

Stat 411/511

MORE ON THE ANOVA

Feb 22 2012

Charlotte Wickham

stat511.cwick.co.nz

Midterm

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

TRUE or FALSE, the probability the true mean lies between 5 and 10 is 0.95. **FALSE**

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

TRUE or FALSE, the probability the true mean lies between 5 and 10 is 0.95. **FALSE**

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

TRUE or FALSE, the probability the true mean lies between 5 and 10 is 0.95. **FALSE**

TRUE or FALSE, the probability (under repeated sampling) that the true mean lies in a 95% CI, is 0.95. **TRUE**

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

TRUE or FALSE, the probability the true mean lies between 5 and 10 is 0.95. **FALSE**

TRUE or FALSE, the probability (under repeated sampling) that the true mean lies in a 95% CI, is 0.95. **TRUE**

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

Final will be about the same level of difficulty but much longer

Midterm

The probability the true population mean falls inside a 95% confidence interval is 0.95.

A researcher finds a 95% CI for the mean of 5 to 10.

TRUE or FALSE, the probability the true mean lies between 5 and 10 is 0.95. **FALSE**

TRUE or FALSE, the probability (under repeated sampling) that the true mean lies in a 95% CI, is 0.95. **TRUE**

Blackboard updated

I will have the midterms in help hours

TAs will have them in lab Friday

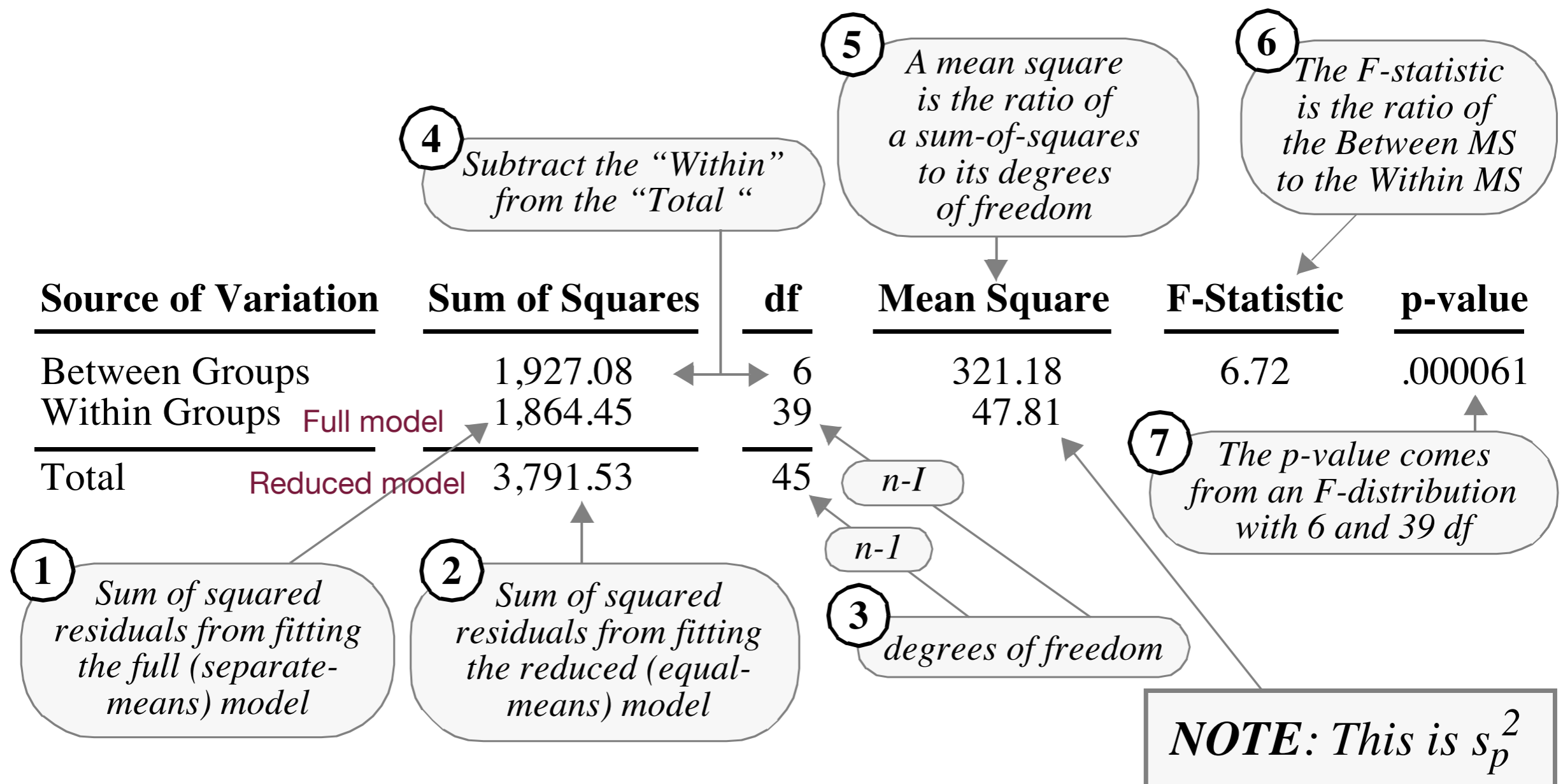
Final will be about the same level of difficulty but much longer

