Chi-squared test for association
and
Analysis of Variance

Overview

Review of chi-squared tests for goodness-of-fit

Chi-squared test for association

One-way analysis of variance (ANOVA)

- Central concepts behind running a one-way ANOVA
- Randomization test using an F-statistic

Review of chi-squared test for goodness-of-fit

Chi-square Goodness-of-Fit Test

To test a hypothesis about the proportions of a categorical variable based on a table of observed counts in *k* cells:

H_0: Specifies proportions π_i, for each cell

H_A: At least one π_i is not as specified in H_0

Compute the expected counts for each cell using $n * \pi_i$

Compute the value of the chi-square statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

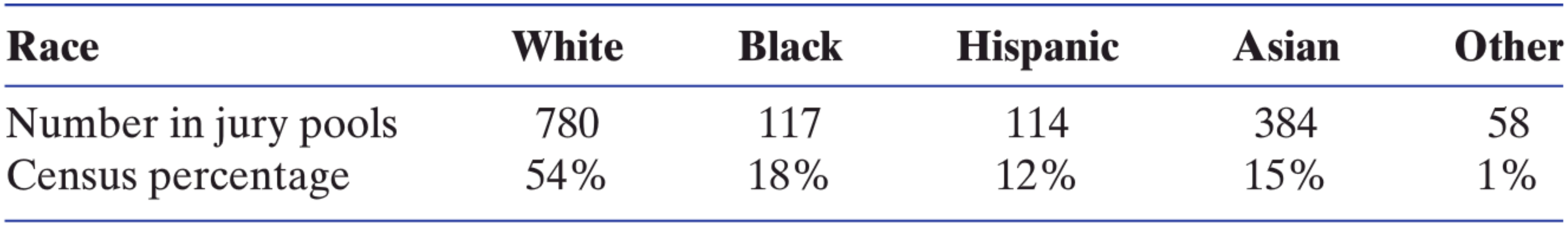
Find the p-value for χ^2 using the upper tail of a chi-square distribution with $k - 1$ degrees of freedom

The chi-square distribution is appropriate if the sample size is large enough that each of the expected counts is at least 5

Alameda County Jury Pools

Prospective jurors are supposed to be drawn at random from eligible adults

The ACLU conducted an analysis of jury pools for 10 trials in Almada County, CA to see if the racial makeup of these jury pools was consistent with the population



Race	White	Black	Hispanic	Asian	Other
Number in jury pools	780	117	114	384	58
Census percentage	54%	18%	12%	15%	1%

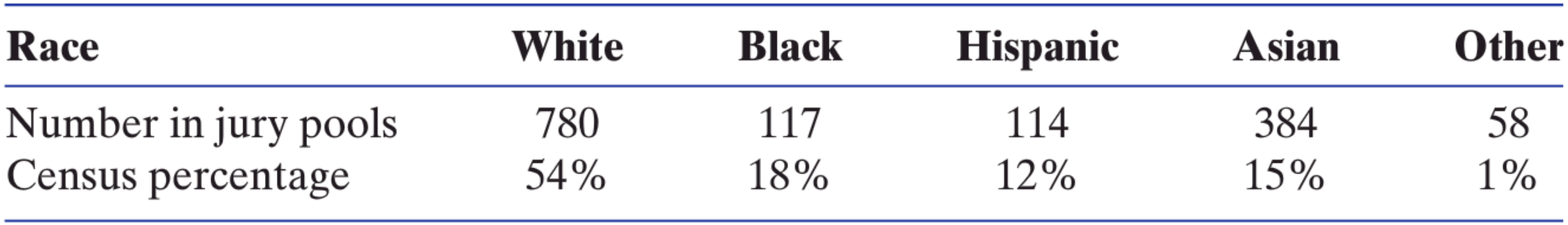
Write down the null and alternative hypotheses

There is a total of n = 1,453 people in this sample
What is the expected number of people in each racial group?

Alameda County Jury Pools

H_0: $\pi_w = .54$, $\pi_b = .18$, $\pi_h = .12$, $\pi_a = .15$, $\pi_o = .01$

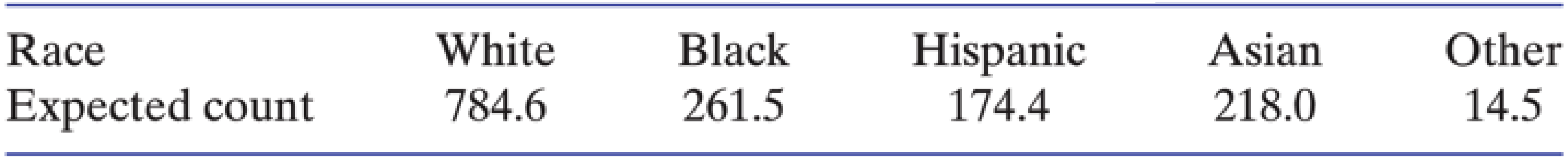
H_A: Some π_i is not as specified in H_0



Race	**White**	**Black**	**Hispanic**	**Asian**	**Other**
Number in jury pools	780	117	114	384	58
Census percentage	54%	18%	12%	15%	1%

Expected number for group *i* is: $n * \pi_i$

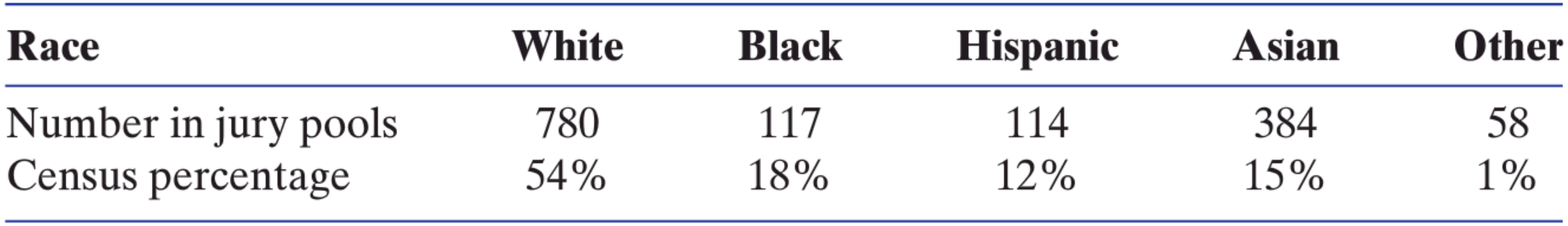
- Expected count white



Race	White	Black	Hispanic	Asian	Other
Expected count	784.6	261.5	174.4	218.0	14.5

