STA305/1004 - Class 11

February 13, 2017

Today's class

- ANOVA table

Today's class

- ANOVA table
- ANOVA identity

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table
- Geometry of ANOVA

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table
- Geometry of ANOVA
- Two estimates of the population variance

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table
- Geometry of ANOVA
- Two estimates of the population variance
- Mean squares

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table
- Geometry of ANOVA
- Two estimates of the population variance
- Mean squares
- F statistic

Today's class

- ANOVA table
- ANOVA identity
- Degrees of freedom and ANOVA table
- Geometry of ANOVA
- Two estimates of the population variance
- Mean squares
- F statistic
- Assumptions

Comparing more than two treatments

If interest is in designing an experiment to compare more than two treatments then the previous designs will need to modified.

- A clinical trial comparing three drugs A, B, C to reduce duration of intubation for patients on mechanical ventilation.

What are the null and alternative hypotheses in these two scenarios?

Comparing more than two treatments

If interest is in designing an experiment to compare more than two treatments then the previous designs will need to modified.

- A clinical trial comparing three drugs A, B, C to reduce duration of intubation for patients on mechanical ventilation.
- Coagulation time of blood samples for animals receiving four different diets A, B, C, D.

What are the null and alternative hypotheses in these two scenarios?

Blood Coagulation Study

- 24 animals were randomized to four treatments with 6 animals in each group.

Blood Coagulation Study

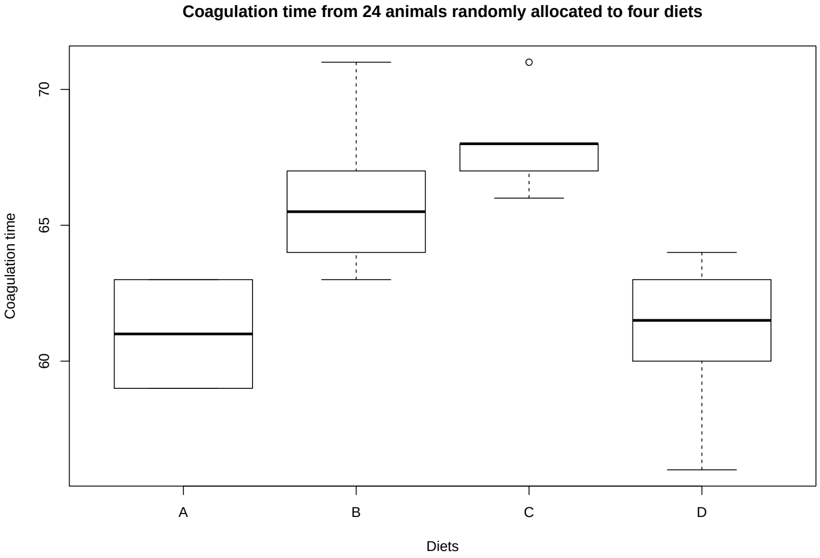
- 24 animals were randomized to four treatments with 6 animals in each group.
- How many possible treatment assignments?

Blood Coagulation Study

- The data for coagulation times for blood samples drawn from 24 animals receiving four different diets A, B, C, and D are shown below.

A	B	C	D	
60	65	71	62	
63	66	66	60	
59	67	68	61	
63	63	68	64	
62	64	67	63	
59	71	68	56	
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

Blood Coagulation Study



Do the boxplots show evidence of a difference between diets?

Analysis of Variance (ANOVA)

- An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.

Analysis of Variance (ANOVA)

- An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.
- Fisher introduced the method in his 1925 book "Statistical Methods for Research Workers".

Analysis of Variance (ANOVA)

- An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.
- Fisher introduced the method in his 1925 book "Statistical Methods for Research Workers".
- The statistical procedure enables experimenters to answer several questions at once.

Analysis of Variance (ANOVA)

- An idea due to Fisher is to compare the variation in mean coagulation times *between* the diets to the variation of coagulation times *within* a diet. These two measures of variation are often summarized in an analysis of variance (ANOVA) table.
- Fisher introduced the method in his 1925 book "Statistical Methods for Research Workers".
- The statistical procedure enables experimenters to answer several questions at once.
- The prevailing method at the time was to test one factor at a time in an experiment.

