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. Interest is in whether there is variance among groups, above and beyond variance within groups.

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 effects. The categories we used were either fixed by intervention (as with an experiment) or fixed by choice of units (as with size classes of fathers, or choice of species of limpets). We can also analyze our response variable with respect to random effects. We do not intervene nor do we choose levels of our categorical variable. Why would we do this? Sometimes we have no choice. We know too little about our experimental units to make a choice by an attribute that drives the response variable in a known way. Often, we are interested only in repeatability. We carry out an experiment at one location, then repeat it at another location just to see if we get the same result regardless of location. Location is by convenience, or haphazard, and so it is a random factor.

A random factor has categories that are considered a sample from some larger population of units (such as plots or fields in an agricultural an experiment). A fixed factor has levels that are the only ones of interest (as in the analysis of sleep data in relation to drug). The decision on whether to consider a categorical variable as a random or fixed effect is not always clear cut. However, 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	

Treatments, drugs, *etc*. are fixed effects because we intervene to set them.

Age is considered a fixed factor because ages are ordered, we do not consider the ages we use to be randomly taken from some larger population, and we expect differences among age groups.

Age is usually a fixed factor because we expect changes with age.

Time is a fixed factor when we expect some change with time, usually directional. However, on short time scales, we might consider time a random factor if we repeat an experiment at some later time to see if we get the same result.

Sex is usually a fixed factor if we expect differences among sexes. It *could* be a random factor if all we want to know is whether there are differences among sexes.
Location is a fixed factor if we choose locations by some characteristic.
It is a random factor if we make no choice.

Subjects, litters, and observers are usually considered random because we wish to infer to a larger population of potential subjects, litters, or observers.

Application of GLM. One way ANOVA, Random Effects.

Example. Generic recipe applied to Box 9.1 of Sokal and Rohlf 1995, p. 210.
Does tick size, as measured by scutum width, differ among hosts (rabbits)?

1. Construct model.

State a verbal model of explanatory variable *H* relative to W_{scut} VERBAL

"scutum width W_{scut} depends on host identity H

Draw a picture of W_{scut} as a function of *H* GRAPHICAL

Draw 4 stacks of data points,
one stack for each rabbit on X axis

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

See next page ---->

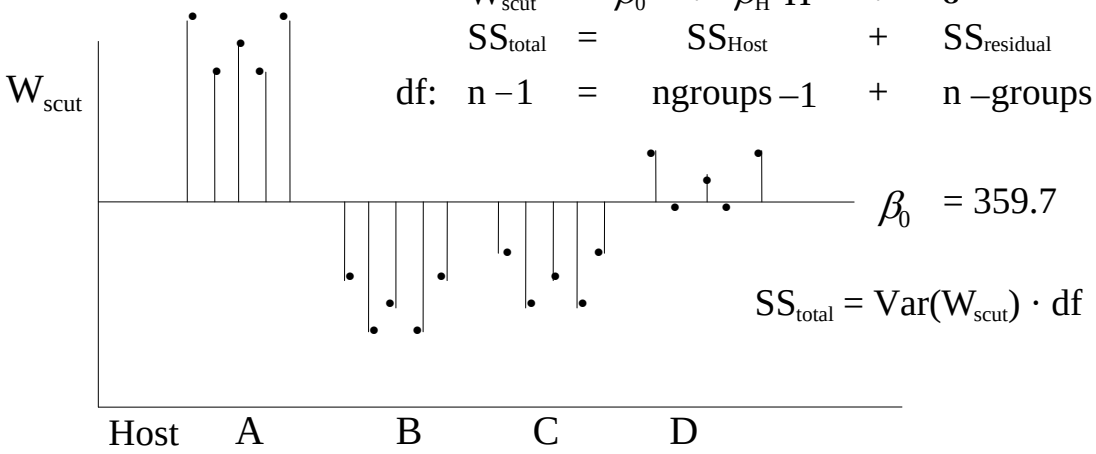
Response variable is scutum width of ticks, W_{scut} = microns

