# Analysis of Variance (ANOVA)

MATH 5910

# ANOVA

What is it?

- Linear model (as in regression)
  - Continuous response.
  - Discrete independent variables.
- How different from regression?
  - Presentation (ANOVA table).
  - Interpretation.

# One-Way ANOVA

Word model - similar to simple regression

$$\mathtt{Y} = \mathtt{X}$$

where $\mathtt{Y}$ is the (continuous) response and $\mathtt{X}$ is the independent variable as before BUT is now discrete.

Formally...

# One-Way ANOVA

Two representations.

- Means model:

$$Y_{ij} = \mu_i + e_{ij}$$

  where

$$i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

- Effects model:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

  where

$$i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

  so that

$$\mu_i = \mu + \alpha_i$$

# One-Way ANOVA

- Note $n = \sum_{i=1}^{I} n_i$
- Assume $e_{ij} \sim$ i.i.d. $N(0, \sigma^2)$.

# One-Way ANOVA

Hypotheses.

- Means model:

$$H_0 : \mu_1 = \cdots = \mu_I$$

  versus $H_A$ : at least one $\mu_i$ different.

- Effects model:

$$H_0 : \alpha_1 = \cdots = \alpha_I$$

  versus $H_A$ : at least one $\alpha_i$ different.

Perform F-test for either hypothesis.

# One-Way ANOVA

In either case, we have the ANOVA table (corrected):

| Source | d.f. | SS | MS | F |
|---|---|---|---|---|
| Treatment | $I-1$ | SS$_{Treat}$ | MS$_{Treat}$ | MS$_{Treat}$/MSE |
| Residual | $n-I$ | SSE | MSE | |
| Total | $n-1$ | SST | | |

SS$_{Treat}$: Sum of squares for treatment.
SSE: Sum of squares for error (residual), same as RSS
SST: Sum of squares total.

And the MS is the mean squares (SS divided by d.f.).

# One-Way ANOVA

$$\text{SS}_{Treat} = \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$$

$$\text{SST} = \mathbf{Y}^T\mathbf{Y} - n\overline{Y}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2$$

where

$$\overline{Y}_{\cdot\cdot} = \overline{Y} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}}{n} \quad \text{and} \quad \overline{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

# One-Way ANOVA

- The “Treatment” row is referred as “Between Group” because it looks at variation between levels of a treatment (groups)
- The “Residual” row is referred as “Within Group” because it looks at error (residual) variation; recall that $\hat{\sigma}^2 = \text{MS}_{Resid} = \text{MSE}$

# One-Way ANOVA

- Note that in regression, we had $\text{MS}_{Resid}$ which is the same as MSE.
- In addition, we had $\text{SS}_{Reg}$ instead of $\text{SS}_{Treat}$ in regression.
- It can be seen that

$$\text{SS}_{Treat} + \text{SSE} = \text{SST}$$

# Estimation

- Can compute $\hat{\mu}_i$ or $\hat{\mu}$ and $\hat{\alpha}_i$.
- However, there are different ways to compute them.
  - Set-to-zero, sum-to-zero, etc.
- Estimation not important here.
- Instead, the F-test more important.

# Example 1

First example

```
a <- c(1,1,1,1,2,2,3,3,3,3,3)
y <- c(3,4,5,5,3,2,9,12,5,8,5)
```

Fit a model

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

where

$$i = 1, 2, 3, \quad j = 1, \ldots, n_i$$

$$n_1 = 4,\ n_2 = 2,\ n_3 = 5$$

so that $n = 11$.

# Example 1

We may try `aov()` function, with the following

```
> aov(y~a)
Call:
   aov(formula = y ~ a)

Terms:
                          a Residuals
Sum of Squares     30.39054  58.33673
Deg. of Freedom           1         9

Residual standard error: 2.54595
Estimated effects may be unbalanced
```

See anything(s) odd?

# Example 1

We will need a fix: with `factor()`

```
> aov(y~factor(a))
Call:
   aov(formula = y ~ factor(a))

Terms:
                factor(a) Residuals
Sum of Squares   50.67727  38.05000
Deg. of Freedom         2         8

Residual standard error: 2.180883
Estimated effects may be unbalanced
```

Much better.

