

Stat 411/511

## COMPARISON OF MANY GROUPS

Feb 14 2012

Charlotte Wickham

[stat511.cwick.co.nz](http://stat511.cwick.co.nz)

# More groups

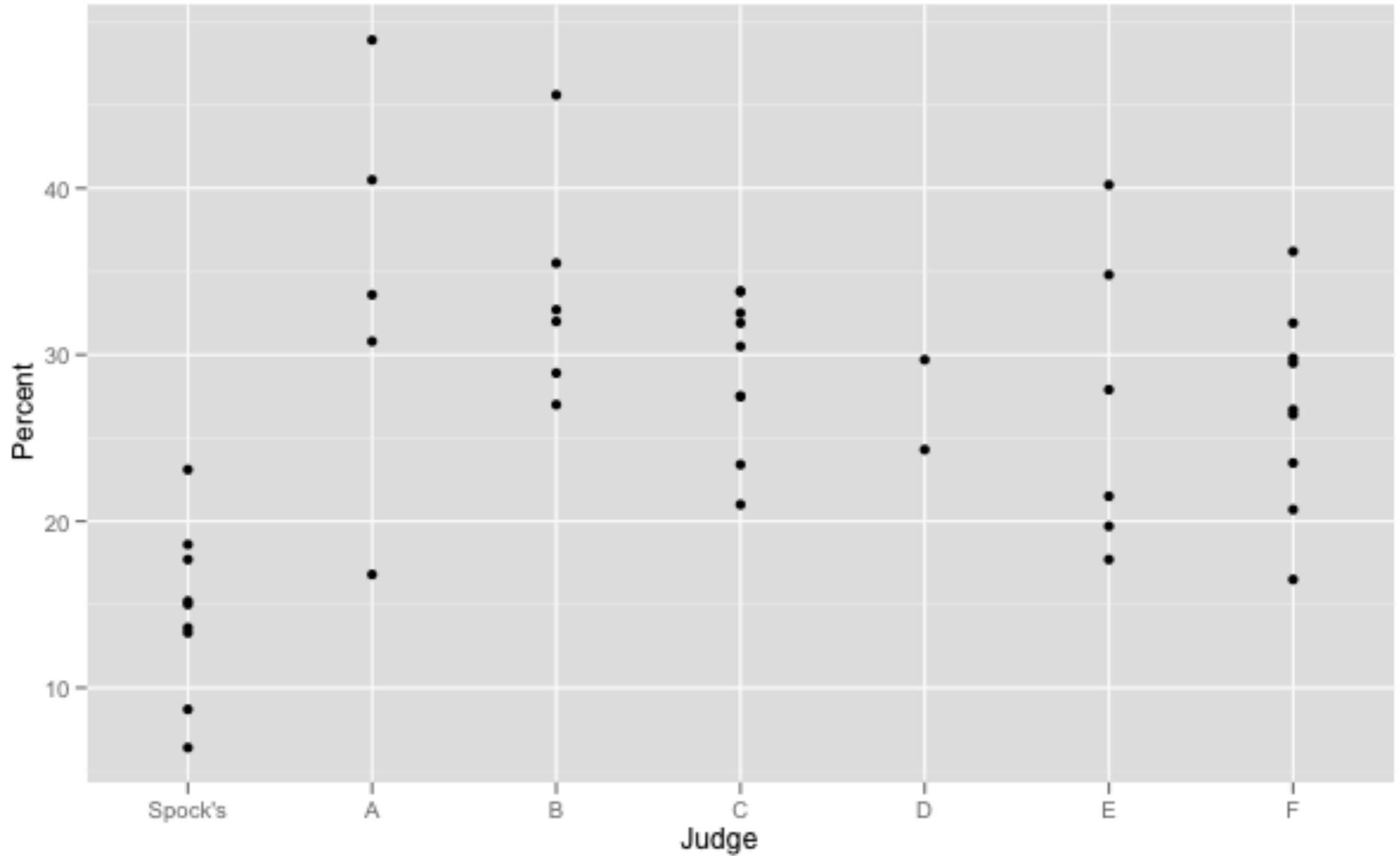
So far we have only looked at **two groups**, now we'll look at **multiple groups**.