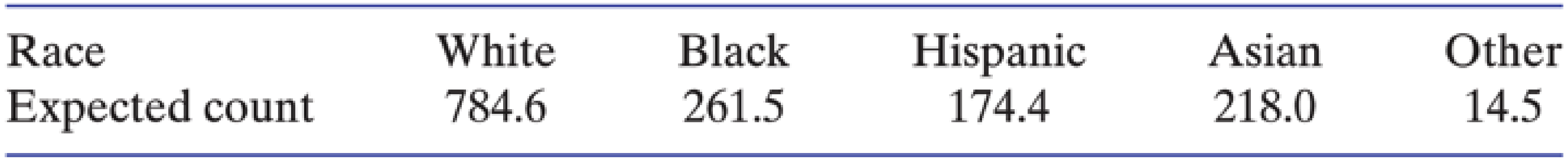
Assess whether Alameda County juries are representative of the population

Observed



Race	White	Black	Hispanic	Asian	Other
Number in jury pools	780	117	114	384	58
Census percentage	54%	18%	12%	15%	1%

Expected



Race	White	Black	Hispanic	Asian	Other
Expected count	784.6	261.5	174.4	218.0	14.5

$$\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Can get the p-value in R using:

```
pchisq(chi_squared_stat, df, lower.tail = FALSE)
```

Let’s try it in R…

Chi-squared test for an association between categorical variables

Testing for an association between categorical variables

One of my former students, Thomas, was interested in knowing whether people who believe in God have different levels of life satisfaction compared to people who do not believe in God.

To address this question Thomas conducted a survey where he asked people whether they believe in God, and also how satisfied they were with their lives. Below are the results from the survey:

Life satisfaction	Does not believe in God	Believes in God	Totals
≤ 5	11	16	**27**
6	11	16	**27**
7	13	20	**33**
≥ 8	11	30	**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

What are the null and alternative hypotheses?

Life satisfaction	Does not believe in God	Believes in God	Totals
≤ 5	11	16	**27**
6	11	16	**27**
7	13	20	**33**
≥ 8	11	30	**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

What are the null and alternative hypotheses?

H_0: There is no an association between happiness and belief in God

H_A: There is an association between happiness and belief in God

Life satisfaction	**Does not believe in God**	**Believes in God**	**Totals**
≤ 5	11	16	**27**
6	11	16	**27**
7	13	20	**33**
≥ 8	11	30	**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

Life satisfaction	Does not believe in God	Believes in God	Totals
≤ 5			**27**
6			**27**
7			**33**
≥ 8			**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

If there is no association, the probability of each cell is:

Pr(R, C) = Pr(R) * Pr(C)

Life satisfaction	**Does not believe in God**	**Believes in God**	**Totals**
≤ 5			**27**
6			**27**
7			**33**
≥ 8			**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

If there is no association, the probability of each cell is:

$$\Pr(R, C) = \Pr(R) * \Pr(C)$$

Life satisfaction	**Does not believe in God**	**Believes in God**	**Proportions**
≤ 5			**27/128 = .21**
6			**27 = .21**
7			**33 = .26**
≥ 8			**41 = .32**
Proportions	**46/128 = .36**	**82/128 = .64**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

If there is no association, the probability of each cell: is

$$Pr(R, C) = Pr(R) * Pr(C)$$

Life satisfaction	**Does not believe in God**	**Believes in God**	**Proportions**
≤ 5	.076	.134	**27/128 = .21**
6	.076	.134	**27 = .21**
7	.093	.166	**33 = .26**
≥ 8	.115	.205	**41 = .32**
Proportions	**46/128 = .36**	**82/128 = .64**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

The counts in each cell are then just: Pr(R, C) * n

Life satisfaction	Does not believe in God	Believes in God	Proportions
≤ 5	.076	.134	**27/128 = .21**
6	.076	.134	**27 = .21**
7	.093	.166	**33 = .26**
≥ 8	.115	.205	**41 = .32**
Proportions	**46/128 = .36**	**82/128 = .64**	**128**

Is there a relationship between belief in God and life satisfaction?

If the null hypothesis was true and there was no association, what would the expected counts in the table look like?

The counts in each cell are then just: Pr(R, C) * n

Life satisfaction	**Does not believe in God**	**Believes in God**	**Totals**
≤ 5	9.70	17.30	**27**
6	9.70	17.30	**27**
7	11.86	21.14	**33**
≥ 8	14.73	26.27	**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

Note: the counts in each cell are: Pr(R, C) * n

= Pr(R) * Pr(C) * n

= row total/n * col total/n * n

= (row total * col total)/n

(27 * 46)/128 = 9.70

Life satisfaction	Does not believe in God	Believes in God	Totals
≤ 5	9.70	17.30	**27**
6	9.70	17.30	**27**
7	11.86	21.14	**33**
≥ 8	14.73	26.27	**41**
Totals	**46**	**82**	**128**

Is there a relationship between belief in God and life satisfaction?

Expected counts can be found by: ***(row total * col total)/sample size***

Life satisfaction	**Does not believe in God**	**Believes in God**	**Totals**
≤ 5	9.70	17.30	**27**
6	9.70	17.30	**27**
7	11.86	21.14	**33**
≥ 8	14.73	26.27	**41**
Totals	**46**	**82**	**128**

Calculating the observed statistic

Observed

Life satisfaction	Does not believe in God	Believes in God	
≤ 5	11	16	**27**
6	11	16	**27**
7	13	20	**33**
≥ 8	11	30	**41**
Totals	**46**	**82**	**128**

Expected

Life satisfaction	Does not believe in God	Believes in God	
≤ 5	9.70	17.30	**27**
6	9.70	17.30	**27**
7	11.86	21.14	**33**
≥ 8	14.73	26.27	**41**
Totals	**46**	**82**	**128**

With these two tables we can then compute a chi-squared statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

$\chi^2 = 2.19$

Degrees of freedom

We know our null distribution for our χ^2 statistic comes from a χ^2 distribution