ANOVA for Null #1 from Monday

Display 5.10

p. 127

Analysis of variance table: a test for equal mean percents of women in venires of seven judges; Spock data

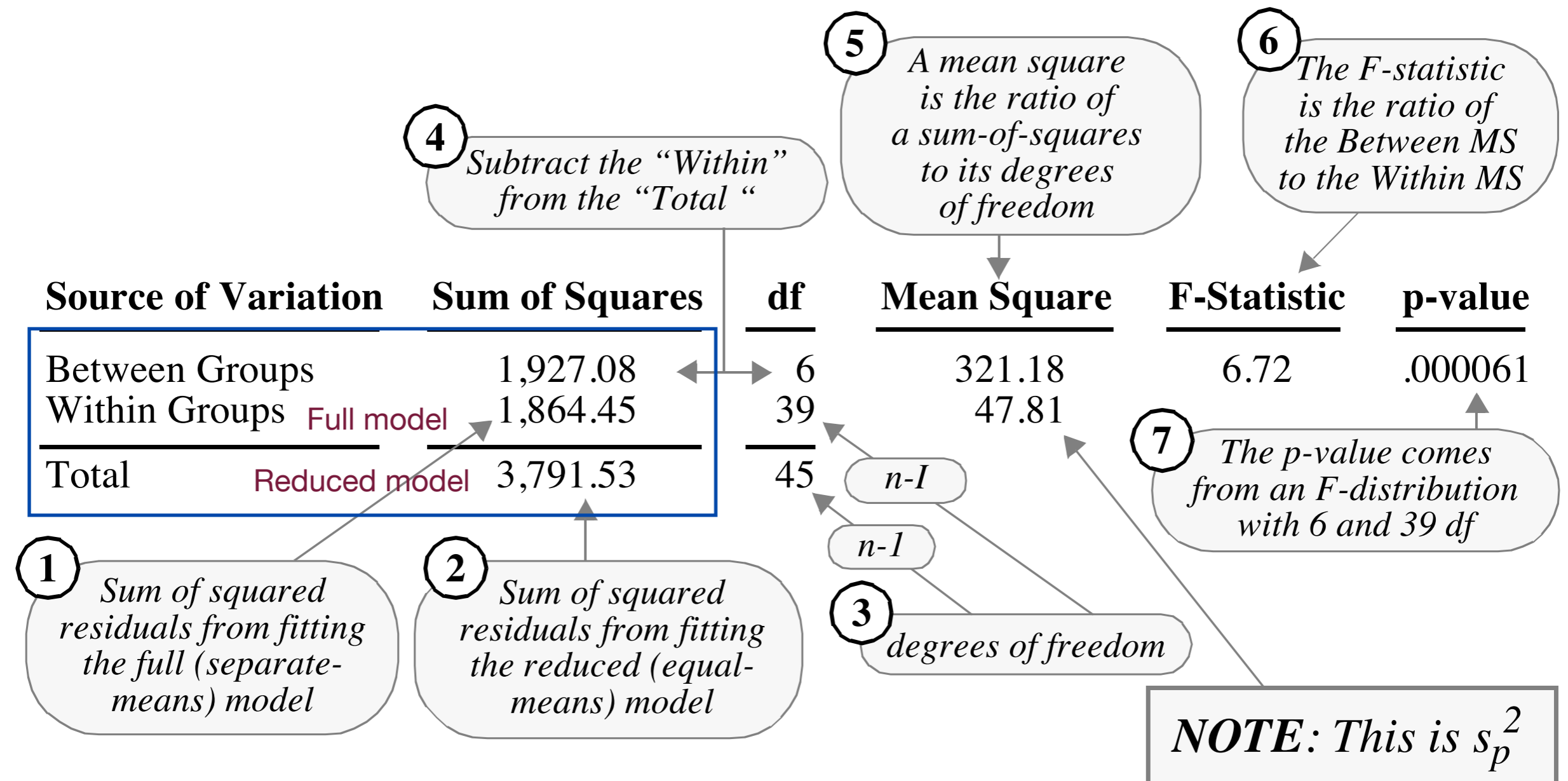


ANOVA for Null #1 from Monday

Display 5.10

p. 127

Analysis of variance table: a test for equal mean percents of women in venires of seven judges; Spock data



Another way to look at the Sums of Squares

Total sum of squares = Between group sum of squares + Within group sum of squares