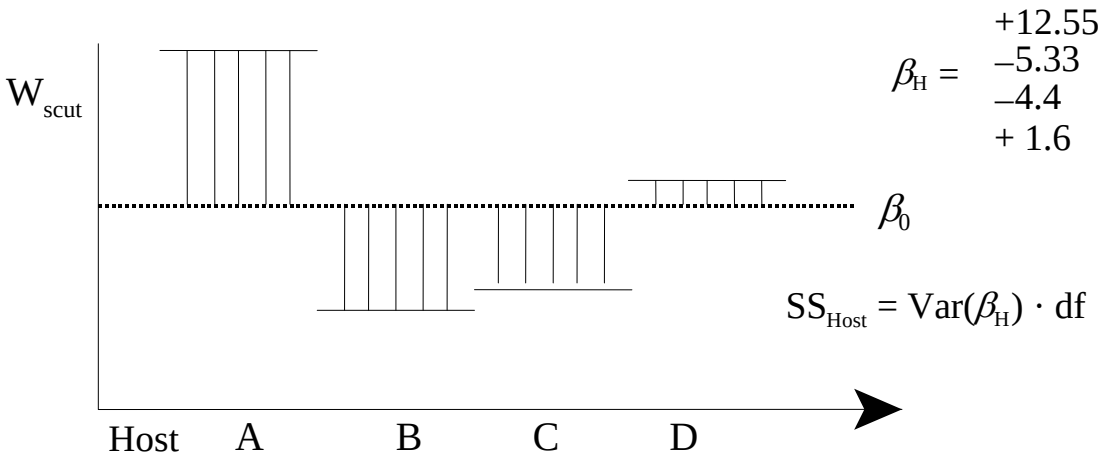
Explanatory variable is host, *H* = Rabbit A, Rabbit B, Rabbit C, Rabbit D

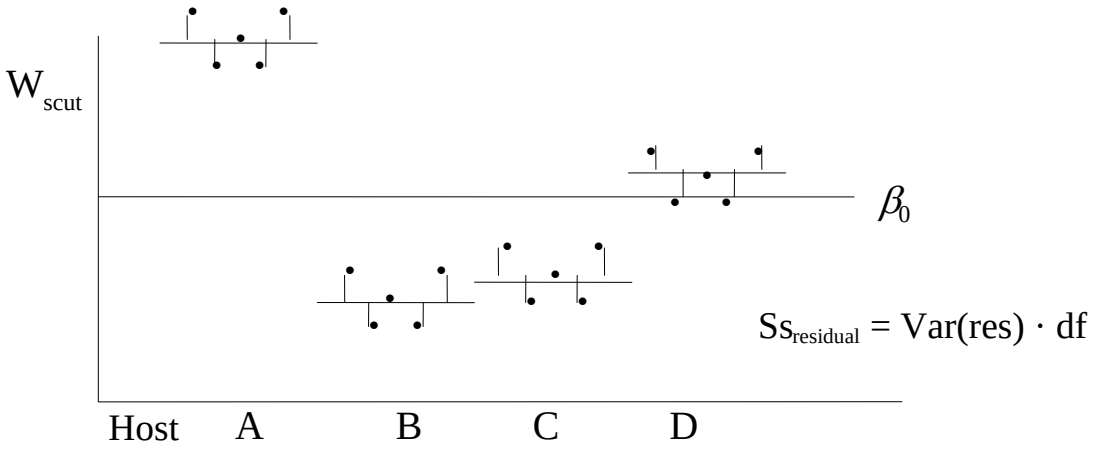
$$W_{scut} = \beta_0 + \beta_H \cdot H + \varepsilon$$

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

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







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 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.
```

The fitted values are the means in each of the four groups.

Fits and residuals from

model statement output of fitted values and residuals (as above)

direct calculation of parameters (four means)

parameters reported by GLM routine

The parameters of the model can be slopes, means, odds ratios (as in G-tests). These estimates are generally accomplished by statistical packages, with little discussion of how the estimates are made. In general, the widely used statistical packages provide good all-around estimates of parameters, based on widely accepted procedures.

There are whole courses in statistics devoted to the matter of estimating parameters. Very little time will be devoted to the matter of estimation in this course. It is sufficient to know that parameters (usually represented by greek letters) are the value for the entire population, while estimates of this true value (which cannot be known without sampling the entire population) are made from samples.

2. Execute analysis.

For ANOVA, the parameters are the means in each group.

For ANOVA in the GLM format, parameters are β_o the grand mean, and β_H the means in each group, expressed as deviations from β_o

To estimate β_o

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

$\hat{\beta}_o = 359.7$

2. Execute analysis.

To estimate β_H

```
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 parameteric 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}.

