Analysis of Variance

Overview

Review and continuation of hypothesis tests for multiple proportions using the chi-square statistic

- Randomization test using the chi-square statistic
- Parametric chi-square statistic test for goodness-of-fit

One-way analysis of variance (ANOVA) for comparing multiple means

- Running a one-way ANOVA
- If there is time: Randomization test using an F-statistic

Randomization test to compare multiple proportions

Testing more than two categories

A Yale professor was interesting in examining whether Yale students were equally likely to be born in each month

Q_1: What is the null and alternative hypotheses he could use to test this hypothesis?

Testing more than two categories

The hypotheses:

H_0: $\pi_{Jan} = \pi_{Feb} = \ldots = \pi_{Dec} = 1/12$

H_A: At least one of the proportions π_i is different

Testing more than two categories

Q_2: Any ideas how we could test this?

One solution: we could do 12 hypothesis tests for each $\pi_i = 1/12$

Problem: multiple comparisons would lead to higher type I error rate

Chi-square goodness-of-fit test

If we want to test proportions for k > 2 categories, we can use:

1. A randomization test
2. A parametric "*chi-square goodness-of-fit"* test

These test:

H_0: $\pi_1 = a$, $\pi_2 = b$, … $\pi_k = z$

H_A: Some π_i is not as specified in H_0

The tests don't specify which proportion differs from what is specified in the null hypothesis, just that at least one proportion does differ

Chi-square statistic

The **chi-square statistic**, denoted χ^2, is found by comparing the **observed counts** from a sample with the **expected counts** derived from a null hypothesis and is computed as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Note this is a Greek symbol even though it is a statistic ☹

Birth months from 198 Yale students

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	**Total**
Observed	20	17	21	24	11	18	12	17	13	14	17	14	**198**

Q: What is the expected value for each month?

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} = 10.12$$

Q: Is the observed statistic beyond what we would expect if H_0 was true?

Creating a null distribution

Any ideas how we could create a null distribution using randomization methods?

We could use randomization method:

- Roll k-sided weighted die
- Probability of getting each side is equal to the proportions specified in the null hypothesis
- Roll the die n times to simulate one experiment
  - Calculate χ^2 statistic based on these rolls
- Repeat many times to get a null distribution

Creating a null distribution

For the Yale birth month example…

How many sides would the die have?

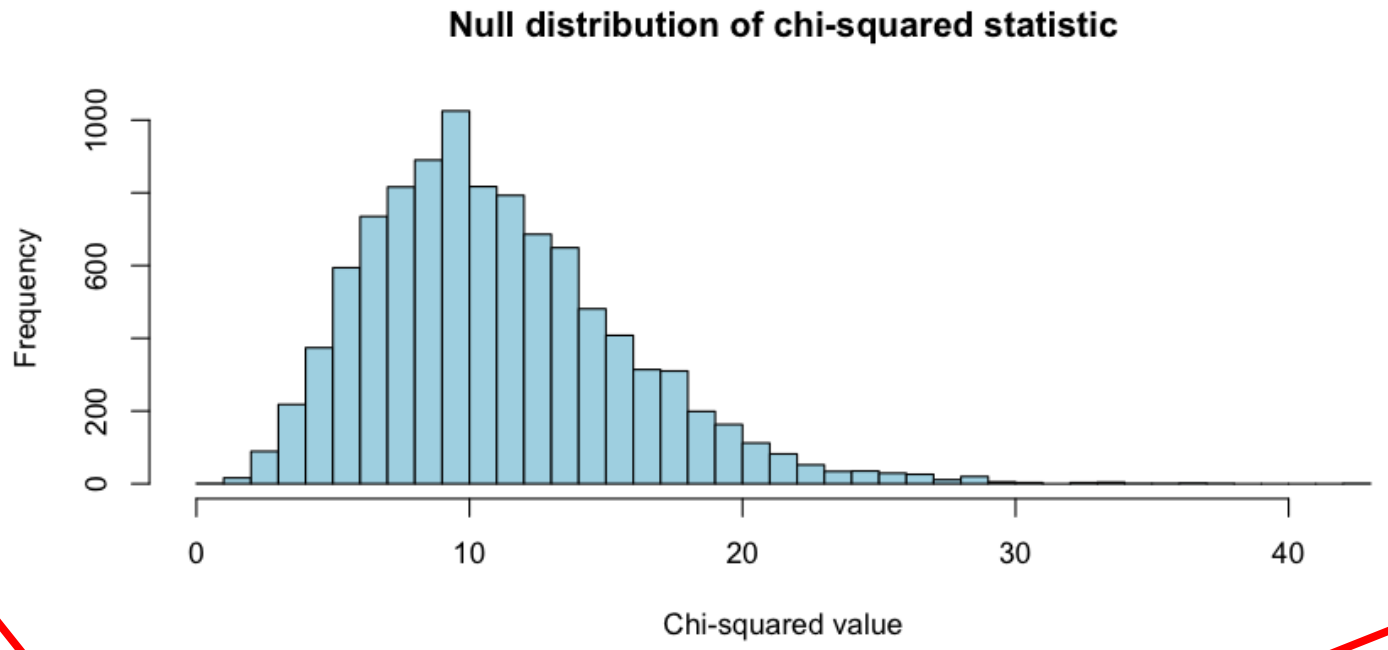
What would the probability be of getting each side?

How many times would we roll the die to simulate one data set?

How many times would we repeat this process?

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Randomization null distribution



Vector of randomly generated counts in each category

"observed" data consistent with H_0

Vector of expected proportions for each category level

Vector length is the number of categories k

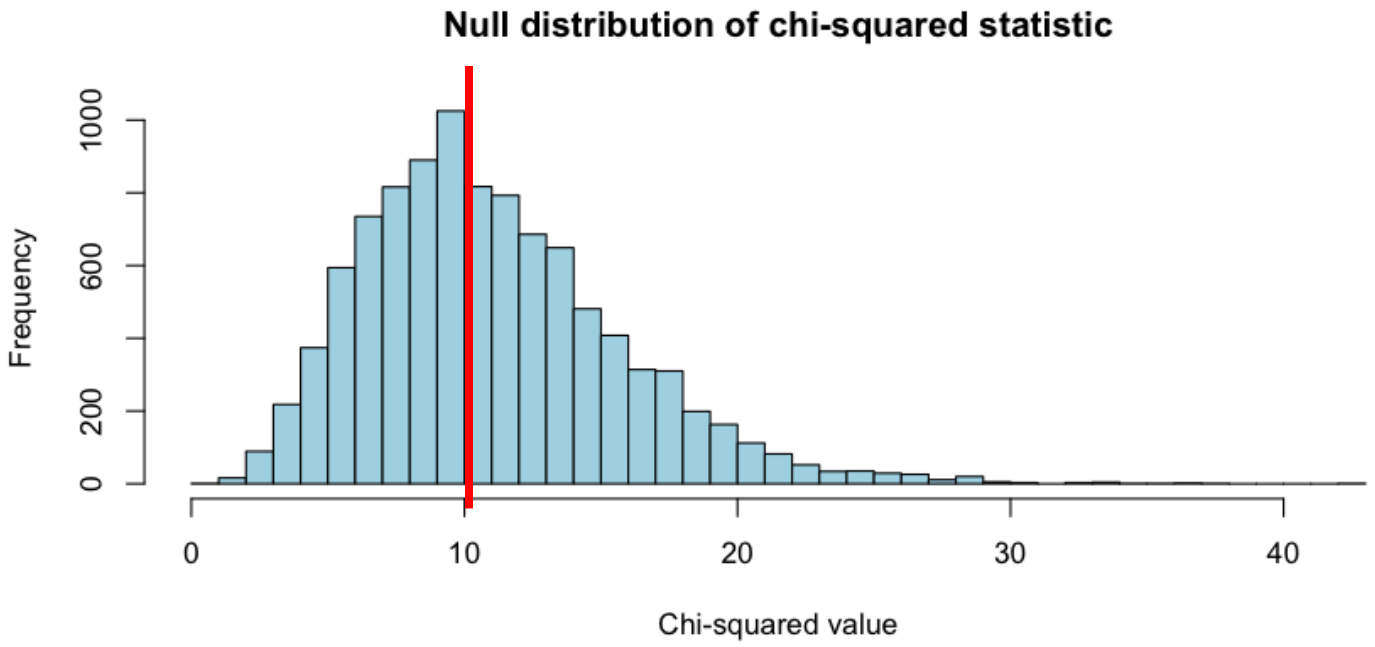
simulated_counts <- rmultinom(1, n, expected_proportions)

simulated_counts <- rmultinom(1, 198, rep(1/12, 12))