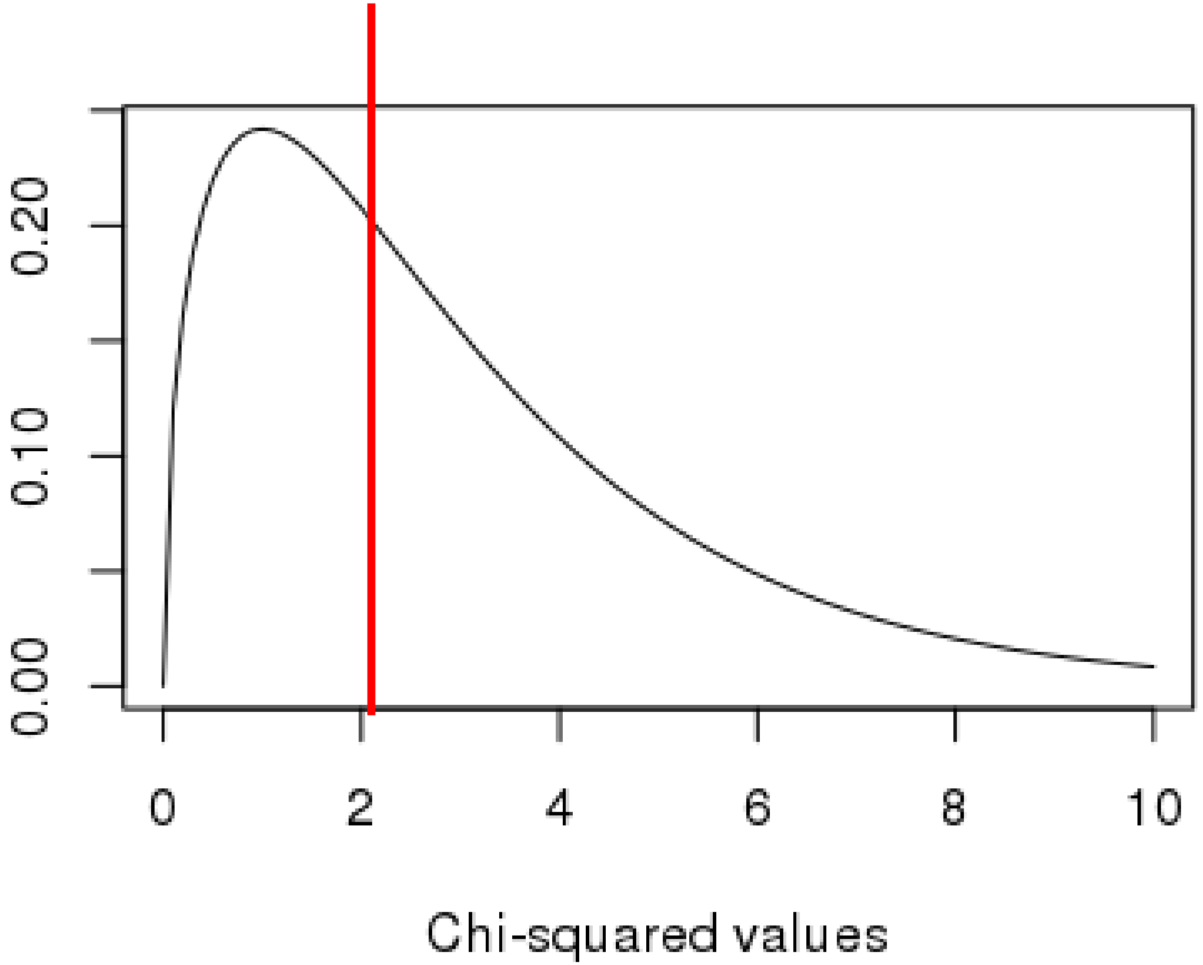
In order to find the p-value we need to know the degrees of freedom

For two-way tables, the degrees of freedom are:

(number of rows – 1) * (number of columns – 1)

For this problem we have: (4 – 1) * (2 - 1)

The null distribution



$\chi^2 = 2.19$

p-value = 0.534

Summary: Chi-square Test for Association

To test for an association between two categorical variables A and B based on a two-way table:

Setup hypotheses:

- H_0: There is no association between A and B
- H_A: There is an association between A and B

Compute the expected count for each cell in the table:

- Expected count = ((row total) * (column total))/sample size

Compute the value of the chi-squared statistic using: $$\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Find the p-value using the upper tail of a chi-square distribution with (row – 1) * (column – 1) degrees of freedom

This test can be used if all cells have at least 5 entries in them

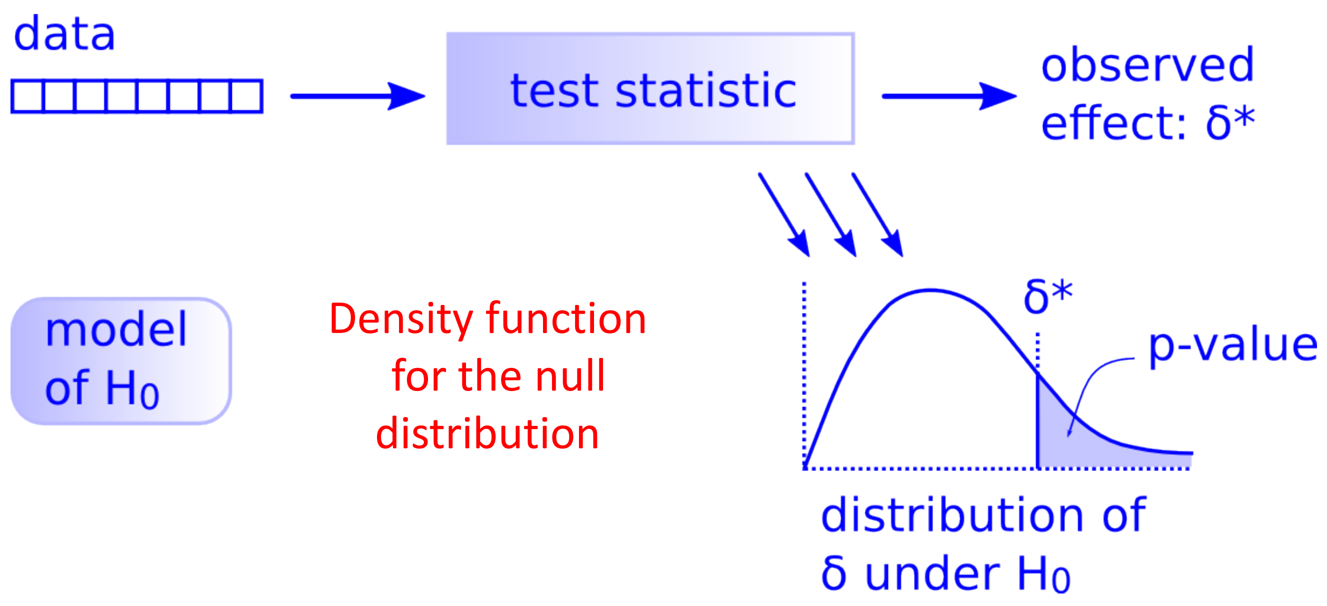
Let’s try it in R…

Parametric test for comparing more than one mean

One-way analysis of variance (ANOVA)

