

Inferentials with Infer

Week 12







Packages needed and a Note about Icons

Please load up the following packages. Remember to first install the ones you don't have.

```
library(tidyverse)
library(infer)
library(patchwork)
```

You may come across the following icons. The table below lists what each means.

Icon	Description
	Indicates that an example continues on the following slide.
	Indicates that a section using common syntax has ended.
	Indicates that there is an active hyperlink on the slide.
	Indicates that a section covering a concept has ended.



Load up General Social Survey

```
## # A tibble: 6 × 11
##   year  age sex  college partyid hompop hours income class finrela weight
##   <dbl> <dbl> <fct> <fct> <fct> <dbl> <dbl> <ord> <fct> <fct> <dbl>
## 1  2014   36 male  degree  ind     3     50 $25000 ... middl... below a... 0.896
## 2  1994   34 female no degree rep     4     31 $20000 ... worki... below a... 1.08
## 3  1998   24 male  degree  ind     1     40 $25000 ... worki... below a... 0.550
## 4  1996   42 male  no degree ind     4     40 $25000 ... worki... above a... 1.09
## 5  1994   31 male  degree  rep     2     40 $25000 ... middl... above a... 1.08
## 6  1996   32 female no degree rep     4     53 $25000 ... middl... average 1.09
```

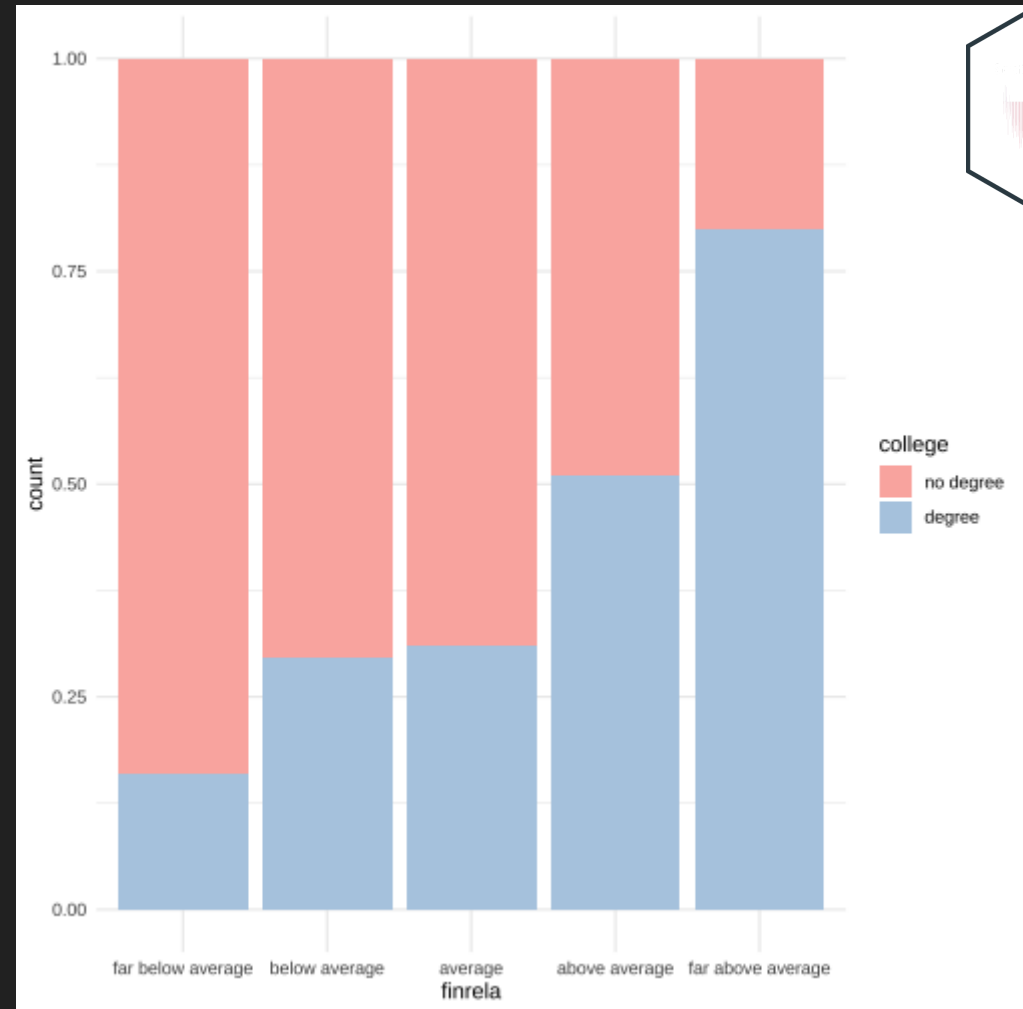
Check the unique values for each

```
## [1] degree      no degree
## Levels: no degree degree
```

```
## [1] $25000 or more $20000 - 24999 $15000 - 19999 $8000 to 9999 $10000 - 14999
## [6] $5000 to 5999 $6000 to 6999 $4000 to 4999 $1000 to 2999 $7000 to 7999
## [11] $3000 to 3999 lt $1000
## 12 Levels: lt $1000 < $1000 to 2999 < $3000 to 3999 < ... < $25000 or more
```

```
## [1] below average      above average      average              far below average
## [5] DK                  far above average
## 6 Levels: far below average below average average ... DK
```

```
gss %>%
  filter(finrela != "DK") %>%
  ggplot() +
  aes(x = finrela, fill = college) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "Pastel1") +
  theme(axis.text.x = element_text(size = 12,
                                     angle = 45,
                                     vjust = 0.5),
        theme_minimal())
```



```
gss %>%  
  specify(college ~ finrela) %>% # compare  
  hypothesise(null = "independence") %>% #  
  calculate(stat = "Chisq") # tell infer w/  
## Response: college (factor)  
## Explanatory: finrela (factor)  
## Null Hypothesis: independence  
## # A tibble: 1 × 1  
##   stat  
##   <dbl>  
## 1  30.7
```



