Statistical Science.

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

Data from Sokal and Rohlf
Tick scutum width on 4 hosts

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.

New Concept: Random effects.
Today, we start with a new concept, random effects. Until today we have been analyzing our response variable relative to fixed effect factors. We can also analyze our response variable with respect to random effect factors. We do not intervene nor do we choose levels of our categorical variable. Why would we do this? We are interested only statistical control for random variation in our experimental units. For example we would use such a study to plan for an experiment in plots on the landscape where we cannot control for spatial variation by forcing plots to uniformity.

A random factor has categories that are considered a sample from some larger population of units, such as plots or fields in an agricultural experiment. A fixed factor has levels that are the only ones of interest, as in the analysis of sleep data in relation to drug. Here are some guidelines (from D.G. Kleinbaum and L.L. Kupper. 1978. Applied Regression Analysis. Boston: Duxbury Press).

Random	Fixed	Either, depending on situation
Subjects	Sex (M F)	Locations
Litters	Age (age groups)	Time
Observers	Drugs, Treatments	

With a fixed factor we are inferring only to the levels at hand. With a random factor we assume that the levels are a sample from a larger population of levels or units, which vary. With a random factor we infer to a larger population of levels or units that might have been chosen. The choice between random and fixed depends on how we define the population. 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 intertidal locations. Then repeats the experiment two more times, so that each location is exposed to each nutrient level on all three occasions. The nutrient factor is clearly fixed. The location factor is usually random. However, the location factor could be taken as fixed, if the biologist had chosen the locations to represent the range of possible conditions in the study area. The three occasions can also be taken as either random or fixed. They would be random if known sourced of temporal variation, such as season and time of day were held the same. The become fixed if occasions differ with respect to a known source of temporal variation.

One way ANOVA, Random Effects.

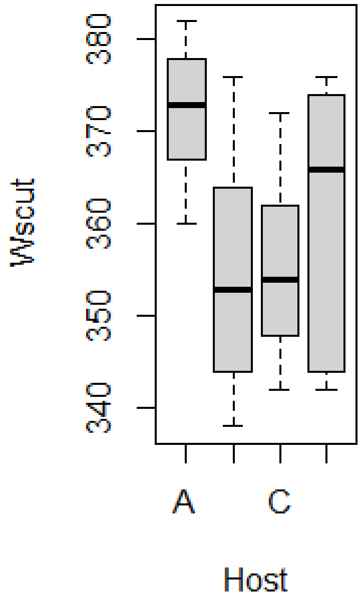
Example. Data from Box 9.1 of Sokal and Rohlf 2012, p. 209.
Does the variance in tick size among hosts (rabbits) exceed variance within?

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 dispersion around 4 means, one for each rabbit.



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

Are the explanatory variables categorical?
Yes.

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
In general, it is best to avoid using numeric labels for categorical variables.
These are too easily taken as ratio scale rather than nominal scale (categorical) variables.

	Wscut	HostNumber	HostLabel
1	360	1	A
…			
8	382	1	A
9	338	2	B
10	342	2	B

2. Execute analysis.

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.
```

```
TSizeRanMod<lm(Wscut~(1|Host), )
TSizeRanMod$fitted.values
TSizeRanMod$residuals
```

The ANOVA code on the left worked before graphics interfaces became available.
It still works.
The random effects code (1 |Host) on the right worked in 2019.
Four years later this code no longer works.

