Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 10.4 One way ANOVA, Random Effects

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9)
10.1 Single Sample t-test
10.2 Two Sample t-test
10.3 One way ANOVA, Fixed Effects
10.4 One way ANOVA, Random Effects
Fixed versus random effects
Example: Scutum widths

SRBX9_1.out

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) 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 with the explanatory variable X fixed and with the explanatory measured with error.

ReCap (Ch 10) ANOVA is another special case of the general linear model.
The relation of the response to explanatory variable is expressed as set of means. When classes within a factor are fixed by experimental design, it is natural to investigate which classes are responsible for significant variation. *A priori* (planned) comparisons are based on our knowledge of the reasons for collecting the data. These are more informative than *a posteriori* (after the fact) comparisons.

Today: ANOVA as a special case of the GLM.
Single Factor ANOVA - Random Effects

Wrap-up. GLM. ANOVA. Explanatory variable on nominal scale.
Random factor. Inference to a population of units instead of inference to fixed factor categories.

Introduction. Random effects.
Until today we have been analyzing our response variable relative to fixed effect factors. A fixed factor has categories, or levels, that we set to certain values in an experiment or levels that we choose in an observational study. We infer only to those levels. Experimental examples are treated versus untreated (control) units, before versus after treatment of an experimental unit. Examples from observational studies include day versus night, habitat types, and insect stages (larval, adult),
A random factor has categories that we have not chosen, or that vary even after we make them as uniform as possible. Examples are tanks in aquaculture, plots in agriculture, and individual organisms. Inference is usually to similar units although inference can also be only to those units.

The choice between random and fixed depends on how we define the contrasts among means. Here is an example. A biologist carries out an experiment on the effects of nutrient enrichment on the growth of marine algae, at three different locations in the field. Then repeats the experiment two more times, so that each location is exposed to each nutrient level. The nutrient factor is clearly fixed if nutrient levels are manipulated. The location factor is usually random. However, the location factor could be taken as fixed, if the biologist restricts inference only to the locations in the study. Time is fixed if we expect time-dependent variation. In this case we compare results from time 2 to time 1, from time 3 to time 2, etc. Time can be random if we expect no trends and do not expect time-dependent variation

Example.
Data from Box 9.1 of Sokal and Rohlf 2012, p. 209.
Does tick size, as measured by scutum width, differ among hosts (rabbits)?
The purpose of the study is to measure the proportion of variation in tick size attributable to host. Studies such as this are used to design manipulative experiments. If variation among hosts is small, then we can use relatively few hosts in a future manipulative study. If variation is large, than we need to control this by increasing the number of host rabbits in a future manipulative study.

1. Construct model.

~~What is the best test?~~
What model do we use to analyze this data?

Verbal model.

Scutum width W_{scut} varies among hosts H (4 rabbits)

Graphical model

Plot showing W_{scut} as a function of H
Model consists of 4 means, one for each rabbit.

```
MTB> plot  'width'  'host'
```

```
R> boxplot(Wscut~Host, data=TickSize)
```

1. Construct model.

What are the response and explanatory variables?
Response variable is scutum width of tick larvae *Haemaphysalis leporispalustris*, W_{scut} = microns
Explanatory variable is host, H = Rabbit A, Rabbit B, Rabbit C, Rabbit D

Are the explanatory variables covariates (regression) or categorical ?
Categorical.

Are the categorical variables random or fixed?
Rabbits were a 'random sample of the population of host individuals'
(Sokal and Rohlf 2012, p211).

The data appear to be symmetrically distributed around the model (the means) so we will use a normal error model.

Formal model $W_{scut} = \beta_o + \beta_H \cdot H + \varepsilon$

2. Execute analysis.

Place data in model format:
Column with response variable, scutum width W_{scut}.
Column with explanatory variable, Rabbit Host = 0 or 1 or 2 or 3
These are labels (categories), not numbers on ratio scale.

Code the model statement in statistical package according to the GLM
$W_{scut} = \beta_o + \beta_H \cdot H + \varepsilon$

```
MTB> ANOVA 'Wscut' = 'Host'
MTB> GLM 'Wscut' = 'Host';
SUBC> fits c4;
SUBC> res c5.
```

```
R> TSizeModel<- lm(Wscut~(1|Host),
         data=TickSize)
```

(1|Host) denotes random factor

The fitted values are the means in each of the four groups.
The residuals are calculated from the observed and fitted values.