The observed χ^2 statistic is 30.6825231. Now, we want to compare this statistic to a null distribution, generated under the assumption that these variables are not actually related, to get a sense of how likely it would be for us to see this observed statistic if there were actually no association between education and income.



```
gss %>%  
  specify(college ~ finrela) %>%  
  assume(distribution = "Chisq")
```

```
## A Chi-squared distribution with 5 degrees of freedom.
```

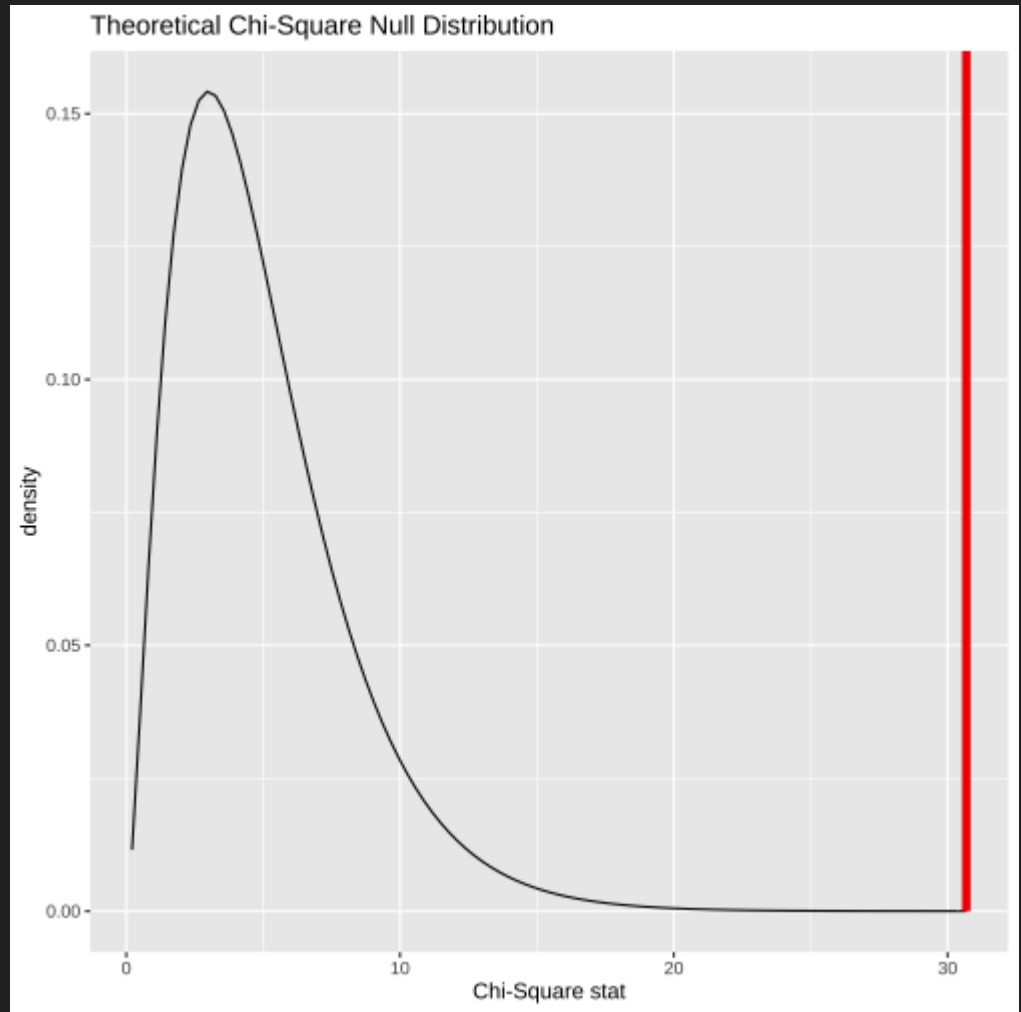



```
gss %>%  
  specify(college ~ finrela) %>%  
  hypothesise(null = "independence") %>%  
  generate(reps = 1000, type = "permute") %>%  
  assume(distribution = "Chisq")
```

```
## A Chi-squared distribution with 5 degrees of freedom.
```



```
null_distribution %>%  
  visualize() +  
  shade_p_value(observed_indep_statistic,  
                direction = "greater")
```