Calculate statistic χ^2 and repeat 10,000 times

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Randomization null distribution



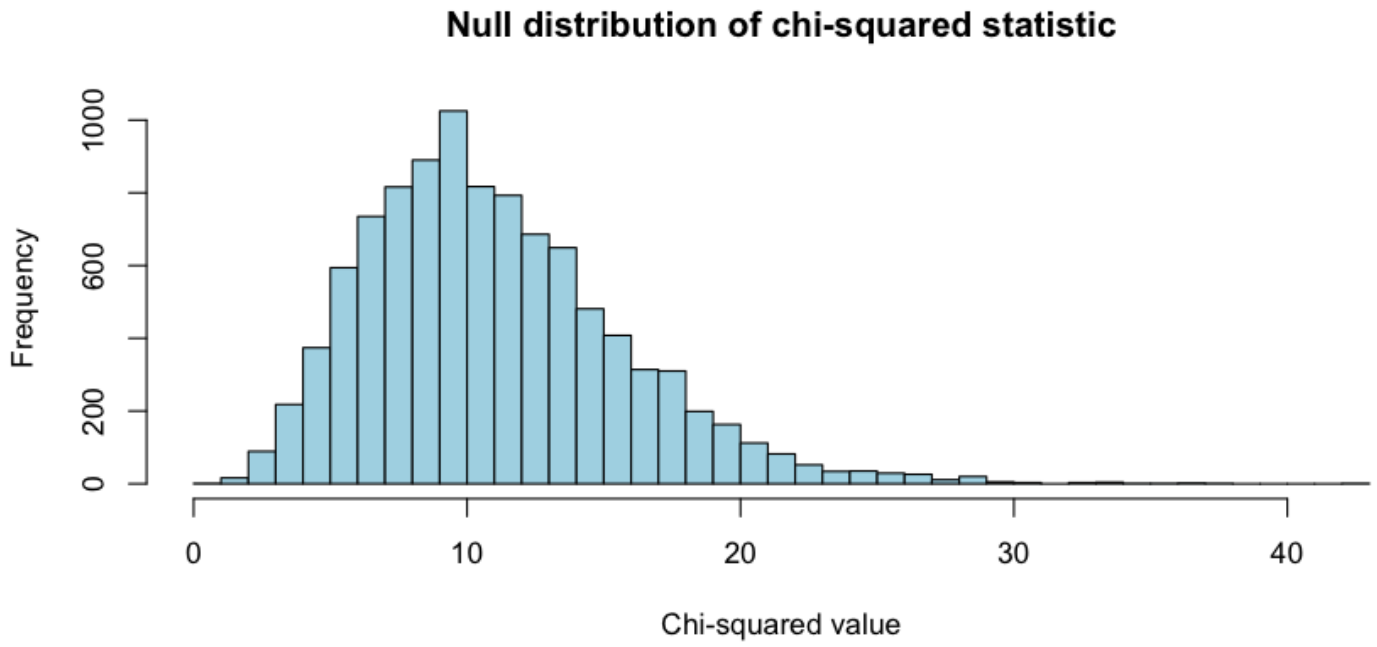
$\chi^2 = 10.12$

p-value = …?

Let's try it in R!

χ^2 test for goodness-of-fit to compare multiple proportions

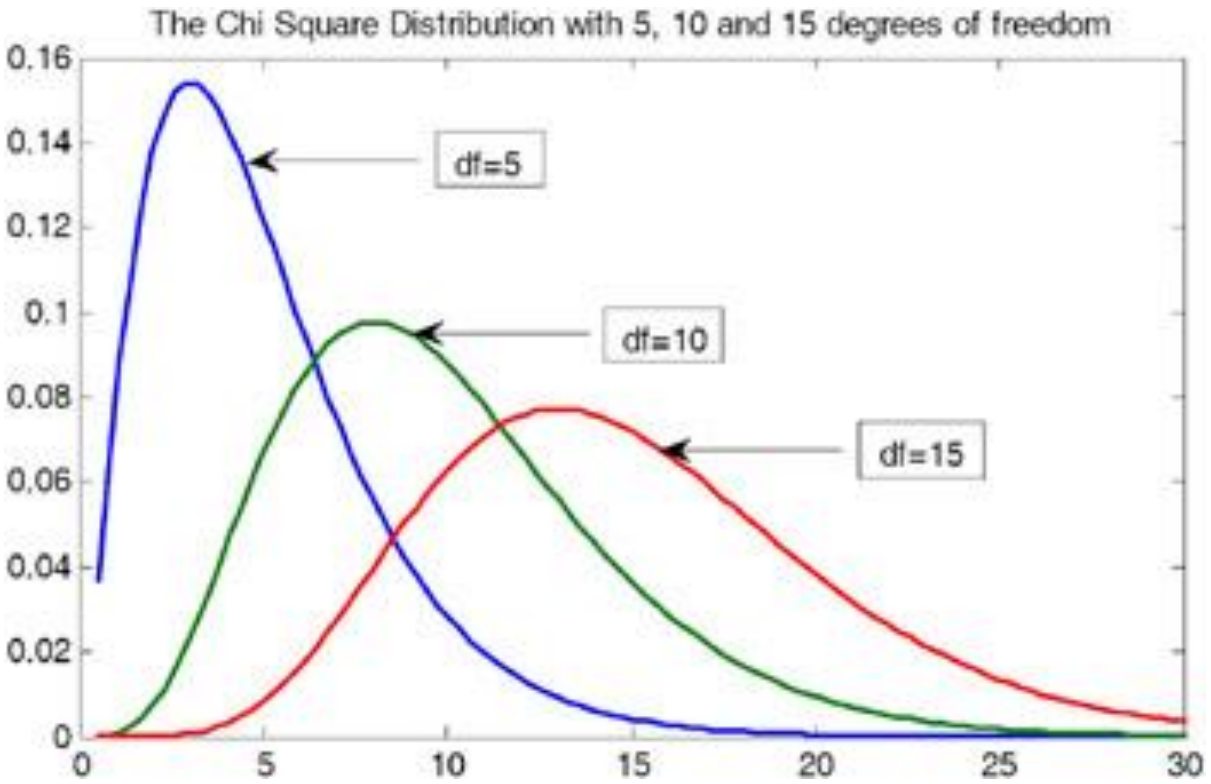
Parametric null distribution



Is there a parametric null distribution we could use instead of simulations?

Yes! The χ^2 distribution!