Observation - overall mean

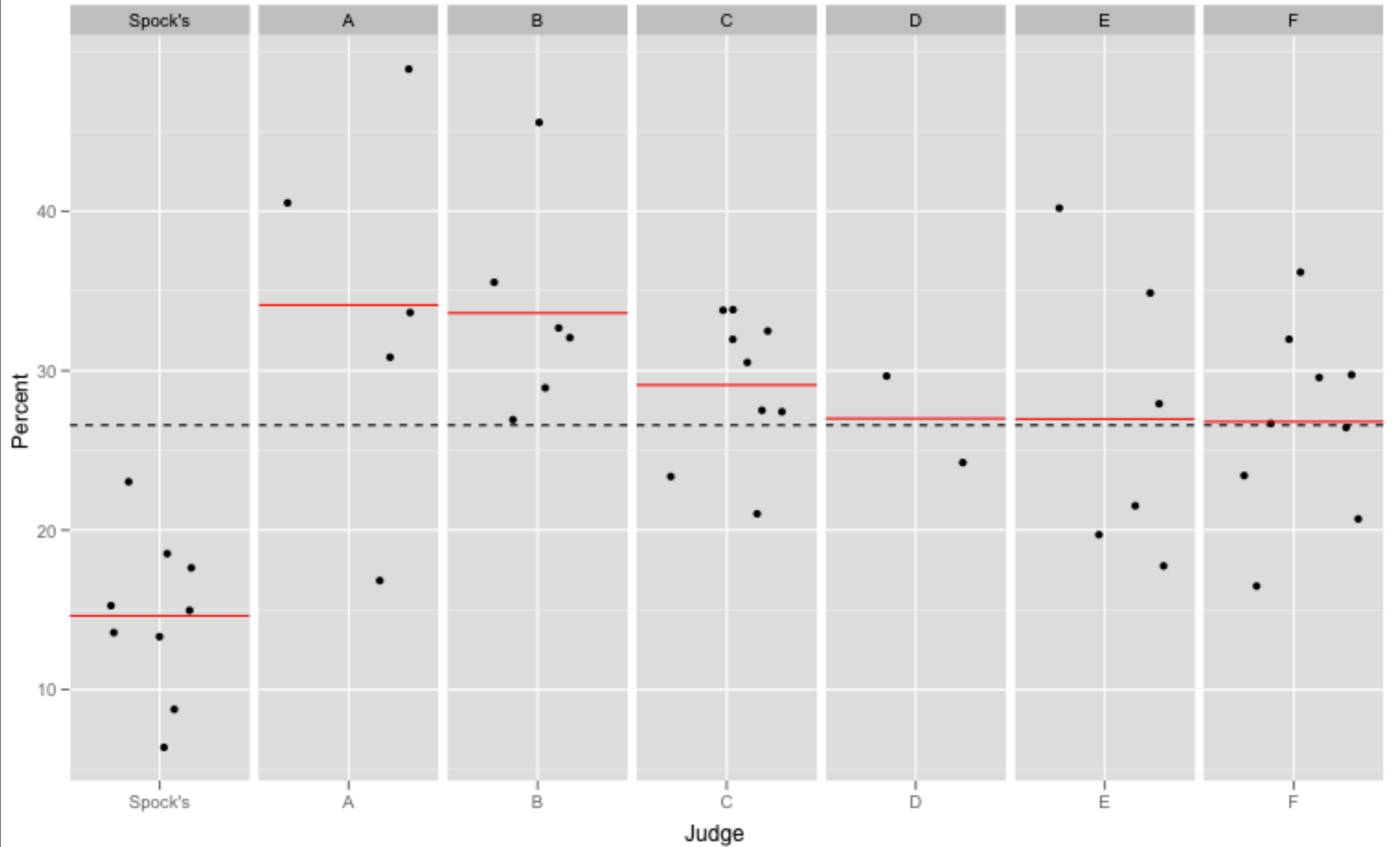
**Restricted
model**

group mean - overall mean

Observation - group mean

**Full
model**

Null #1: sum of squares illustration



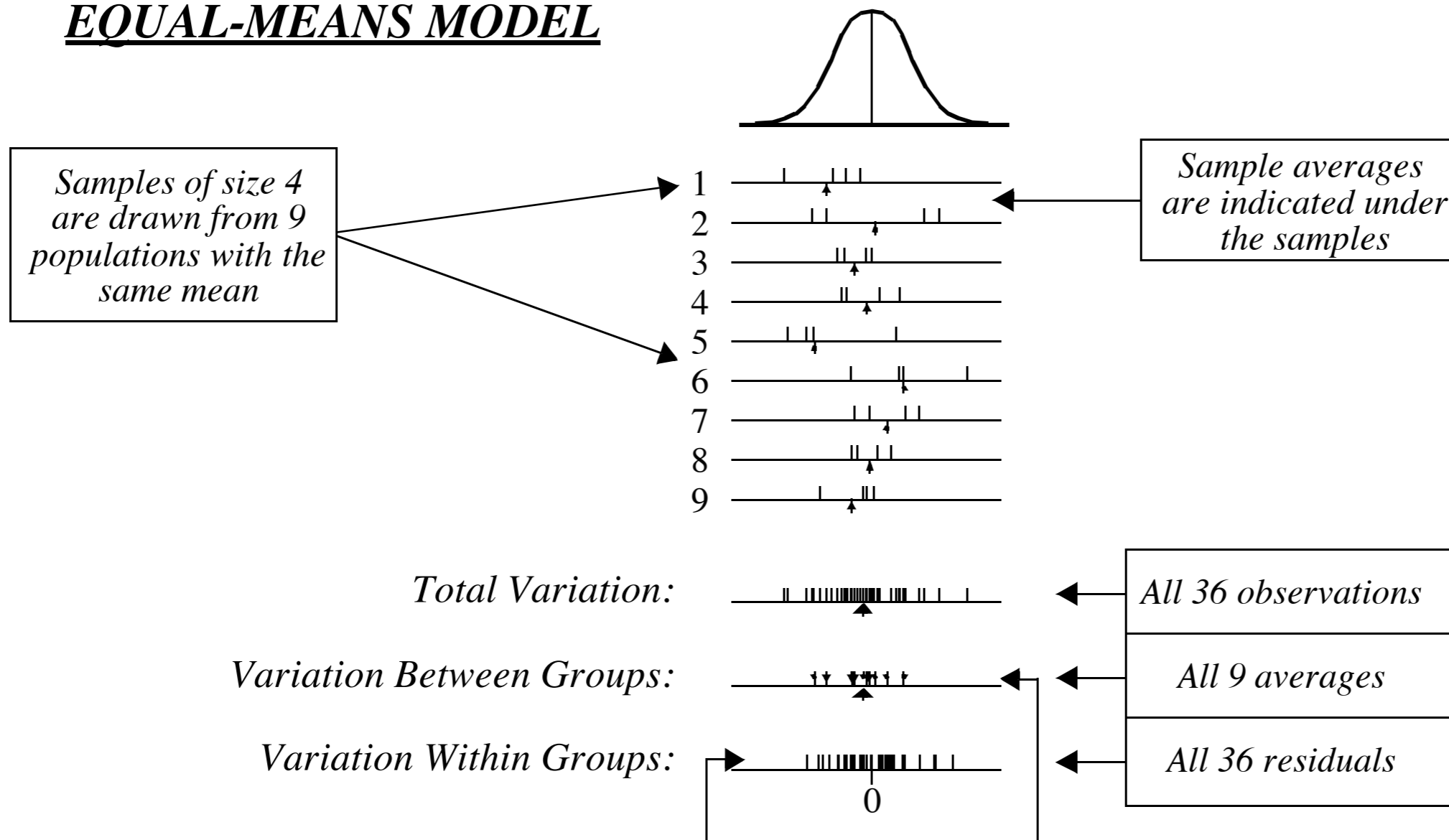
The F-statistic

Compares the between group variation to the within group variation.

If between group variation is large compared to within group variation, we have evidence against the equal means model.