χ^2 statistic



```
## Warning: The chisq_stat() wrapper has been deprecated in favor of the more general  
## observe(). Please use that function instead.
```

```
## X-squared  
## 30.68252
```

```

gss %>%
  specify(response = finrela) %>%
  hypothesise(null = "point",
    p = c("far below average" = 1
          "below average" = 1/6,
          "average" = 1/6,
          "above average" = 1/6,
          "far above average" = 1
          "DK" = 1/6)) %>%
  calculate(stat = "Chisq")
## Response: finrela (factor)
## Null Hypothesis: point
## # A tibble: 1 × 1
##   stat
##   <dbl>
## 1  488.

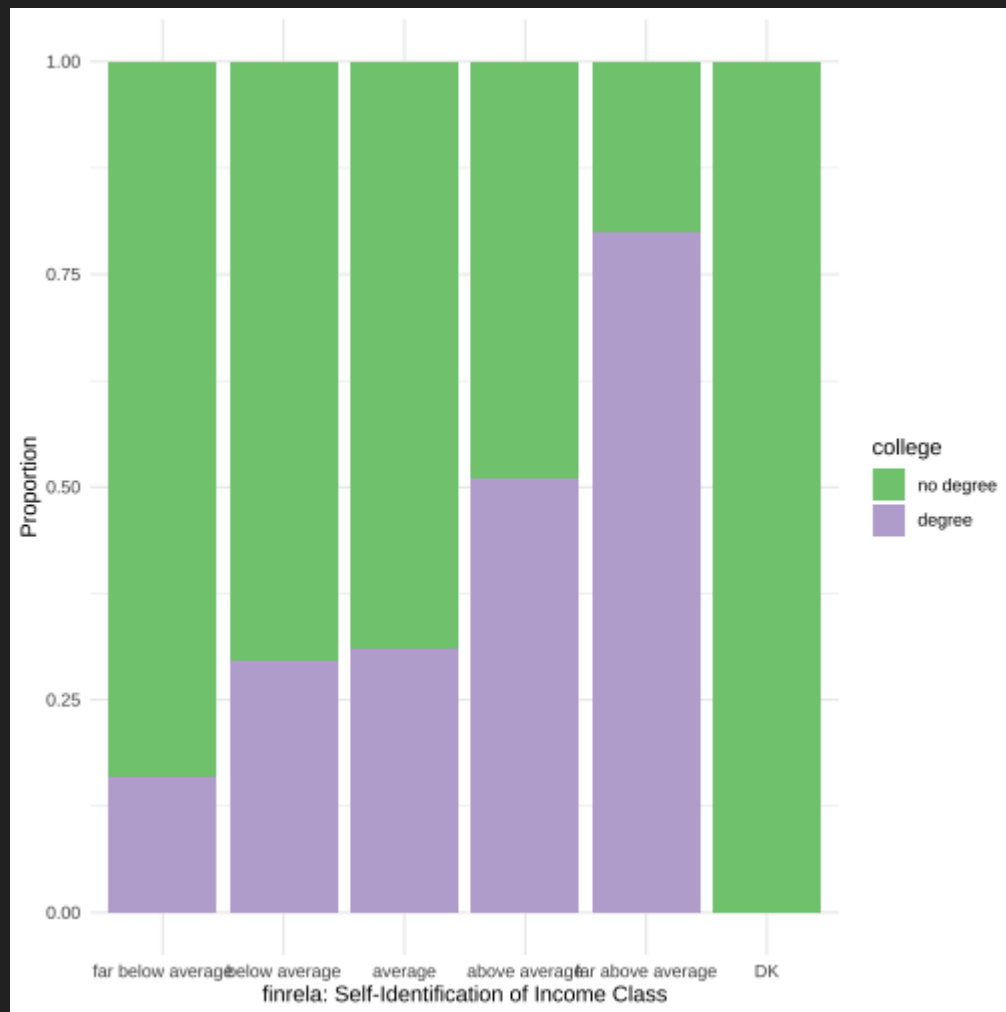
```

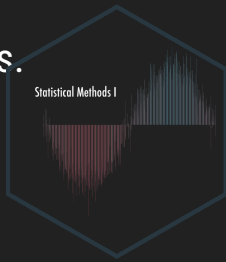


```
observed_gof_statistic <-  
  gss %>%  
  specify(response = finrela) %>%  
  hypothesise(null = "point",  
              p = c("far below average" = 1/6,  
                    "below average" = 1/6,  
                    "average" = 1/6,  
                    "above average" = 1/6,  
                    "far above average" = 1/6,  
                    "DK" = 1/6)) %>%  
  calculate(stat = "Chisq")
```



```
gss %>%
  ggplot() +
  aes(x = finrela, fill = college) +
  geom_bar(position = "fill") +
  scale_fill_brewer(type = "qual") +
  theme(axis.text.x = element_text(angle =
                                     vjust =
                                     labs(x = "finrela: Self-Identification