One test to rule them all

There is only one hypothesis test!



data

test statistic

observed effect: δ*

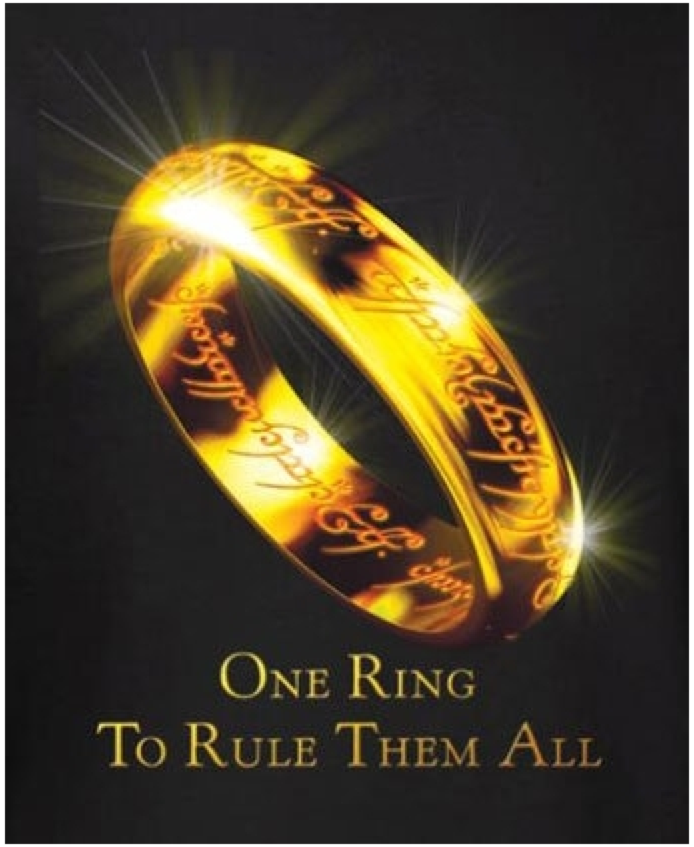
model of H_0

Density function for the null distribution

δ*

p-value

distribution of δ under H_0



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

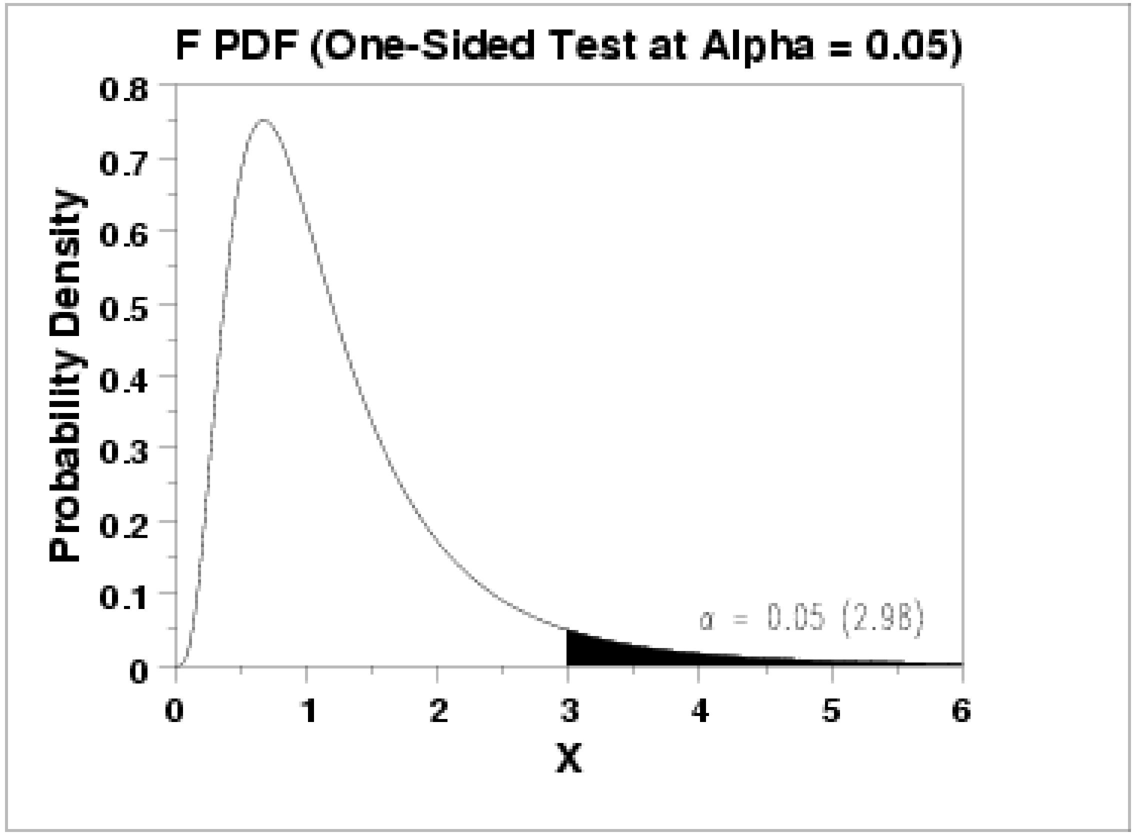
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