In this chapter two big questions:

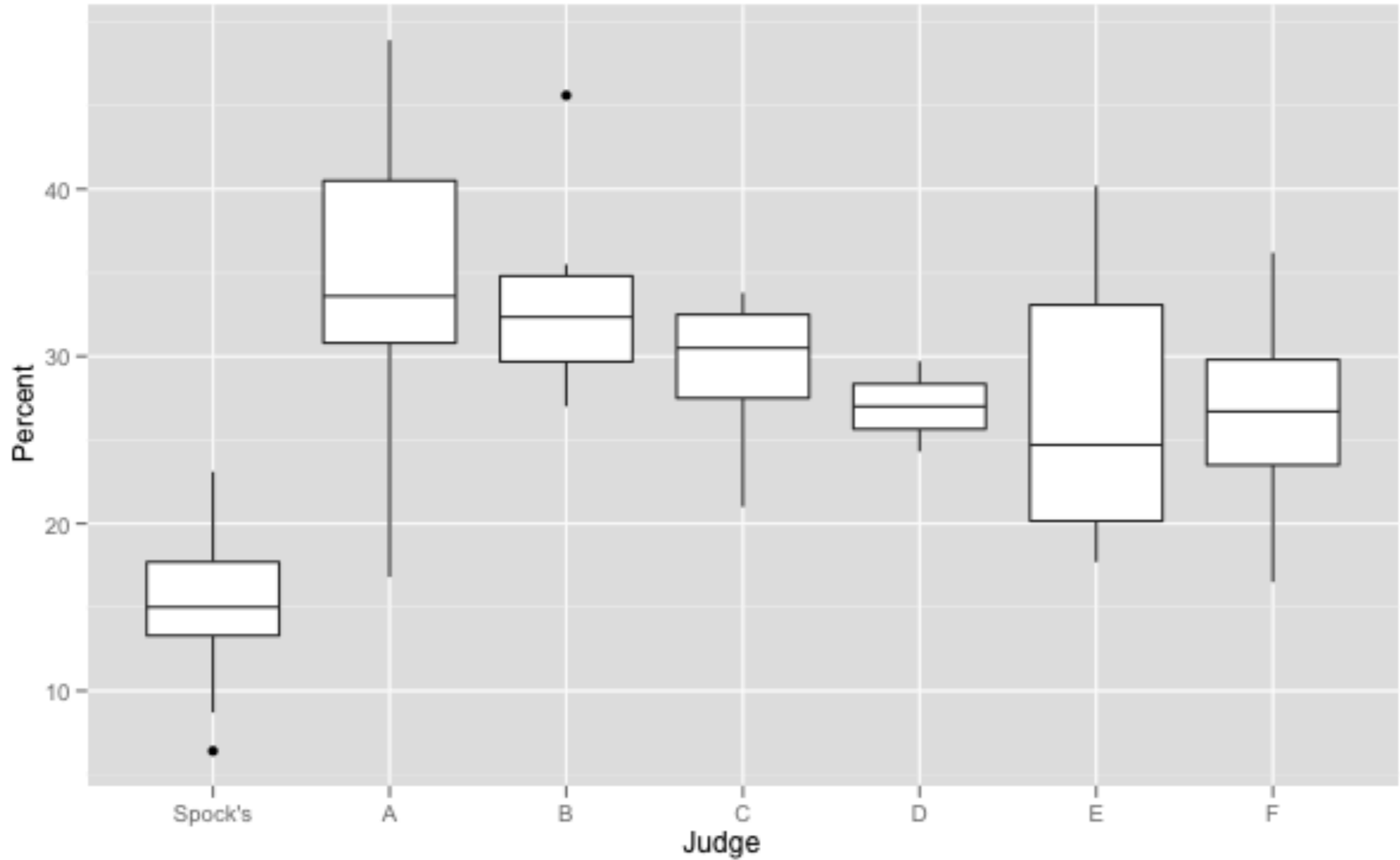
1. How can we compare two groups in particular? back to the *t*-test
2. Are **any** of the group means different?

