Decoding Sums of Squares

David Gerard

2018-12-07

Objectives

- Demonstrate how sums of squares F-test.

Full F-test

Model

- Model: $Y_{ij} = \mu_i + \epsilon_{ij}$
- Y_{ij}: Percent women in venire j of judge i.
- μ_i: Mean percent women for judge i.
- ϵ_{ij}: Individual-specific noice for venire j of judge i. Assumed to have mean 0 and variance σ^2.
- σ^2 is assumed to be the **same** for all venires of all judges.

Hypotheses

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ (all judges have the same mean percent women).
- H_A: As least some $\mu_i \neq \mu_j$ (there are at least two judges with different mean percent women).

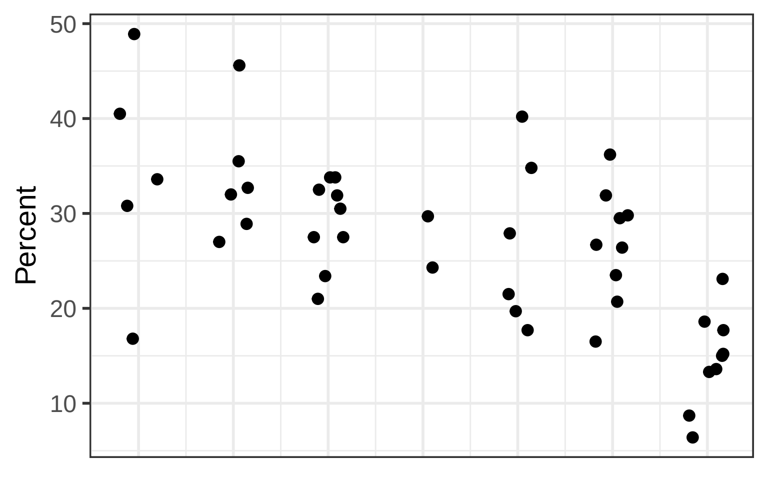
Estimates

We estimate the means within each groups differently according to the full and reduced models

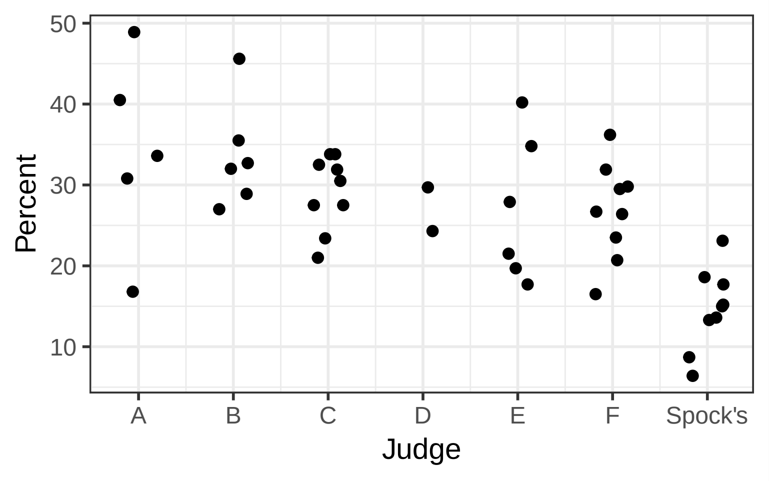
Group	1	2	3	4	5	6	7
Full	$\bar{Y}_{1\bullet}$	$\bar{Y}_{2\bullet}$	$\bar{Y}_{3\bullet}$	$\bar{Y}_{4\bullet}$	$\bar{Y}_{5\bullet}$	$\bar{Y}_{6\bullet}$	$\bar{Y}_{7\bullet}$
Reduced	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$

- $\bar{Y}_{i\bullet} = \frac{1}{n_i}(Y_{i1} + Y_{i2} + \cdots Y_{in_i})$
- $\bar{Y}_{\bullet\bullet}$ = Average of all values in dataset.

