Single Factor ANOVA Models

Corresponds to Chapter 12 of Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT)
with some slides by Jacqueline Telford
(Johns Hopkins University).

Chapter 8: How to compare two treatments

Chapter 12: How to compare more than two treatments (or just two).

Example: yields of several varieties of barley.

Variety is the treatment factor (predictor)

Yield is the response

Experimental Designs

	Two Treatments	More than two treatments
Independent Samples	Independent Samples	Completely Randomized Design
Dependent Samples	Matched Pair Design	Randomized Block Design

S-Plus barley data set (observation 13:30)

```
> barley.small
      yield     variety   year   site
13  35.13333  Svansota  1931   University Farm
14  47.33333  Svansota  1931   Waseca
15  25.76667  Svansota  1931   Morris
16  40.46667  Svansota  1931   Crookston
17  29.66667  Svansota  1931   Grand Rapids
18  25.70000  Svansota  1931   Duluth
19  39.90000  Velvet    1931   University Farm
20  50.23333  Velvet    1931   Waseca
21  26.13333  Velvet    1931   Morris
22  41.33333  Velvet    1931   Crookston
23  23.03333  Velvet    1931   Grand Rapids
24  26.30000  Velvet    1931   Duluth
25  36.56666  Trebi     1931   University Farm
26  63.83330  Trebi     1931   Waseca
27  43.76667  Trebi     1931   Morris
28  46.93333  Trebi     1931   Crookston
29  29.76667  Trebi     1931   Grand Rapids
30  33.93333  Trebi     1931   Duluth
```

Completely Randomized Design Notation

See Table 12.1, page 458 in the course textbook.

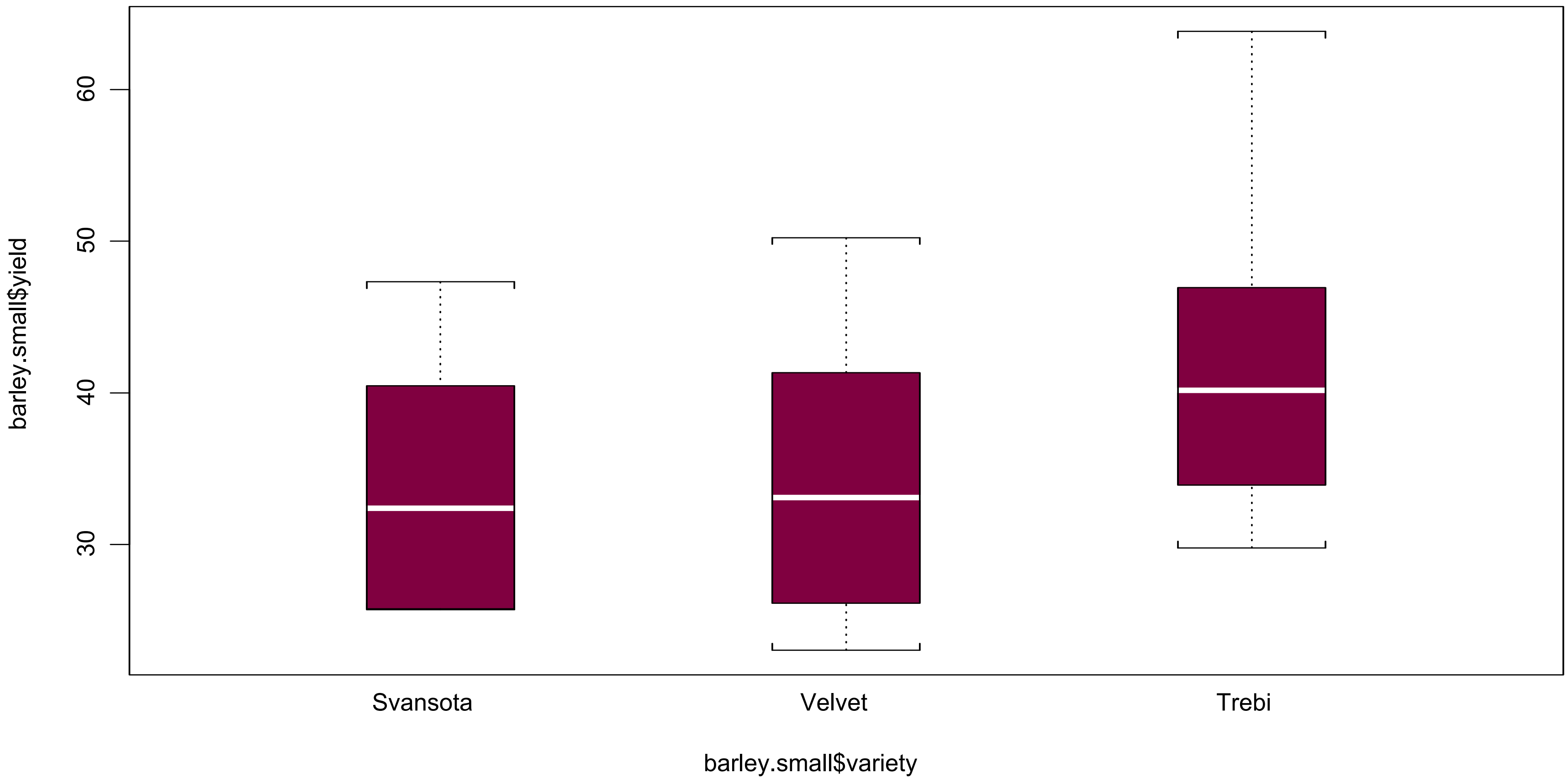
If the sample sizes are equal the design is **balanced**; otherwise the design is **unbalanced**

