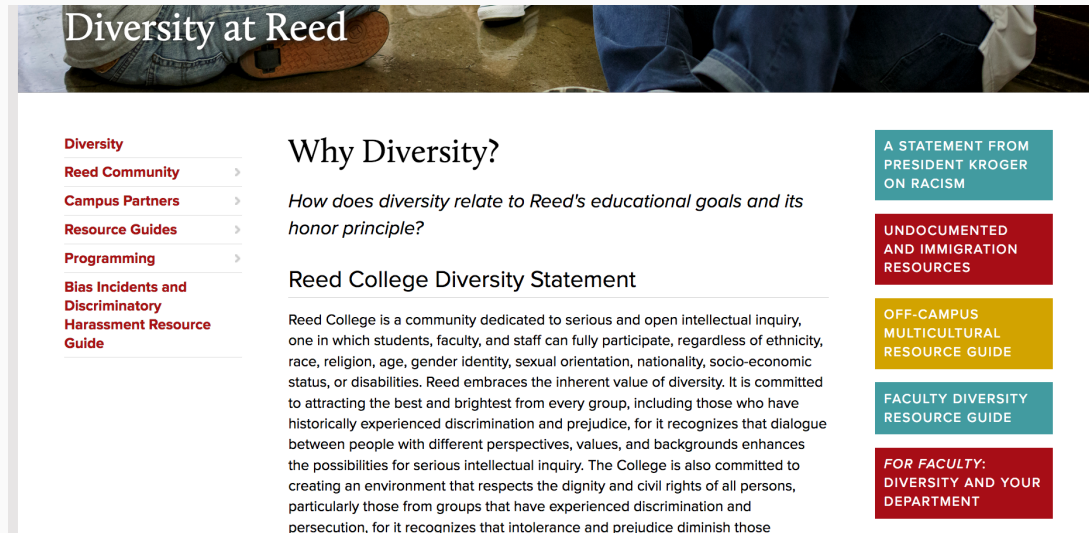


# Chi-squared Goodness of Fit

## Ex: Diversity at Reed



The screenshot shows the Reed College Diversity website. At the top is a banner image with the text "Diversity at Reed". Below the banner is a navigation menu on the left with links: Diversity, Reed Community, Campus Partners, Resource Guides, Programming, Bias Incidents and Discriminatory Harassment Resource Guide. The main content area has a heading "Why Diversity?" followed by a paragraph: "How does diversity relate to Reed's educational goals and its honor principle?". Below this is a heading "Reed College Diversity Statement" followed by a paragraph: "Reed College is a community dedicated to serious and open intellectual inquiry, one in which students, faculty, and staff can fully participate, regardless of ethnicity, race, religion, age, gender identity, sexual orientation, nationality, socio-economic status, or disabilities. Reed embraces the inherent value of diversity. It is committed to attracting the best and brightest from every group, including those who have historically experienced discrimination and prejudice, for it recognizes that dialogue between people with different perspectives, values, and backgrounds enhances the possibilities for serious intellectual inquiry. The College is also committed to creating an environment that respects the dignity and civil rights of all persons, particularly those from groups that have experienced discrimination and persecution, for it recognizes that intolerance and prejudice diminish those". On the right side of the main content area are five colored boxes with text: "A STATEMENT FROM PRESIDENT KROGER ON RACISM", "UNDOCUMENTED AND IMMIGRATION RESOURCES", "OFF-CAMPUS MULTICULTURAL RESOURCE GUIDE", "FACULTY DIVERSITY RESOURCE GUIDE", and "FOR FACULTY: DIVERSITY AND YOUR DEPARTMENT".