Analysis of Variance (ANOVA) table

- The between treatments variation and within treatment variation are two components of the total variation in the response.

$$y_{ij} - \bar{y}_{\cdot\cdot} = \underbrace{(y_{i\cdot} - \bar{y}_{\cdot\cdot})}_{\text{treatment deviation}} + \underbrace{(y_{ij} - \bar{y}_{i\cdot})}_{\text{residual deviation}}$$

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

Analysis of Variance (ANOVA) table

- The between treatments variation and within treatment variation are two components of the total variation in the response.
- In the coagulation study data we can break up each observation's deviation from the grand mean into two components: treatment deviations; and residuals within treatment deviations.

$$y_{ij} - \bar{y}_{\cdot\cdot} = \underbrace{(y_{i\cdot} - \bar{y}_{\cdot\cdot})}_{\text{treatment deviation}} + \underbrace{(y_{ij} - \bar{y}_{i\cdot})}_{\text{residual deviation}}$$

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

Analysis of Variance (ANOVA) table

- The between treatments variation and within treatment variation are two components of the total variation in the response.
- In the coagulation study data we can break up each observation's deviation from the grand mean into two components: treatment deviations; and residuals within treatment deviations.
- Let y_{ij} be the *jth* $(j = 1, ..., 6)$ observation taken under treatment $i = 1, 2, 3, 4$.

$$y_{ij} - \bar{y}_{\cdot\cdot} = \underbrace{(y_{i\cdot} - \bar{y}_{\cdot\cdot})}_{\text{treatment deviation}} + \underbrace{(y_{ij} - \bar{y}_{i\cdot})}_{\text{residual deviation}}$$

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

Analysis of Variance (ANOVA) model

- Let y_{ij} be the *jth* observation taken under treatment $i = 1, ..., a$.

$$E(y_{ij}) = \mu_i = \mu + \tau_i,$$

and $Var(y_{ij}) = \sigma^2$ and the observations are mutually independent.

Analysis of Variance (ANOVA) model

- Let y_{ij} be the *jth* observation taken under treatment $i = 1, ..., a$.

$$E(y_{ij}) = \mu_i = \mu + \tau_i,$$

  and $Var(y_{ij}) = \sigma^2$ and the observations are mutually independent.
- The parameter τ_i is the *ith* treatment effect.

Analysis of Variance (ANOVA) model

- Let y_{ij} be the *jth* observation taken under treatment $i = 1, ..., a$.

$$E(y_{ij}) = \mu_i = \mu + \tau_i,$$

  and $Var(y_{ij}) = \sigma^2$ and the observations are mutually independent.
- The parameter τ_i is the *ith* treatment effect.
- The parameter μ is the overall mean.

Analysis of Variance (ANOVA) model

We are interested in testing if the a treatment means are equal.

$$H_0 : \mu_1 = \cdots = \mu_a \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j,\ i \neq j.$$

There will be n observations under the *ith* treatment.

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n,$$

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N,$$

where $N = an$ is the total number of observations. The "dot" subscript notation means sum over the subscript that it replaces.

The ANOVA identity

The total sum of squares $SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$ can be written as

$$\sum_{i=1}^{a} \sum_{j=1}^{n} [(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2$$

by adding and subtracting $\bar{y}_{i.}$ to SS_T.

It can be shown that

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = \underbrace{n \sum_{i=1}^{a} (y_{i.} - \bar{y}_{..})^2}_{\text{Sum of Squares Due to Treatment}} + \underbrace{\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2}_{\text{Sum of Squares Due to Error}}$$

$$= SS_{Treat} + SS_E.$$

The ANOVA identity

This is sometimes called the analysis of variance identity. It shows how the total sum of squares can be split into two sum of squares: one part that is due to differences between treatments; and one part due to differences within treatments.