$$N = \sum_{j=1}^{a} n_i$$

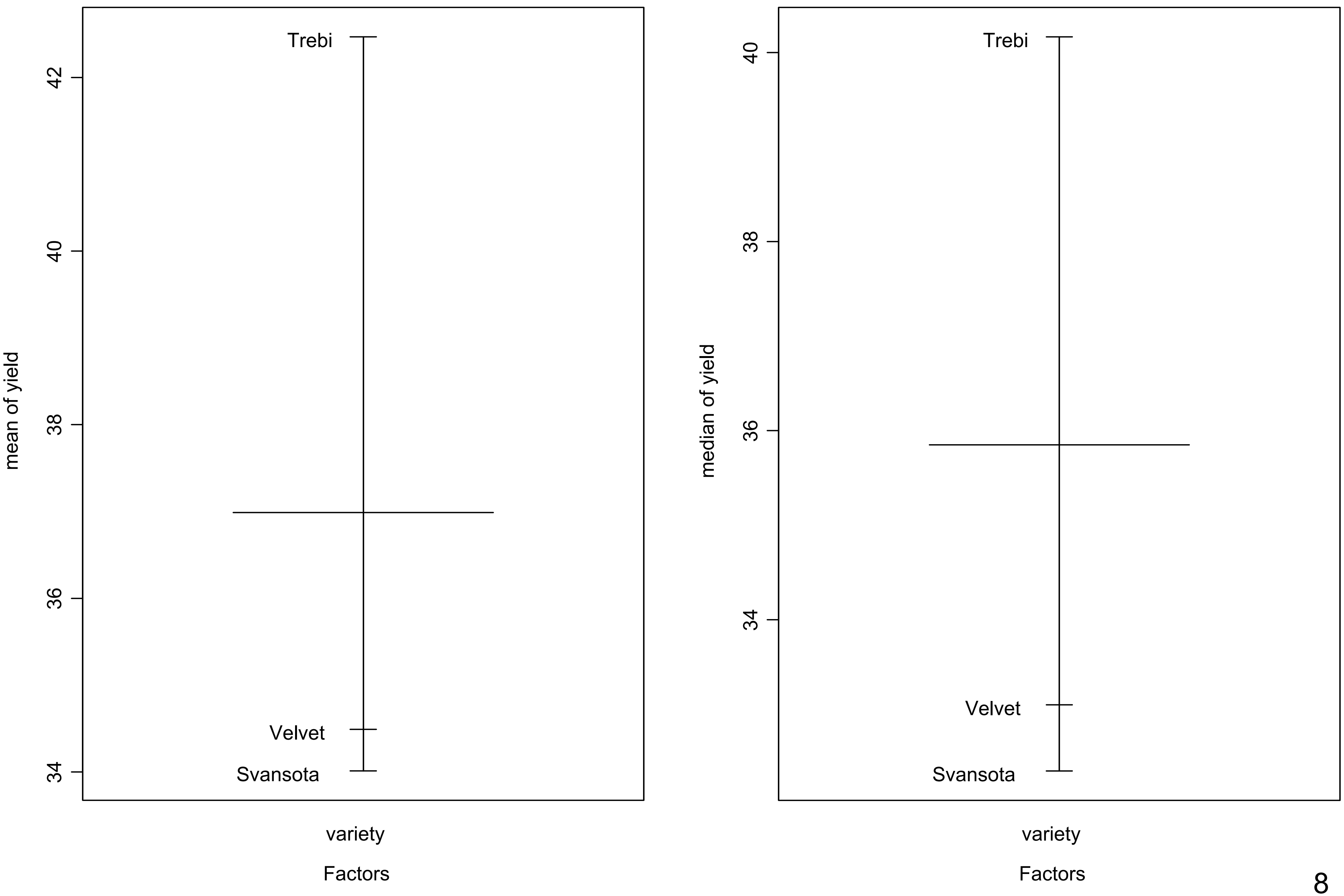
S-Plus barley dataset (observations 13:30)

Variety	Svansota	Velvet	Trebi
	35.13333	39.90000	36.56666
	47.33333	50.23333	63.83330
	25.76667	26.13333	43.76667
	40.46667	41.33333	46.93333
	29.66667	23.03333	29.76667
	25.70000	26.30000	33.93333
Variety Mean	34.01111	34.48889	42.46666

Plot of yield by variety for S-Plus barley data set



S-plus plot.design function



CRD: Model and Estimation (cell means model)

See Section 12.1.1 and Figure 12.2 on page 460 of the course textbook.