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

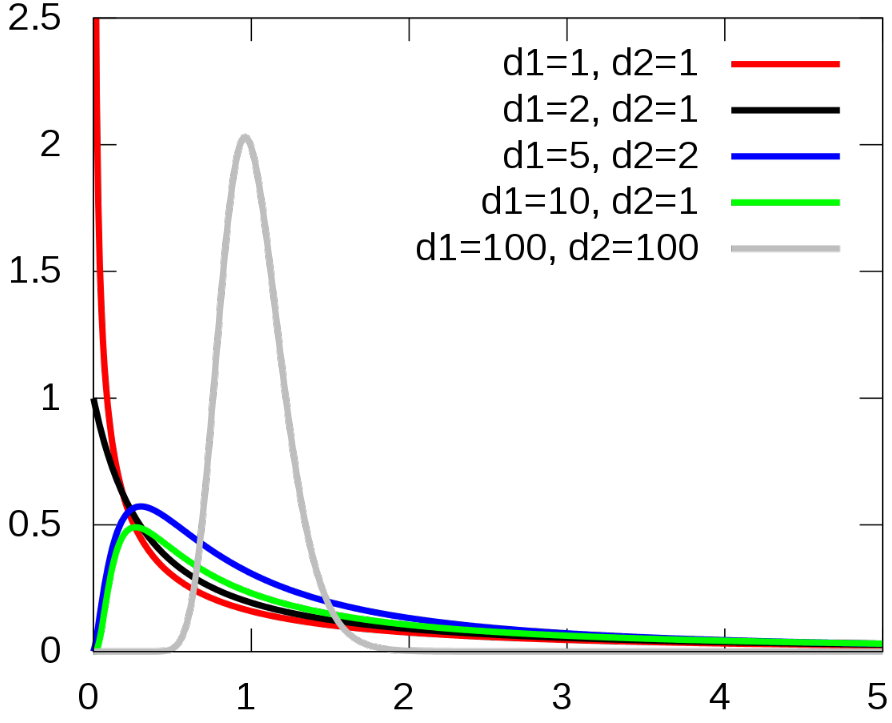


One-way ANOVA – the central idea

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

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

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

What should we do next?

Let’s check if the ANOVA conditions are met first…

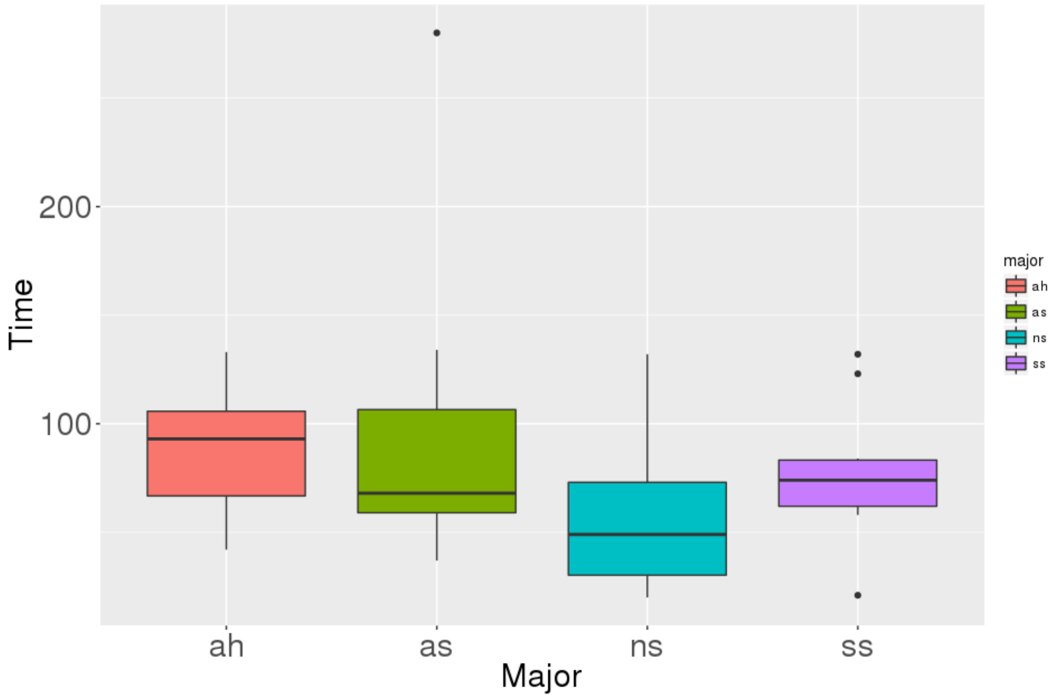
Checking ANOVA conditions (‘assumptions’)

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

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



$s_{ah} = 27.9$ $s_{ns} = 34.39$

$s_{as} = 71.59$ $s_{as} = 31.89$

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 jth 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 jth student’s time from the ith 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 SDS100 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(SDS100)

download_data("MajorPuzzle.txt")

sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)
```

Let’s try this analysis in R…

```
# get the data
library(SDS100)
download_class_data("MajorPuzzle.txt")
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