# Example 1

Better yet,

```
> summary(aov(y~factor(a)))
            Df Sum Sq Mean Sq F value Pr(>F)
factor(a)    2  50.68  25.339   5.327 0.0338 *
Residuals    8  38.05   4.756
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We now get a familiar ANOVA table.

Note that “Total” row is suppressed.

# Example 1

Can also do

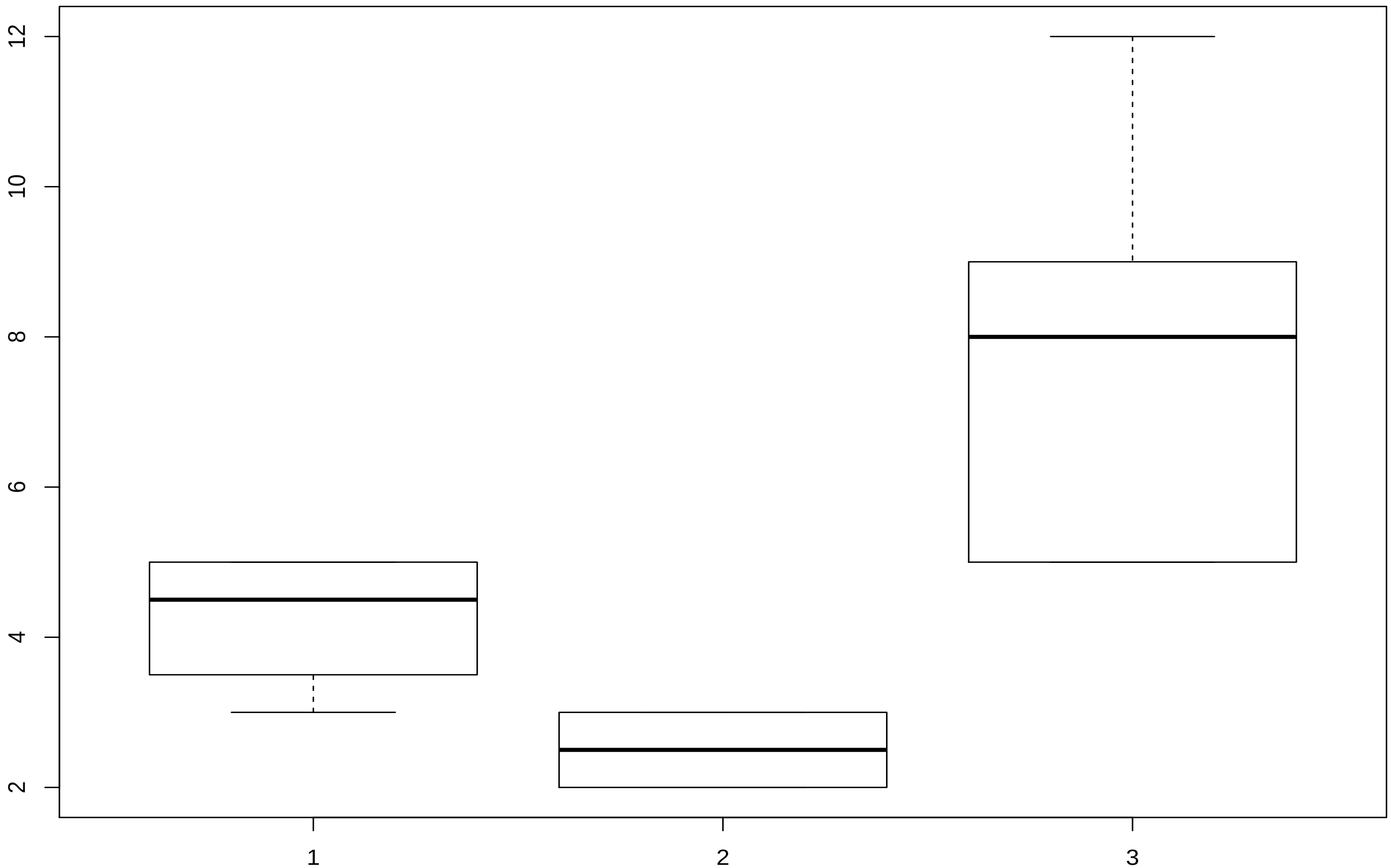
```
> anova(lm(y~factor(a)))

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
factor(a)  2 50.677 25.3386  5.3274 0.03382 *
Residuals  8 38.050  4.7562
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# Example 1

Look at the box plot: `boxplot(y~factor(a))`



# Example 2

Another example.