### Diversity at Reed

- Diversity
- Reed Community
- Campus Partners
- Resource Guides
- Programming
- Bias Incidents and Discriminatory Harassment Resource Guide

### Why Diversity?

*How does diversity relate to Reed's educational goals and its honor principle?*

### Reed College Diversity Statement

Reed College is a community dedicated to serious and open intellectual inquiry, one in which students, faculty, and staff can fully participate, regardless of ethnicity, race, religion, age, gender identity, sexual orientation, nationality, socio-economic status, or disabilities. Reed embraces the inherent value of diversity. It is committed to attracting the best and brightest from every group, including those who have historically experienced discrimination and prejudice, for it recognizes that dialogue between people with different perspectives, values, and backgrounds enhances the possibilities for serious intellectual inquiry. The College is also committed to creating an environment that respects the dignity and civil rights of all persons, particularly those from groups that have experienced discrimination and persecution, for it recognizes that intolerance and prejudice diminish those

- A STATEMENT FROM PRESIDENT KROGER ON RACISM
- UNDOCUMENTED AND IMMIGRATION RESOURCES
- OFF-CAMPUS MULTICULTURAL RESOURCE GUIDE
- FACULTY DIVERSITY RESOURCE GUIDE
- FOR FACULTY: DIVERSITY AND YOUR DEPARTMENT

In terms of ethnic diversity, how does the first year student body compare to the general population of Oregon?

# Facts about Reed

BROUGHT TO YOU BY INSTITUTIONAL RESEARCH

[Institutional Research Home](#)

[Students](#)

[Alumni](#)

[Graduation Rates](#)

[Outcomes](#)

[Institutional](#)

[Faculty](#)

[Finance](#)

[Other Statistics](#)

[IR Grant Activities](#)

[Contact us](#)

## First-year Students Ethnicity–2019

	Asian	Black	Hispanic	Internat'l	Native Amer	Pacific Islander	White	Unknown	Total	Percent
Women	29	6	19	17	8	0	126	4	209	53%
Men	29	7	9	23	3	0	113	1	185	47%
Total	58	13	28	40	11	0	239	5	394	100%
Percent	15%	3%	7%	10%	3%	0%	61%	1%	100%	

Note: Updated September 23, 2019.

## Oregon

Want more? [Browse data sets for Oregon](#)

People QuickFacts	Oregon	USA
<i>i</i> Population, 2014 estimate	3,970,239	318,857,056
<i>i</i> Population, 2010 (April 1) estimates base	3,831,073	308,758,105
<i>i</i> Population, percent change - April 1, 2010 to July 1, 2014	3.6%	3.3%
<i>i</i> Population, 2010	3,831,074	308,745,538
<i>i</i> Persons under 5 years, percent, 2014	5.8%	6.2%
<i>i</i> Persons under 18 years, percent, 2014	21.6%	23.1%
<i>i</i> Persons 65 years and over, percent, 2014	16.0%	14.5%
<i>i</i> Female persons, percent, 2014	50.5%	50.8%
<hr/>		
<i>i</i> White alone, percent, 2014 (a)	87.9%	77.4%
<i>i</i> Black or African American alone, percent, 2014 (a)	2.0%	13.2%
<i>i</i> American Indian and Alaska Native alone, percent, 2014 (a)	1.8%	1.2%
<i>i</i> Asian alone, percent, 2014 (a)	4.3%	5.4%
<i>i</i> Native Hawaiian and Other Pacific Islander alone, percent, 2014 (a)	0.4%	0.2%
<i>i</i> Two or More Races, percent, 2014	3.6%	2.5%
<i>i</i> Hispanic or Latino, percent, 2014 (b)	12.5%	17.4%
<i>i</i> White alone, not Hispanic or Latino, percent, 2014	77.0%	62.1%

## The data

<b>Ethnicity</b>	<b>Asian</b>	<b>Black</b>	<b>Hispanic</b>	<b>White</b>	<b>Other</b>	<b>Total</b>
Reed count	58	13	28	239	51	394
Oregon %	.043	.02	.125	.77	.042	1

If the students at Reed were drawn from a population with these proportions, how many *counts* would we expect in each group?

$$\text{exp. count} = n \times p_i$$

## The data

<b>Ethnicity</b>	<b>Asian</b>	<b>Black</b>	<b>Hispanic</b>	<b>White</b>	<b>Other</b>	<b>Total</b>
Obs. count	58	13	28	239	51	394
Exp. count	16.94	7.88	49.25	303.38	16.548	394

- Some sampling variability is expected, but how far from expected is too far?