CRD: Treatment Effects Model

Alternative Formulation of the Model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \ \ (i = 1, 2, ..., a; \ j = 1, 2, ..., n_i)$$

Formula from 12.1.1, page 460 in the course textbook.

CRD parameter estimates

$\mu = \textit{grand}$ mean, estimated by $\bar{\bar{\text{y}}} = (1'\text{y})/\text{n}$

$\mu_\text{i} =$ mean of i^{th} treatment, estimated by $\bar{\text{y}}_\text{i} = (1'\text{y}_\text{i})/\text{n}_\text{i}$

$\hat{y} =$ vector of fitted values $=$ treatment means

$\text{e} =$ error $= \text{y} - \hat{y}$

σ^2 estimated by $\text{s}^2 = \text{e}'\text{e}/(\text{n} - \text{a})$

Fitted values and residuals for barley example

```
> cbind(barley.small[,1:2],fitted(tmp),resid(tmp))
       yield    variety     fitted      resid
13  35.13333  Svansota  34.01111   1.122218
14  47.33333  Svansota  34.01111  13.322218
15  25.76667  Svansota  34.01111  -8.244442
16  40.46667  Svansota  34.01111   6.455558
17  29.66667  Svansota  34.01111  -4.344442
18  25.70000  Svansota  34.01111  -8.311112
19  39.90000    Velvet  34.48889   5.411113
20  50.23333    Velvet  34.48889  15.744443
21  26.13333    Velvet  34.48889  -8.355557
22  41.33333    Velvet  34.48889   6.844443
23  23.03333    Velvet  34.48889 -11.455557
24  26.30000    Velvet  34.48889  -8.188887
25  36.56666     Trebi  42.46666  -5.900000
26  63.83330     Trebi  42.46666  21.366640
27  43.76667     Trebi  42.46666   1.300010
28  46.93333     Trebi  42.46666   4.466670
29  29.76667     Trebi  42.46666 -12.699990
30  33.93333     Trebi  42.46666  -8.533330
```

X matrix?

1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1

Model.matrix in S-Plus

```
> round(model.matrix(barley.small.aov),3)
   (Intercept) variety.L variety.Q
13           1    -0.707     0.408
14           1    -0.707     0.408
15           1    -0.707     0.408
16           1    -0.707     0.408
17           1    -0.707     0.408
18           1    -0.707     0.408
19           1     0.000    -0.816
20           1     0.000    -0.816
21           1     0.000    -0.816
22           1     0.000    -0.816
23           1     0.000    -0.816
24           1     0.000    -0.816
25           1     0.707     0.408
26           1     0.707     0.408
27           1     0.707     0.408
28           1     0.707     0.408
29           1     0.707     0.408
30           1     0.707     0.408
```

Model Coefficients

- > summary.lm(barley.small.aov)

- Call: aov(formula = yield ~ variety, data = barley.small)
- Residuals:
-     Min   1Q Median   3Q   Max
- -12.7 -8.294 -1.611 6.194 21.37

- Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	36.9889	2.5207	14.6741	0.0000
variety.L	5.9790	4.3660	1.3695	0.1910
variety.Q	3.0619	4.3660	0.7013	0.4939

- Residual standard error: 10.69 on 15 degrees of freedom
- Multiple R-Squared: 0.1363
- F-statistic: 1.184 on 2 and 15 degrees of freedom, the p-value is 0.3332

- Correlation of Coefficients:

	(Intercept)	variety.L
variety.L	0	
variety.Q	0	0

S-plus model.tables command gives treatment means or effects

```
> model.tables(barley.small.aov,type="mean")
Warning messages:
  Model was refit to allow projection in: model.tables(tmp, type =
    "mean")

Tables of means
Grand mean

 36.989

 variety
 Svansota   Velvet    Trebi
   34.011   34.489   42.467
```

S-plus model.tables command gives treatment means or effects

```
> model.tables(barley.small.aov)
Warning messages:
  Model was refit to allow projection in:
    model.tables(barley.small.aov)

Tables of effects

 variety
 Svansota  Velvet    Trebi
 -2.9778   -2.5000   5.4778
```

Analysis of Variance (ANOVA)