- From R help file.

```
> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)

> group <- gl(2,10,20, labels=c("Ctl","Trt"))
> group
 [1] Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl Ctl Trt Trt Trt Trt Trt Trt Trt T
Levels: Ctl Trt

> weight <- c(ctl, trt)
> weight
 [1] 4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14 4.81 4.17 4.41 3.5
```

# Example 2

Perform one-way ANOVA with 2 levels (use `anova()` function).

```
> anova(lm.D9 <- lm(weight ~ group))
Analysis of Variance Table

Response: weight
          Df Sum Sq Mean Sq F value Pr(>F)
group      1 0.6882 0.68820  1.4191  0.249
Residuals 18 8.7292 0.48496
```

Note again that “Total” row is suppressed.

# Example 2

What if you do `summary()`?

```
> summary(lm.D9)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.2202  22.850 9.55e-15 ***
groupTrt     -0.3710     0.3114  -1.191    0.249
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared: 0.07308,Adjusted R-squared: 0.02158
F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249
```

Estimates value for `Trt` in group, but not for `Ctl` (why?).

# T-test

- Notice that p-values for both F-test and t-test are the same (0.249).
- Are they related somehow?
- Let's find out. . .

# T-test

Can use original data: `ctl,trt`.

```
> t.test(ctl,trt,var.equal=T)
Two Sample t-test

data:  ctl and trt
t = 1.1913, df = 18, p-value = 0.249
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -0.2833003  1.0253003
sample estimates:
mean of x mean of y
    5.032     4.661
```

# T-test

- Since $t = 1.1913$ (previous page) and $F = 1.491$
- And

```
> 1.1913^2
[1] 1.419196
```

- You see that F is a square of t (subject to round-off error).

# Sum of Squares

For computing the sum of squares "by hand" (NOT done here). Recall

$$\text{SS}_{Treat} = \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2, \quad \text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$$

$$\text{SST} = \mathbf{Y}'\mathbf{Y} - n\overline{Y}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2$$

where

$$\overline{Y}_{\cdot\cdot} = \overline{Y} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}}{n} \quad \text{and} \quad \overline{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

# Sum of Squares

Possible to compute (using the current data)

- $\overline{Y}_{..}$ - This is simply `mean(weight)`
- $\overline{Y}_{i.}$ - Here, we have `tapply(weight,group,mean)`
- $n_i$ - Similarly, this is `tapply(weight,group,length)`

All others quantities are just straight forward applications (although could be tedious).

# Estimates

Recall

```
> summary(lm.D9)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.2202  22.850 9.55e-15 ***
groupTrt     -0.3710     0.3114  -1.191    0.249
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared: 0.07308,Adjusted R-squared: 0.02158
F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249
```

# Estimates

Match the estimate numbers of `summary(lm.D9)`. To start, set up a design matrix

```
> X<-cbind(rep(1,20),rep(c(1,0),each=10),
           rep(c(0,1),each=10))
```

Any Problems?

# Estimates

To fix this, R imposes set-to-zero constraint with first estimate set at 0 (i.e., $\alpha_1 = 0$).

To set this with the design matrix, do the following:

```
> X1<-cbind(rep(1,20),rep(c(0,1),each=10))
```

# Estimates

Then

```
> y<-weight
> beta.hat<-solve(t(X1)%*%X1)%*%t(X1)%*%y
> beta.hat
        [,1]
[1,]  5.032
[2,] -0.371
```

As desired.