                                           y = "Proportion") +
  theme_minimal()
```





- If there were no relationship, we would expect to see the purple bars reaching to the same height, regardless of income class.
- Are the differences we see in the plot just due to random noise?
- We can *generate* the null distribution in one of two ways
 - using randomization: approximates the null distribution by permuting the response and explanatory variables, so that each person's educational attainment is matched up with a random income from the sample in order to break up any association between the two
 - theory-based methods: the general approximation



```
gss %>%
  specify(college ~ finrela) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")

## Response: college (factor)
## Explanatory: finrela (factor)
## Null Hypothesis: independence
## # A tibble: 1,000 × 2
##   replicate  stat
##   <int>    <dbl>
## 1         1  8.21
## 2         2 12.4
## 3         3  3.22
## 4         4  2.97
## 5         5  8.03
## 6         6 10.5
## 7         7  0.629
## 8         8  0.399
## 9         9  3.46
## 10        10  3.17
## # ... with 990 more rows
```

```
null_dist_sim <-  
  gss %>%  
  specify(college ~ finrela) %>%  
  hypothesize(null = "independence") %>%  
  generate(reps = 1000, type = "permute") %>%  
  calculate(stat = "Chisq")
```



```
gss %>%  
  specify(college ~ finrela) %>%  
  assume(distribution = "Chisq")
```