```
TSizeRanMod<lmer(Wscut~(1|Host), )
TSizeRanMod$residuals  [null]
TSizeMod<lm(Wscut~(1|Host), )
TSizeMod$residuals  [7.75 3.75 …..
```

The lmer() call in R uses 1|Host.
It does not produce residuals.
It cannot be used for model checking of this analysis.

In this analysis the parameters (means) are no longer of interest.
The variance among rabbits, relative to within, is of interest.
The partitioning of the total variance (among versus within host) is of interest.

Digression on notation.
There are several different symbols for estimates of variance and standard deviation.

Placing a hat over the greek symbol $\hat{\sigma}_W$ for the standard deviation.

Using a roman letter s_W for estimate of σ_W

Subscript notation becomes cumbersome for the variance $\sigma^2{}_W$
It becomes more cumbersome for the estimate of the variance $\hat{\sigma}^2{}_W$
An alternative is to use a function for the estimate: stdev(W), var(W).

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 more 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 categories.

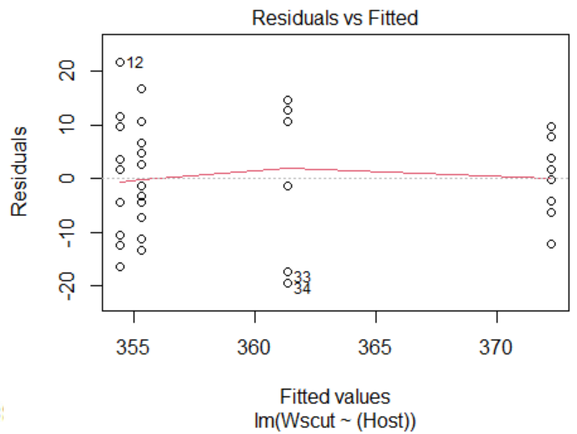
3. Evaluate model.

3a Structural model.

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

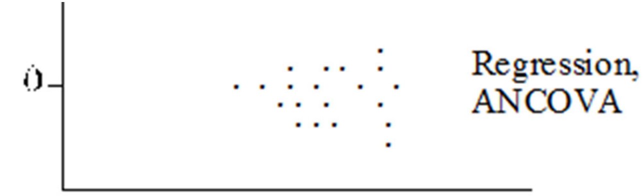
3b Error model. Homogeneity?

Plot residuals versus fitted values.

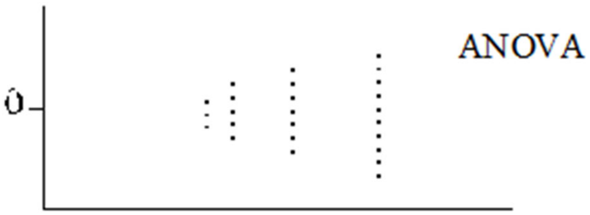


Residual versus fit plot shows vertical distribution of residuals to be about the same in all four groups. So residuals are judged homogeneous.

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.



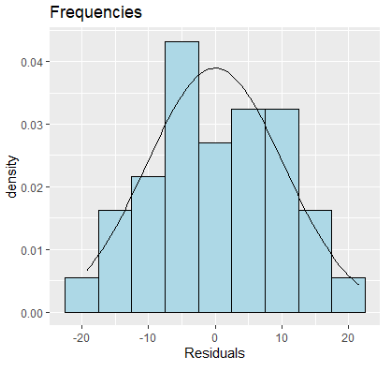
For ANOVA, there are a limited number of fitted values, hence the plot is present at only a few points long the x-axis. The fan pattern is the same in both plots, but vertical swaths are missing from the plot with categorical variables.

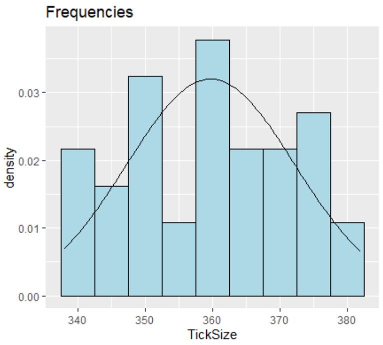


Residuals normal ?

The residuals are close to normal. The scutum widths are less so.

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



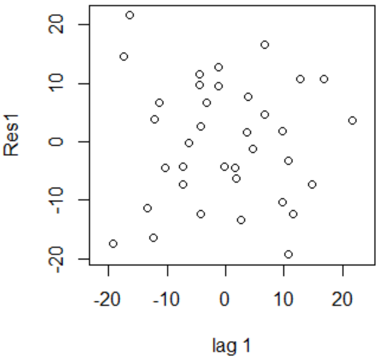


3b Evaluate error model.

Residuals independent?

Yes.

Residuals in order of data presented in text book show no upward or downward trend.



Conclusion.

Normal error model acceptable.

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.

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

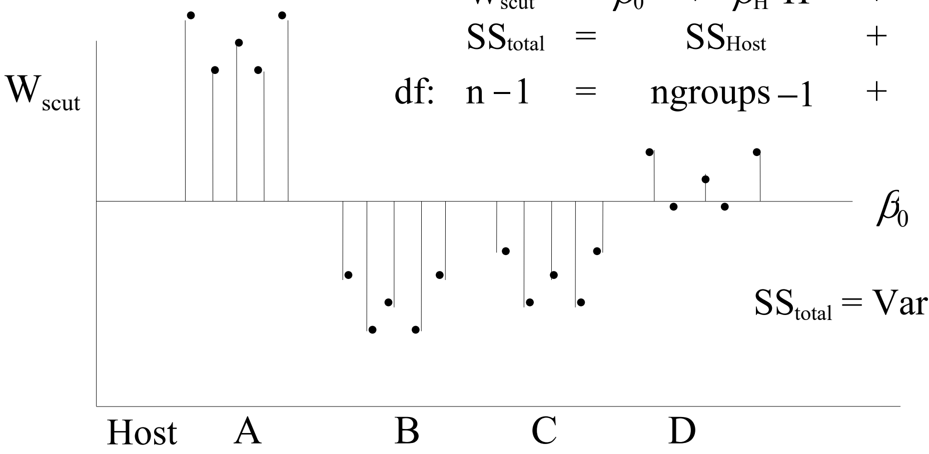
1. $SS_{tot} = (n - 1) * Var(W_{scut}) = 36 * 155.2 = 5586$
2. $SS_{tot} = \Sigma W_{scut}^2 - n^{-1} (\Sigma W_{scut})^2 = 4792797.3 - 37^{-1} \cdot 13308.9^2 = 5586$