# Estimates

Alternatively, use the means model:

```
> summary(lm(weight ~ group-1))

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
groupCtl   5.0320     0.2202   22.85 9.55e-15 ***
groupTrt   4.6610     0.2202   21.16 3.62e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared:  0.9818,Adjusted R-squared:  0.9798
F-statistic: 485.1 on 2 and 18 DF,  p-value: < 2.2e-16
```

# Estimates

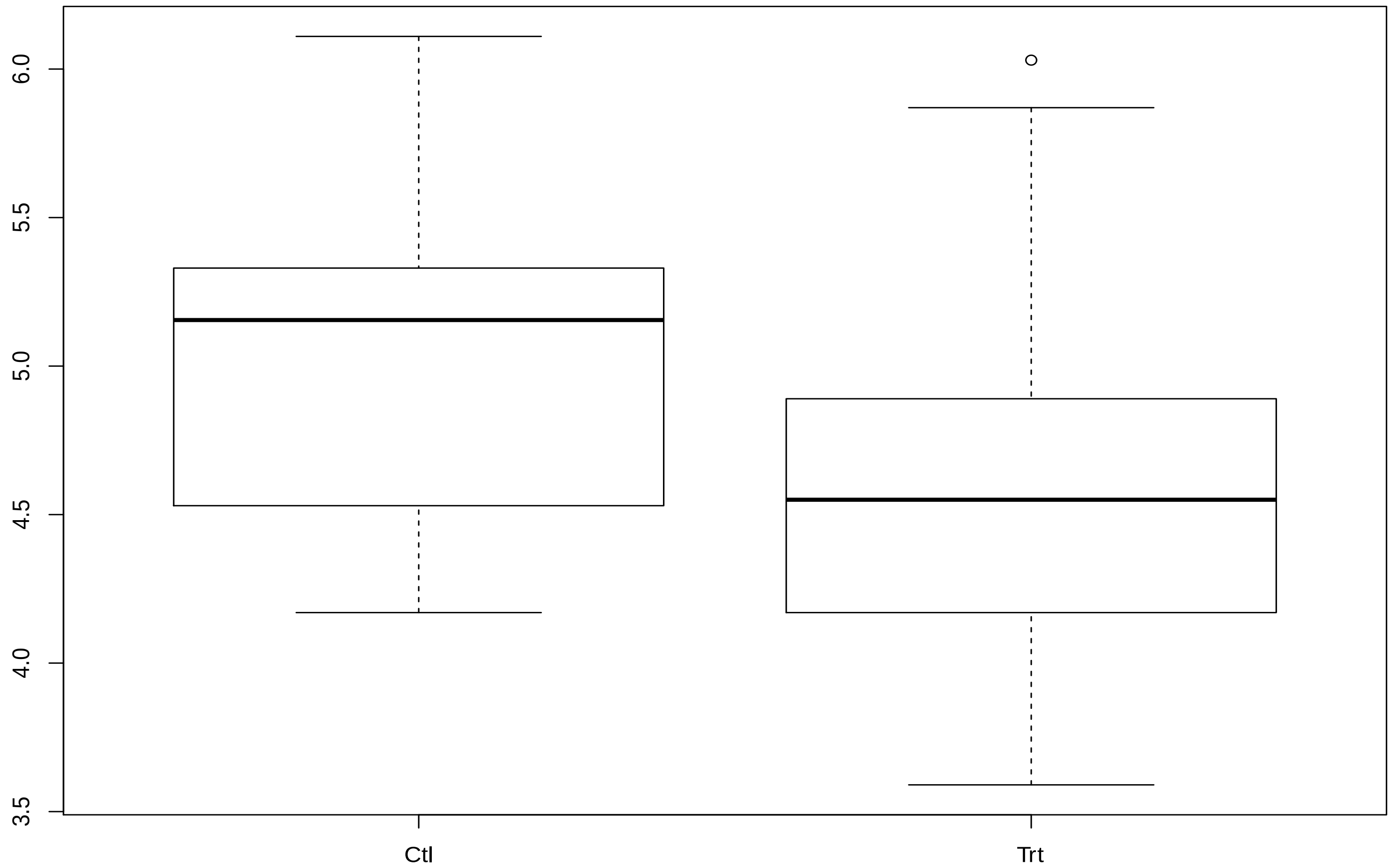
Check:

```
> X2<-cbind(rep(c(1,0),each=10),
            rep(c(0,1),each=10))
> mu.hat<-solve(t(X2)%*%X2)%*%t(X2)%*%y
> mu.hat
      [,1]
[1,] 5.032
[2,] 4.661
```

As expected.