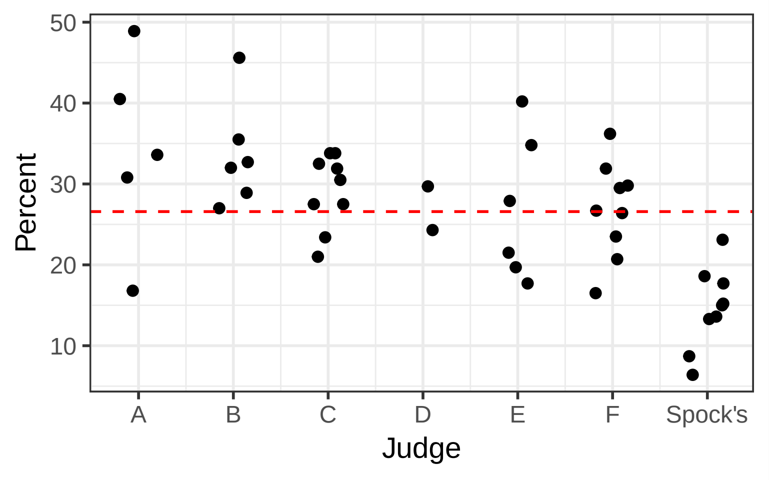
Spock Data: Null Model



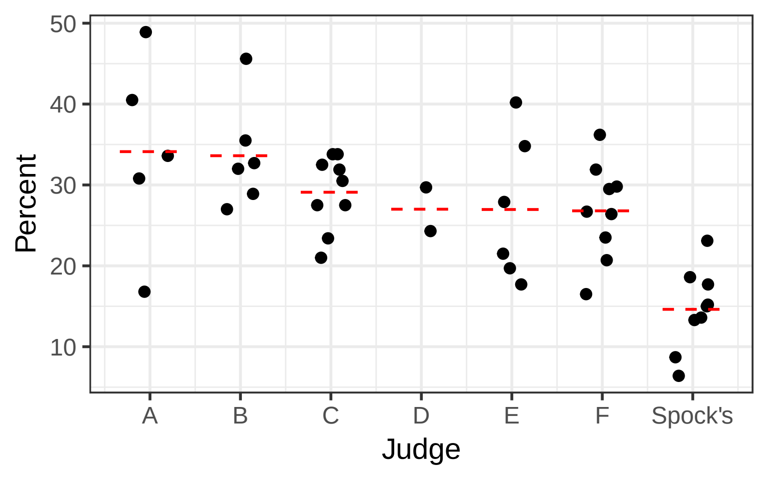
Spock Data: Alternative Model



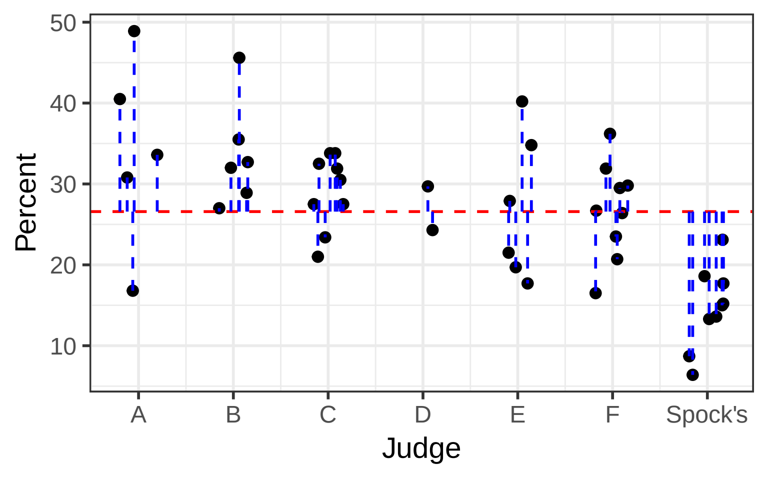
Spock Data: Null Model Estimates



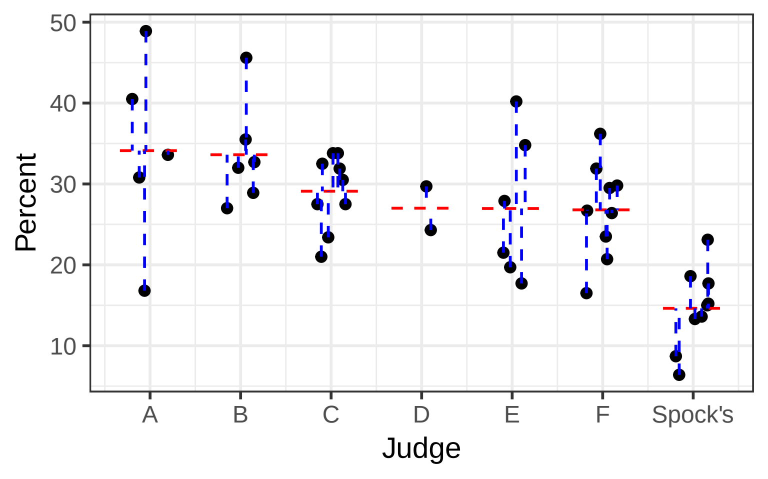
Spock Data: Alternative Model Estimates



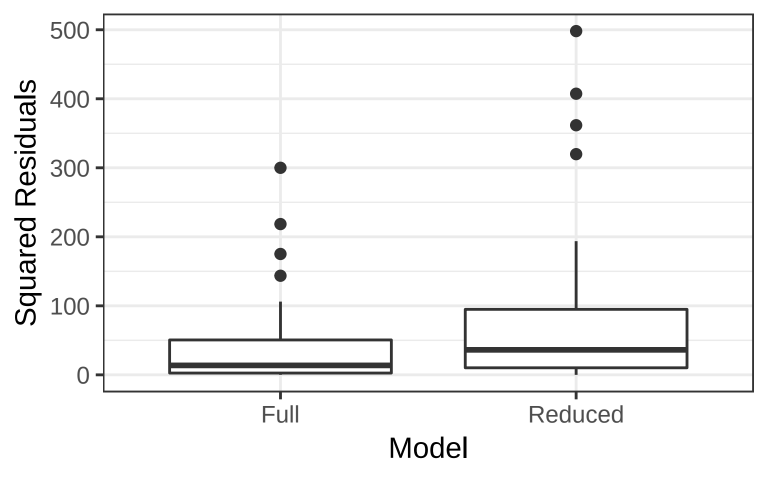
Spock Data: Null Model Residuals



Spock's Data: Alternative Model Residuals



Squared Residuals



Sum of Squared Residuals

- $RSS_{full} = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = 1864.4452$ (Sum of squared residuals in the full model)
- $df_{full} = \#\text{obs - }\#\{\text{parameters in full model}\} = n - 7$ (degrees of freedom in the full model)

Sum of Squared Residuals

- $RSS_{full} = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = 1864.4452$ (Sum of squared residuals in the full model)
- $df_{full} = \#\text{obs - }\#\{\text{parameters in full model}\} = n - 7$ (degrees of freedom in the full model)
- $RSS_{reduced} = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = 3791.5261$ (Sum of the squared residuals in the reduced model)
- $df_{reduced} = \#\text{obs - }\#\{\text{parameters in reduced model}\} = n - 1$ (degrees of freedom in the reduced model)

Sum of Squared Residuals

- $RSS_{full} = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 = 1864.4452$ (Sum of squared residuals in the full model)
- $df_{full} = \#\text{obs - }\#\{\text{parameters in full model}\} = n - 7$ (degrees of freedom in the full model)
- $RSS_{reduced} = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = 3791.5261$ (Sum of the squared residuals in the reduced model)
- $df_{reduced} = \#\text{obs - }\#\{\text{parameters in reduced model}\} = n - 1$ (degrees of freedom in the reduced model)
- Extra sum of squares $= ESS = RSS_{reduced} - RSS_{full}$. (how much larger is the sum of squared residuals in the reduced model compared to that in the full model)
- $df_{extra} = df_{reduced} - df_{full}$

Nota Bene

- $RSS_{reduced}$ will **always** be bigger than RSS_{full}
- So ESS will **always** be positive.