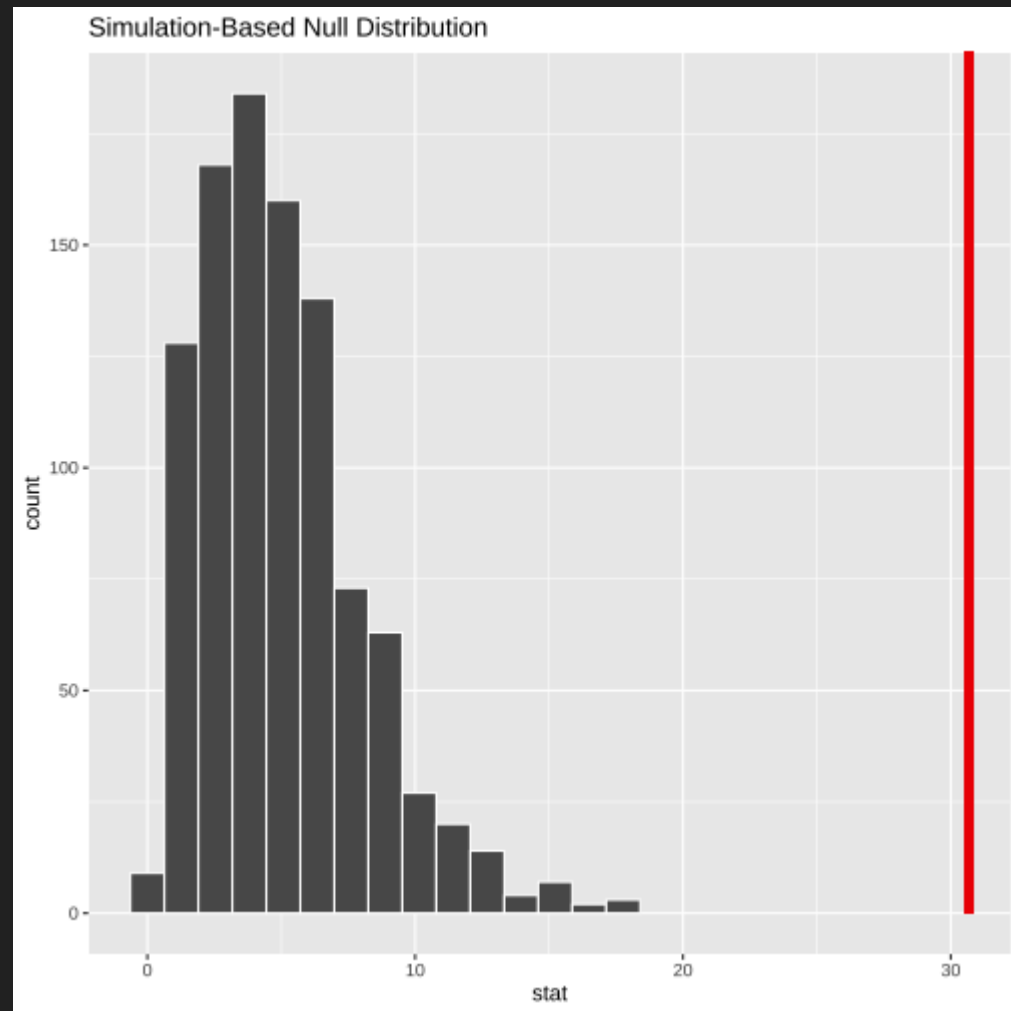
```
## A Chi-squared distribution with 5 degrees of freedom.
```