# Box Plot

Let us look at the box plot: `boxplot(weight ~ group)`



# Design Consideration

- Because ANOVA F-test and t-test are related (in one-way, 2-level case).
- ANOVA needs to follow the t-test assumptions.
- From $e_{ij} \sim$ i.i.d. $N(0, \sigma^2)$
  - Data $Y_{ij}$ must be normal, which follows from model.
  - Data must be independent within and between groups, which is required in linear models.
  - Constant variance assumption must be satisfied as well.

# Design Consideration

- In particular, the assignment of treatments to groups must be random.
- In other words, we must have CRD (completely randomized design) for correct analysis of one-way ANOVA.
- More design revelations in higher-way ANOVA. . .

# Two-Way ANOVA

- How to deal with 2 (or more) factors?
- More complications than one-way model?

# Two-Way ANOVA

Additive model (no interaction).

- Means model:

  $$Y_{ijk} = \mu_{ij} + e_{ijk}$$

  where

  $$i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n_{ij}$$

- Effects model:

  Replace

  $$\mu_{ij} = \mu + \alpha_i + \beta_j$$

  above.

# Two-Way ANOVA

So

- Additive model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

  or

```
Y = A + B
```

# Two-Way ANOVA

ANOVA Table

| Source | d.f. | SS | MS | F |
|---|---|---|---|---|
| Treatment A | $I-1$ | SSA | MSA | MSA/MSE |
| Treatment B | $J-1$ | SSB | MSB | MSB/MSE |
| Residual | $n-I-J+1$ | SSE | MSE | |
| Total | $n-1$ | SST | | |

Skip the SS formula. Also, quite messy if unbalanced.

# Two-Way ANOVA

Tests:

- For factor A

$$H_0 : \alpha_1 = \cdots = \alpha_I$$

- For factor B

$$H_0 : \beta_1 = \cdots = \beta_J$$

- Alternatives: at least one level different.

Both are F-tests.

# Example 3

Data.

```
N <- c(0,1,0,1,1,1,0,0,0,1,1,0,1,1,0,0,1,0,1,0,1,1,0,0)
K <- c(1,0,0,1,0,1,1,0,0,1,0,1,0,1,1,0,0,0,1,1,1,0,1,0)
yield <- c(49.5,62.8,46.8,57.0,59.8,58.5,55.5,56.0,62.8,
           55.8,69.5,55.0,62.0,48.8,45.5,44.2,52.0,51.5,
           49.8,48.8,57.2,59.0,53.2,56.0)
```

```
> length(yield)
[1] 24
> table(N,K)
   K
N    0 1
  0 6 6
  1 6 6
```

# Example 3

ANOVA table.

```
> anova(lm(yield~factor(N)+factor(K)))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
factor(N)  1 189.28 189.282  6.7157 0.01703 *
factor(K)  1  95.20  95.202  3.3778 0.08027 .
Residuals 21 591.88  28.185
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# Example 3

Get p-values based on F, manually.

```
> pf(6.7157, 1, 21, lower.tail=F)
[1] 0.01703116

> pf(3.3778, 1, 21, lower.tail=F)
[1] 0.08027043
```

Know how to do this for other distributions

# Interaction

Recall

- Additive model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

or

```
Y = A + B
```

# Interaction

- What is an interaction?
- How to set up the ANOVA model and determine interaction analytically?

# Interaction

Between factors (between A and B, for example).

Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where

$$i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n_{ij}$$

so the $\gamma_{ij}$ is an interaction term.

# Interaction

Alternatively,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

Word model:

```
Y = A + B + AB
```

# Interaction

ANOVA Table

| Source | d.f. | SS | MS | F |
|---|---|---|---|---|
| Treatment A | $I-1$ | SSA | MSA | MSA/MSE |
| Treatment B | $J-1$ | SSB | MSB | MSB/MSE |
| Interaction | $(I-1)(J-1)$ | SSAB | MSAB | MSAB/MSE |
| Residual | $n-IJ$ | SSE | MSE | |
| Total | $n-1$ | SST | | |

# Interaction

Tests:

- For factor A

$$H_0 : \alpha_1 = \cdots = \alpha_I$$

- For factor B

$$H_0 : \beta_1 = \cdots = \beta_J$$

- For interaction

$$H_0 : (\alpha\beta)_{ij} = 0 \ \text{ for all } i, j.$$

- Alternatives: at least one different.

All are F-tests.

# Interaction

Interpretation.

- **Interaction** - When the "effect" of one factor (`A`) on the response is the same at different levels of another factor (`B`), we say that there is no interaction; otherwise, we say that there an interaction between `A` and `B`.
- Easier to understand by "interaction plot."

# Example

Same data as before; recall

```
N <- c(0,1,0,1,1,1,0,0,0,1,1,0,1,1,0,0,1,0,1,0,1,1,0,0)
K <- c(1,0,0,1,0,1,1,0,0,1,0,1,0,1,1,0,0,0,1,1,1,0,1,0)
yield <- c(49.5,62.8,46.8,57.0,59.8,58.5,55.5,56.0,62.8,
           55.8,69.5,55.0,62.0,48.8,45.5,44.2,52.0,51.5,
           49.8,48.8,57.2,59.0,53.2,56.0)
```

# Example

ANOVA table.

```
> anova(lm(yield~factor(N)+factor(K)+factor(N):factor(K)))
Analysis of Variance Table

Response: yield
                    Df Sum Sq Mean Sq F value  Pr(>F)
factor(N)            1 189.28 189.282  6.7752 0.01702 *
factor(K)            1  95.20  95.202  3.4077 0.07975 .
factor(N):factor(K)  1  33.14  33.135  1.1860 0.28908
Residuals           20 558.75  27.937
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# Example