Homogeneity Hypothesis:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_a \;\; vs. \;\; H_1: Not\; all\; the\; \mu_i\; are\; equal.$$

$$H_0: \tau_1 = \tau_2 = \ldots = \tau_a \;\; vs. \;\; H_1: At\; least\; some\; \tau_i \neq 0.$$

Variation Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Treatments (A)	$\sum n_i(\bar{y}_i - \bar{\bar{y}})^2$	$a-1$	$\frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Error (E)	$\sum \sum (y_{ij} - \bar{y}_i)^2$	$N-a$	$\frac{SSE}{N-a}$	
Total (T)	$\sum \sum (y_{ij} - \bar{\bar{y}})^2$	$N-1$		

Note SSR=SSA=Treatment sums of squares

ANOVA table for model with 3 varieties of barley, year 1

```
> summary(aov(yield~variety,barley.small))
            Df  Sum of Sq  Mean Sq  F Value     Pr(F)
   variety  2   270.739    135.3694 1.183614 0.3332005
Residuals 15  1715.544   114.3696
```

ANOVA table for model with all 10 varieties of barley, year 1

```
> summary(aov(yield~variety,barley1))
            Df  Sum of Sq  Mean Sq   F Value    Pr(F)
   variety  9    646.262     71.8069 0.5963671 0.793823
Residuals 50  6020.357   120.4071
>
```

F-statistic for One-way ANOVA

$$F = \frac{MSA}{MSE} \sim F_{a-1,n-a}$$

$$E(MSE) = \sigma^2$$

$$E(MSA) = \sigma^2 + \frac{\sum_{i=1}^{a} n_i \tau_i^2}{a-1}$$

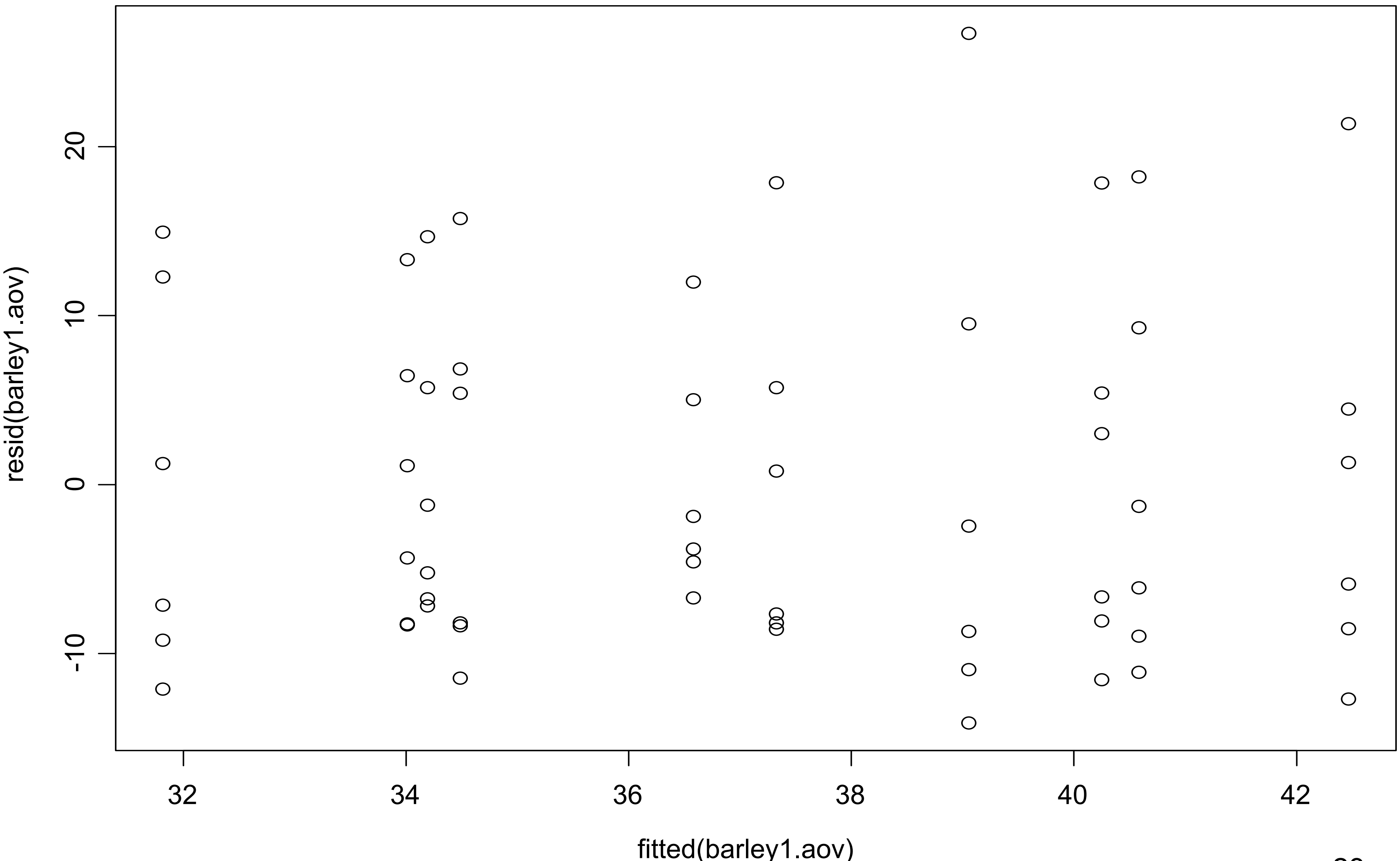
Fitting model with continuous vs. character predictor

```
> summary(aov(barley.small$yield~varnum))
          Df Sum of Sq  Mean Sq F Value     Pr(F)
   varnum  1   214.489   214.4889 1.93692   0.1830502
Residuals 16  1771.794 110.7371

> summary(aov(barley.small$yield~as.factor(varnum)))
                   Df Sum of Sq  Mean Sq  F Value     Pr(F)
as.factor(varnum)  2    270.739  135.3694  1.183614 0.3332005
        Residuals 15  1715.544 114.3696
```