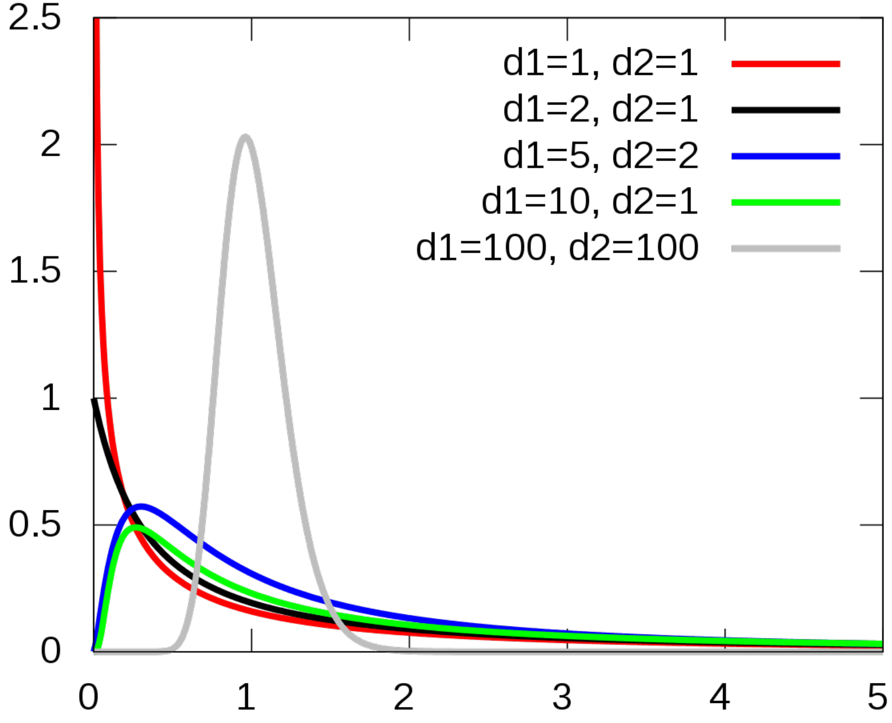
3. Plot the null distribution

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

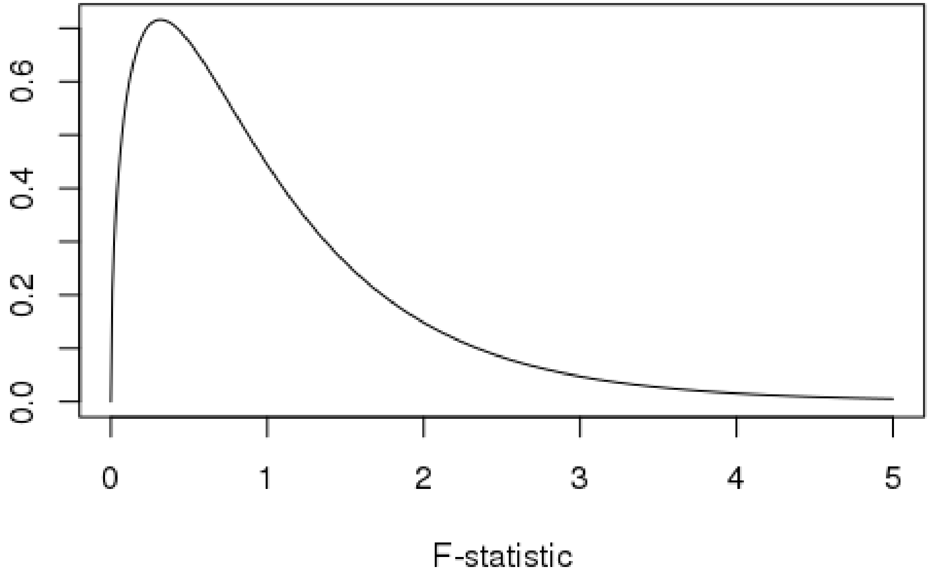
In R we can plot the density function for an F-distribution using the function…

- 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, by = .01)
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…

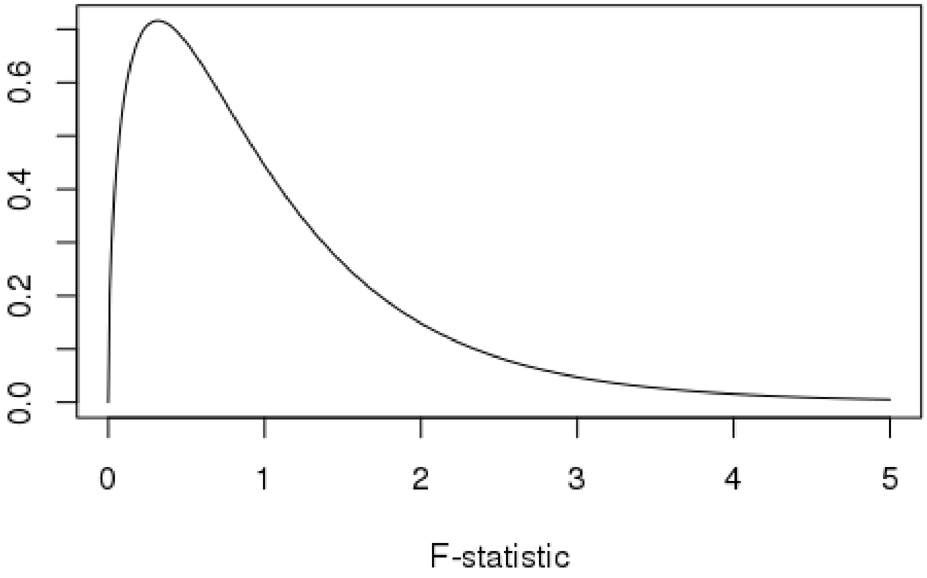
- df(x_vals, df1, df2)

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…

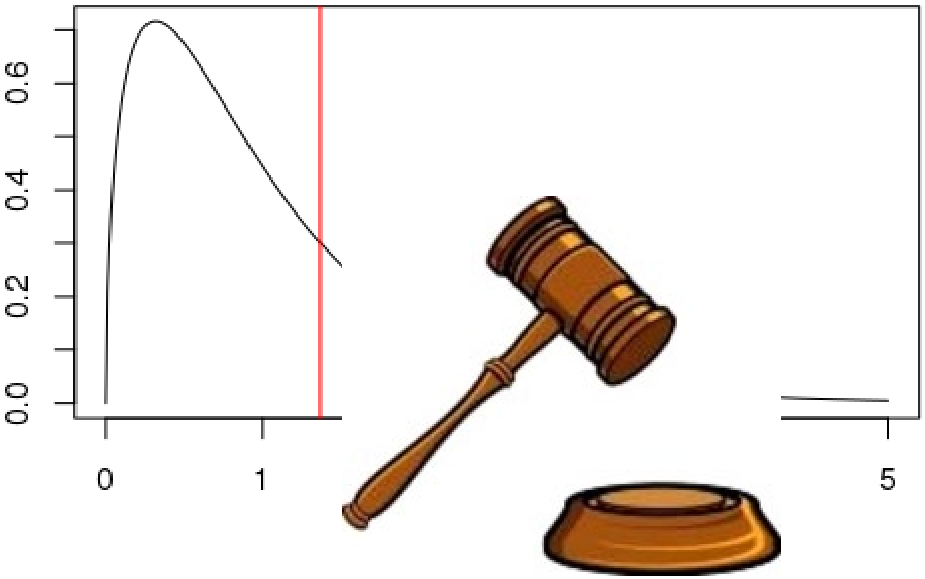
- df(x_vals, df1, df2)

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



Permutation test comparing multiple means using the F-Statistic

Sudoku by field

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

Thoughts on the statistic of interest?

Comparing multiple means

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

1. <u>Group range statistic:</u>

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

2. <u>Mean absolute difference (MAD):</u>

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

3. <u>F statistic:</u>

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

The MAD statistic vs. the F-Statistic

Mean absolute difference (MAD):

