurlbert, S. 1984.

McConnell, J. V. (1964). Cannibalism and memory in flatworms. New Scientist, 21, 465--468.

Nester 1996

At this point entertain questions brought in for answering.

Table 11.1 Commonly Used Tests, Based on the General Linear Model.

Analysis	Response Variable	Explanatory Variable	Interaction	Comments
t-test	1 ratio	1 nominal	Absent	compares two means
1-way ANOVA	1 ratio	1 nominal	Absent	compares 3 or more means in 1 category
2-way ANOVA	1 ratio	2 nominal	Present	tests for interactive effects
				compares means in 2 categories, if no interaction
Paired Comparison	1 ratio	2 nominal	Absent if too few df	compares 2 means in 1 category, controlled for 2nd category (blocks or units)
Randomized Blocks	1 ratio	2 nominal	Assumed Absent*	compares 3 or more means in 1 category, controlled for 2nd category (blocks or sampling units)
Hierarchical ANOVA	1 ratio	≥2 nominal	Absent	nested comparisons of means
ANCOVA	1 ratio	≥ 1 ratio	Present	compares two or more slopes
		≥ 1 nominal	Assumed Absent*	compares means, controlled for slopes
Regression	1 ratio	1 ratio	Absent	tests linear relation of response to explanatory
Multiple Regression	1 ratio	≥2 ratio	Assumed Absent*	tests linear relation to 2 or more explanatory variables relation expressed as a plane

*The interaction term is often assumed to be absent. Including the interaction term allows us to check the assumption. This is a good idea provided there are at least as many df in the error term as the interaction term.