# Simulating Oregonian Reedies

```
n <- 354
p <- c(.043, .02, .125, .77, .042)
samp <- sample(c("asian", "black", "hispanic", "white", "other"),
               size = n,
               replace = TRUE,
               prob = p) %>%
  factor(levels = c("asian", "black", "hispanic", "white", "other"))
table(samp)
```

```
## samp
##      asian      black hispanic      white      other
##         20          9         39        274         12
```

```
obs <- c(58, 13, 28, 239, 51)
```

# Simulating Oregonian Reedies, again

```
samp <- sample(c("asian", "black", "hispanic", "white", "other"),
              size = n,
              replace = TRUE,
              prob = p) %>%
  factor(levels = c("asian", "black", "hispanic", "white", "other"))
table(samp)
```

```
## samp
##      asian      black hispanic      white      other
##         20          4         34        284         12
```

```
obs <- c(58, 13, 28, 239, 51)
```



# Simulating Oregonian Reedies, again again

```
samp <- sample(c("asian", "black", "hispanic", "white", "other"),
              size = n,
              replace = TRUE,
              prob = p) %>%
  factor(levels = c("asian", "black", "hispanic", "white", "other"))
table(samp)
```

```
## samp
##      asian      black hispanic      white      other
##         16          7         40        271         20
```

```
obs <- c(58, 13, 28, 239, 51)
```

# Simulating Oregonian Reedies

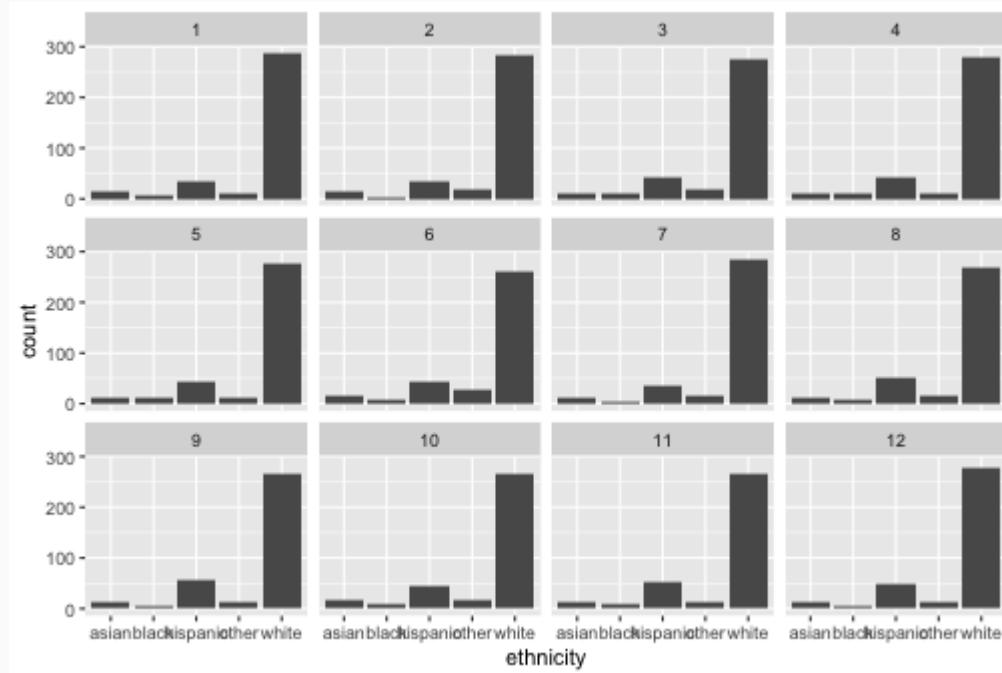
```
sim_reedies <- reed_demos %>%  
  specify(response = ethnicity) %>%  
  hypothesize(null = "point", p = c("asian"      = .043,  
                                     "black"      = .02,  
                                     "hispanic"    = .125,  
                                     "white"       = .77,  
                                     "other"       = .042)) %>%  
  generate(reps = 12, type = "simulate")  
sim_reedies
```

```
## Response: ethnicity (factor)  
## Null Hypothesis: point  
## # A tibble: 4,248 x 2  
## # Groups:   replicate [12]  
##   ethnicity replicate  
##   <fct>         <fct>  
## 1 white         1  
## 2 white         1  
## 3 white         1  
## 4 asian         1  
## 5 white         1  
## 6 white         1  
## 7 white         1  
## 8 black         1  
## 9 white         1  
## 10 white        1
```