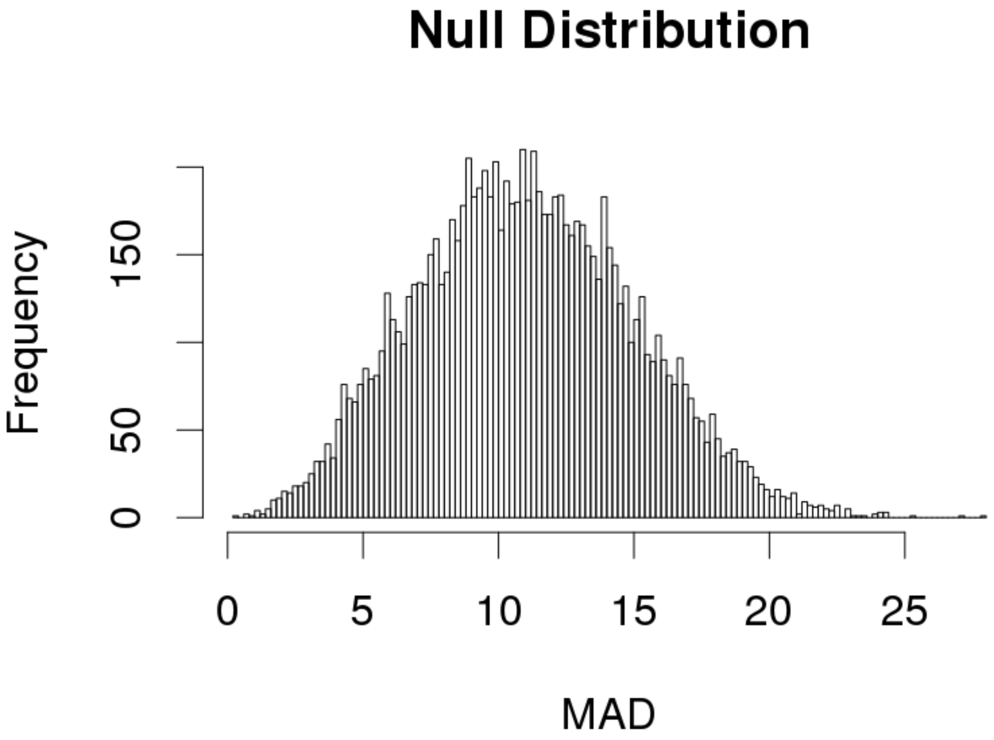
$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