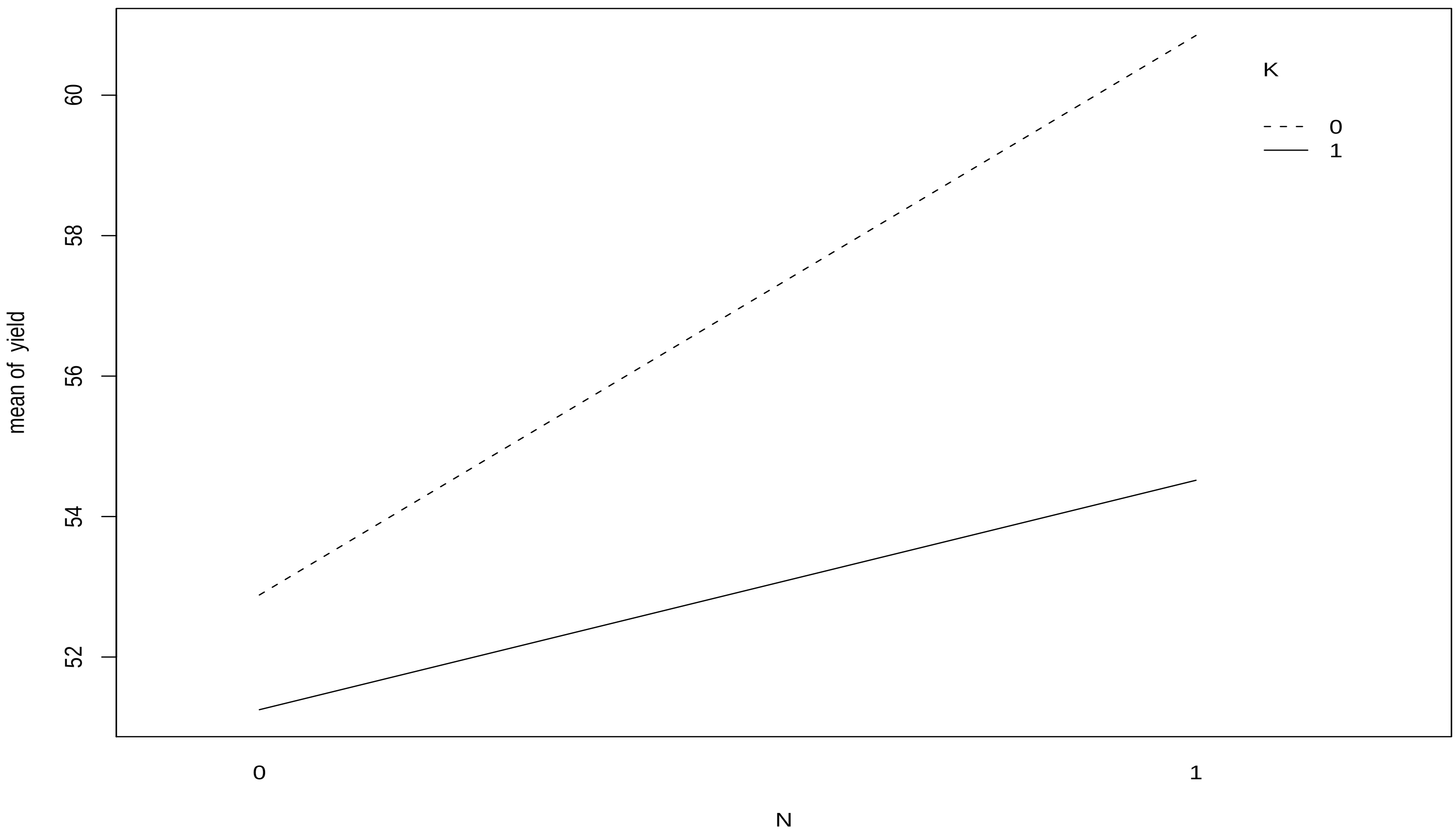
Shortcut.

```
> anova(lm(yield~factor(N)*factor(K)))
Analysis of Variance Table

Response: yield
                    Df Sum Sq Mean Sq F value  Pr(>F)
factor(N)            1 189.28 189.282  6.7752 0.01702 *
factor(K)            1  95.20  95.202  3.4077 0.07975 .
factor(N):factor(K)  1  33.14  33.135  1.1860 0.28908
Residuals           20 558.75  27.937
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

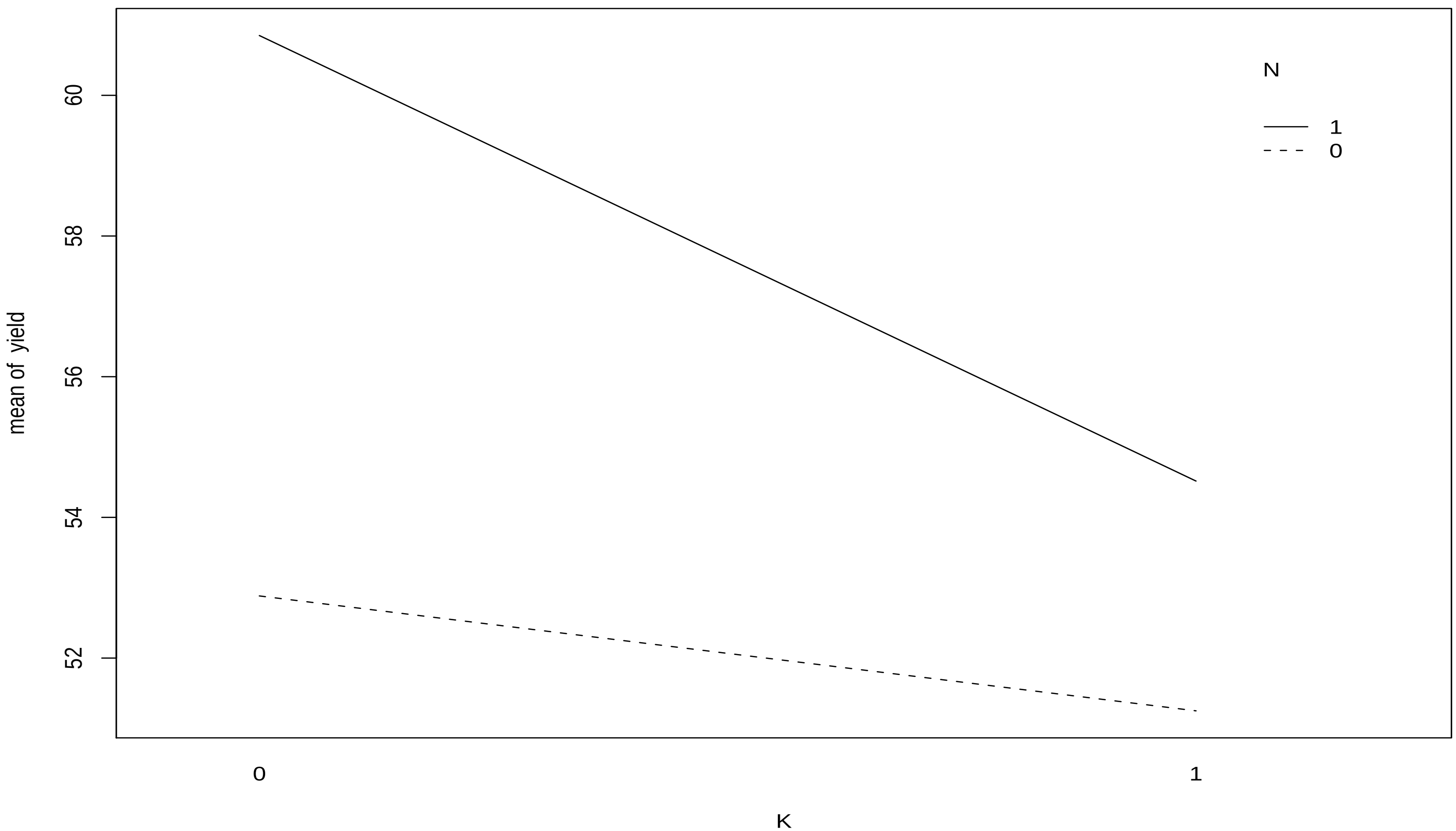
# Interaction Plot