Nota Bene

- $RSS_{reduced}$ will **always** be bigger than RSS_{full}
- So ESS will **always** be positive.
- But if the Null (reduced) model is true, then $RSS_{reduced}$ won't be **a lot** bigger than RSS_{full}.
- So if the Null (reduced) model is true, ESS won't be "very" far from 0.

Nota Bene

- $RSS_{reduced}$ will **always** be bigger than RSS_{full}
- So ESS will **always** be positive.
- But if the Null (reduced) model is true, then $RSS_{reduced}$ won't be **a lot** bigger than RSS_{full}.
- So if the Null (reduced) model is true, ESS won't be "very" far from 0.
- We can quantify what "a lot" and "very" mean using statistical theory.

Null Distribution

- If the Null (reduced) model is correct, then the following F-statistic follows an F distribution.

$$F - \text{statistic} = \frac{ESS/df_{extra}}{s_p^2} = \frac{ESS/df_{extra}}{RSS_{full}/df_{full}}$$

Null Distribution

- If the Null (reduced) model is correct, then the following F-statistic follows an F distribution.

$$F - \text{statistic} = \frac{ESS/df_{extra}}{s_p^2} = \frac{ESS/df_{extra}}{RSS_{full}/df_{full}}$$

- df_{extra} is the number of added parameters and is called the "extra degrees of freedom". It is the number of parameters in the mean for the full model minus the number of parameters in the mean for the reduced model.

Null Distribution

- If the Null (reduced) model is correct, then the following F-statistic follows an F distribution.

$$F - \text{statistic} = \frac{ESS/df_{extra}}{s_p^2} = \frac{ESS/df_{extra}}{RSS_{full}/df_{full}}$$

- df_{extra} is the number of added parameters and is called the "extra degrees of freedom". It is the number of parameters in the mean for the full model minus the number of parameters in the mean for the reduced model.
- s_p^2 is the pooled estimate of the variance. It is equal to RSS_{full}/df_{full}.

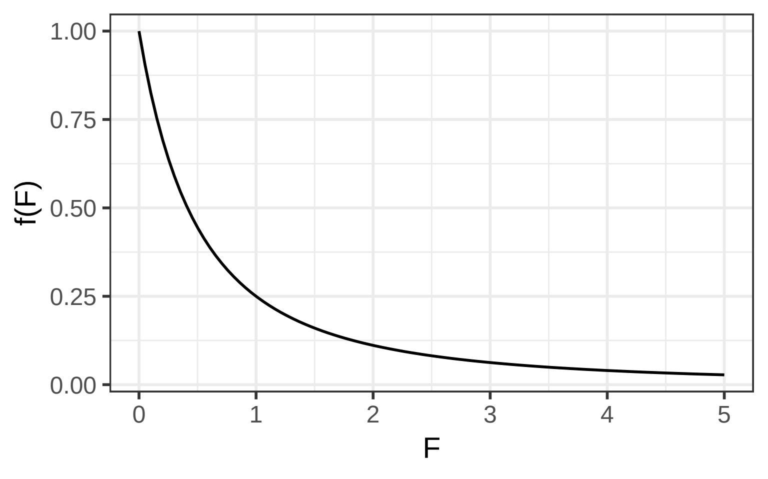
Spock's F-statistic

- $s_p^2 = 47.81$.
- F-stat $=$ ((3791.5261 - 1864.4452) / 6) / 47.81 $=$ 6.7178

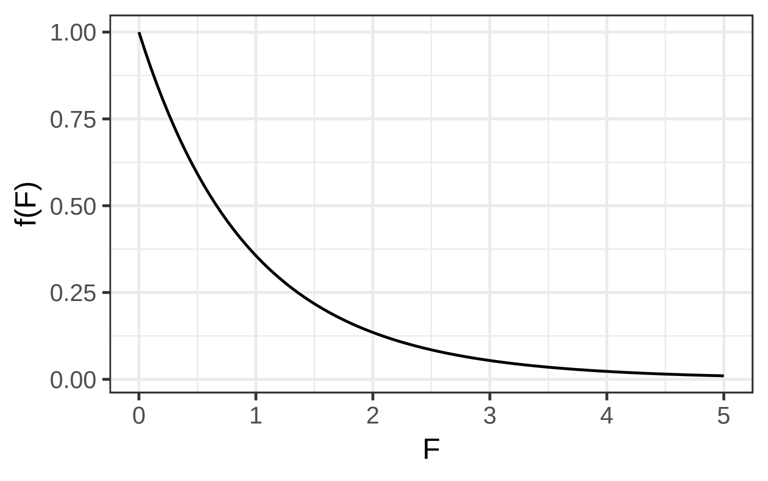
The F-distribution

- Parameterized by two parameters, the *numerator degrees of freedom* (the extra degrees of freedom) and the *denominator degrees of freedom* (degrees of freedom in the full model).
- Interact with `df()`, `pf()`, `qf()`, `rf()` in R.
- Only need upper tail probabilities for p-values (because only large values are extreme).