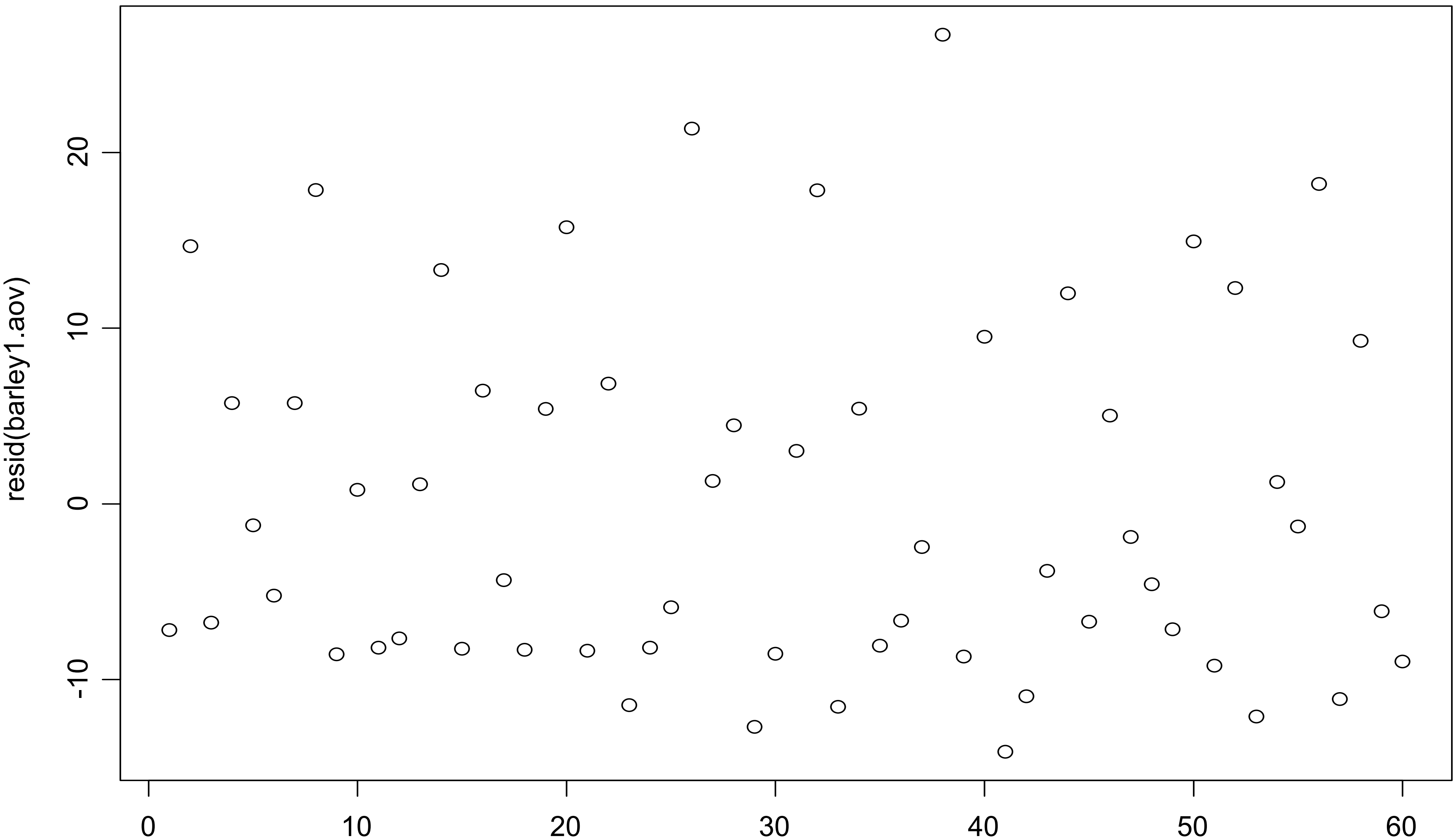
Equivalence of T test and ANOVA for model with single factor with 2 levels

```
> t.test(y[1:6],y[7:12])

    Standard Two-Sample t-Test

data:  y[1:6] and y[7:12]
t = -1.194, df = 10, p-value = 0.26
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.864726   6.909179
sample estimates:
 mean of x mean of y
  34.48889  42.46666

> summary(aov(yield~variety,barley.vsmall))
          Df Sum of Sq  Mean Sq   F Value     Pr(F)
  variety  1   190.935  190.9346  1.425727   0.2600178
Residuals 10  1339.209  133.9209
```