2. Execute analysis.

For ANOVA, the parameters are contrasts among the group means.
In this example there are 3 contrasts relative to the mean of the reference group.

2. Execute analysis.

For ANOVA in the GLM format, parameters are β_o the mean of the reference group, and β_H for the contrasts with the other 3 groups.

To estimate β_o

```
MTB > describe 'width'
                 N      MEAN    MEDIAN    TRMEAN     STDEV    SEMEAN
width  1       37    359.7
```

$\hat{\beta}_o = 359.7$ To estimate the mean for each group

```
MTB > describe 'width' ;
SUBC> by 'host' .

    host   N    MEAN   MEDIAN   TRMEAN   STDEV   SEMEAN
width  1     8    372.25   373.00   372.25    7.36          2.60
       2    10   354.40   353.00   353.75    11.92        3.77
       3    13   355.31   354.00   355.00    8.92          2.47
       4     6   361.33   366.00   361.33    15.27         6.23
```

$\hat{\beta}_o + \hat{\beta}_H \cdot H =$ 372.25 Hence $\hat{\beta}_H =$ +12.55

354.40 – 5.33

355.31 – 4.4

361.33 + 1.6

There are several different symbols for estimates.

Placing a hat over the greek symbol $\hat{\beta}_H$

Placing a bar over the symbol for the quantity, in the case of the mean μ_{scut}

Using a roman letter (use b_1 for estimate of β_H)

> The symbol μ_W is also used for the parametric mean of the quantity W. This notation is difficult to use with symbols having subscripts, such as W_{scut} for scutum width. Similarly, the symbol σ_W^2 is used for the parameteric variance of the quantity W. The estimate (derived from a sample) is s_W^2. This is another example of cumbersome notation that is difficult to use with subscripted symbols such as W_{scut}.

The same information is reported relative to one of the means, taken as the intercept.

```
R> summary(TSizeModel)
```

```
              Estimate    Std. Error
(Intercept)   372.250       3.783
HostB         -17.850       5.075
HostC         -16.942       4.808
HostD         -10.917       5.779
```

3. Evaluate model.

Structural model.

No regression lines estimated in ANOVA so no need to check straight line

Error model. Homogeneity?

Plot residuals versus fitted values.

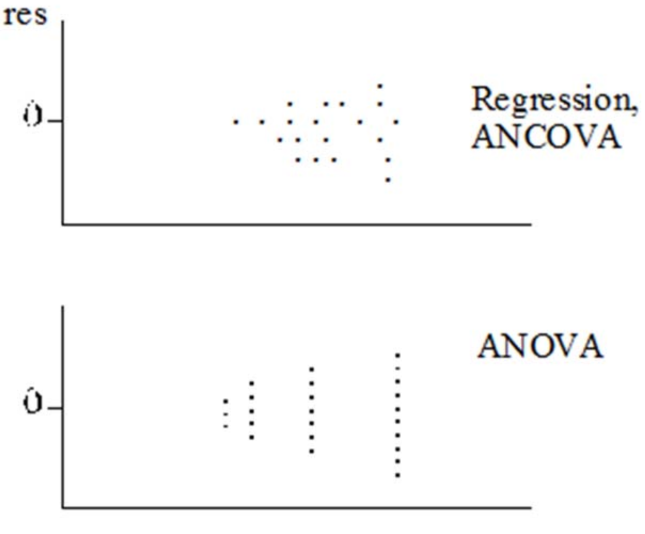
```
MTB > plot 'res' 'fits'

res     -
        -    *
        -      *
      15+                *
        -    * *                2
        -    *                                                  2
        -      3
        -    2 *                                                2
       0+      *                *                               *
        -    2 2                                                *
        -      2                                                *
        -    *
        -    * 2                                                *
     -15+    *
        -                       2
        -
        -
          +---------+---------+---------+---------+---------+------fits
      353.5     357.0     360.5     364.0     367.5     371.0
```