The ANOVA identity

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

- The decomposition of the first observation $y_{11} = 60$ in diet A is

$$
\begin{aligned}
y_{11} - \bar{y}_{\cdot\cdot} &= (y_{1\cdot} - \bar{y}_{\cdot\cdot}) + (y_{11} - \bar{y}_{1\cdot}) \\
60 - 64 &= (61 - 64) + (60 - 61) \\
-4 &= -3 + -1
\end{aligned}
$$

The ANOVA identity

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

- The decomposition of the first observation $y_{11} = 60$ in diet A is

$$\begin{aligned} y_{11} - \bar{y}_{\cdot\cdot} &= (y_{1\cdot} - \bar{y}_{\cdot\cdot}) + (y_{11} - \bar{y}_{1\cdot}) \\ 60 - 64 &= (61 - 64) + (60 - 61) \\ -4 &= -3 + -1 \end{aligned}$$

- If each observation is decomposed in this manner then there will be three tables of residuals: total residuals; between treatment residuals; and within treatment residuals.

Example - Blood coagulation study (SS_T)

The deviations from the grand average $(y_{ij} - \bar{y}_{\cdot\cdot})$ are in the table below:

A	B	C	D
-4	1	7	-2
-1	2	2	-4
-5	3	4	-3
-1	-1	4	0
-2	0	3	-1
-5	7	4	-8

The total sum of squares is obtained by squaring all the entries in this table and summing: $SS_T = (-4)^2 + (-1)^2 + \cdots + (-8)^2 = 340$.

Example - Blood coagulation study (SS_{Treat})

The between treatment deviations $(y_{i\cdot} - \bar{y}_{\cdot\cdot})$ are in the table below:

A	B	C	D
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3

The sum of squares due to treatment is obtained by squaring all the entries in this table and summing: $SS_{Treat} = (-3)^2 + (2)^2 + \cdots + (-3)^2 = 228$.

Example - Blood coagulation study (SS_E)

The within treatment deviations $(y_{ij} - \bar{y}_{i\cdot})$ are in the table below:

A	B	C	D
-1	-1	3	1
2	0	-2	-1
-2	1	0	0
2	-3	0	3
1	-2	-1	2
-2	5	0	-5

The sum of squares due to error $(y_{ij} - \bar{y}_{i\cdot})$ is obtained by squaring the entries in this table and summing: $SS_E = (-1)^2 + (2)^2 + \cdots + (-5)^2 = 112$.

$$\underbrace{340}_{SS_T} = \underbrace{228}_{SS_{Treat}} + \underbrace{112}_{SS_E}.$$

Which illustrates the ANOVA identity for the blood coagulation study.

ANOVA - degrees of freedom

The deviations

- SS_{Treat} is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).

ANOVA - degrees of freedom

The deviations

- SS_{Treat} is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).
- There are $an = N$ total observations. So SS_T has $N - 1$ degrees of freedom.

ANOVA - degrees of freedom

The deviations

- SS_{Treat} is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).
- There are $an = N$ total observations. So SS_T has $N - 1$ degrees of freedom.
- There are a treatment levels so SS_{Treat} has $a - 1$ degrees of freedom.

ANOVA - degrees of freedom

The deviations

- SS_{Treat} is called the sum of squares due to treatments (i.e., between treatments), and SS_E is called the sum of squares due to error (i.e., within treatments).
- There are $an = N$ total observations. So SS_T has $N - 1$ degrees of freedom.
- There are a treatment levels so SS_{Treat} has $a - 1$ degrees of freedom.
- Within each treatment there are n replicates with $n - 1$ degrees of freedom. There are a treatments. So, there are $a(n - 1) = an - a = N - a$ degrees of freedom for error.

Geometry and the ANOVA Table

A	B	C	D
-4	1	7	-2
-1	2	2	-4
-5	3	4	-3
-1	-1	4	0
-2	0	3	-1
-5	7	4	-8

A	B	C	D
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3
-3	2	4	-3

A	B	C	D
-1	-1	3	1
2	0	-2	-1
-2	1	0	0
2	-3	0	3
1	-2	-1	2
-2	5	0	-5

Geometry and the ANOVA Table

- Let a be the vector of deviations from the grand mean,

$$
\begin{aligned}
a &= (\text{-}4, \text{-}1, \text{-}5, \text{-}1, \text{-}2, \text{-}5, 1, 2, 3, \text{-}1, 0, 7, 7, 2, 4, 4, 3, 4, \text{-}2, \text{-}4, \text{-}3, 0, \text{-}1, \text{-}8),\\
b &= (\text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3),\\
c &= (\text{-}1, 2, \text{-}2, 2, 1, \text{-}2, \text{-}1, 0, 1, \text{-}3, \text{-}2, 5, 3, \text{-}2, 0, 0, \text{-}1, 0, 1, \text{-}1, 0, 3, 2, \text{-}5).
\end{aligned}
$$