Chi-square distribution

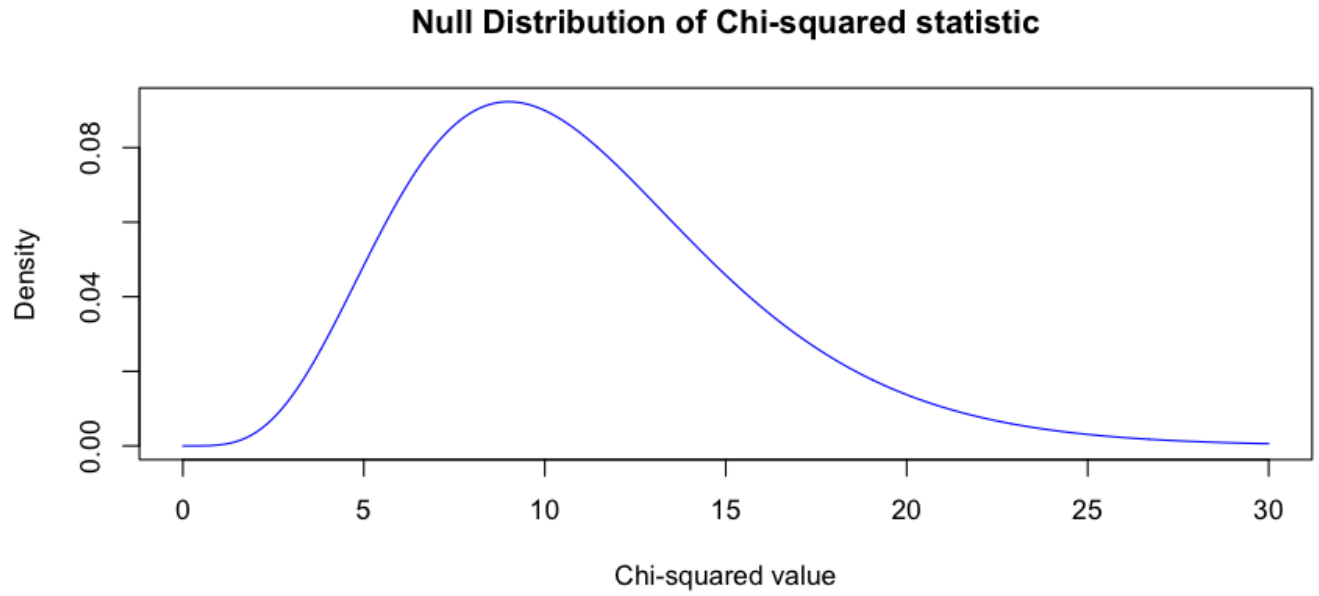


k is the number of groups

The χ^2 has one parameter called 'degrees of freedom', which is equal to k - 1

χ^2 distribution can be used as a null distribution as long as there are at least 5 expected counts in every condition

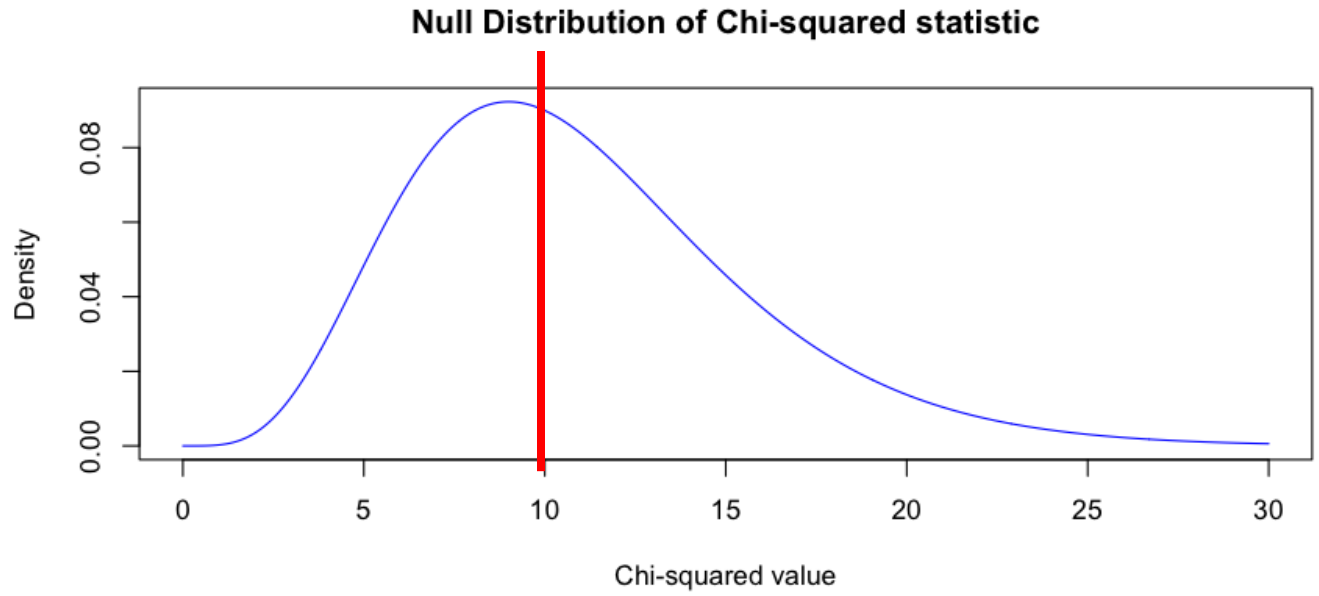
Chi-square distribution



To plot the chi-squared density we can use: y_vals <- dchisq(x_vals, df)

To get a p-value we can use: pchisq(chi_stat, df, lower.tail = FALSE)

Chi-square distribution



For Yale birth month example, since there 12 months there are 11 degrees of freedom (df = 11)

P-value based on the chi-squared distribution = …?

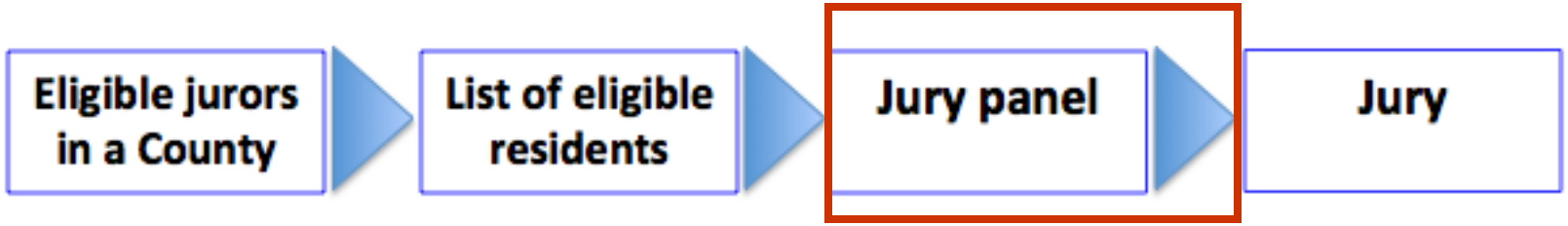
Let's try it in R!

Another example

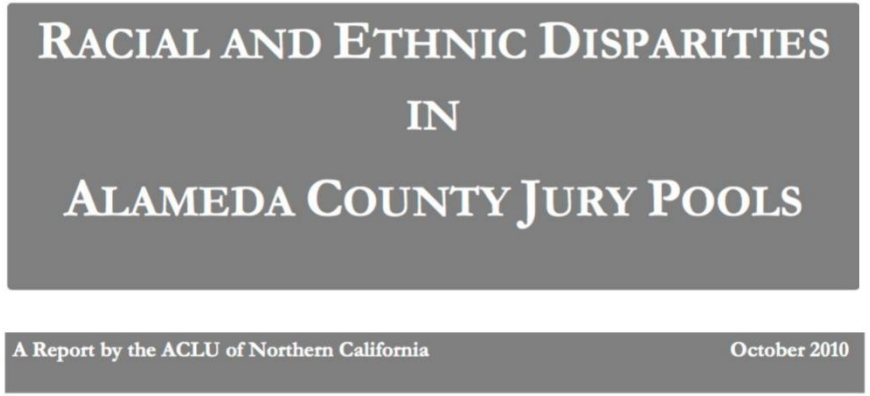
Jury selection in Alameda county

Section 197 of California's Code of Civil Procedure says:

" All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court."



In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report that concluded that certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.



Jury selection in Alameda county data

The ACLU compiled data on the composition of 1453 people who were on jury panels from in the years 2009 and 2010.

The demographics and jury panel proportions were:

	Asian	Black	Latino	White	Other
Eligible	0.15	0.18	0.12	0.54	0.01
Panel	0.26	0.08	0.08	0.54	0.04

Question: were ethnicities selected to be on the panel representative of what would be expected from the underlying population?