# Visualizing our Simulated Oregonian Reedies

```
sim_reedies %>%  
  ggplot(aes(x = ethnicity)) +  
  geom_bar() +  
  facet_wrap(vars(replicate))
```

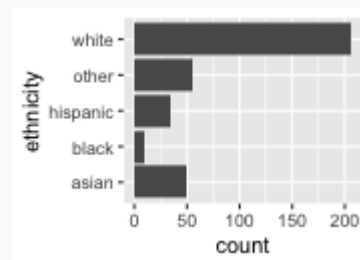
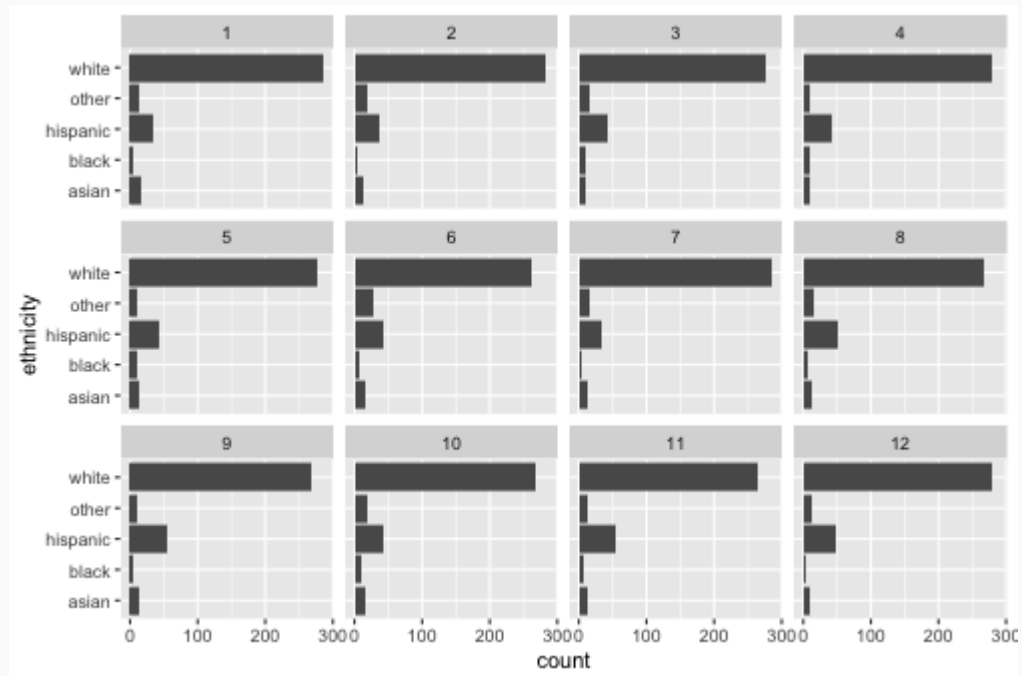
# Visualizing our Simulated Oregonian Reedies