Geometry and the ANOVA Table

- Let a be the vector of deviations from the grand mean,
- Let b be the vector of deviations of treatment deviations

$$
\begin{aligned}
a =& (\text{-}4, \text{-}1, \text{-}5, \text{-}1, \text{-}2, \text{-}5, 1, 2, 3, \text{-}1, 0, 7, 7, 2, 4, 4, 3, 4, \text{-}2, \text{-}4, \text{-}3, 0, \text{-}1, \text{-}8),\\
b =& (\text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3, \text{-}3),\\
c =& (\text{-}1, 2, \text{-}2, 2, 1, \text{-}2, \text{-}1, 0, 1, \text{-}3, \text{-}2, 5, 3, \text{-}2, 0, 0, \text{-}1, 0, 1, \text{-}1, 0, 3, 2, \text{-}5).
\end{aligned}
$$

Geometry and the ANOVA Table

- Let a be the vector of deviations from the grand mean,
- Let b be the vector of deviations of treatment deviations
- Let c be the vector of within-treatment deviations.

$$a = (\text{-4, -1, -5, -1, -2, -5, 1, 2, 3, -1, 0, 7, 7, 2, 4, 4, 3, 4, -2, -4, -3, 0, -1, -8}),$$
$$b = (\text{-3, -3, -3, -3, -3, -3, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, -3, -3, -3, -3, -3, -3}),$$
$$c = (\text{-1, 2, -2, 2, 1, -2, -1, 0, 1, -3, -2, 5, 3, -2, 0, 0, -1, 0, 1, -1, 0, 3, 2, -5}).$$

Geometry and the ANOVA Table

- The dot product of b and c, $b \cdot c$, is

```
b*c
```

```
 A  B  C  D
 3 -2 12 -3
-6  0 -8  3
 6  2  0  0
-6 -6  0 -9
-3 -4 -4 -6
 6 10  0 15
```

```
sum(b*c)
```

```
[1] 0
```

Geometry and the ANOVA Table

- The dot product of b and c, $b \cdot c$, is

```
b*c
```

```
  A  B  C  D
  3 -2 12 -3
 -6  0 -8  3
  6  2  0  0
 -6 -6  0 -9
 -3 -4 -4 -6
  6 10  0 15
```

```
sum(b*c)
```

```
[1] 0
```

- Therefore, the vectors b and c are orthogonal.

Geometry and the ANOVA Table

- The dot product of b and c, $b \cdot c$, is

```
b*c
```

```
 A  B  C  D
 3 -2 12 -3
-6  0 -8  3
 6  2  0  0
-6 -6  0 -9
-3 -4 -4 -6
 6 10  0 15
```

```
sum(b*c)
```

```
[1] 0
```

- Therefore, the vectors b and c are orthogonal.
- Thus, the vector a is the hypotenuse of a right triangle with sides b and c.

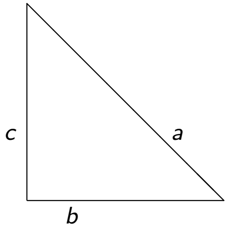
Geometry and the ANOVA Table

Pythagoras' theorem in n dimensions is $|a|^2 = |b|^2 + |c|^2$, where $|a| = \sqrt{a_1^2 + \cdots + a_n^2}$.

The ANOVA identity can be seen using Pythagoras' theorem since

$$|a|^2 = SS_T, |b|^2 = SS_{Treat}, |c|^2 = SS_E.$$

If there were only three observations then the vectors would be as shown below.



The degrees of freedom are the dimensions in which the vectors are free to move given the constraints.