Jury selection in Alameda county data

The ACLU compiled data on the composition of 1453 people who were on jury panels from in the years 2009 and 2010.

The demographics and jury panel proportions were:

	Asian	Black	Latino	White	Other
Eligible	0.15	0.18	0.12	0.54	0.01
Panel	0.26	0.08	0.08	0.54	0.04

Question: What are the null and alternative hypotheses?

H_0: $\pi_A = .15$, $\pi_B = .18$, $\pi_L = .12$, $\pi_W = .54$, $\pi_o = .01$

H_A: Some π_i is not as specified in H_0

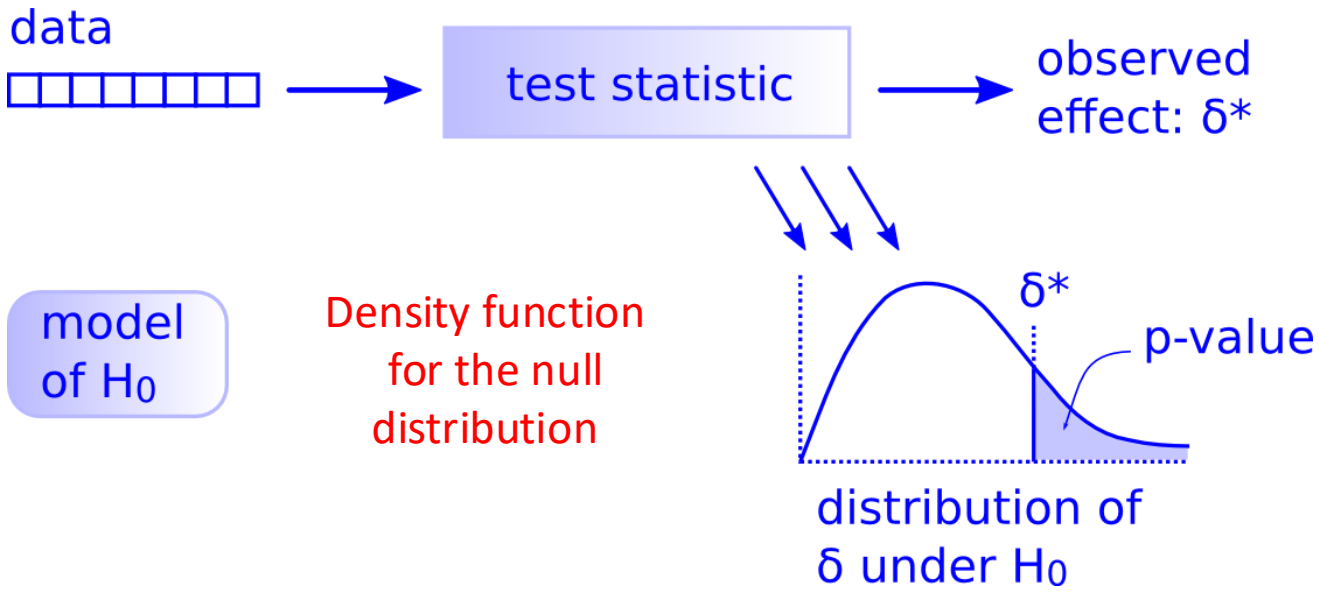
Try this at home!

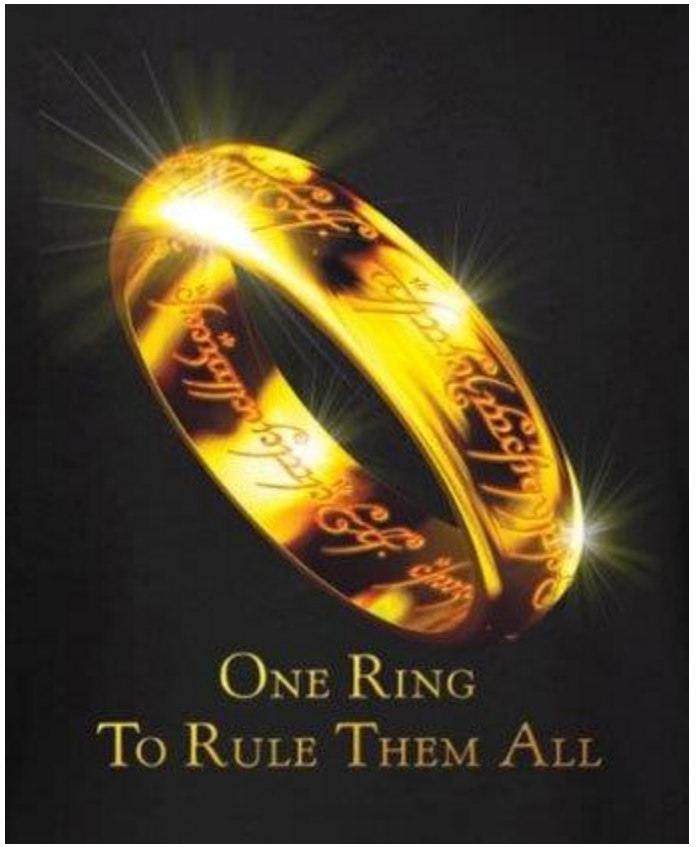
One-way analysis of variance (ANOVA):

Parametric test for comparing more than two means

One test to rule them all

There is only one [hypothesis test](): !





Just follow the 5 hypothesis tests steps!

One-way ANOVA

An Analysis of Variance (ANOVA) is a parametric hypothesis test that can be used to examine if a set of means are all the same

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A: $\mu_i \neq \mu_j$ for some i, j

The statistic we use for a one-way ANOVA is the F-statistic

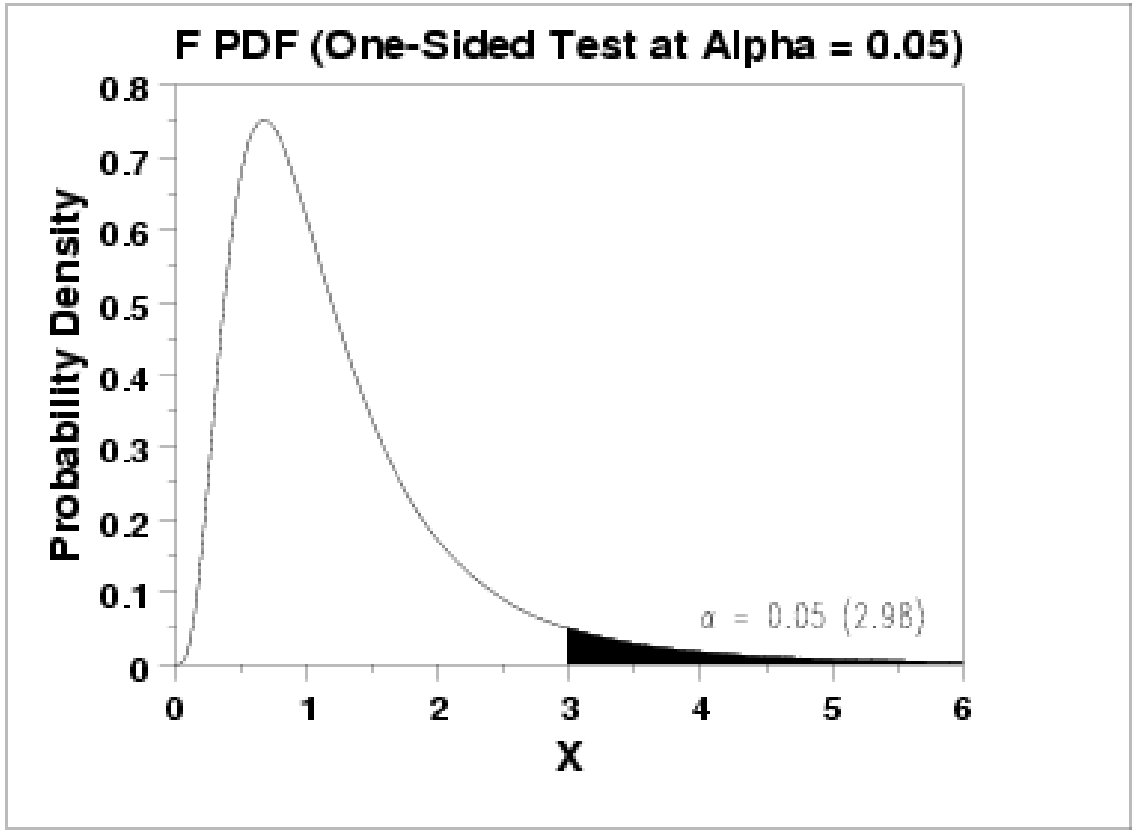
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