```
MTB> let k2 = mean('width')
MTB> let k3 = SSQ('width' - k2) )
MTB> print k3
MTB> let k1 = (37-1)*stdev('width')*stdev('width')
MTB> print k1  (should be same as k3)
```

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. Partition SS according to the model

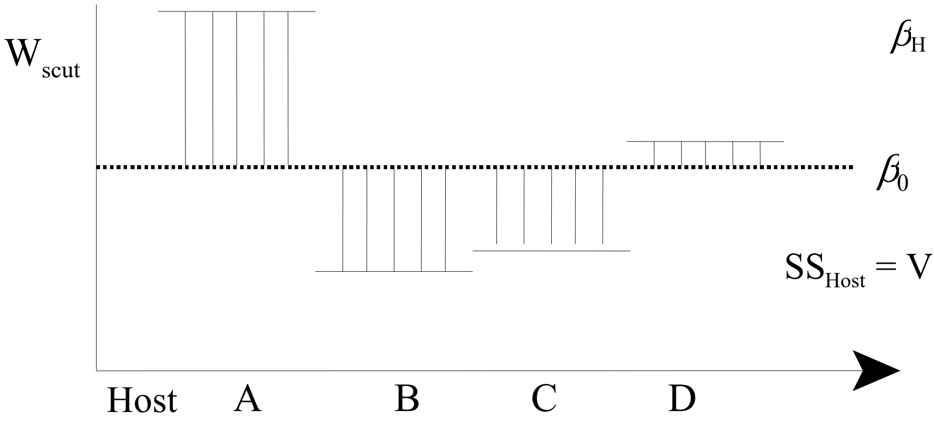
$$\begin{array}{rlclcl} & W_{scut} & = & \beta_0 \; + \; \beta_H \cdot H & + & \varepsilon \\ & SS_{total} & = & SS_{Host} & + & SS_{residual} \\ \text{df:} & n-1 & = & \text{ngroups}-1 & + & n-\text{groups} \end{array}$$



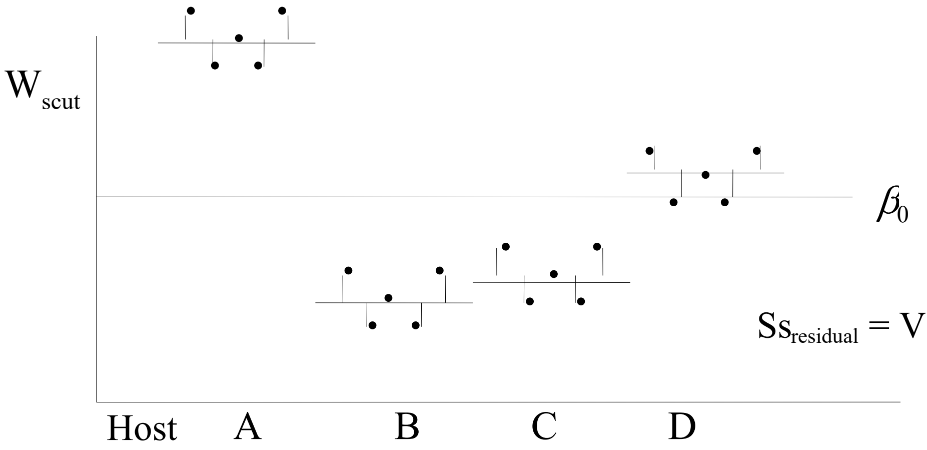
$\beta_0 = 359.7$

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

$$\beta_H = \begin{array}{l} +12.55 \\ -5.33 \\ -4.4 \\ +1.6 \end{array}$$



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



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

4. Calculate likelihood ratio for omnibus model

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

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

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

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

LR = $(5586)^{37/2}$ / $(3778)^{37/2}$ / = 10994

The alternative model (variance due to hosts) is over 10^5 times more likely than the null model, no variance due to hosts.

4. State model pair.

The focus of the random effects analysis is the variance in parasite size among rabbits. The focus differs from fixed effect factors, where the H_A/H_o pair is stated as contrasts among means.

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."

5. State the population and whether the sample is representative.

Text examples present data, rather than data situations. In practice most data is collected in a situation where there is considerably more known than just the numerical values of each quantity. This information can be used to judge a reasonable target of inference. Statistical inference is a procedure for making statements about populations based on samples. The statement about a population is valid if (1) the sample is representative of the population and (2) logical statistical procedures are used.

The conditions for taking the sample are important.
Hypothetical populations are used in many applications. Here we assume that the results can be inferred to any future study carried out according to the same protocol.
Enumerable populations are sometimes used. We can enumerate all possible units, sample randomly from these units, and from this assume that the sample represents the larger population of units.

For this example (Scutum widths) we are going to infer to a hypothetical population of rabbits similar to those in this sample.
Conclusions by statistical inference apply to any study that uses the same measurement protocol, including the method used to sample rabbits.

5. State the population. Fixed versus random effect factors.

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

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

In this case a measure of weight of evidence would be sufficient. At the same time, inference to a population is valid because we can define a population based on the experimental protocol. We will report the likelihood ratio as a measure of evidence, without hypothesis testing.

Here are the data equations.

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

  ROW     width      fits          res

    1       380      372.250      7.7500
    2       376      372.250      3.7500
    3       360      372.250     -12.2500
    4       368      372.250      -4.2500
    5       372      372.250      -0.2500
    6       366      372.250      -6.2500
    7       374      372.250       1.7500
    8       382      372.250       9.7500
    9       350      354.400      -4.4000
   10       356      354.400       1.6000
   11       358      354.400       3.6000
   12       376      354.400      21.6000
   13       338      354.400     -16.4000
   14       342      354.400     -12.4000
   15       366      354.400      11.6000
   16       350      354.400      -4.4000
   17       344      354.400     -10.4000
   18       364      354.400       9.6000
   19       354      355.308      -1.3077
   20       360      355.308       4.6923
   21       362      355.308       6.6923
   22       352      355.308      -3.3077
   23       366      355.308      10.6923
   24       372      355.308      16.6923
   25       362      355.308       6.6923
   26       344      355.308     -11.3077
   27       342      355.308     -13.3077
   28       358      355.308       2.6923
   29       351      355.308      -4.3077
   30       348      355.308      -7.3077
   31       348      355.308      -7.3077
   32       376      361.333      14.6667
   33       344      361.333     -17.3333
   34       342      361.333     -19.3333
   35       372      361.333      10.6667
   36       374      361.333      12.6667
   37       360      361.333      -1.3333
```