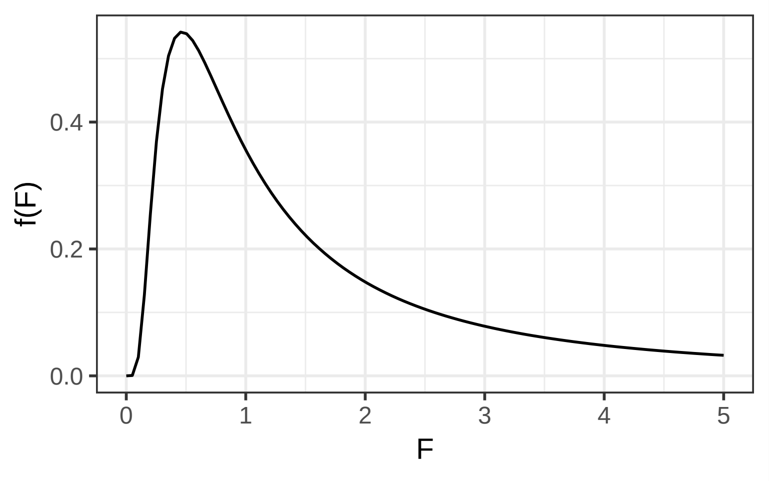
$F_{2,2}$ distribution



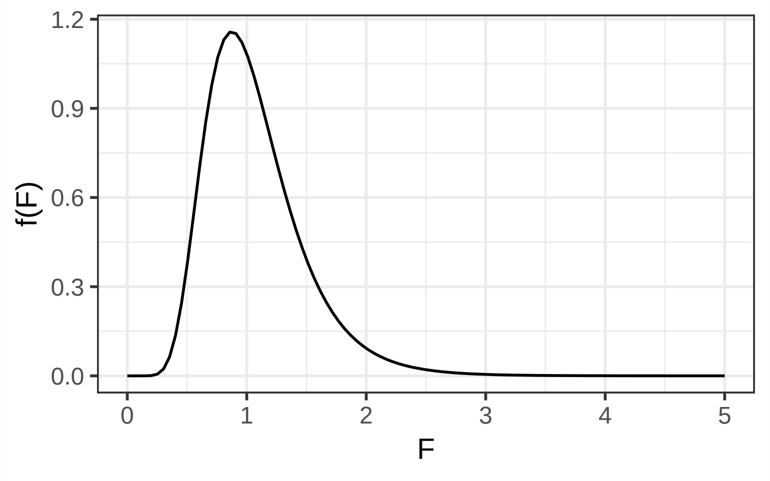
$F_{2,30}$ distribution



$F_{30,2}$ distribution



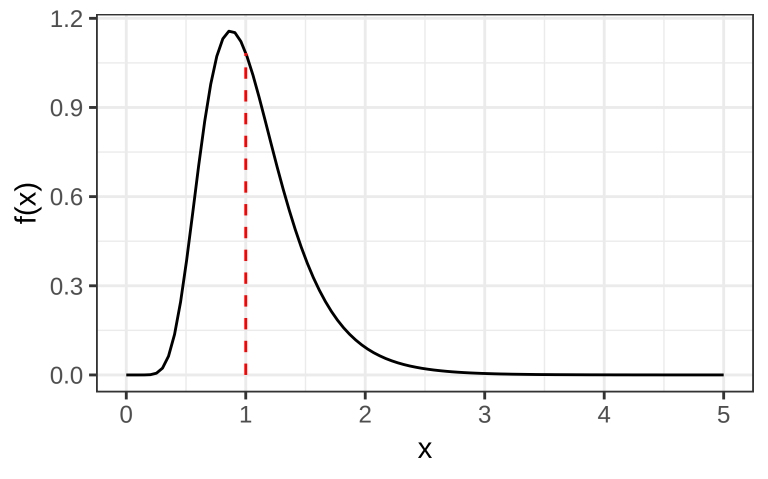
$F_{30,30}$ distribution



Density Function

```r
df(x = 1, df1 = 30, df2 = 30)
```

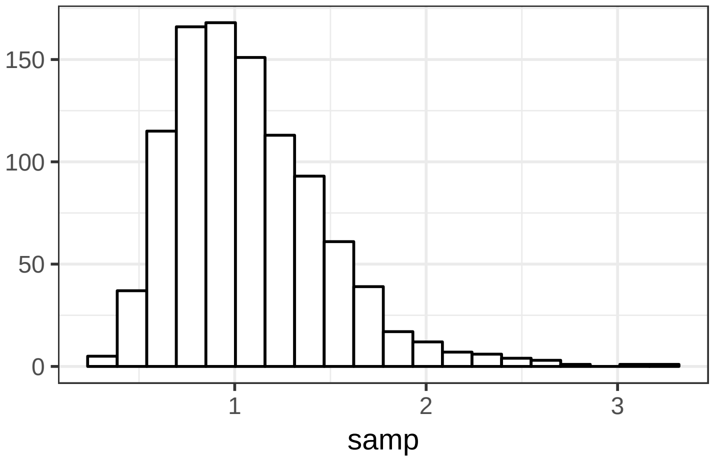
```
## [1] 1.083
```



Random Generation

```r
samp <- rf(n = 1000, df1 = 30, df2 = 30)
head(samp)
```

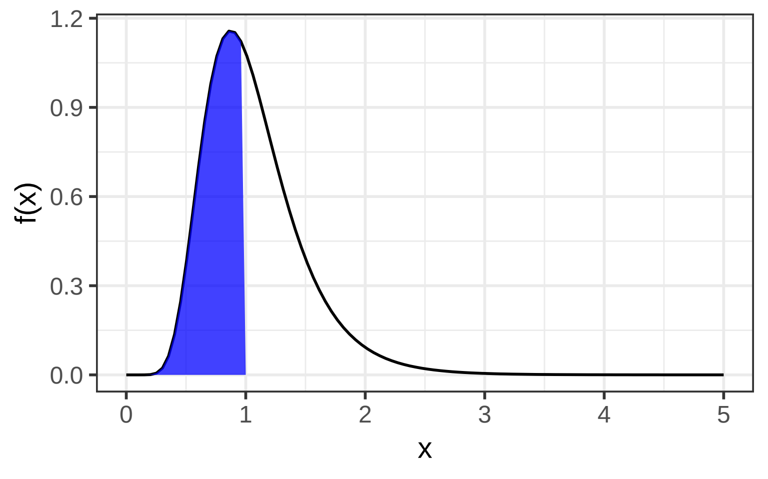
```
## [1] 0.7094 0.5605 1.3322 1.4533 1.0763 0.8123
```



Cumulative Distribution Function

```r
pf(q = 1, df1 = 30, df2 = 30)
```

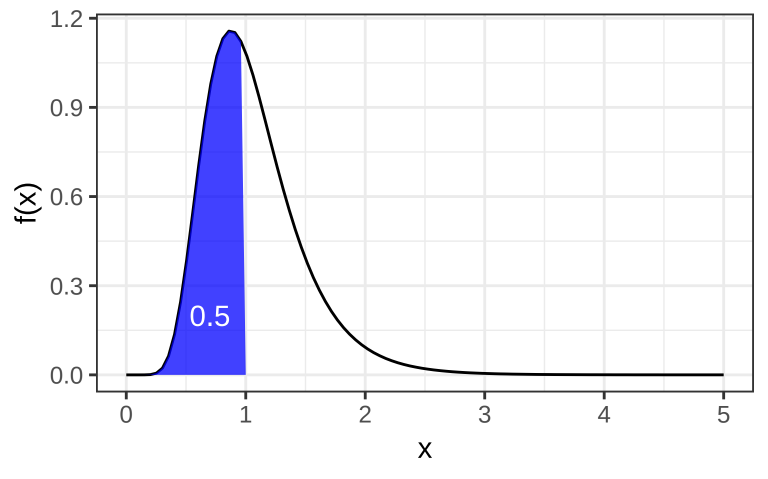
```
## [1] 0.5
```



Quantile Function

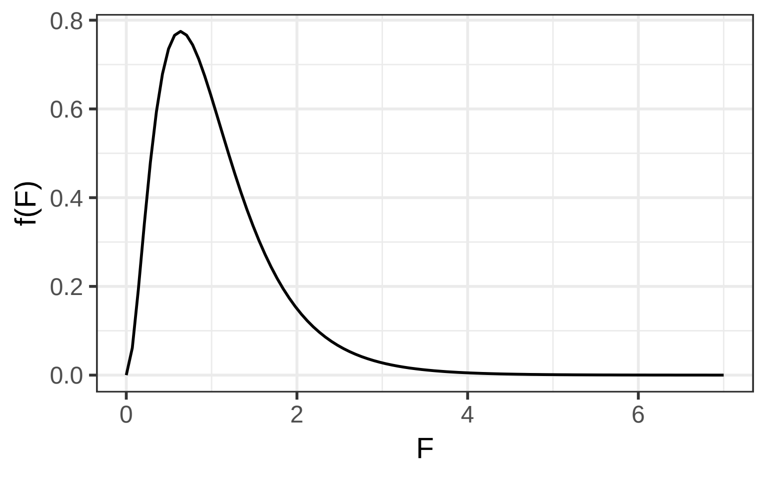
```
qf(p = 0.5, df1 = 30, df2 = 30)
```

```
## [1] 1
```



Spock Example

- Spock's F follows an $F_{I-1,n-I} = F_{6,39}$ distribution *under* H_0.
- How rare is our observed F-stat $= 6.7$, if H_0 were true?



Spock Example

- The p-value is found with

```
pf(q = 6.718, df1 = 6, df2 = 39, lower.tail = FALSE)
```

```
## [1] 6.099e-05
```

Submodel

Another Test

- Suppose we are interested in testing $H_0 : \mu_2 = \mu_3 = \cdots = \mu_7$ against the alternative that at least one mean is different from some other mean.
- We could do the full F-test on the subset of the data that excludes group 1.
- But we would **lose degrees of freedom** because we wouldn't be using group 1 to improve our estimate of the variance.

Model

- Model: $Y_{ij} = \mu_i + \epsilon_{ij}$
- Y_{ij}: Percent women in venire j of judge i.
- μ_i: Mean percent women for judge i.
- ϵ_{ij}: Individual-specific noice for venire j of judge i. Assumed to have mean 0 and variance σ^2.
- σ^2 is assumed to be the **same** for all venires of all judges.

Hypotheses

- H_0: $\mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$ (judges 2 through 7 have the same mean percent women, but judge 1 is allowed to have a different mean).
- H_A: As least some $\mu_i \neq \mu_j$ for judges 2 through 7 (there are at least two judges with different mean percent women, among judges 2 through 7).