Extra Sum of Squares F-test (ANOVA)

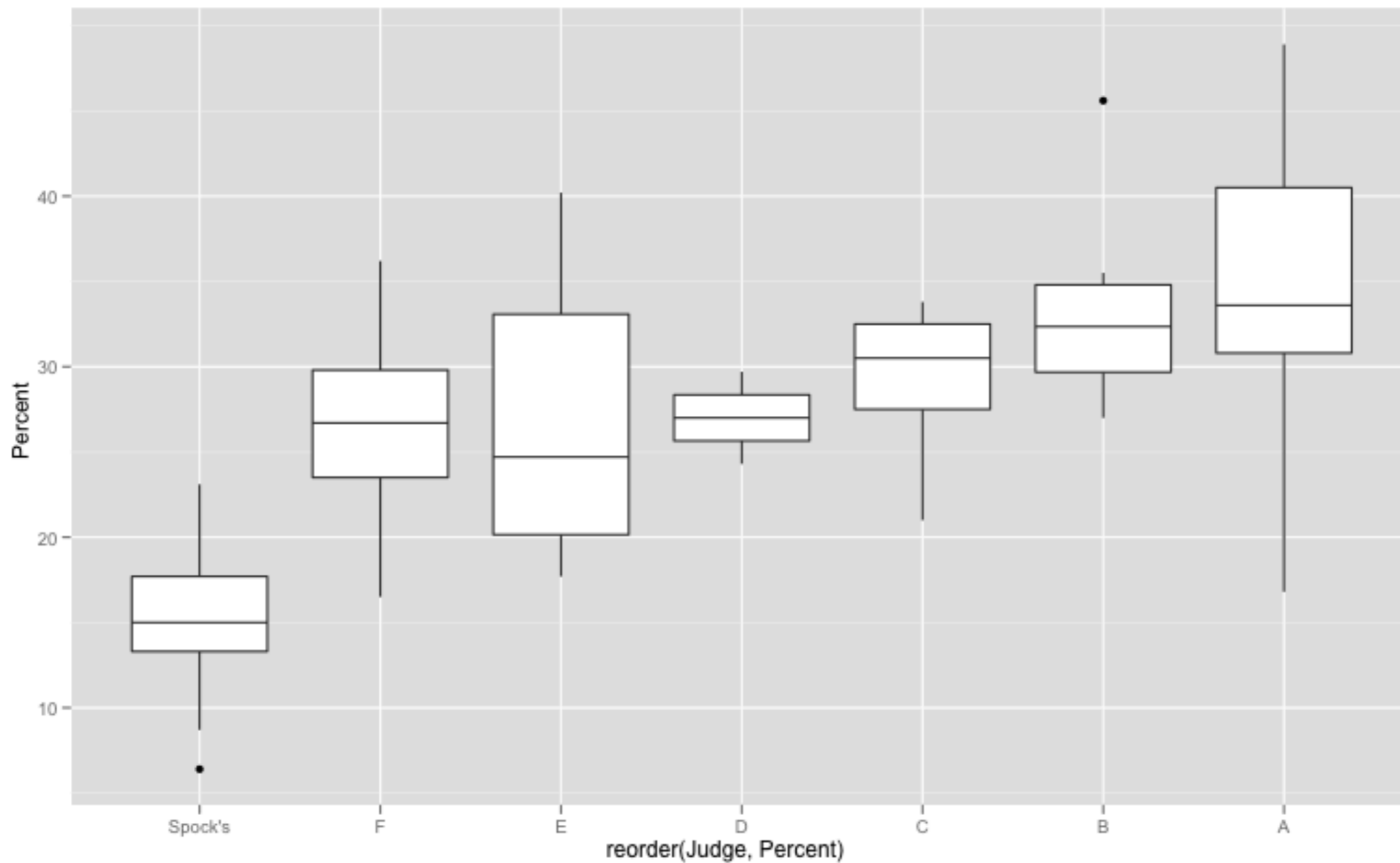
```
library(Sleuth2)
library(ggplot2)
qplot(Judge, Percent, data = case0502)
```



```
qplot(Judge, Percent, data = case0502, geom = "boxplot")
```



```
qplot(reorder(Judge, Percent), Percent, data = case0502, geom = "boxplot")
```



```

qplot(Percent, data = case0502)
  + facet_wrap(~ Judge, ncol = 1)

```

## Some notation

$I$  = number of groups

$\mu_1$  = population mean of population 1.

$\sigma_1$  = population standard deviation of population 1.

$\bar{Y}_1$  = sample average of group 1.

$s_1$  = sample standard deviation of group 1.

$\mu_2$  = population mean of population 2.

$\sigma_2$  = population standard deviation of population 2.

$\bar{Y}_2$  = sample average of group 2.

$s_2$  = sample standard deviation of group 2.

