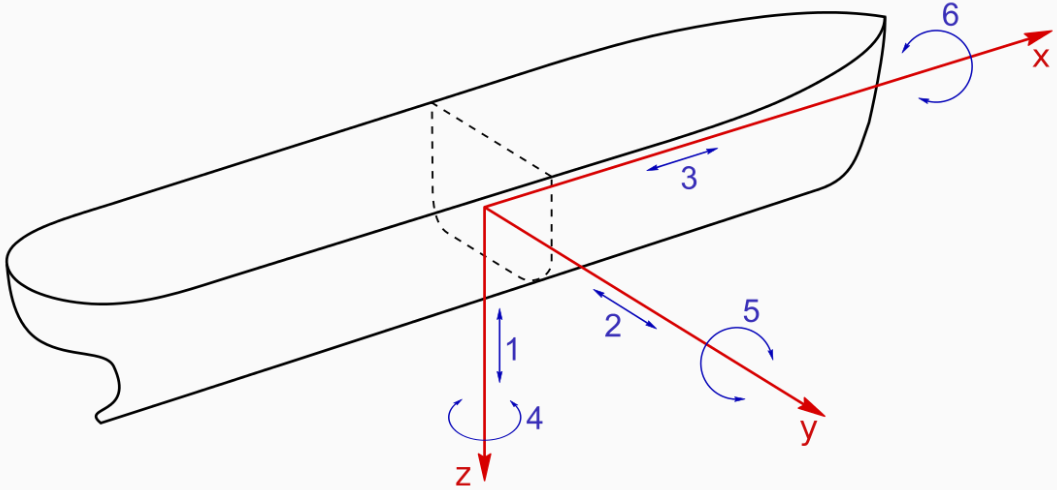
More with Chi-squared

Degrees of Freedom (physics)

The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it. (wikipedia)



Degrees of Freedom (statistics)

The number of parameters that are free to vary, without violating any constraint imposed on it.

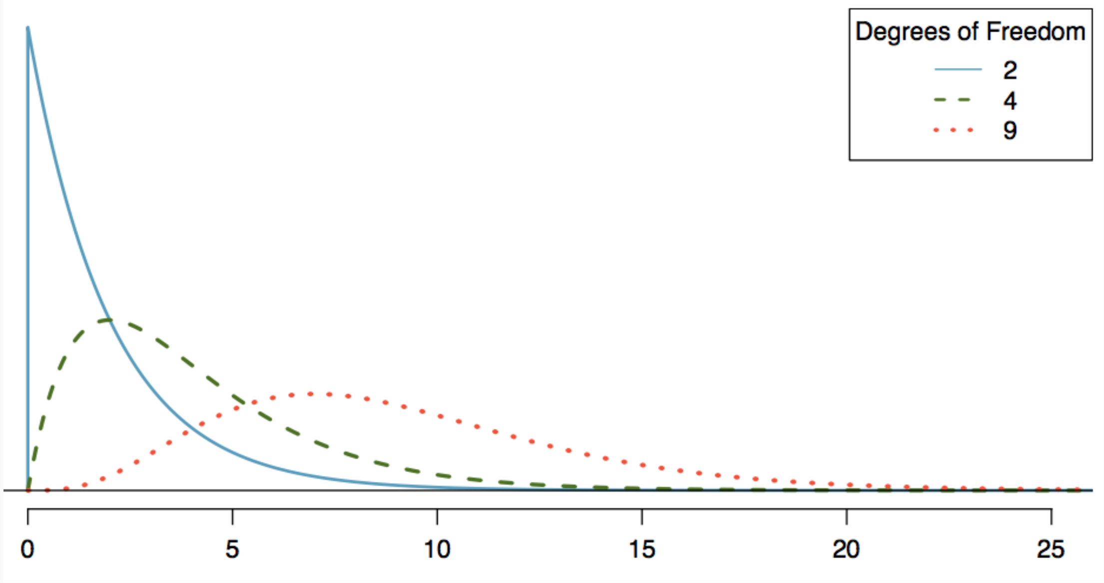
Parameters

$p_{acu}, p_{sham}, p_{trad}$

Since $\sum_{i=1}^{k} p_i = 1$, one of our parameters is contrained, leaving $k - 1$ that are free to vary. Generally:

$$df = (number\ of\ rows - 1) \times (number\ of\ columns - 1)$$

Shape of the χ^2



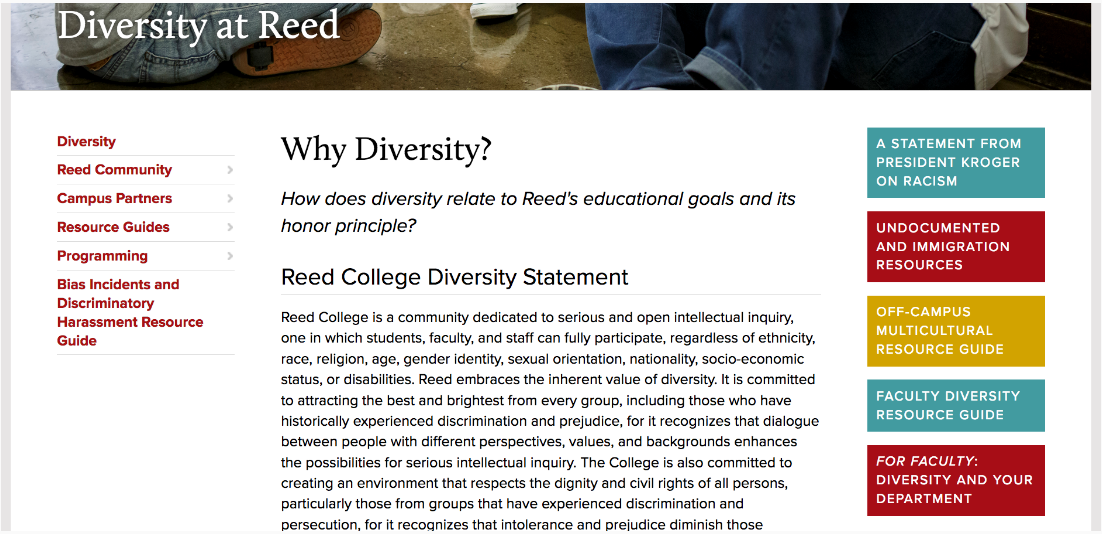
Confidence Intervals for χ^2

These make no sense. Why not?

- We don't really care what the *true* χ^2 parameter value is.
- A two-sided interval wouldn't make sense
  - Distribution is bounded on the left by zero
  - Only "extreme" values are in the right tail

Chi-squared Goodness of Fit

Ex: Diversity at Reed



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

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

The data

Ethnicity	Asian	Black	Hispanic	White	Other	Total
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

Ethnicity	Asian	Black	Hispanic	White	Other	Total
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

```r
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
```

```r
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

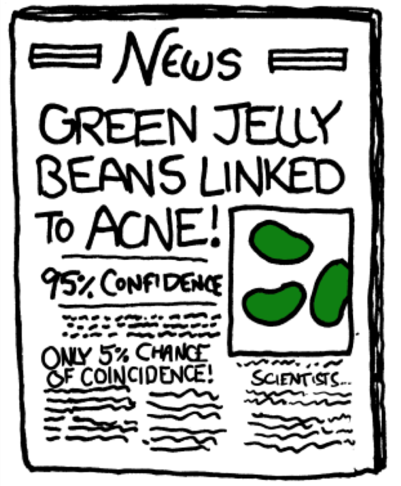
```
reed_demos %>%
  specify(response = ethnicity) %>%
  hypothesize(null = "point", p = c("asian"    = .043,
                                    "black"    = .02,
                                    "hispanic" = .125,
                                    "white"    = .77,
                           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

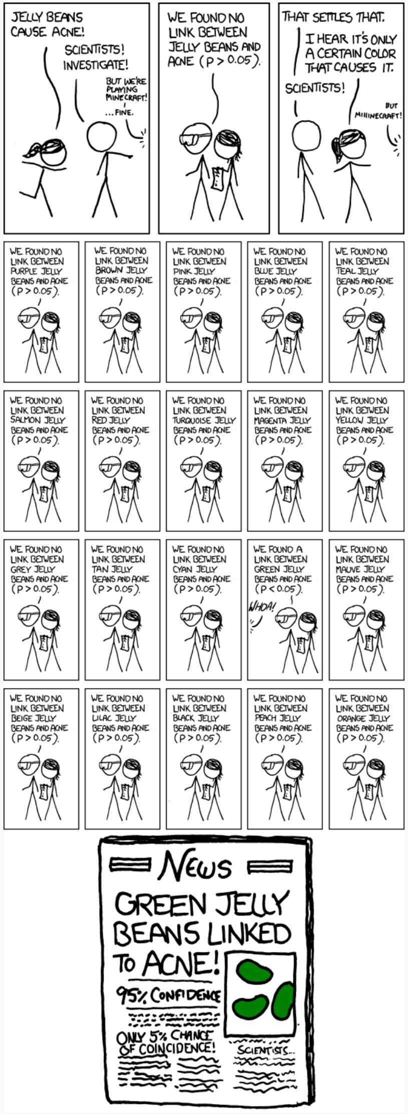
```
## Response: ethnicity (factor)
## Null Hypothesis: point
## # A tibble: 177,000 x 2
## # Groups:   replicate [500]
##    ethnicity replicate
##    <fct>     <fct>
##  1 white     1
##  2 white     1
##  3 white     1
##  4 asian     1
##  5 white     1
##  6 white     1
##  7 white     1
##  8 black     1
##  9 white     1
## 10 white     1
## # … with 176,990 more rows
```

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!




JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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

$$
\begin{aligned}
{}^2_{asian} &= (49 - 15.22)^2/15.22 = 74.97 \\
Z^2_{black} &= (10 - 7.08)^2/7.08 = 1.20 \\
Z^2_{hispanic} &= (34 - 51.5)^2/51.5 = 5.95 \\
Z^2_{white} &= (206 - 272.58)^2/272.58 = 16.26 \\
Z^2_{other} &= (55 - 14.87)^2/14.87 = 108.30
\end{aligned}
$$

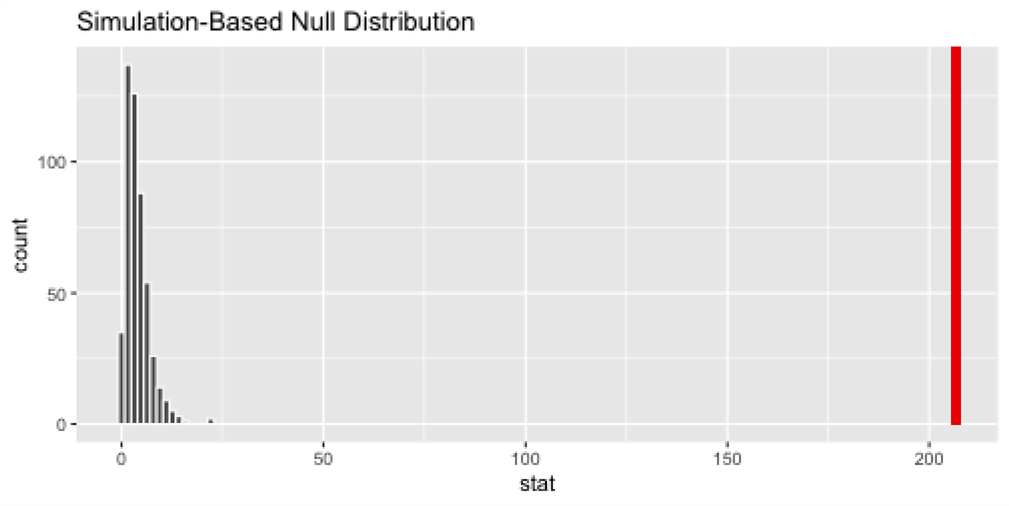
$$
{}^2_{asian} + Z^2_{black} + Z^2_{hispanic} + Z^2_{white} + Z^2_{other} = 206.68 = \chi^2_{obs}
$$

Simulating χ^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          0.395
##  2 2          0.986
##  3 3          5.90
##  4 4          1.16
##  5 5          0.676
##  6 6          3.77
##  7 7          7.53
##  8 8          6.80
##  9 9          5.02
## 10 10        12.7
## # … 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.

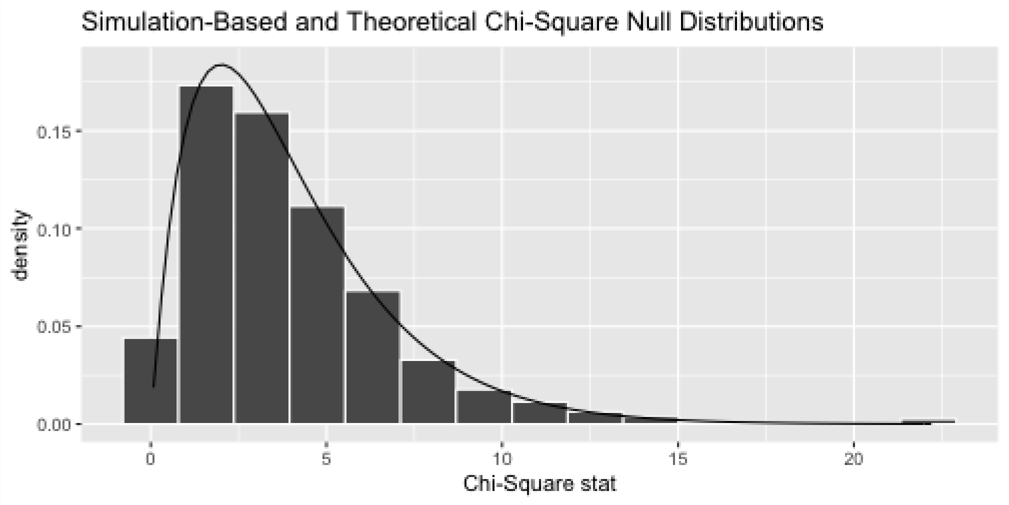
An alternate path to the null

If...

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

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

The null distribution




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

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