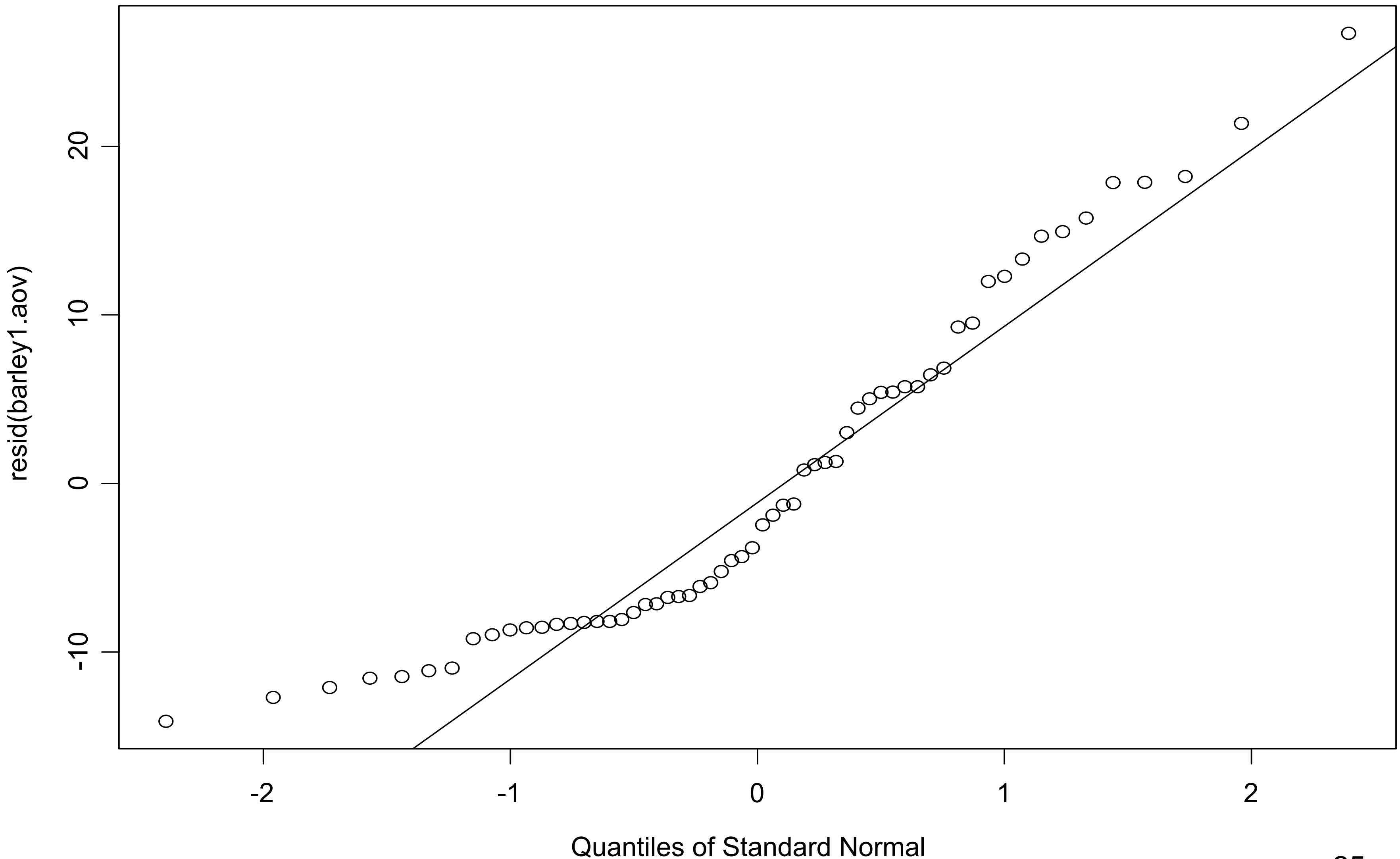
Model Diagnostics, residual vs. fitted value (all 10 varieties, year 1)