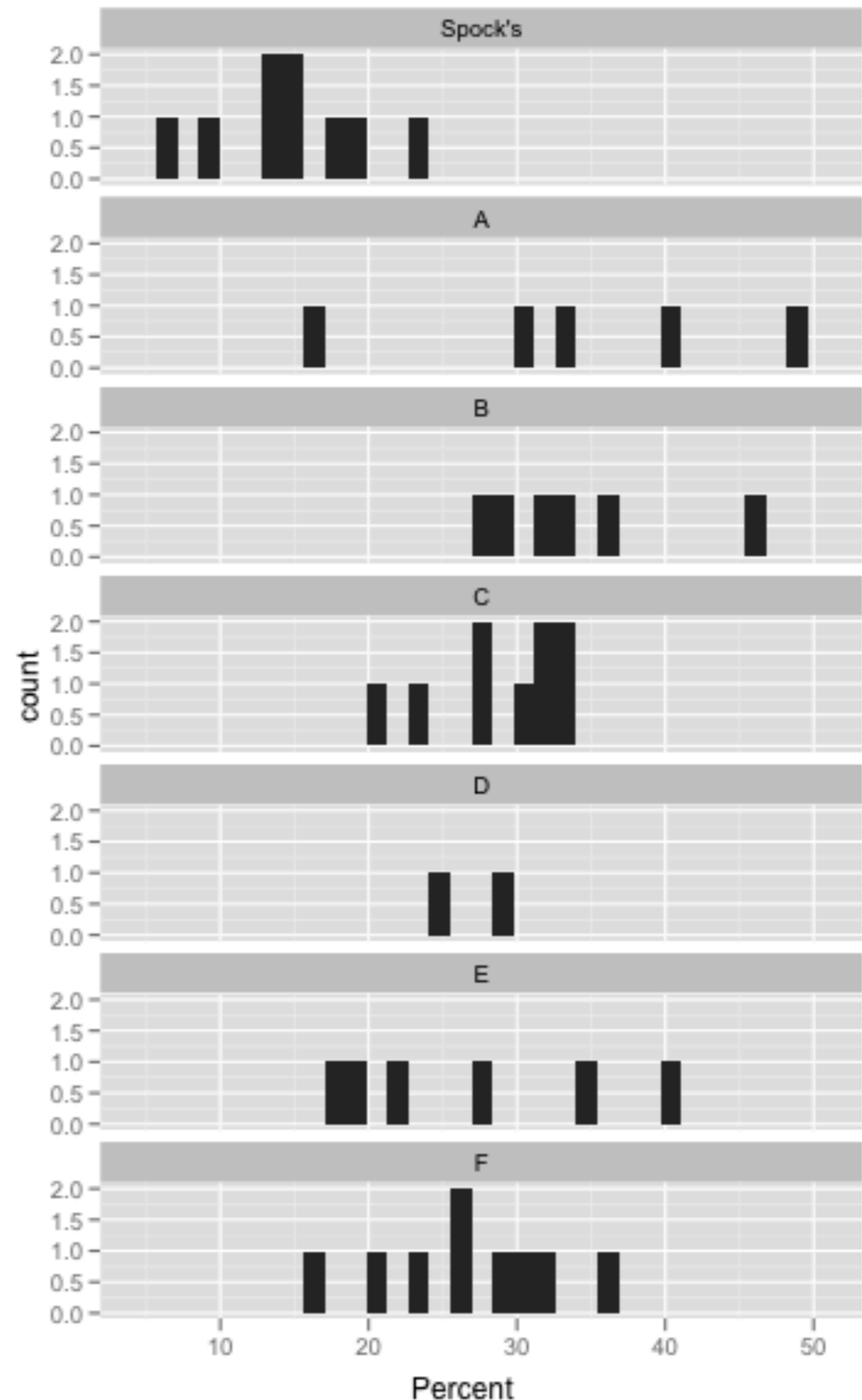
and so on...

$\mu_I$  = population mean of population  $I$ .

$\sigma_I$  = population standard deviation of population  $I$ .

$\bar{Y}_I$  = sample average of group  $I$ .

$s_I$  = sample standard deviation of group  $I$ .



# Assumptions

Normally distributed populations.

Equal population standard deviations,

$$\sigma_1 = \sigma_2 = \dots = \sigma_I = \sigma.$$

Independence of subjects between  
and within groups.