One-way ANOVA – the central idea

If H_0 is true, the F-statistic we compute from our data will come from an F-distribution if these conditions are met:

- The data in each group should follow a normal distribution
- The variances in each group should be approximately equal

We can get a p-value by finding the probability we will get a F-statistic larger than the observed F-statistic

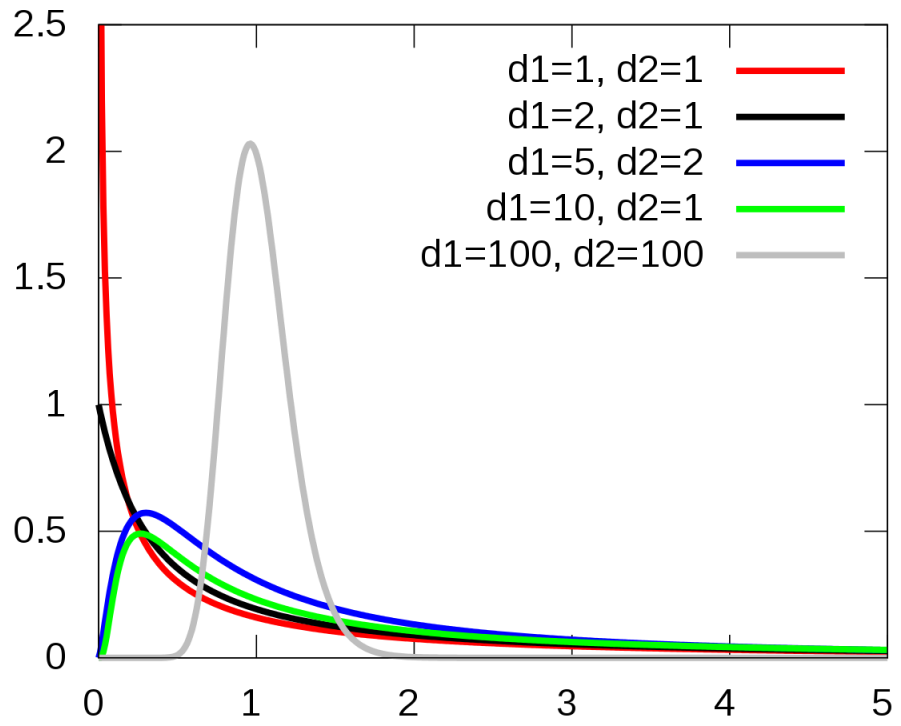


One-way ANOVA – the central idea

The F-distribution is a family of distributions that have two parameters: df_1 and df_2

When using the F-distribution as a null distribution for our F-statistic, the appropriate parameter values are:

- $df_1 = K - 1$
- $df_2 = N - K$



Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

They grouped majors into four categories

- Applied science (as)
- Natural science (ns)
- Social science (ss)
- Arts/humanities (ah)

What is the first step of hypothesis testing?

1. State the null and alternative hypotheses

H_0: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

H_A: $\mu_i \neq \mu_j$ for one pair of fields of study

What should we do next?

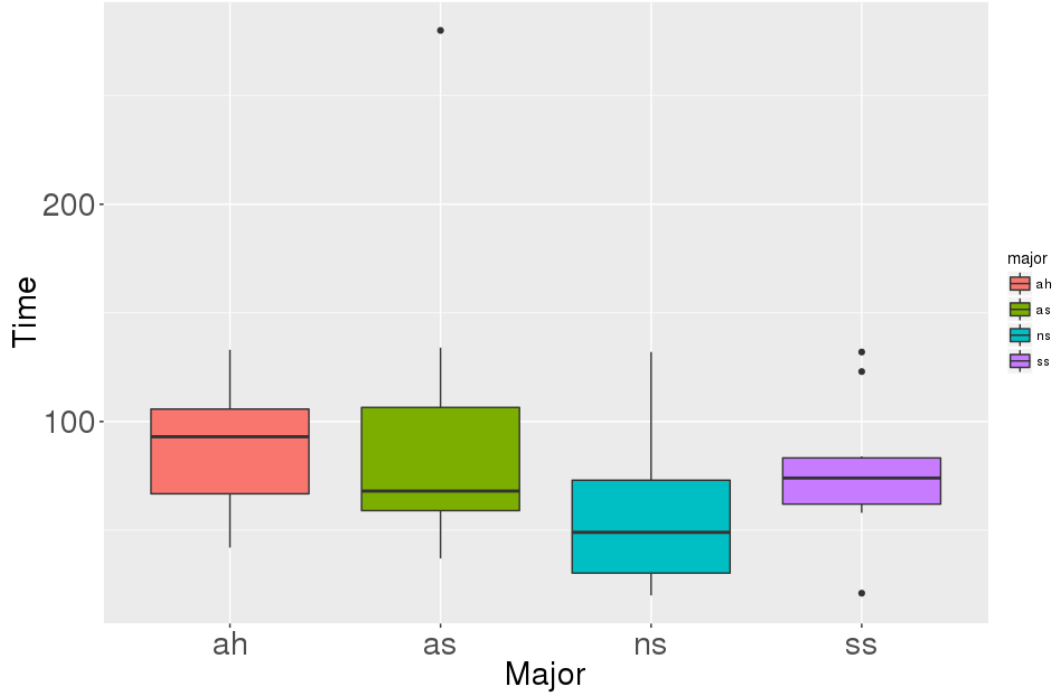
Checking ANOVA conditions ('assumptions')

We can check if the data in each group is relatively normal by creating boxplots and seeing:

- Is the data very skewed?
- Are there are many outliers?

We can check the equal variance condition by seeing if the ratio of the largest to smallest standard deviation is greater than 2

- $s_{max}/s_{min} < 2$



$s_{ah} = 27.9$

$s_{as} = 71.59$

$s_{ns} = 34.39$

$s_{as} = 31.89$

2. Calculating the observed F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

K: the number of groups

N: total number of points

$\bar{x}_{tot}$: the mean across all the data

$\bar{x}_i$: the mean of group i

n_i: the number of points in group i

x_{ij} : the j^{th} data point from group i

K = 4 different majors here

N = 40 total students in the full data set

$\bar{x}_{tot}$: the mean across Sudoku times

$\bar{x}_i$: the means for ah, as, ns, and ss

n_i = 10 students in each major

x_{ij} : the j^{th} student's time from the i^{th} major

2. Calculating the observed F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Fortunately, there is a function in the SDS1000 package that will calculate this for us!

get_F_stat(data_vector, grouping_vector)

- data_vector: a vector of quantitative data
- grouping_vector: a vector indicating which group the quantitative data is in

Let’s try this analysis in R…

```
# get the data
library(SDS1000)
sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)

# Extract vectors from the data frame  (how do we do this?)
completion_time <- sudoku_data$time
major <- sudoku_data$major
```

Let's try this analysis in R…

We can get the F statistic using the get_F_stat() function

get_F_stat(data_vector, grouping_vector)

- data_vector: a vector of quantitative data
- grouping_vector: a vector indicating which group the quantitative data is in

Can you get the F statistic for the sudoku data?

```
obs_stat <- get_F_stat(completion_time, major)
```