Three sources of variation for data simulated from the equal-means model

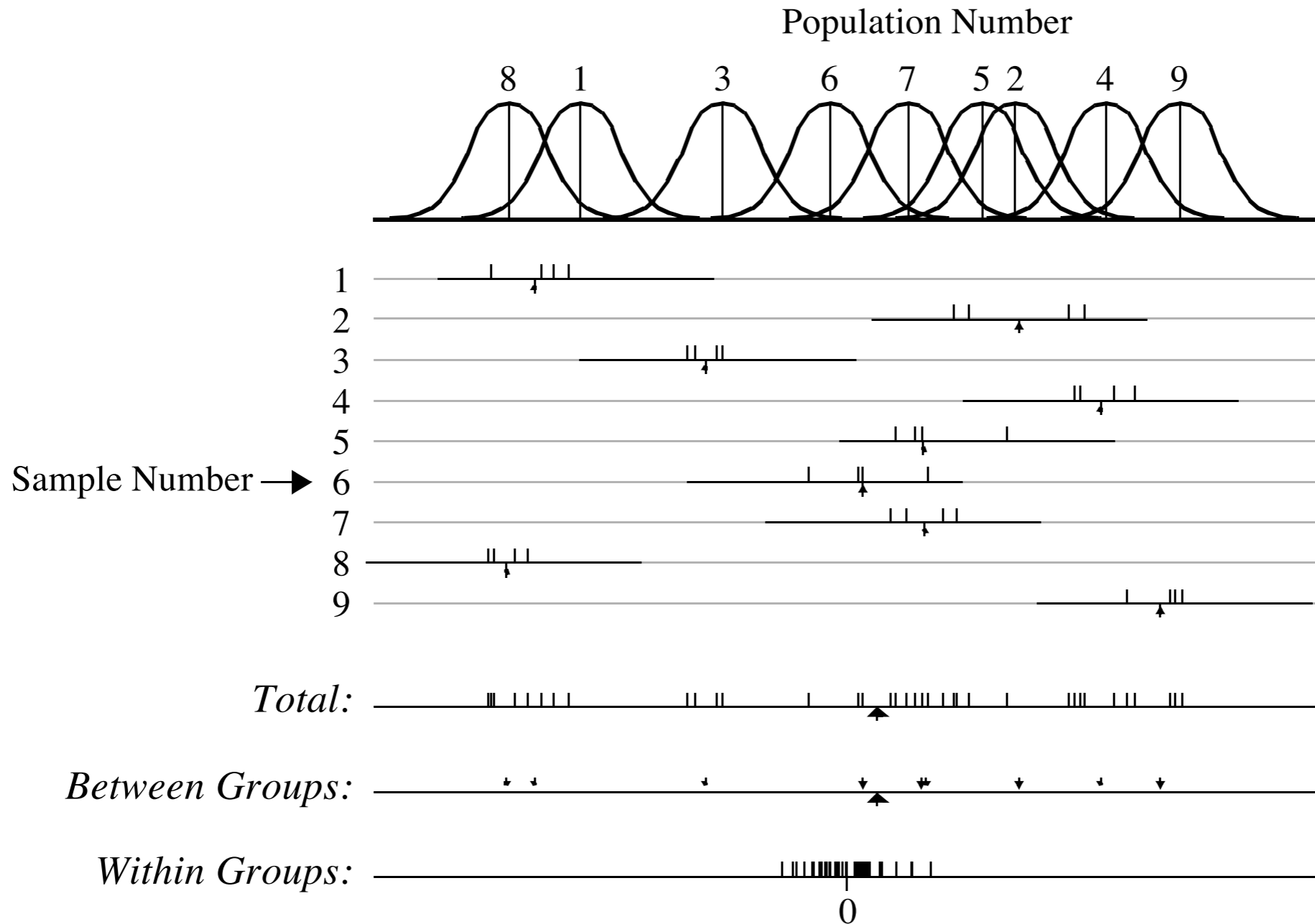
EQUAL-MEANS MODEL



Variation between groups a bit smaller than the variation within groups

Variations in the several group problem for data simulated from the separate-means model

SEPARATE MEANS MODEL



Variation between groups much larger than the variation within groups

Anova in R

```
anova(lm(Percent ~ Judge, data = case0502))
```

```
specifies full model
```

```
Analysis of Variance Table
```

```
Response: Percent
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Judge	6	1927.1	321.18	6.7184	6.096e-05 ***
Residuals	39	1864.5	47.81		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This will test the full model against the equal means model

Anova in R

To test the full model against another model