# Summarising many groups

```
> mean(subset(case0502, Judge == "Spock's")$Percent)
[1] 14.62222
> mean(subset(case0502, Judge == "A")$Percent)
[1] 34.12
> mean(subset(case0502, Judge == "B")$Percent)
[1] 33.61667
> mean(subset(case0502, Judge == "C")$Percent)
[1] 29.1
> mean(subset(case0502, Judge == "D")$Percent)
[1] 27
> mean(subset(case0502, Judge == "E")$Percent)
[1] 26.96667
> mean(subset(case0502, Judge == "F")$Percent)
[1] 26.8
```

# Shortcut

```
> with(case0502, tapply(Percent, Judge, mean)) sample averages
Spock's      A      B      C      D      E      F
14.62222 34.12000 33.61667 29.10000 27.00000 26.96667 26.80000
```

```
> with(case0502, tapply(Percent, Judge, sd)) sample sds
Spock's      A      B      C      D      E      F
5.038794 11.941818 6.582223 4.592929 3.818378 9.010142 5.968878
```

```
> with(case0502, tapply(Percent, Judge, length)) sample sizes
Spock's      A      B      C      D      E      F
          9      5      6      9      2      6      9
```

```
averages <- with(case0502, tapply(Percent, Judge, mean))
sds <- with(case0502, tapply(Percent, Judge, sd))
ns <- with(case0502, tapply(Percent, Judge, length))
```

**save for later...**

# Your turn

```
> head(case0501)
  Lifetime Diet
1      35.5  NP
2      35.4  NP
3      34.9  NP
4      34.8  NP
5      33.8  NP
6      33.5  NP
```

How would you get a sample average for each Diet group?

How would you get a sample median for each Diet group?

# Comparing two groups

# Comparing two groups

Did Spock's judge have less women in his venire than judge A?

**Null:**  $\mu_{\text{spock}} = \mu_A$ ,

$\mu$  = mean percent of women.

Two sample  $t$ -statistic:

$$\bar{Y}_{\text{spock}} - \bar{Y}_A / \text{SE}_{\bar{Y}_{\text{spock}} - \bar{Y}_A} \quad \text{same as before}$$

$$\text{SE}_{\bar{Y}_2 - \bar{Y}_1} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{same as before}$$

involves the pooled standard deviation

# Pooled standard deviation

Use **all** the groups (even if you only want to compare two)

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_I - 1)}} \quad \text{different}$$

$$= \sqrt{\frac{\sum_{i=1}^I (n_i - 1)s_i^2}{\sum_{i=1}^I (n_i - 1)}}$$

# pooled standard deviation

`sqrt(sum((ns - 1) * sds^2) / sum(ns - 1))`

degrees of freedom =

different

total number of subject - number of groups =

$n - I$

Did Spock's judge have less women in his venire than judge A?

Two sample *t*-statistic:

$$\bar{Y}_{\text{spock}} - \bar{Y}_A / SE_{\bar{Y}_{\text{spock}} - \bar{Y}_A}$$

```
(means[1] - means[2] ) / (sp * sqrt(1/ns[1] + 1/ns[2]))
```

```
-5.055741
```

Under the **null**: Has a Student's *t*-distribution with *n*-1 degrees of freedom.

```
pt(t.stat, sum(ns) - length(ns))
```

```
5.25123e-06
```

p-value

# Comparing two groups

When you have multiple groups:

Use all groups to get the pooled standard deviation

The degrees of freedom are  $n - I$ .

Extra sum of squares F-test

Are any of the means different?

**Null:** the population means are the same (OR there are no treatment effects).

$$\mu_1 = \mu_2 = \dots = \mu_I = \mu$$

**Alternative:** At least one of the population means is different.

**Null:** the population mean percentage of women are the same for all the judges.

$$\mu_1 = \mu_2 = \dots = \mu_I = \mu$$

**Alternative:** At least one of the judges has a different population mean percentage of women.

# Two models

**Full model:** a model that fully describes the data.

All the means are different

**Restricted model:** a restriction of the full model imposed by the null hypothesis.

All the means are the same.

# Fit both models

The Extra Sum of Squares F-test, compares how well the models fit, by comparing the sum of squared **residuals** of each model.