= 1.370

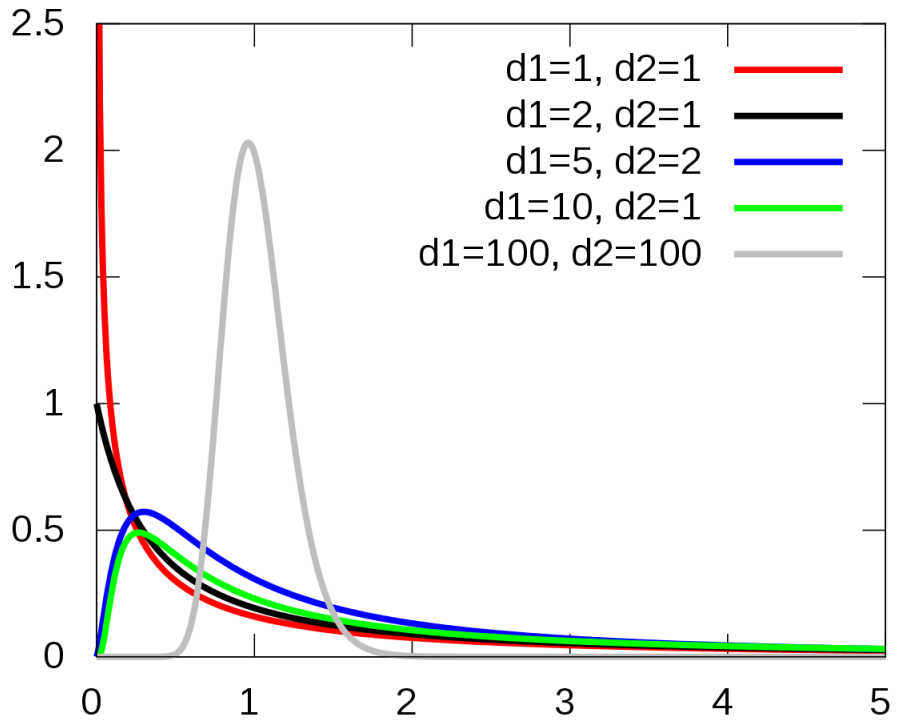
3. Plot the null distribution

If a few conditions (assumptions) are met, the null distribution for our F-statistic will be an F-distribution

The F-distribution is a family of distributions that have two parameters: df_1 and df_2

When using the F-distribution as a null distribution for our F-statistic, the appropriate parameter values are:

- $df_1 = K - 1$
- $df_2 = N - K$



3. Plot the null distribution

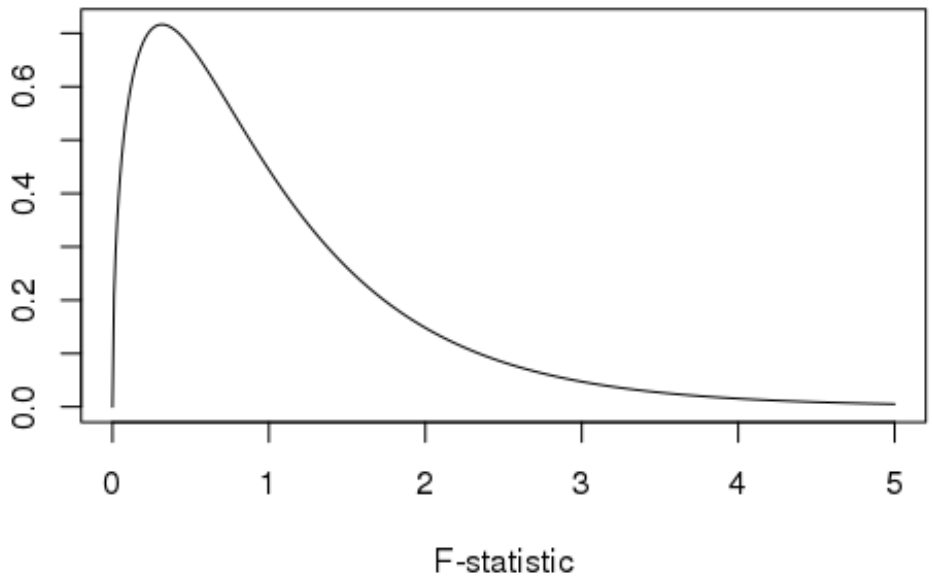
In R we can plot the density function for an F-distribution using the function…

- df(x_vals, df1, df2)

When using the F-distribution as a null distribution for our F-statistic, the appropriate parameter values are:

- $df_1 = K - 1 = 4 - 1 = 3$
- $df_2 = N - K = 40 - 4 = 36$

```
x_vals <- seq(0, 5, length.out = 1000)
y_vals <- df(x_vals, 3, 36)
plot(x_vals, y_vals, type = 'l')
```



4. Calculate the p-value

In R we can plot the density function for an F-distribution using the function…

- df(x_vals, df1, df2)

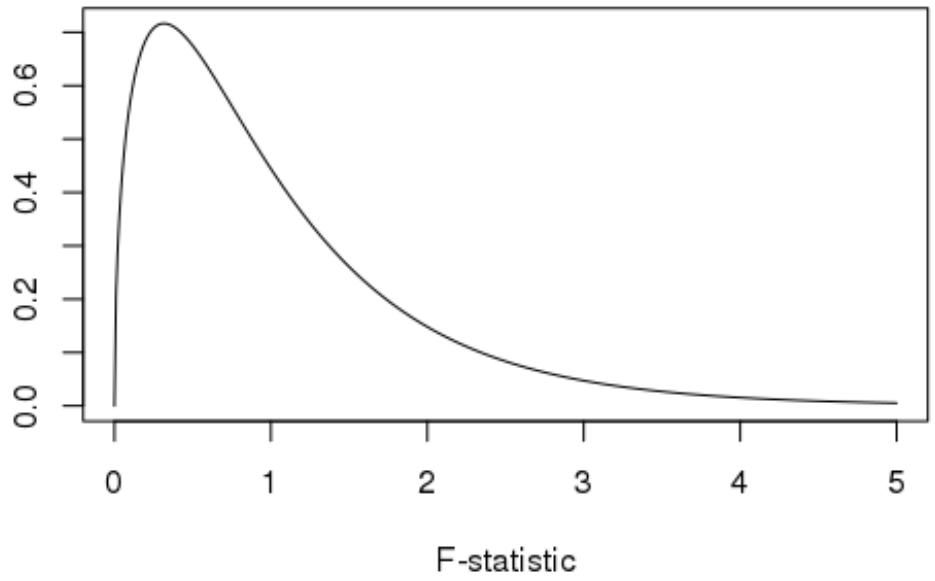
When using the F-distribution as a null distribution for our F-statistic, the appropriate parameter values are:

- $df_1 = K - 1 = 4 - 1 = 3$
- $df_2 = N - K = 40 - 4 = 36$

abline(v = obs_stat, col = "red")

pf(obs_stat, 3, 36, lower.tail = FALSE)

= 0.267433



5. Conclusion?

In R we can plot the density function for an F-distribution using the function…

- df(x_vals, df1, df2)