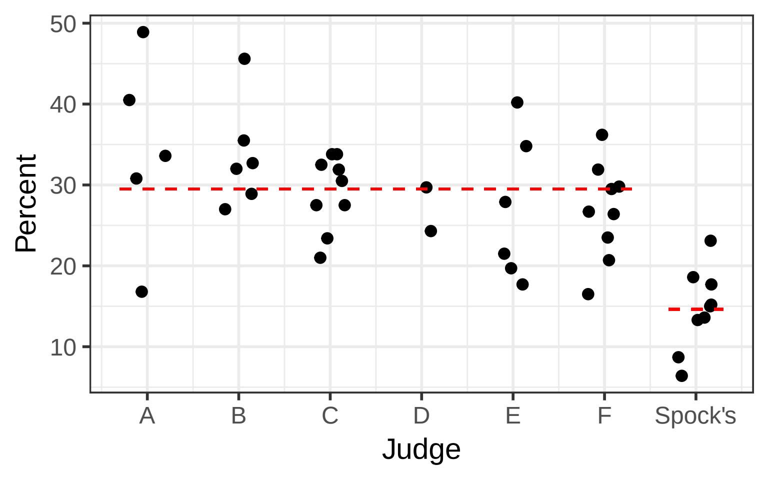
Estimates

We estimate the means within each groups differently according to the full and reduced models

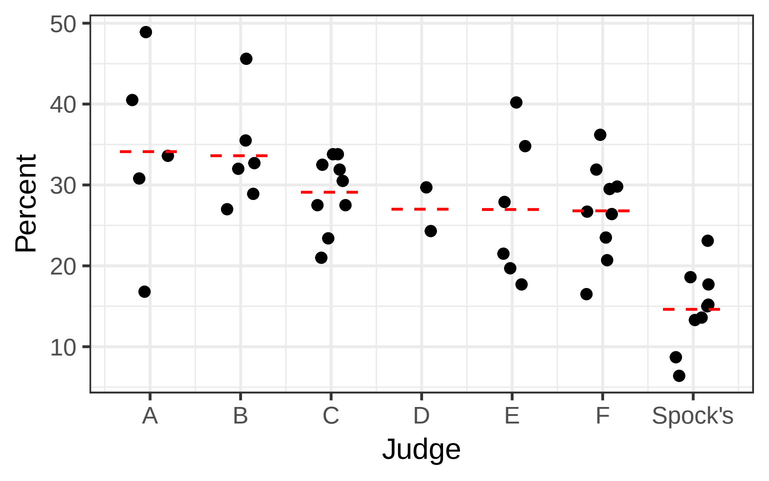
Group	1	2	3	4	5	6	7
Full	$\bar{Y}_{1\bullet}$	$\bar{Y}_{2\bullet}$	$\bar{Y}_{3\bullet}$	$\bar{Y}_{4\bullet}$	$\bar{Y}_{5\bullet}$	$\bar{Y}_{6\bullet}$	$\bar{Y}_{7\bullet}$
Reduced	$\bar{Y}_{1\bullet}$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$

- $\bar{Y}_{i\bullet} = \frac{1}{n_i}(Y_{i1} + Y_{i2} + \cdots Y_{in_i})$
- $\bar{Y}_0 =$ Average of all values in judges 2 through 7.

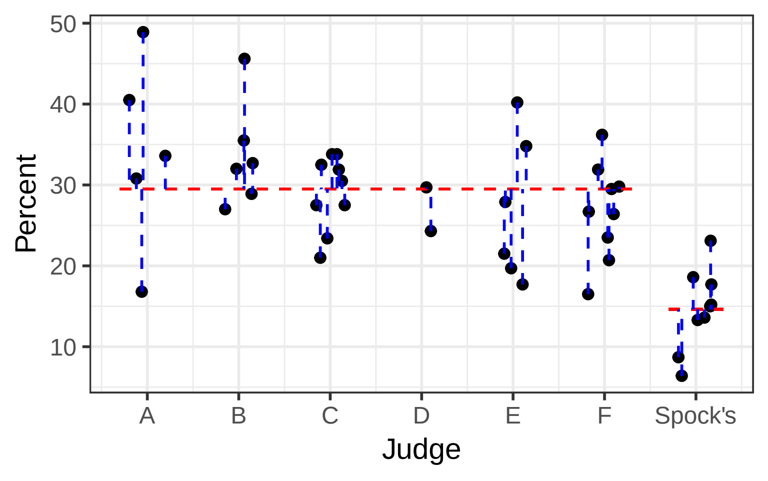
Estimate under Null Model



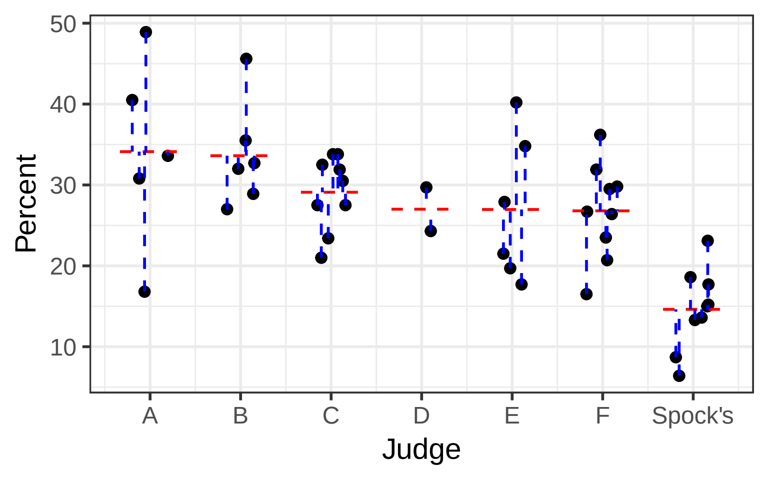
Estimate Under Full Model



Residuals Under Reduced Model



Residuals Under Full Model



F-test

- $RSS_{full} = 1864.4452$ (same as before).
- $df_{full} = n - I = 46 - 7 = 39$.