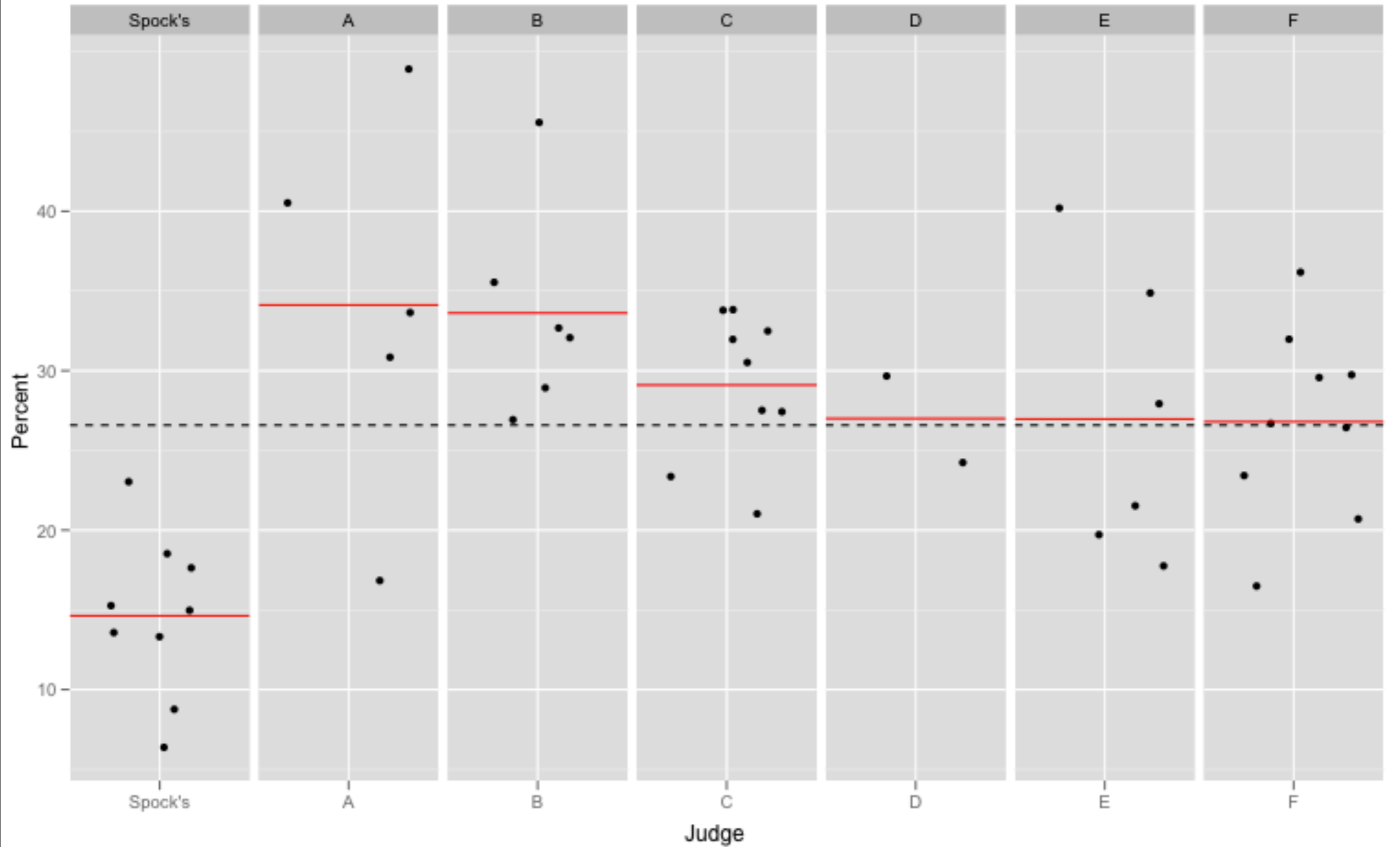
Residual =  $Y_i$  - Estimate of  $\mu_i$

Restricted model residuals:  $Y_i$  - Overall average

Full model residuals:  $Y_i$  - Group average

Residual Sum of Squares: Square each residual and add them up

# sum of squares illustration



```
> case0502$group_average <- with(case0502, ave(Percent, Judge))
> case0502$overall_average <- with(case0502, mean(Percent))
```

```
> head(case0502)
```

	Percent	Judge	group_average	overall_average
1	6.4	Spock's	14.62222	26.58261
2	8.7	Spock's	14.62222	26.58261
3	13.3	Spock's	14.62222	26.58261
4	13.6	Spock's	14.62222	26.58261
5	15.0	Spock's	14.62222	26.58261
6	15.2	Spock's	14.62222	26.58261

```
> with(case0502, sum((Percent - overall_average)^2))
[1] 3791.526
```

One mean

```
> with(case0502, sum((Percent - group_average)^2))
[1] 1864.445
```

I means

Is this a big enough "reduction" to give evidence  
the against the null hypothesis?

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