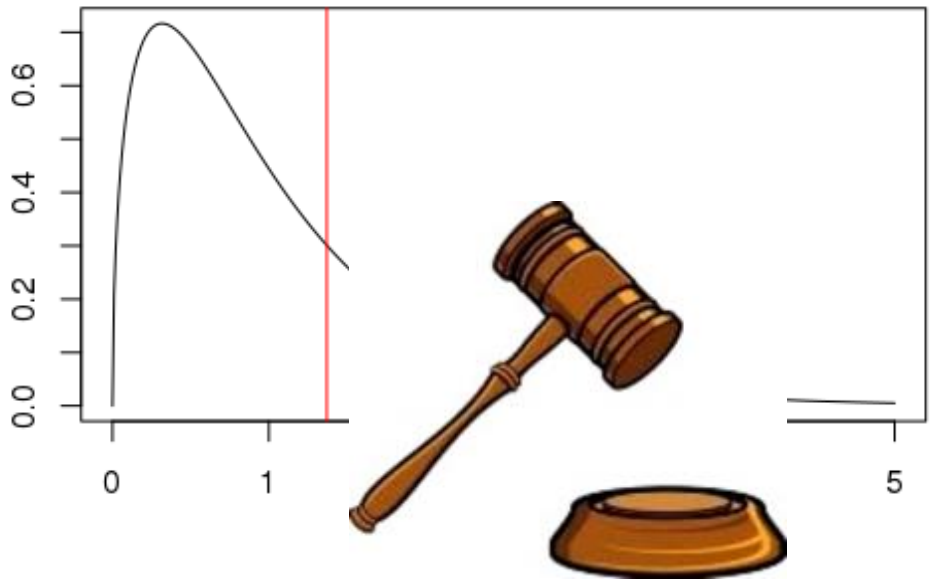
When using the F-distribution as a null distribution for our F-statistic, the appropriate parameter values are:

- $df_1 = K - 1 = 4 - 1 = 3$
- $df_2 = N - K = 40 - 4 = 36$

abline(v = obs_stat, col = "red")

pf(obs_stat, 3, 36, lower.tail = FALSE)

= 0.267433



Let's try it in R!

Brief mention/review: Permutation test comparing multiple means using the F-statistic

Sudoku by field

1. State the null and alternative hypotheses!

$\mathbf{H_0}$: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

$\mathbf{H_A}$: $\mu_i \neq \mu_j$ for one pair of fields of study

Thoughts on the statistic of interest?

Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:

   $\max \bar{x} - \min \bar{x}$

2. Mean absolute difference (MAD):

   $(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

3. F statistic:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

The MAD statistic vs. the F-Statistic

Mean absolute difference (MAD):

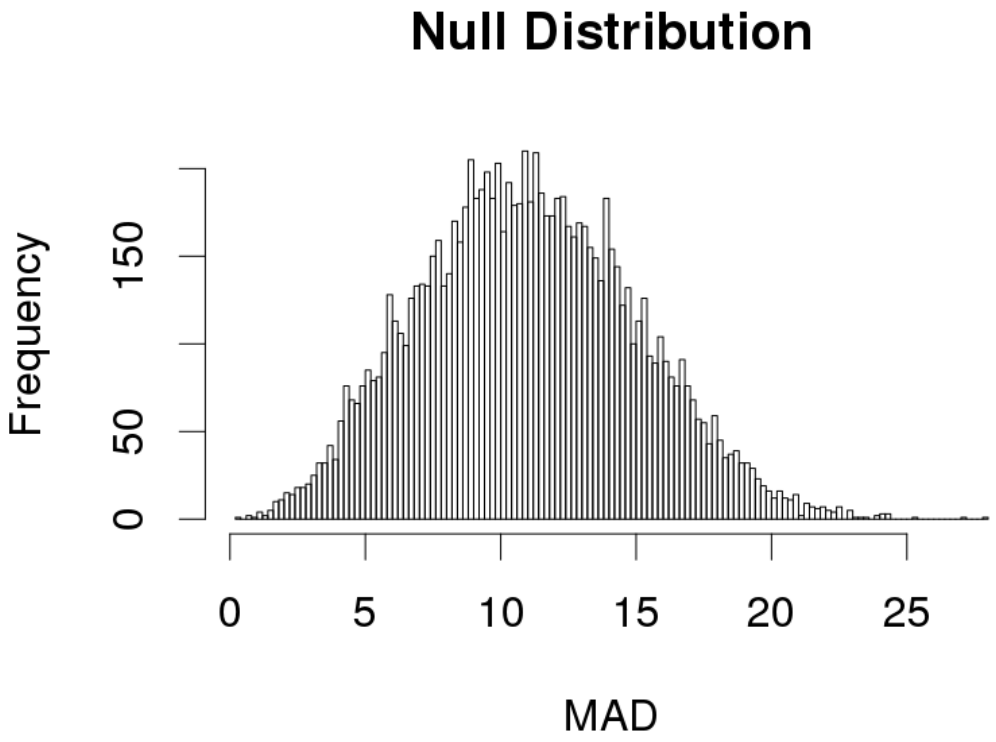
$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