```
interaction.plot(N,K,yield)
```



# Interaction Plot

```
interaction.plot(K,N,yield)
```



# Interaction

For this example, since the interaction term is not significant, our final model will not include the interaction term.

$$\text{Yield} = \text{N}$$

or

$$\text{Yield} = \text{N} + \text{K}$$

Note: If the interaction is significant, then all main effects need to be left in the model.

# Higher-Way ANOVA

ANOVA for more than 2 factors

- Possible
- Much more complicated, especially with interactions.

Example 3 continued:

- Add another factor to previous Example

```
P <- c(1,1,0,0,0,1,0,1,1,1,0,0,0,1,0,1,1,0,0,1,0,1,1,0)
```

Fit:

$$\text{Yield} = \text{N} + \text{P} + \text{K} + \text{Interactions}$$

# Higher-Way ANOVA

## Then

```
> anova(lm(yield~factor(N)*factor(P)*factor(K)))
Analysis of Variance Table

Response: yield
                              Df Sum Sq Mean Sq F value  Pr(>F)
factor(N)                      1 189.28 189.282  6.1608 0.02454 *
factor(P)                      1   8.40   8.402  0.2735 0.60819
factor(K)                      1  95.20  95.202  3.0986 0.09746 .
factor(N):factor(P)            1  21.28  21.282  0.6927 0.41750
factor(N):factor(K)            1  33.14  33.135  1.0785 0.31448
factor(P):factor(K)            1   0.48   0.482  0.0157 0.90192
factor(N):factor(P):factor(K)  1  37.00  37.002  1.2043 0.28870
Residuals                     16 491.58  30.724
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

# Higher-Way ANOVA

- Note that there are 2-way and 3-way interactions here.
- If 3-way interaction significant, then all terms need to be left in the model, significant or not.
- Similarities to polynomial regression?