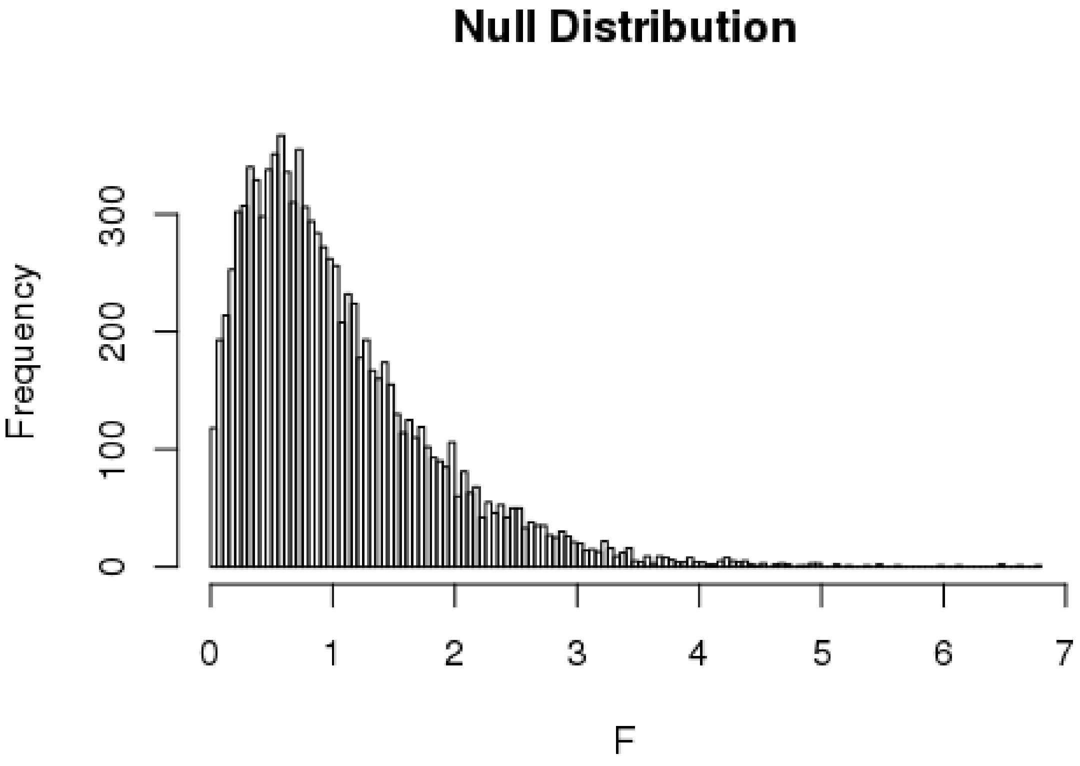
Observed MAD statistic value = 13.92

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

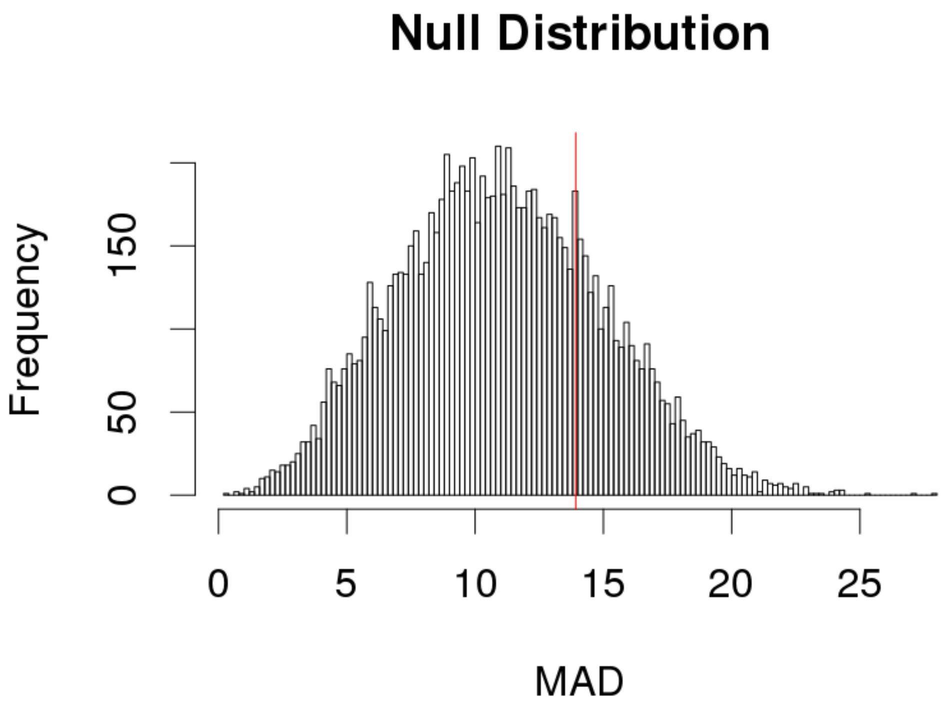
Observed F statistic value = 1.370

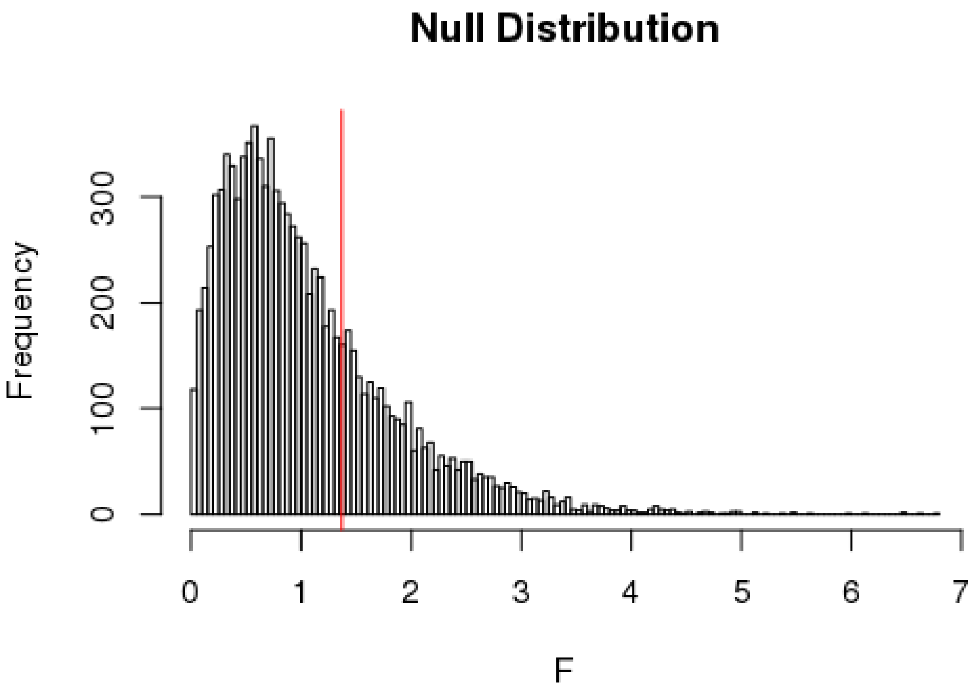
Null distributions





P-value

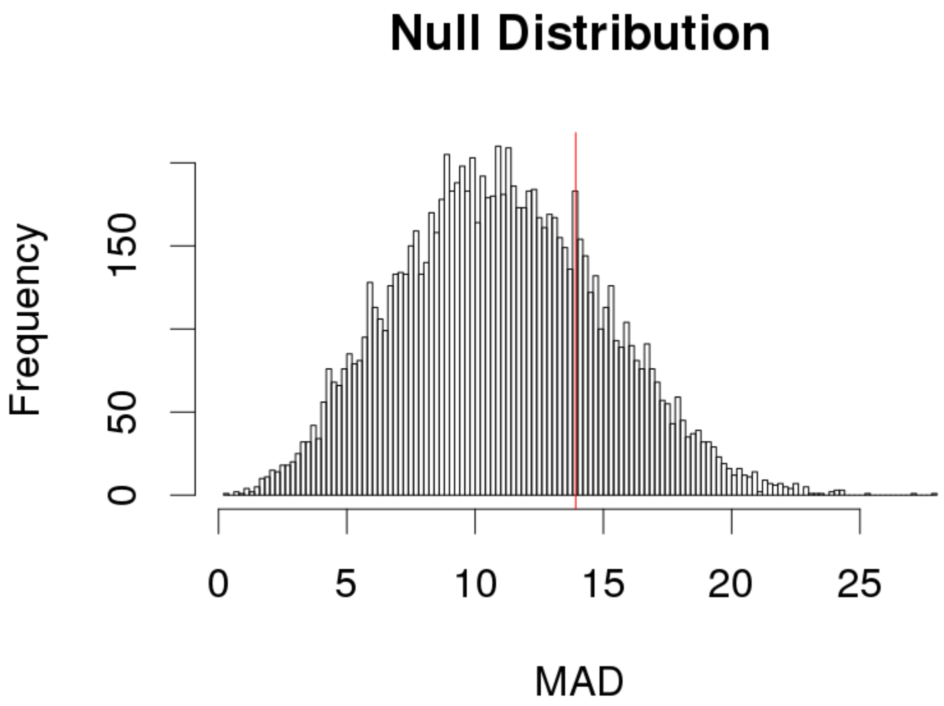


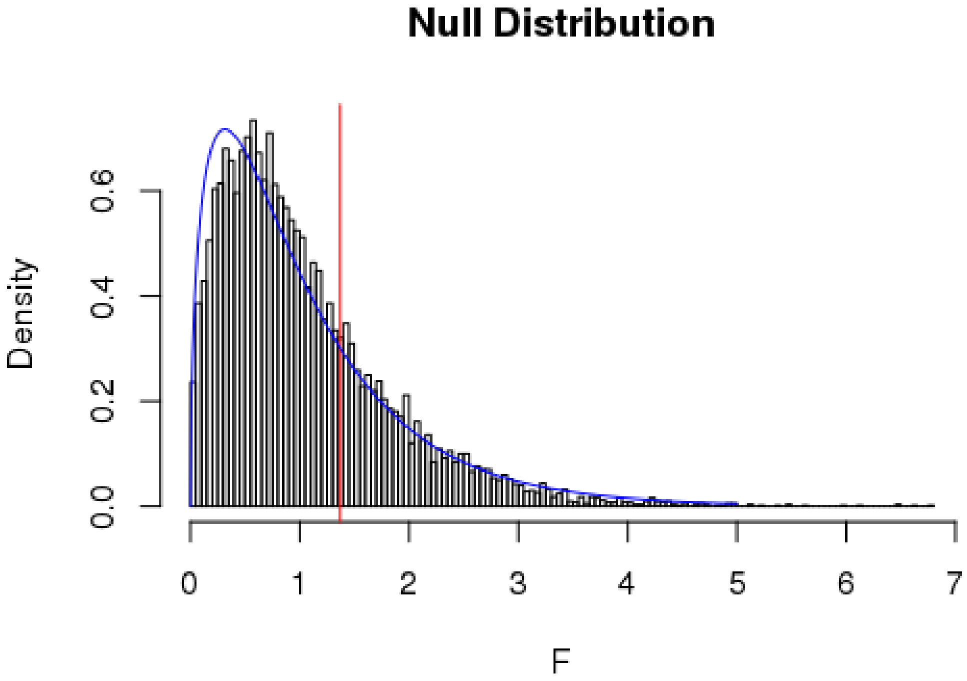


p-value = .4682

p-value = 0.2653

P-value





p-value = .4682

p-value = 0.2674

Conclusions?