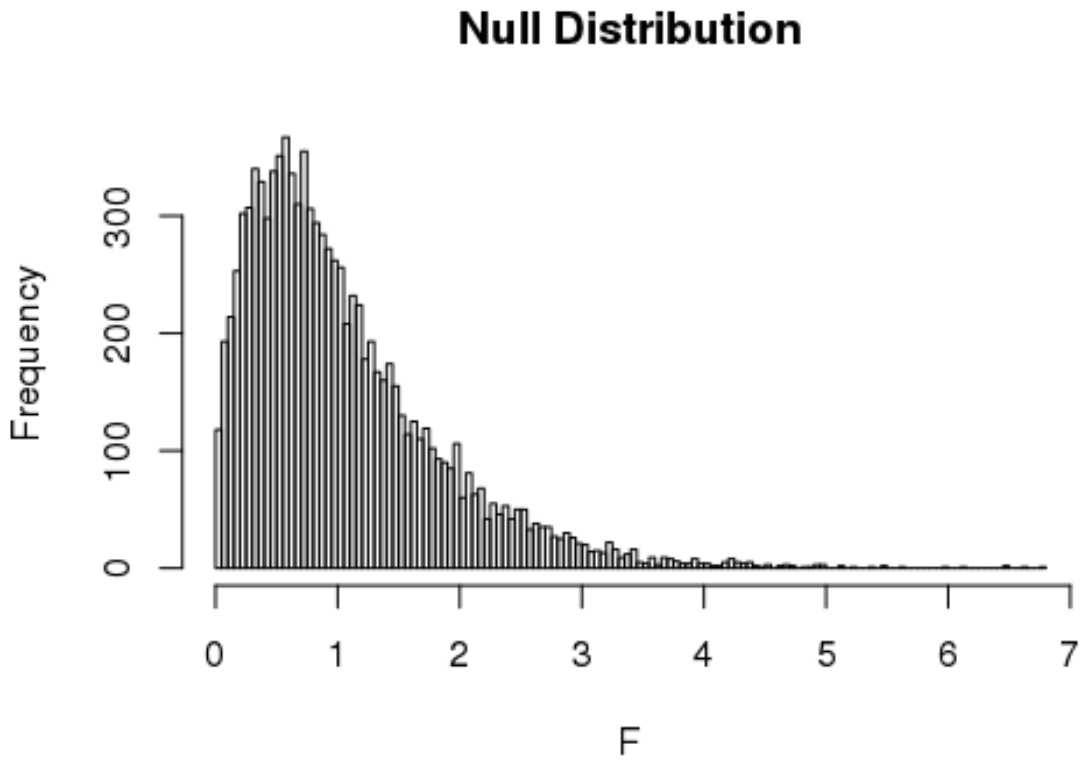
Observed MAD statistic value = 13.92

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

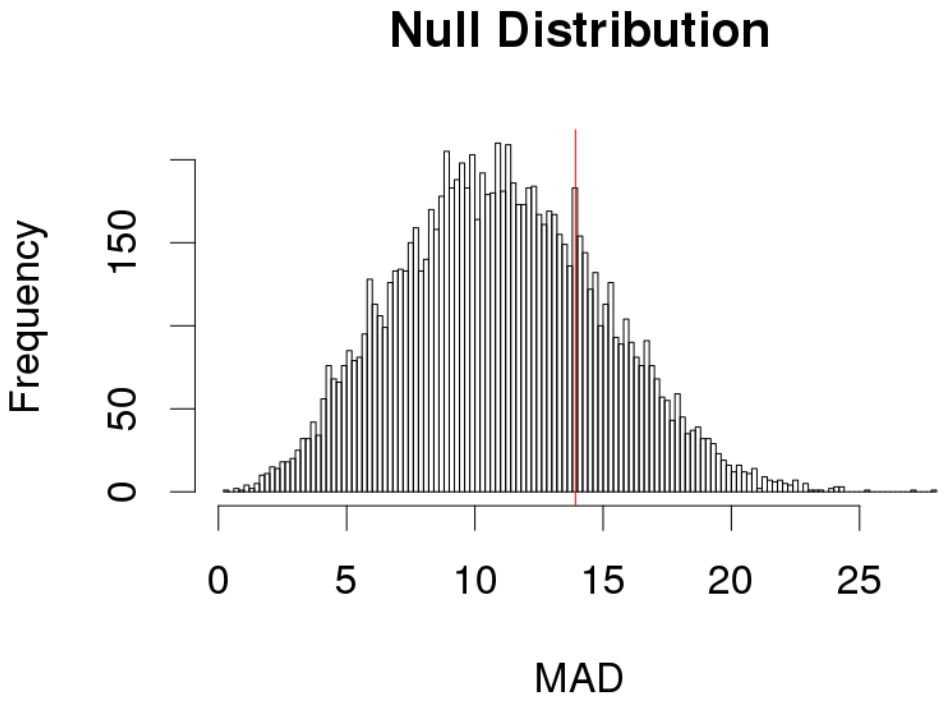
Observed F statistic value = 1.370

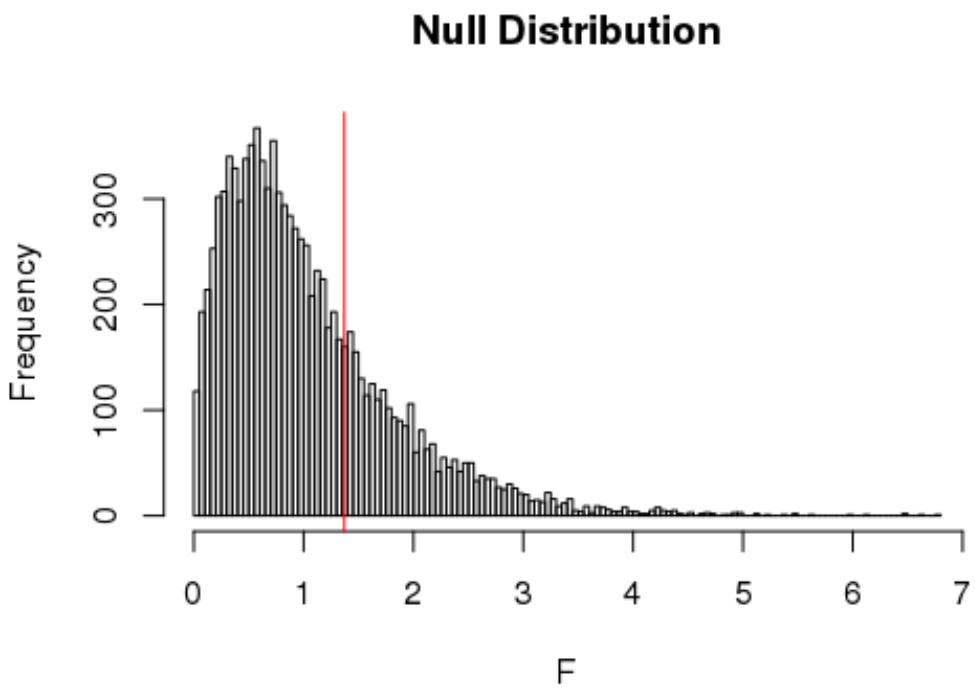
Null distributions





P-value

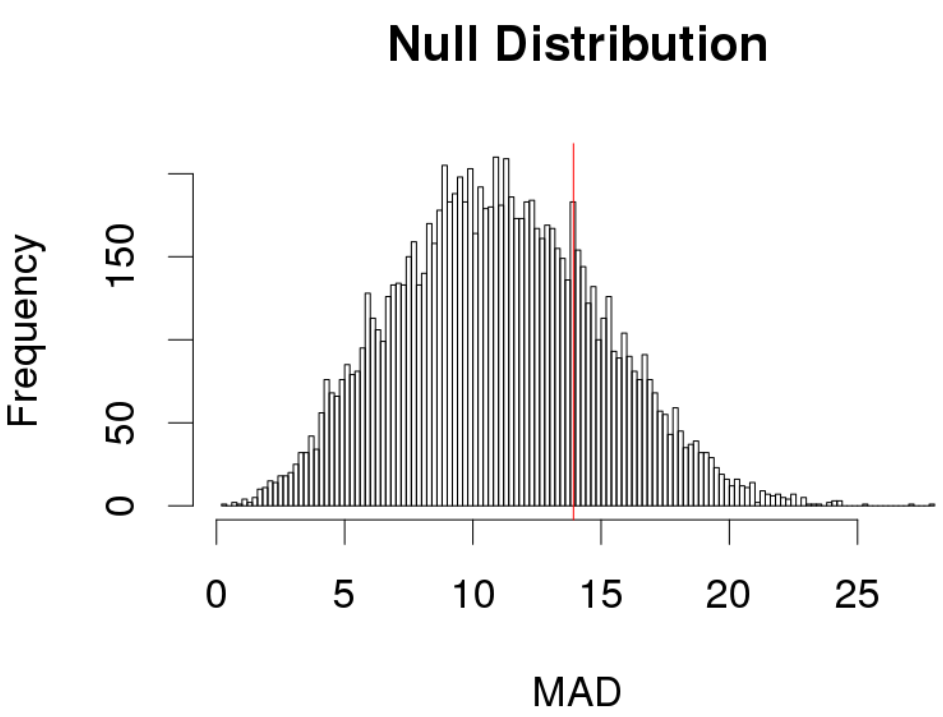


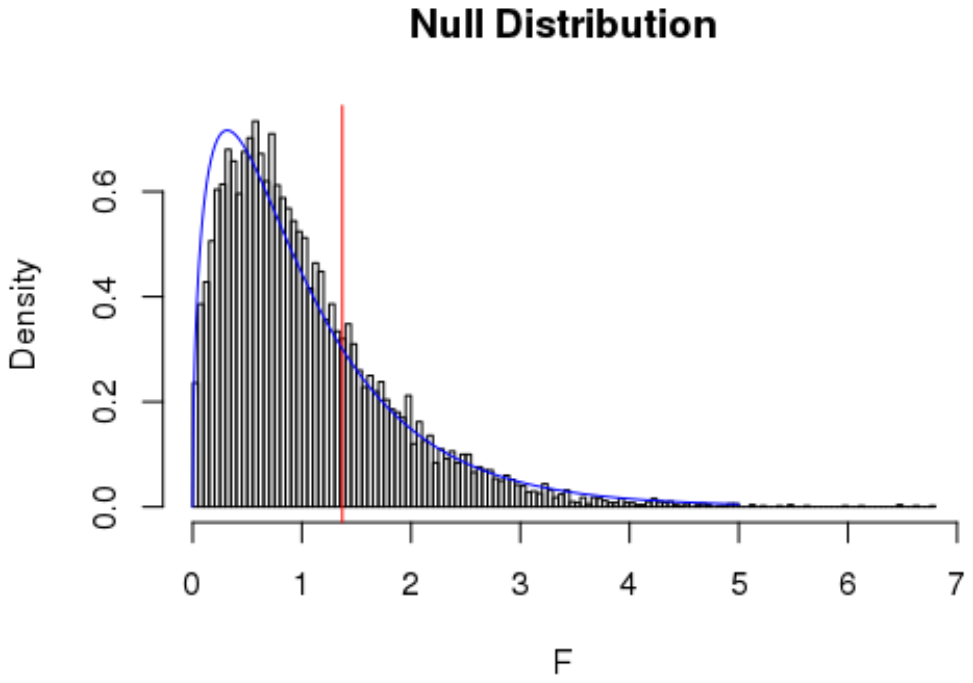


p-value = .4682

p-value = 0.2653

P-value





p-value = .4682

p-value = 0.2674

Conclusions?