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.

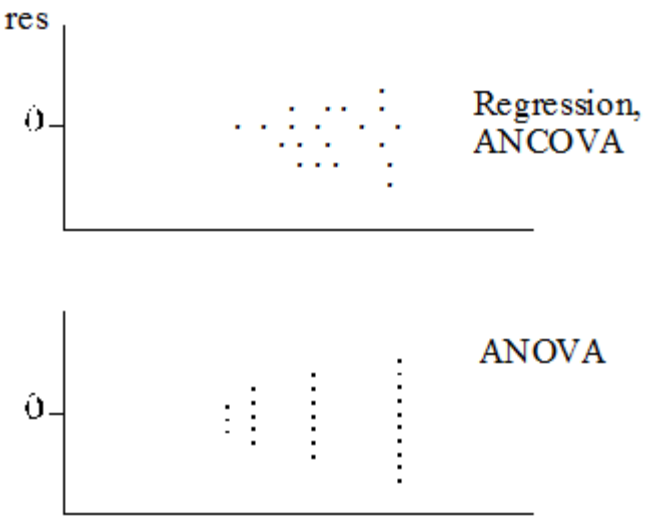
```
MTB > plot 'res' 'fits'

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 same vertical distribution of residuals to be about the same in all four groups. So residuals are homogeneous.

When this assumption is not met, the plot of residuals versus fits will often show left or right facing cones 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 cone pattern is the same in both plots, but vertical swaths have been removed from the ANOVA plot.

3. Evaluate model.

Error model

Residuals normal ?

The residuals meet the normality assumption even though the response variable does not.

```
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 would erroneously conclude that the normality assumption was not met, when in fact it was.

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

Text examples rarely state the population or conditions for taking measurements. For the sake of brevity 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. Statistical methods are a way of making statements about <u>populations</u> based on <u>samples</u>. The statement about a population is valid if (1) the sample is representative of the population and (2) appropriate statistical procedures are used.

The conditions for taking the sample are important.

<u>Hypothetical populations</u> 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.

<u>Enumerable populations</u> are sometimes used. Here 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. Conclusions by statistical inference should apply to any study that uses the same measurement protocol.

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

Fixed versus random effects.

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 fixed effects model. However, usual practice is to treat the rabbits as a random sample of all possible rabbits. If we do this, then wwe have a random effects model. Choice of random or fixed factor will determine how we construct our H_A and H_o pair.

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

In this case it is appropriate because we want to know if there is an among host component of variance in scutum widths, in addition to the variance within a host. We are not interested in parameter values or effect sizes.

6 State H_A / H_o pair, test statistic, distribution, tolerance for Type I error.

There is one term in the model, Hosts. This is a random factor. The interest lies in whether there is an additional component of variability in tick size, due to their host. The H_A / H_o pair thus concerns the variance among means.

H_A: $\text{Var}(\beta_H \cdot H) > 0$

"The true group means deviate from the true grand mean, hence there is variance in the collection of values, $\beta_H \cdot H$

H_o: $\text{Var}(\beta_H \cdot H) = 0$

H_o includes everything not covered by H_A.
Hence another reason for stating the H_A first.

This statement of the H_A / H_o pair differs from the fixed effects model, such as we saw with the oneway ANOVA on the pea growth data.

State test statistic	F-ratio
Distribution of test statistic	F-distribution
Tolerance for Type I error	5%

The tolerance for Type I error will be set at the conventional level of 5%

6 State H_A / H_o pair, test statistic, distribution, tolerance for Type I error.

The tolerance for Type I error is called α, which is conventionally set at 5% in biology.
There is no reason that α has to be 5%, one can set this at 1% or at 10%, depending on whether one is worried about Type II error. Setting α at a low value (*e.g.* α = 0.001) increases Type II error, the chance of rejecting a true effect.
If many tests are to be made, it is advisable to set tolerance for Type I error at α/n, where n = number of tests. This is called a Bonferroni criterion, or Bonferroni tests. It takes into account the fact that the Type I error for multiple tests is not the same as for a single test. If you set α = 5% and perform 20 tests, then you expect one "significant" result even when there is no significant effect.
Because α should generally be set <u>before</u> one carries out a test, α should be stated in the methods section. If different α levels are used in a single study, the reasons for the different levels should be stated.