$sd^2 = 12.46^2 \quad 7.09^2 \quad 10.24^2$

$sd^2 \cdot 36 = 155.25 \quad 50.27 \quad 104.86$

$SS = 5589 \quad 1809 \quad 3775$

$$F = \frac{SS_{fits} / df_{fits}}{SS_{res} / df_{res}}$$

$$F = \frac{1808 / 3}{3778 / 33}$$

$$F = 5.26$$

7. Complete the ANOVA table, showing % variance at each level

Source	df	SS	% SStotal	Source
Host	3	1808	32.4	SS among
Res	33	3778	67.6	SS within
Total	36	5586		

8. Recompute the p-value by randomization if assumptions are not met.
Not necessary. Assumptions met. No estimates of parameters or p-values.

9. Report statistical conclusion.

$LR = L(\beta_o\ , \beta_H \mid\ W_{scut}\ )\ \ /\ \ L(\beta_o \mid\ W_{scut}\ )\ = 10^5$

The full model is10^5 times more likely than the reduced (null) model.
The variance among hosts accounts for 32% of the variance.

10. Report science conclusions.

The variance among the means is of interest. How large is the variance among groups, compared to the total variance across all ticks? This information will be used in planning further experiments.

The among unit SS = 1808 / 5586 = 32%

The among rabbit variability is far from 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.