# Visualizing our Simulated Oregonian Reedies

```
sim_reedies %>%  
  ggplot(aes(x = ethnicity)) +  
  geom_bar() +  
  facet_wrap(vars(replicate)) + coord_flip()
```

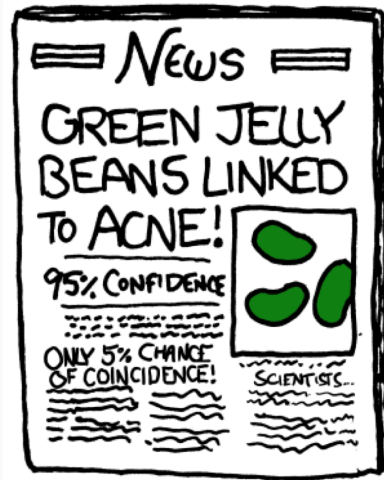
# Simulated vs Observed

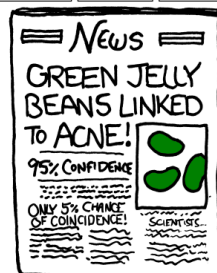
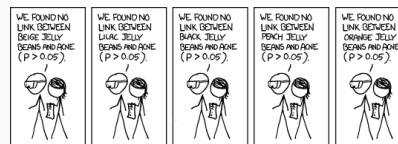
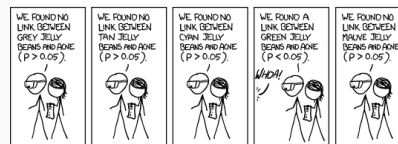
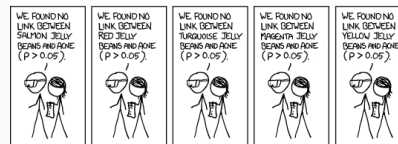
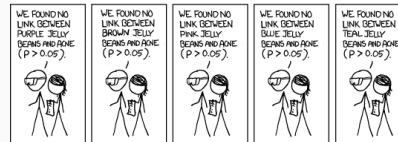
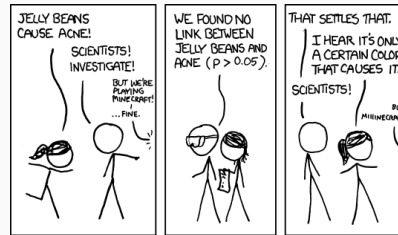


# Inference on many ps

We *could* do a tests/CIs on  $p_{reed} - p_{oregon}$  for each group, however:

- We have the whole population of Oregon.
- Beware of multiple comparisons!