7. ANOVA - Compute and partition the df 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 4 classes (*e.g.* treatment and 3 types of fertilizer) requires 3 degrees of freedom (df = number of classes – 1). As a rule if explanatory variable X is:

ratio scale	df = 1
nominal scale (factor) with *n* classes	df = *n* – 1

In addition, 1 df is lost in estimating the grand mean.

7. **ANOVA - Compute and partition the variance in the response variable 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. $SS_{tot} = \Sigma W_{scut}^{2} - n^{-1} (\Sigma W_{scut})^{2} = 4792797.3 - 37^{-1} \cdot 13308.9^{2} = 5586$

```
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

We arrange Source, df, and SS in an ANOVA table. The moving from left to right, we
-compute MS from SS and df in ANOVA table
-compute F from MS
-compute p-value from F-ratio.

Source	df	SS	MS	F	---->	p
Host	3	1808	602.6	5.26		0.004
Res	33	3778	114.5			
Total	36	5586				

Computer packages produce these tables. However, it is important to learn how one quantity is computed from another in this table, in order to understand the table. It is also important to write the model out, before executing the analysis. Writing the model, and the list of explanatory variables, then calculating the degrees of freedom, is useful in making sure the computer executed the analysis you had in mind, rather than something else.

7. ANOVA - Compute SS, MS, F-ratio in table

In learning to apply the GLM, it helps greatly to work out the degrees of freedom by hand and fill in the table, rather than relying on the computer for this. In this way you can check that the computer has produced the partitioning that you wanted. Partitioning df_{tot} is easy, partitioning SS_{tot} is hard to do by hand.

Here are the data equations, to show the relation of variance components to 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      7.09^2     10.24^2
sd^2 *36 = 155.25    50.27     104.86
SS       = 5589     1809      3775
```

Minitab can be used to calculate each data equation, then calculate SS MS F from data equations.

$SS_{tot} = Var(W_{scut}) * df_{tot}$

$SS_{fits} = Var(fits) * df_{tot}$

$SS_{res} = Var(res) * df_{tot}$

```
MTB> let k1 = Stdev('width')*36
MTB> let k2 = Stdev('fits')*36
MTB> let k3 = Stdev('res')*36
```

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

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

$$F = 5.26$$

7. ANOVA - Compute p-value for the observed F-ratio
The observed *F*-ratio was 5.26, the p-value is
p = 0.004, as calculated from the cdf for an *F*-distribution with 3 and 33 *df*.

8. Recompute p-value by randomization if necessary.
Not necessary, $n > 30$ and residuals homogeneous and normal.

9. Declare and report statistical decision, with evidence

$0.004 = p < \alpha = 0.05$ so reject H_o and accept H_A
reject $H_o : var(\beta_H) = 0$
accept $H_A: var(\beta_H) > 0$

The variance among hosts exceeds variance within hosts.
There is additional component of variability due to hosts.
$F_{3,33} = 5.26$ $p = 0.004$
Note that the exact p-value is reported.
The degrees of freedom are reported as subscripts, the numerator df is always listed first.

10. Report and interpret parameters of biological interest.
There will be no analysis of parameters in this example because the single term in the model is a random factor.

In this example the interest was in whether there was variance among the hosts.
There was no stated interest in which hosts differed, or by how much.

Model I versus Model II ANOVA

Fixed versus random effects.

Model I ANOVA. Explanatory variable is fixed treatment.
This is written $Y = \mu + \alpha + \varepsilon$
Our interest is in contrast among means.
A priori contrasts are used in confirmatory analysis.
A posteriori contrasts are more exploratory in nature.

Model II ANOVA. Explanatory variable is random.
This is written $Y = \mu + A + \varepsilon$
Our interest is in variance in *Y* due to classification, not in the means.