```
> # full model
> m1 <- lm(Percent ~ Judge, data = case0502)
> # all other judges equal
> m2 <- lm(Percent ~ two_groups, data = case0502)
> # gives the anova table for two means versus full model
> anova(m1, m2)
```

A variable I made



Analysis of Variance Table

Model 1: Percent ~ Judge

Model 2: Percent ~ two_groups

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	39	1864.5				
2	44	2190.9	-5	-326.46	1.3658	0.2582

Assumptions

1. Normally distributed populations.
2. Equal population standard deviations, $\sigma_1 = \sigma_2 = \dots = \sigma_l = \sigma$.
3. Independence of subjects between and within groups.

Same as the two-sample t-test,
but now many samples.

Robustness

1. Normality

Like the t-tools, the ANOVA is robust to the normal population assumption with large sample sizes.

2. Equal population standard deviations.

Unlike the t-tools, the ANOVA is not robust to the equal standard deviation assumption with equal sample sizes.

Display 5.13

p. 131

Success rates for 95% confidence intervals for $\mu_1 - \mu_2$ from samples simulated from normal populations with possibly different SDs

n_1	n_2	n_3	$\sigma_2 = \sigma_1$			$\sigma_2 = 2\sigma_1$		
			$\sigma_3 = \sigma_1$	$\sigma_3 = 2\sigma_1$	$\sigma_3 = 4\sigma_1$	$\sigma_3 = \sigma_1$	$\sigma_3 = 2\sigma_1$	$\sigma_3 = 4\sigma_1$
10	10	10	95.4	98.9	99.9	91.9	96.8	99.6
20	10	10	95.5	98.7	99.8	84.8	91.7	98.9
10	20	10	94.1	98.7	99.9	97.0	98.8	99.8
10	10	20	95.6	99.6	99.9	90.4	97.5	99.9

3. Independence of subjects between and within groups.

As always, this assumption is crucial.
The ANOVA is not robust to this
assumption (ever!).

Using residuals to check assumptions

It's often easier to check the assumptions using the residuals, rather than the observations.

Plot:

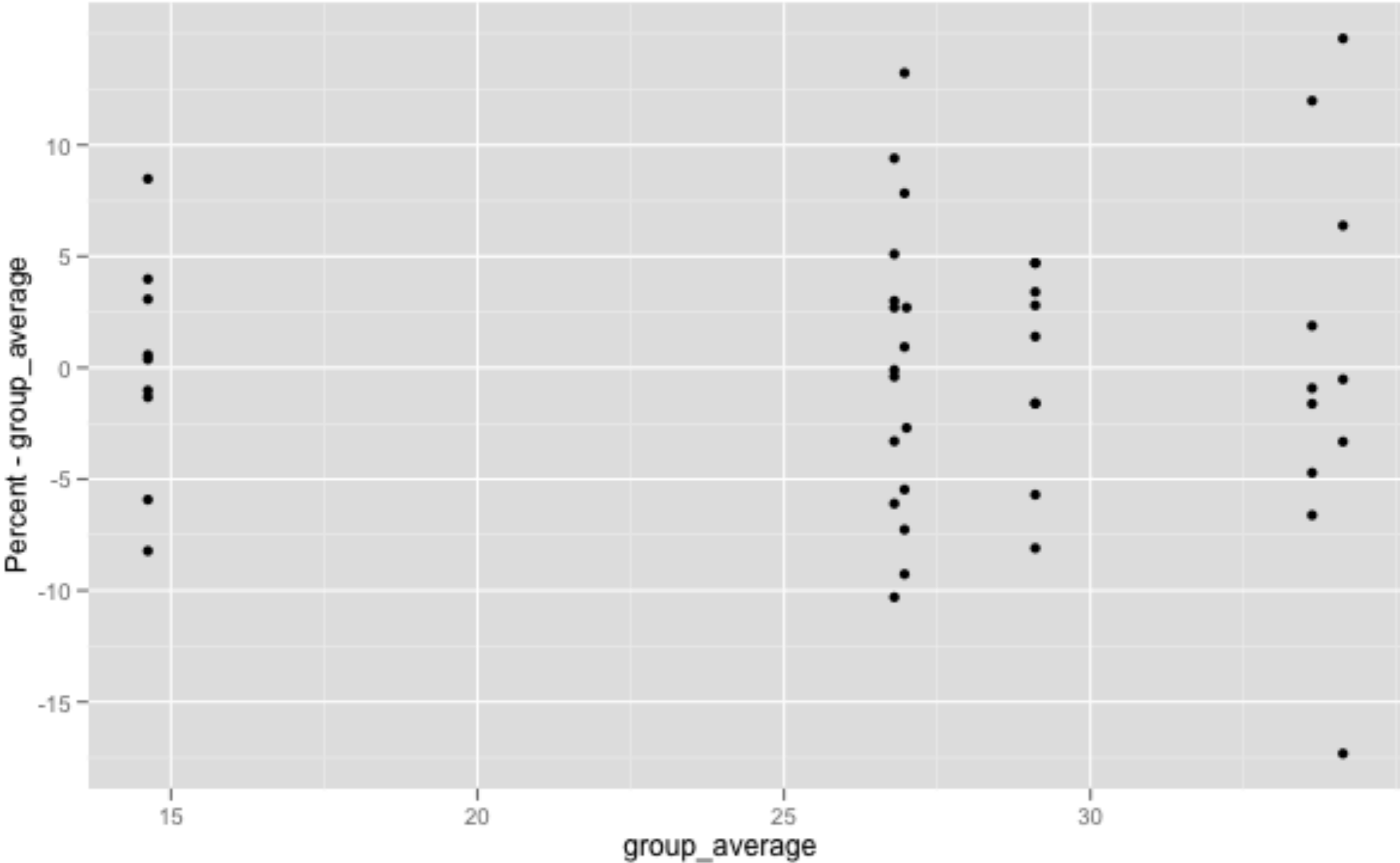
Residuals by group

Residuals against fitted means

Residuals against other variables (like time)

Should be centered around zero (in y direction) and roughly equal spread.

```
qplot( group_average, Percent - group_average, data = case0502)
```



Looks OK

Some important patterns in residual plots