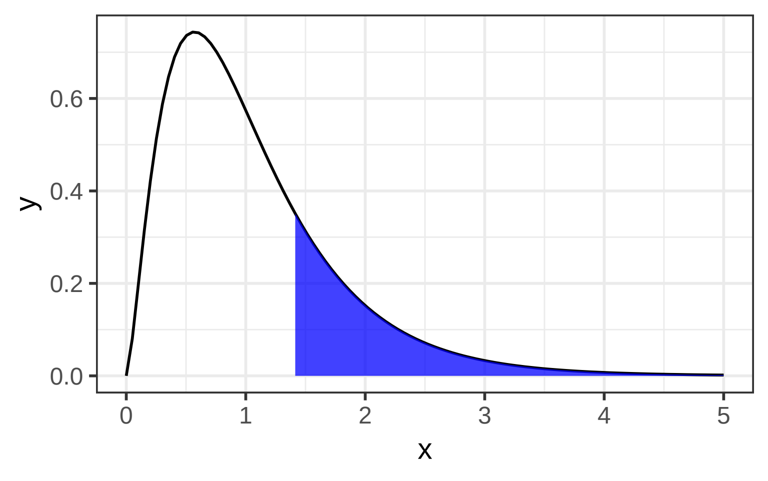
F-test

- $RSS_{full} = 1864.4452$ (same as before).
- $df_{full} = n - I = 46 - 7 = 39.$
- $RSS_{reduced} = 2190.9031$ (smaller than before).
- $df_{reduced} = n - 2 = 46 - 2 = 44$

F-test

- $RSS_{full} = 1864.4452$ (same as before).
- $df_{full} = n - I = 46 - 7 = 39$.
- $RSS_{reduced} = 2190.9031$ (smaller than before).
- $df_{reduced} = n - 2 = 46 - 2 = 44$
- $df_{extra} = df_{reduced} - df_{full} = 44 - 39 = 5$
- $ESS = RSS_{reduced} - RSS_{full} = 326.4579$
- F-stat $= \frac{ESS/df_{extra}}{RSS_{full}/df_{full}} = 1.366$

Compare 1.366 to an $F_{5,39}$ distribution



Compute p-value

```r
pf(q = 1.366, df1 = 5, df2 = 39, lower.tail = FALSE)
```

```
## [1] 0.2581
```

Comparing Submodels

Model

- Model: $Y_{ij} = \mu_i + \epsilon_{ij}$
- Y_{ij}: Percent women in venire j of judge i.
- μ_i: Mean percent women for judge i.
- ϵ_{ij}: Individual-specific noice for venire j of judge i. Assumed to have mean 0 and variance σ^2.
- σ^2 is assumed to be the **same** for all venires of all judges.

Hypotheses

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_7$ (all judges have same mean)
- $H_A : \mu_1 \neq \mu_2 = \mu_3 = \cdots = \mu_7$. (judge 1 is different)

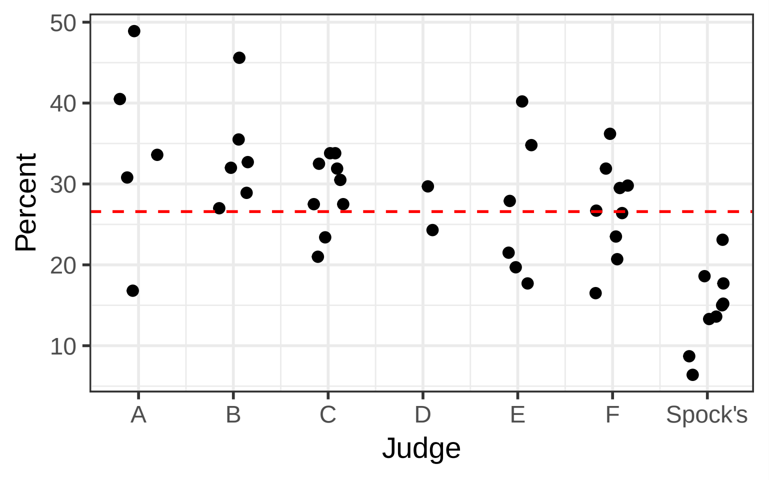
Estimates

We estimate the means within each groups differently according to the full and reduced models

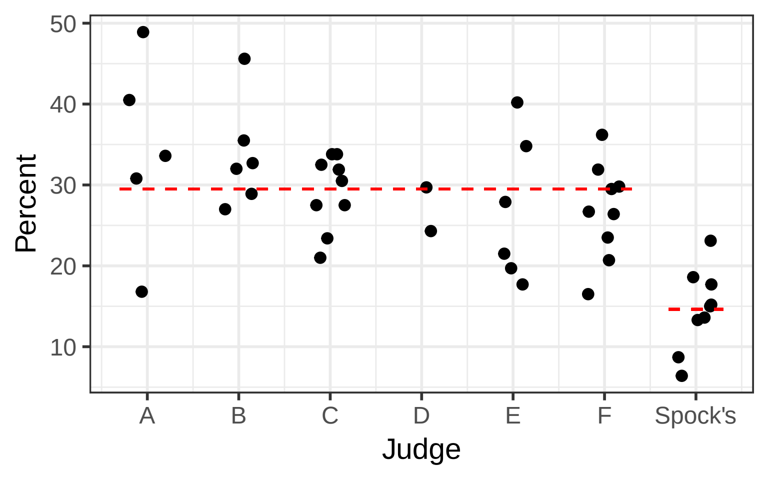
Group	1	2	3	4	5	6	7
Full	$\bar{Y}_{1\bullet}$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$	$\bar{Y}_0$
Reduced	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$	$\bar{Y}_{\bullet\bullet}$

- $\bar{Y}_{i\bullet} = \frac{1}{n_i}(Y_{i1} + Y_{i2} + \cdots Y_{in_i})$
- $\bar{Y}_0 =$ Average of all values in judges 2 through 7.
- $\bar{Y}_{\bullet\bullet} =$ Average of all values in dataset.