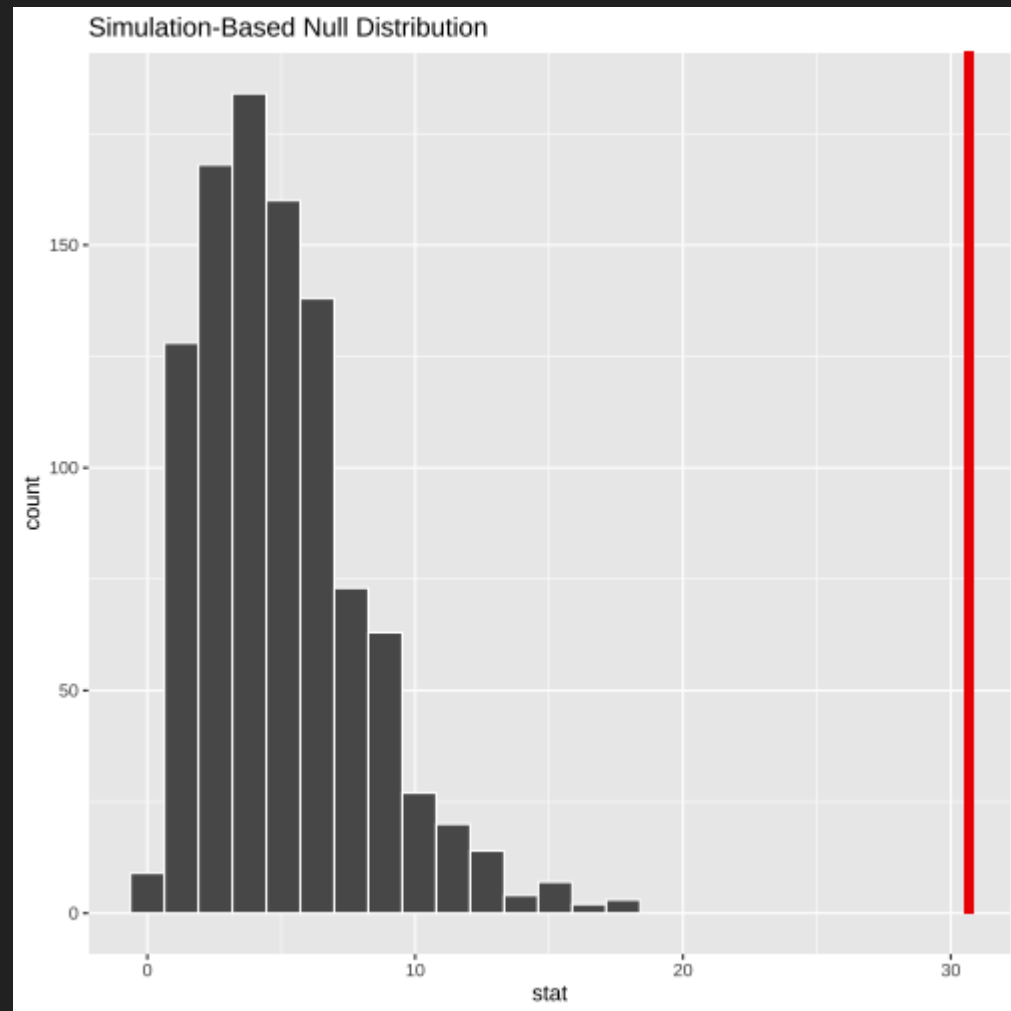
```
null_dist_theory <-  
  gss %>%  
  specify(college ~ finrela) %>%  
  assume(distribution = "Chisq")
```



```
null_dist_sim %>%  
  visualize() +  
  shade_p_value(observed_indep_statistic,  
                direction = "greater")
```