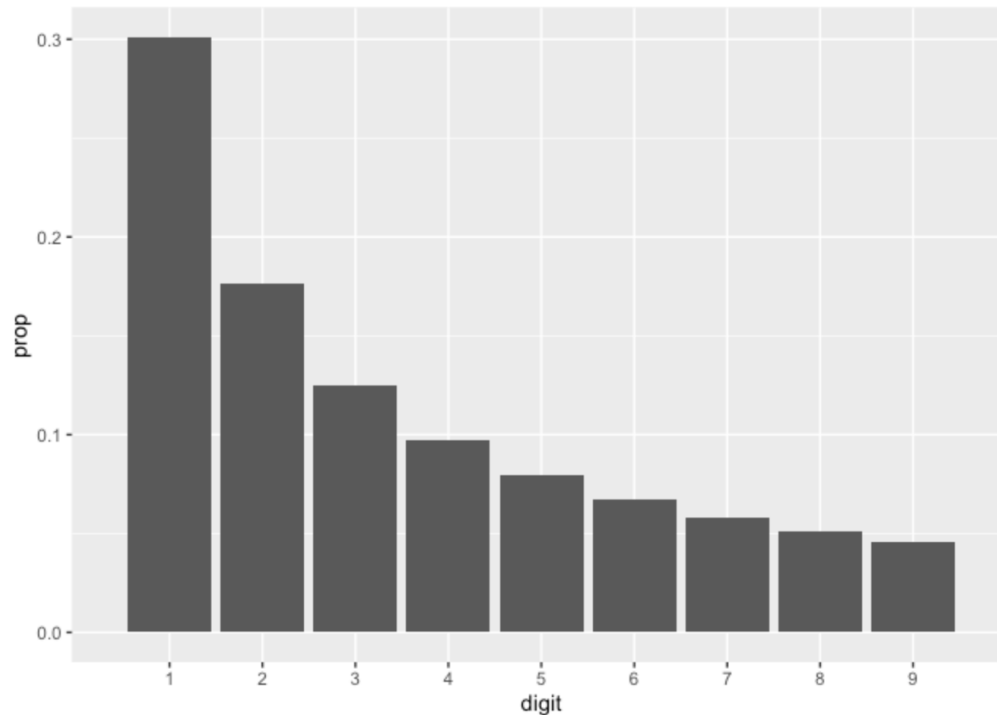






# Sound familiar? Benford's Law!

```
benfords_p <- data.frame(first_digit = 1:9,  
  ben_prop = log10(1 + 1/1:9))
```



6. Formulate your own statistic to measure the distance between the observed proportions ( `obs_prop` ) and those expected by Benford's Law ( `ben_prop` ). There are many many possible choices, but some are more useful than others. Describe this statistic in words (or write out the formula for it if you are comfortable using LaTeX), then calculate it for this data.

# Creating a statistic

# Creating a statistic

For each of  $k$  categories:

1. Calculate the difference between observed and expected counts.
2. Scale each difference by an estimate of the SE (  $\sqrt{exp}$  ).
3. Square the scaled difference to get rid of negatives.

Then add them all up.

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

## Reed Data

Ethnicity	Asian	Black	Hispanic	White	Other	Total
Obs. count	49	10	34	206	55	354
Exp. count	15.22	7.08	44.25	272.58	14.87	354

$$Z_{asian}^2 = (49 - 15.22)^2 / 15.22 = 74.97$$

$$Z_{black}^2 = (10 - 7.08)^2 / 7.08 = 1.20$$

$$Z_{hispanic}^2 = (34 - 51.5)^2 / 51.5 = 5.95$$

$$Z_{white}^2 = (206 - 272.58)^2 / 272.58 = 16.26$$

$$Z_{other}^2 = (55 - 14.87)^2 / 14.87 = 108.30$$

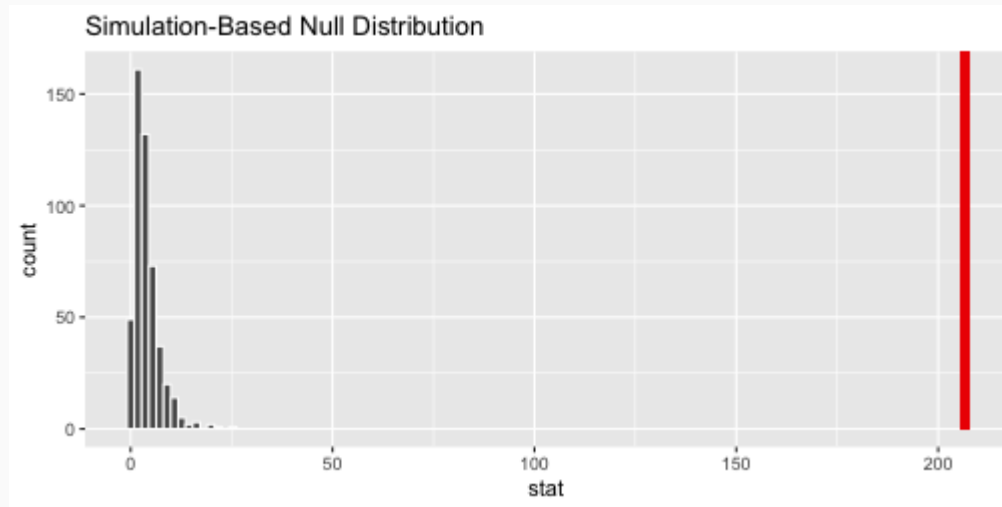
$$Z_{asian}^2 + Z_{black}^2 + Z_{hispanic}^2 + Z_{white}^2 + Z_{other}^2 = 206.68 = \chi_{obs}^2$$