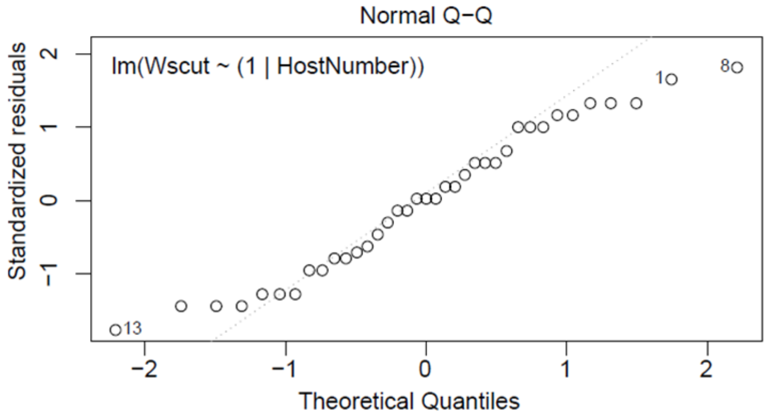
Homogeneity ?

Residual versus fit plot shows vertical distribution of residuals to be about the same in all four groups. So residuals are judged homogeneous.

When this assumption is not met, the plot of residuals versus fits will often show left or right facing fans for any GLM, including regression and ANOVA.



For ANOVA, there are a limited number of fitted values, hence the plot is present at only a few points long the x-axis. The fan pattern is the same in both plots, but vertical swaths are missing from the plot with categorical variables.



3. Evaluate model.

Error model
Residuals normal ?

The residuals deviate slightly from normality

The response variable shows greater deviation from a normal distribution

```
MTB > hist 'res'
MTB > hist 'Wscut'
```

```
MTB > hist c1
Histogram of Wscut   N = 37

Midpoint   Count
     340       4  ****
     345       3  ***
     350       6  ******
     355       2  **
     360       7  *******
     365       4  ****
     370       4  ****
     375       5  *****
     380       2  **
```

```
MTB > hist 'res'
  Histogram of res   N = 37

 Midpoint   Count
      -20       1  *
      -15       3  ***
      -10       4  ****
       -5       8  ********
        0       5  *****
        5       6  ******
       10       6  ******
       15       3  ***
       20       1  *
```

If we evaluate the assumptions before calculating the residuals, we erroneously conclude that the residuals are not normal.

4. Partition df and SS according to the model

GLM	W_{scut}	=	β_o +	$\beta_H \cdot H$	+ ε
Source	Total	=		Host	+ Resid

Compute total degrees of freedom $df_{total} = n - 1 = 37 - 1 = 36$

Partition df_{total} according to model, using rules

4 hosts $df_H = 4 - 1 = 3$

$df_{res} = df_{total} - df_H$ $df_{res} = 36 - 3 = 33$

df denotes the degrees of freedom for each factor.

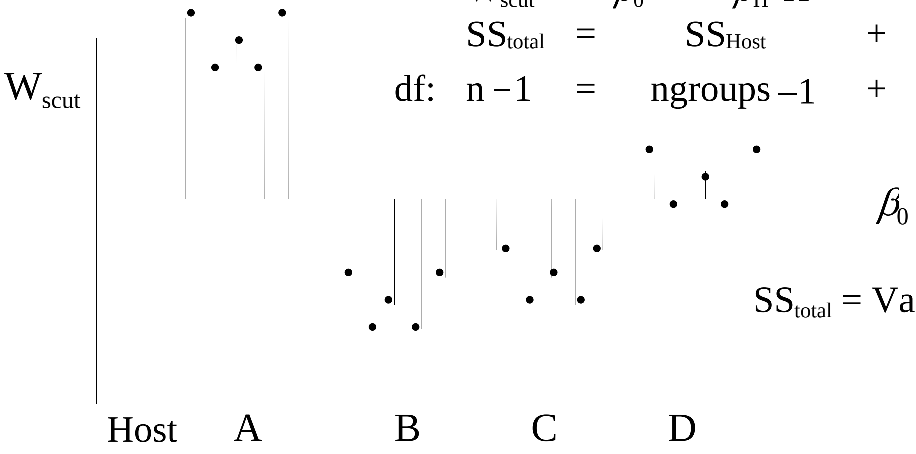
Each parameter that is estimated from the data uses up one degree of freedom. A slope uses up one degree of freedom. An explanatory variable consisting of n classes uses up *n* – 1 df.
1 df is lost in estimating the grand mean.