Geometry and the ANOVA Table

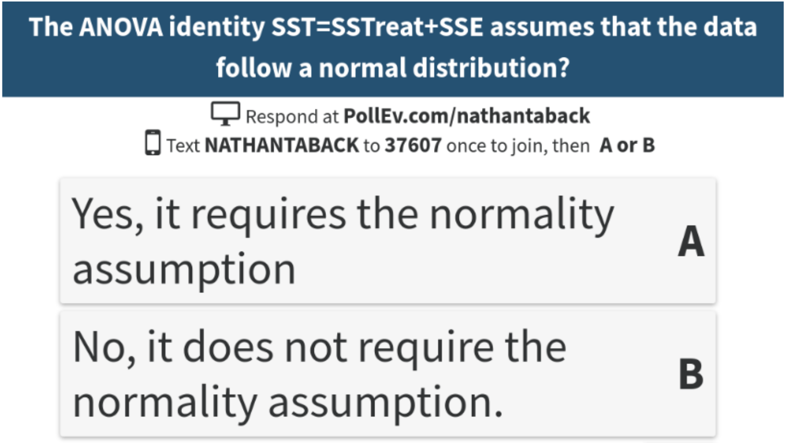


Figure 1:

ANOVAs Two Estimates of the Population Variance (σ^2)

$$SS_E = \sum_{i=1}^{a} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2 \right]$$

If the term inside the brackets is divided by $n-1$ then it is the sample variance for the *ith* treatment

$$S_i^2 = \frac{\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i\cdot})^2}{n-1}, \qquad 1 = 1, ..., a.$$

Combining these a variances to give a single estimate of the common population variance

$$\frac{(n-1)S_1^2 + \cdots + (n-1)S_a^2}{(n-1) + \cdots + (n-1)} = \frac{SS_E}{N-a}.$$

Thus, SS_E is a pooled estimate of the common variance σ^2 within each of the a treatments.

ANOVAs Two Estimates of the Population Variance (σ^2)

If there were no differences between the a treatment means $\bar{y}_{i\cdot}$ we could use the variation of the treatment averages from the grand average to estimate σ^2.

$$\frac{n\sum_{i=1}^{a}\left(y_{i\cdot}-\bar{y}_{\cdot\cdot}\right)^2}{a-1}=\frac{SS_{Treat}}{a-1}$$

is an estimate of σ^2 when the treatment means are all equal.

ANOVAs Two Estimates of the Population Variance (σ^2)

- The analysis of variance identity gives two estimates of σ^2.

ANOVAs Two Estimates of the Population Variance (σ^2)

- The analysis of variance identity gives two estimates of σ^2.
- One is based on the variability within treatments and one based on the variability between treatments.

ANOVAs Two Estimates of the Population Variance (σ^2)

- The analysis of variance identity gives two estimates of σ^2.
- One is based on the variability within treatments and one based on the variability between treatments.
- If there are no differences in the treatment means then these two estimates should be similar.

ANOVAs Two Estimates of the Population Variance (σ^2)

- The analysis of variance identity gives two estimates of σ^2.
- One is based on the variability within treatments and one based on the variability between treatments.
- If there are no differences in the treatment means then these two estimates should be similar.
- If these estimates are different then this could be evidence that the difference is due to differences in the treatment means.

ANOVA - Mean square error

The mean square for treatment is defined as

$$MS_{Treat} = \frac{SS_{Treat}}{a-1}$$

and the mean square for error is defined as

$$MS_E = \frac{SS_E}{N-a}.$$

ANOVA - F statistic

- SS_{Treat} and SS_E are independent.

$$F = \frac{MS_{Treat}}{MS_E} \sim F_{a-1,N-a}.$$

ANOVA - F statistic

- SS_{Treat} and SS_E are independent.
- It can be shown that $SS_{Treat}/\sigma^2 \sim \chi^2_{a-1}$ and $SS_E/\sigma^2 \sim \chi^2_{N-a}$.

$$F = \frac{MS_{Treat}}{MS_E} \sim F_{a-1,N-a}.$$

ANOVA - F statistic

- SS_{Treat} and SS_E are independent.
- It can be shown that $SS_{Treat}/\sigma^2 \sim \chi^2_{a-1}$ and $SS_E/\sigma^2 \sim \chi^2_{N-a}$.
- Thus, if $H_0 : \mu_1 = \cdots = \mu_a$ is true then the ratio

$$F = \frac{MS_{Treat}}{MS_E} \sim F_{a-1,N-a}.$$