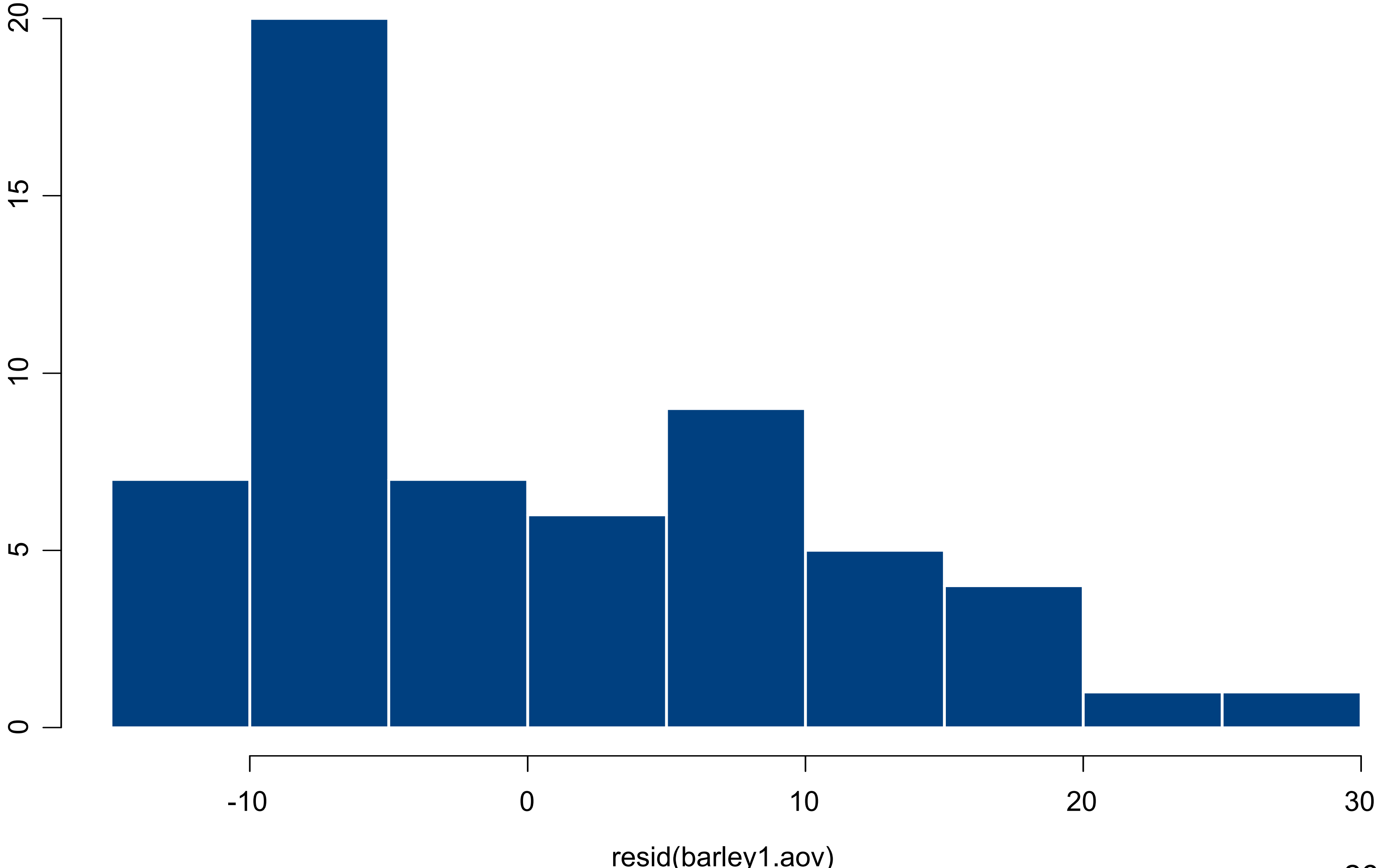
Model Diagnostics, residual vs. observation number
(all 10 varieties, year 1)



Model Diagnostics, normal plot of residuals
(all 10 varieties, year 1)



Model Diagnostics, histogram of residuals
(all 10 varieties, year 1)



Random Effects Model for a One-way Layout

When the treatment levels are determined by the experimenter (or those are the only levels of interest), the design is a fixed effects model.

- Goal is to measure the treatment effects or means ("pick the winner").

When the treatment levels are a random sample from a population of possible treatment levels (e.g. workers in a factory) and the particular levels used in the experiment are not of any interest, the design is a random effects model.

- Goal is to measure the treatment variability (estimate the expected variability among workers).