4. Partition SS according to the model

$$W_{scut} = \beta_0 + \beta_H \cdot H + \varepsilon$$

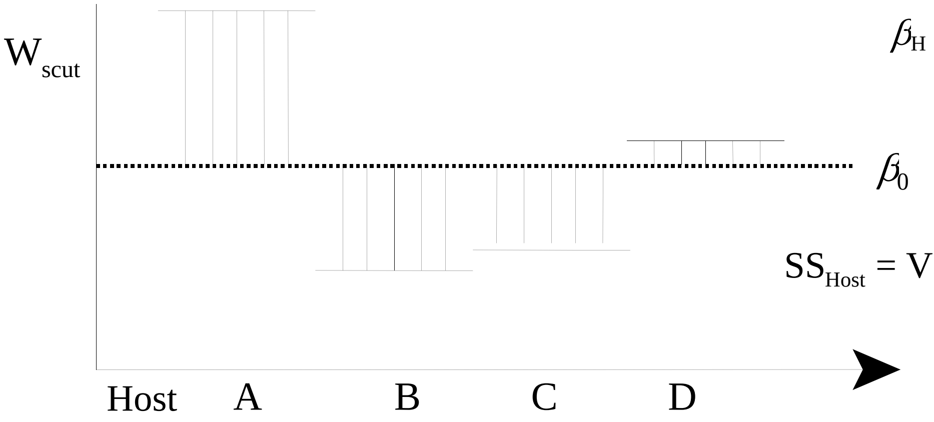
$$SS_{total} = SS_{Host} + SS_{residual}$$

df: $n - 1 = ngroups - 1 + n - groups$



$\beta_0 = 359.7$

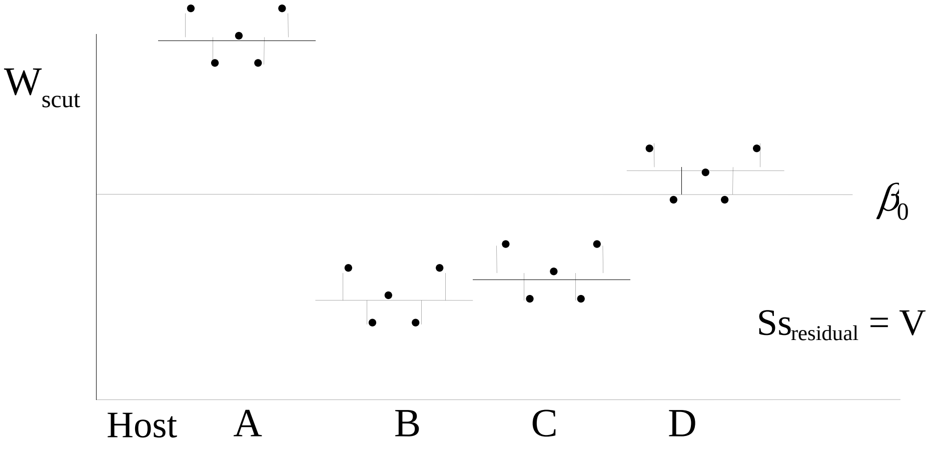
$SS_{total} = Var(W_{scut}) \cdot df$



$\beta_H =$ +12.55
 –5.33
 –4.4
 + 1.6

β_0

$SS_{Host} = Var(\beta_H) \cdot df$



β_0

$Ss_{residual} = Var(res) \cdot df$

4. Partition SS according to the model

Compute $SS_{tot} = Var(W_{scut}) \cdot df_{total}$

1. $SS_{tot} = (n - 1) * Var(W_{scut}) = 36 * 155.2 = 5586$
2. $SS_{tot} = \Sigma W_{scut}^{2} - n^{-1} (\Sigma W_{scut})^{2} = 4792797.3 - 37^{-1} \cdot 13308.9^{2} = 5586$

GLM	W_{scut}	=	β_o +	$\beta_H \cdot H$	+ ε
Source	Total	=		Host	+ error
n	37	=	1 +	3	+ 33
df	36	=		3	+ 33
SS	SS_{tot}	=		SS_{host}	+ SS_{res}
	5586	=		1808	+ 3778

4. Calculate likelihood ratio for omnibus model

How good is the evidence for differences in size of ticks, among rabbits?