# Simulating $\chi^2$ under $H_0$

```
(null <- reed_demos %>%  
  specify(response = ethnicity) %>%  
  hypothesize(null = "point", p = c("asian"    = .043,  
                                     "black"    = .02,  
                                     "hispanic"  = .125,  
                                     "white"    = .77,  
                                     "other"    = .042)) %>%  
  generate(reps = 500, type = "simulate") %>%  
  calculate(stat = "Chisq"))
```

```
## # A tibble: 500 x 2  
##   replicate stat  
##   <fct>      <dbl>  
## 1 1      8.26  
## 2 2      5.66  
## 3 3      1.41  
## 4 4      1.75  
## 5 5      1.37  
## 6 6      1.90  
## 7 7      4.33  
## 8 8      2.03  
## 9 9      5.53  
## 10 10     3.02  
## # ... with 490 more rows
```

# The null distribution



What is the probability of observing our data or more extreme ( $\chi^2 = 206.68$ ) under the null hypothesis that Reedies share the same ethnicity proportions as Oregon?

About zero.

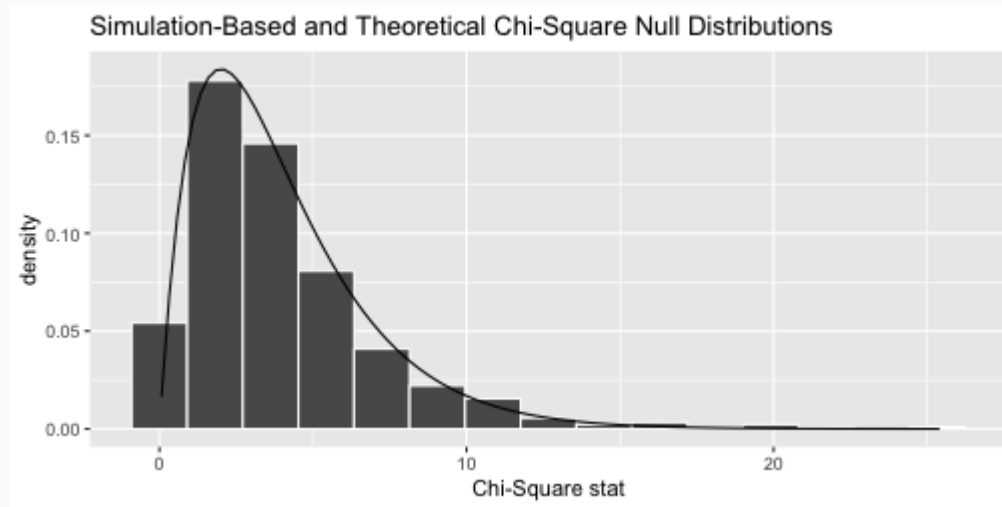
# An alternate path to the null

If...

1. Independent observations
2. Each cell count has a count  $\geq 5$
3.  $k \geq 3$

then our statistic can be well-approximated by the  $\chi^2$  distribution with  $k - 1$  degrees of freedom.

# The null distribution



```
1 - pchisq(206.68, df = 4)
```

```
## [1] 0
```



## Postscript: Great Reed Bake-off 2020