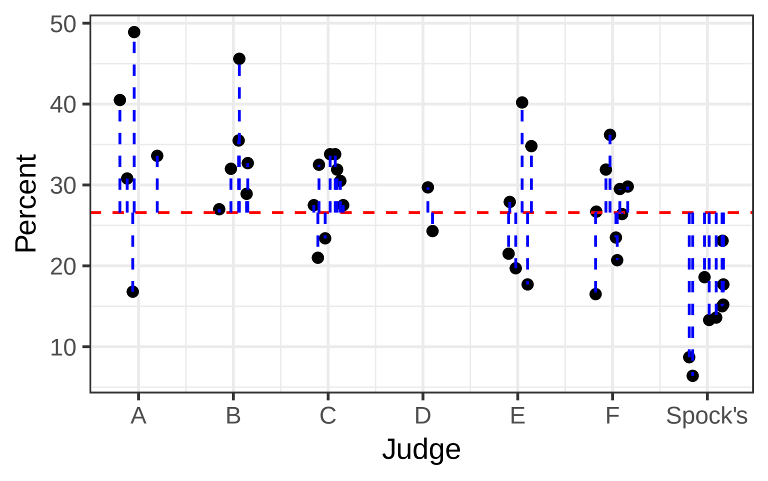
Estimate Under Null Model



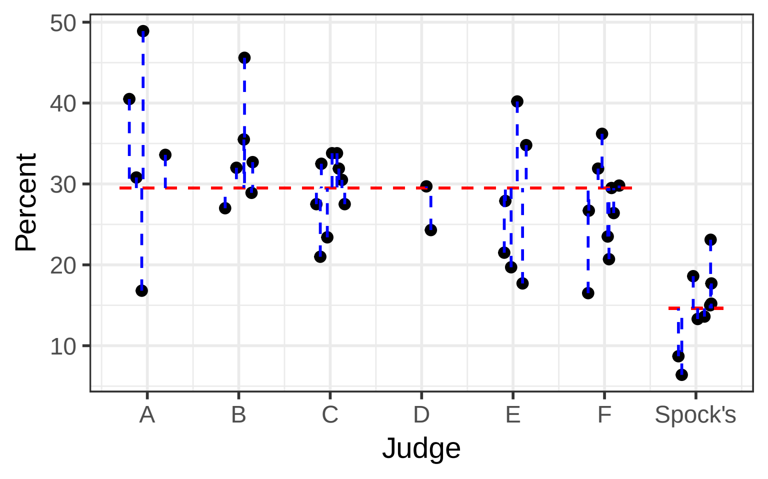
Estimate Under Alternative Model



Residuals Under Null Model



Residuals Under Full Model



F-test

- $RSS_{full} = 2190.9031$ (same as $RSS_{reduced}$ from the **previous** hypothesis test).
- $df_{full} = n - 2 = 46$ - $2 = 44$.

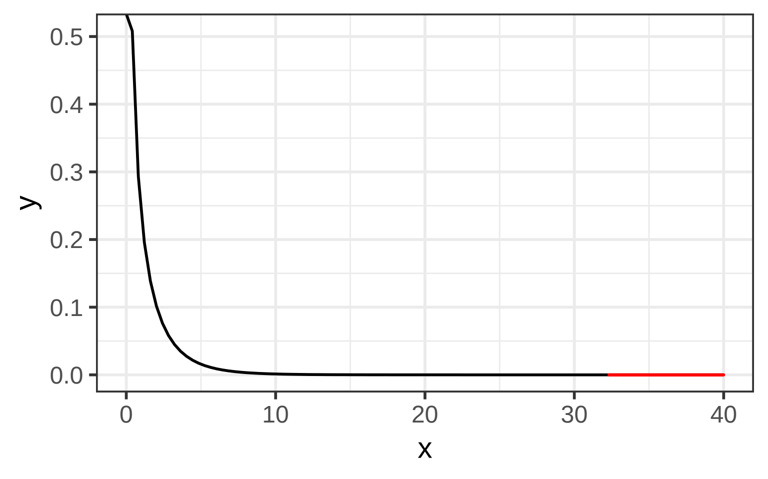
F-test

- $RSS_{full} = 2190.9031$ (same as $RSS_{reduced}$ from the **previous** hypothesis test).
- $df_{full} = n - 2 =$ 46 - 2 $= 44$.
- $RSS_{reduced} = 3791.5261$ (same as $RSS_{reduced}$ from the **first** hypothesis test).
- $df_{reduced} = n - 1 =$ 46 - 1 $= 45$

F-test

- $RSS_{full} = 2190.9031$ (same as $RSS_{reduced}$ from the **previous** hypothesis test).
- $df_{full} = n - 2 =$ 46 - 2 $= 44$.
- $RSS_{reduced} = 3791.5261$ (same as $RSS_{reduced}$ from the **first** hypothesis test).
- $df_{reduced} = n - 1 =$ 46 - 1 $= 45$
- $ESS = RSS_{reduced} - RSS_{full} = 1601$
- $df_{extra} = 1$.
- F-statistic = (1601/1) / (2191/44) = 32.15

Compare to $F_{1,44}$



Compute p-value

```r
pf(q = 32.15, df1 = 1, df2 = 44, lower.tail = FALSE)
```

```
## [1] 1.028e-06
```