Full model: $W_{scut} = \beta_o + \beta_H \cdot H + \varepsilon$

Reduced model: $W_{scut} = \beta_o + \varepsilon$

$LR = L(\beta_o + \beta_H ; W_{scut}) / L(\beta_o ; W_{scut})$

$LR = (3778)^{-37/2} / (5586)^{-37/2} = 1387$

We have strong evidence (LR>1000) for variance in tick size among hosts. The alternative model (variance in means) is over a thousand times more likely than the null model, no variance in means.

5. Target of inference. Fixed versus random effect factors.

For the example of tick scutum widths we are going to infer to a population of rabbits similar to those in this sample. Conclusions from statistical inference apply to any study that uses the same measurement protocol to measure size of *H. leporispalustris* tick larvae on rabbits.

We have data from only four rabbits. We could be very cautious and define the target of inference as "all possible measurement of scutum widths from ticks on these four rabbits only." If we were to do this, then we have a fixed effects model that applies only to these 4 rabbits. Of more interest is a random effects model, where we treat the rabbits as a sample from a population of similar experimental units (rabbits).

5. Choose mode of inference – evidentialist, frequentist, priorist.

The goal of the research was an estimate of variance in size among hosts, relative to the total variance. A measure of evidence along with an estimate of variance among hosts suffices. Priorist inference is groundless in the absence a sound prior probability. Priorist inference is no reelvant; the target of inference is not a revised belief (posterior probability). Similarly we have no need to control Type I error nor any need to declare a decision at some stated level of Type I error.

We will report the likelihood ratio as a measure of evidence. We will report the variance due to hosts as a percentage of the total variance.

Here is a summary of the data equations.

```
MTB > name c3 'fits' c4 'res'
 MTB > print 'width' 'fits' 'res'

  ROW     width       fits           res

    1      380       372.250      7.7500
……
    8      382       372.250       9.7500
    9      350       354.400      -4.4000
,,,,.
   18      364       354.400       9.6000
   19      354       355.308      -1.3077
…..
   31      348       355.308      -7.3077
   32      376       361.333      14.6667
…..
   37      360       361.333      -1.3333

sd^2      =  12.46^2       7.09^2      10.24^2
sd^2 ·36  = 155.25     50.27      104.86
SS        = 5589     1809       3775
```

6. State test statistic, sampling distribution, and use of Type I error if appropriate.

The focus of the random effects analysis is the variance in parasite size among rabbits. This focus differs from fixed effect factors, where the H_A/H_o pair is stated as contrasts among means. Instead of calculating Type I error we will report a measure of evidence, the likelihood ratio.

Full model: Var($\beta_H \cdot H$) > 0
$LR > 1$

“The true group means deviate from the true grand mean, where there is variance in size, among hosts”

Reduced model: Var($\beta_H \cdot H$) = 0
$LR = 1$

“The true group means do not deviate from the grand mean, where there is no among host variance in tick size.”

7. Report Type I error.
Not needed.

8. Recompute Type I error by randomization if assumptions are not met.
Not necessary, the residuals were judged homogeneous and normal.

9. Report statistical conclusions.
$LR = \mathrm{L}(\beta_o\, ,\, \beta_H \mid W_{scut}) \;/\; \mathrm{L}(\beta_o \mid W_{scut}) = 1387$
The full model is1300 times more likely than the reduced (null) model.
The explained variance (variance due to host) is $R^2 =$ 1808 / 5586 = 32%

10. Report science conclusion.
The parameter of interest is the variance among the means. How large is the variance among groups, compared to the total variance across all ticks?

Fixed versus random effects - Notation.

Fixed effects ANOVA. Explanatory variable is fixed treatment.
This is written $Y = \mu + \alpha + \varepsilon$
The fixed factor is shown as a greek letter α
Our interest is in contrast among means.
A priori contrasts are used in confirmatory analysis.
A posteriori contrasts are exploratory in nature.

Random effects ANOVA. Explanatory variable is random.
This is written $Y = \mu + A + \varepsilon$
The random factor is shown as a roman letter A.
Our interest is in variance in *Y* among experimental units.

Among unit SS =
1808 / 5586 = 32%

At 32%, the among rabbit variability is more than negligible.
Sokal and Rohlf (2012) list several biological processes that could generate among host variability:
- the modifying influence of the host on ticks
- ticks on any one host are siblings
- differential selection on size of ticks, among hosts
- different geographic sources of ticks for each host

From the biology of this species of tick, Sokal and Rohlf (2012) consider the genetic explanation (siblings on one host) to be the leading explanation.