Random Effects Model for a One-way Layout

Model: $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$ (looks similar to the fixed effects model), where

$\varepsilon_{ij} \sim N(0,\sigma^2)$
$\mu_i \sim N(\mu,\sigma_A^2)$ or $\tau_i \sim N(0,\sigma_A^2)$ (constants in fixed effects model)

$\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\mu_i) + \mathrm{Var}(e_{ij}) = \sigma_A^2 + \sigma^2$

σ_A^2=variance among, σ^2 = variance within

With balanced one-way layout, n observations per treatment:

$$E(MSE) = \sigma^2$$

$$E(MSA) = \sigma^2 + n\sigma_A^2$$

Can estimate σ_A^2 as (MSA-MSE)/n (if you are lucky!)

Randomized Block Design

See Figure 3.2 on page 99 of the course textbook.

Barley Example
10 varieties, 6 sites

> ym

	University Farm	Waseca	Morris	Crookston	Grand Rapids	Duluth	Variety Mean
Manchuria	27.00000	48.86667	27.43334	39.93333	32.96667	28.96667	34.19445
Glabron	43.06666	55.20000	28.76667	38.13333	29.13333	29.66667	37.32778
Svansota	35.13333	47.33333	25.76667	40.46667	29.66667	25.70000	34.01111
Velvet	39.90000	50.23333	26.13333	41.33333	23.03333	26.30000	34.48889
Trebi	36.56666	63.83330	43.76667	46.93333	29.76667	33.93333	42.46666
No. 457	43.26667	58.10000	28.70000	45.66667	32.16667	33.60000	40.25000
No. 462	36.60000	65.76670	30.36667	48.56666	24.93334	28.10000	39.05556
Peatland	32.76667	48.56666	29.86667	41.60000	34.70000	32.00000	36.58333
No. 475	24.66667	46.76667	22.60000	44.10000	19.70000	33.06666	31.81667
Wisconsin No. 38	39.30000	58.80000	29.46667	49.86667	34.46667	31.60000	40.58333
Site Mean	35.82667	54.34667	29.28667	43.66000	29.05334	30.29333	37.07778

Randomized Block Design (RBD)Method

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \;\; (i = 1, \ldots, a;\; j = 1, \ldots, b)$$

$$\sum_{i=1}^{a} \tau_i = 0 \qquad\qquad \sum_{j=1}^{b} \beta_j = 0$$

a-1 independent treatment effects

b-1 independent block effects

For more information, see 12.4, page 482 in course textbook.

No Interactions Between Treatments and Blocks

$$\mu_{ij} - \mu_{i'j} = (\mu + \tau_i + \beta_j) + (\mu + \tau_{i'} + \beta_j) = \tau_i - \tau_{i'}$$

Formula from page 483 in the course textbook.

RBD: Sums of Squares

See formulas 12.17, 12.18, and 12.19 on pages 484-5 in the course textbook.

ANOVA tables for models for barley data set

```
> summary(aov(yield~variety,barley1))
            Df  Sum of Sq  Mean Sq   F Value    Pr(F)
  variety    9    646.262  71.8069  0.5963671 0.793823
Residuals   50   6020.357 120.4071

> summary(aov(yield~variety+site,barley1))
            Df  Sum of Sq  Mean Sq   F Value        Pr(F)
  variety    9    646.262   71.807   3.67995  0.001612103
     site    5   5142.272 1028.454  52.70610  0.000000000
Residuals   45    878.085   19.513
```

Type 1 and Type 3 Sums of Squares for barley example (balanced design)

```
> summary(barley12.aov)
          Df  Sum of Sq  Mean Sq   F Value       Pr(F)
  variety  9    646.262   71.807   3.67995 0.001612103
     site  5   5142.272 1028.454  52.70610 0.000000000
Residuals 45    878.085   19.513

> summary(barley12.aov,ssType=3)
Type III Sum of Squares
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
  variety  9   646.262   71.807  3.67995 0.001612103
     site  5  5142.272 1028.454 52.70610 0.000000000
Residuals 45   878.085   19.513
```

Degrees of Freedom

Counting the grand mean there are $1+(a-1)+(b-1)=a+b-1$ unknown parameters. (This many degrees of freedom are needed to estimate these parameters.)

There are $N=ab$ observations (total degrees of freedom).

So there are $\nu=ab-(a+b-1)=(a-1)(b-1)$ degrees of freedom for estimating the error variation (degrees of freedom for error).

Effects in barley model

```
> model.tables(barley12.aov,type="effects")
Warning messages:
  Model was refit to allow projection in: model.tables(barley12.aov, type = "effects")

Tables of effects

 variety
 Svanso No. 462  Manch No. 475  Velvet Peatla   Glabron   No. 457  Wisc No. 38   Trebi
 -3.0667  1.9778  -2.8833   -5.2611 -2.5889 -0.4944 0.2500  3.1722      3.5056    5.3889

 site
 Grand Rapids  Duluth University Farm  Morris Crookston  Waseca
  -8.024          -6.784     -1.251             -7.791   6.582      17.269
```