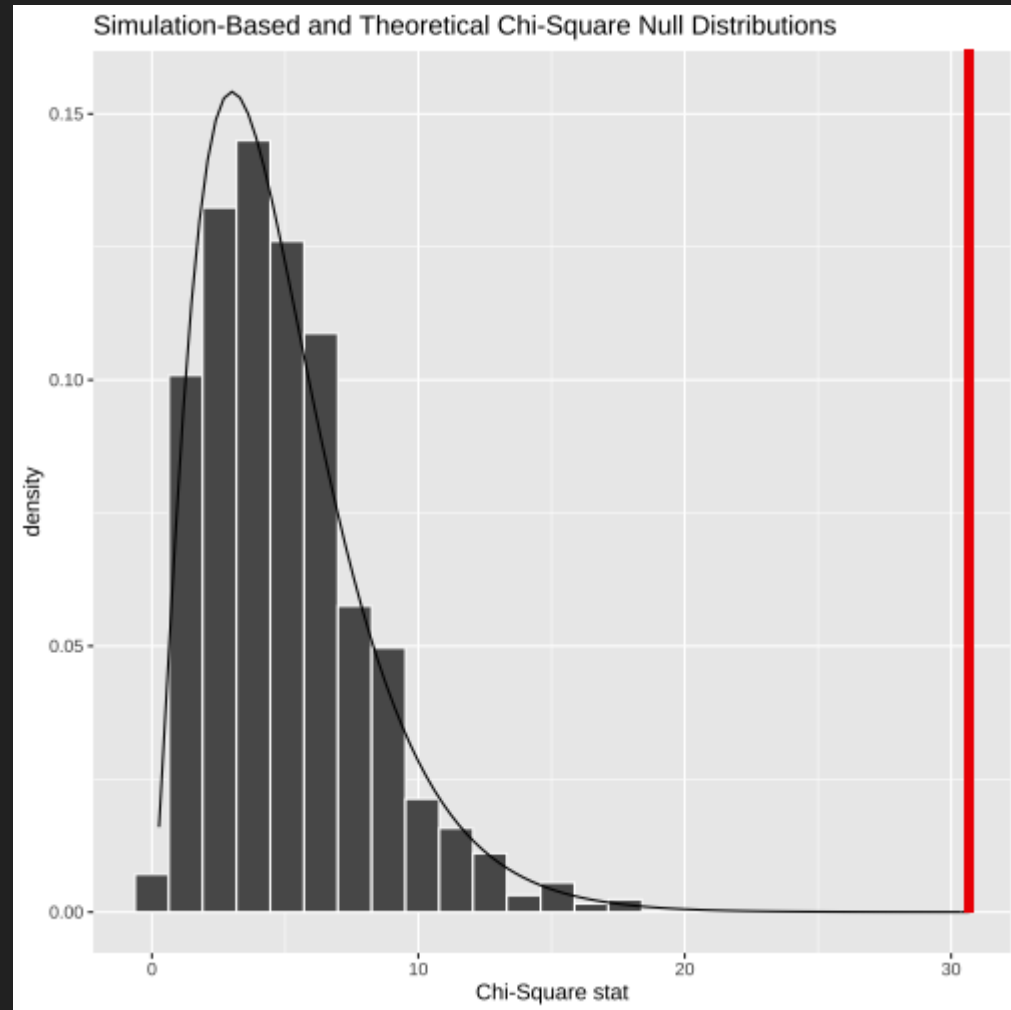


```
null_dist_sim %>%  
  visualize() +  
  shade_p_value(observed_indep_statistic,  
                direction = "greater")
```



```
null_dist_sim %>%  
  visualize(method = "both") +  
  shade_p_value(observed_indep_statistic,  
                direction = "greater")
```

Warning: Check to make sure the conditions have been met for
method. {infer} currently does not check these for you.



```
null_dist_sim %>%  
  get_p_value(obs_stat = observed_indep_sta  
             direction = "greater")  
## Warning: Please be cautious in reporting a p-value of 0. T  
## approximation based on the number of `reps` chosen in the  
## get_p_value()` for more information.  
## # A tibble: 1 × 1  
##   p_value  
##   <dbl>  
## 1      0
```



If there were really no relationship between education and income, our approximation of the probability that we would see a statistic as or more extreme than 30.6825231 is approximately 0



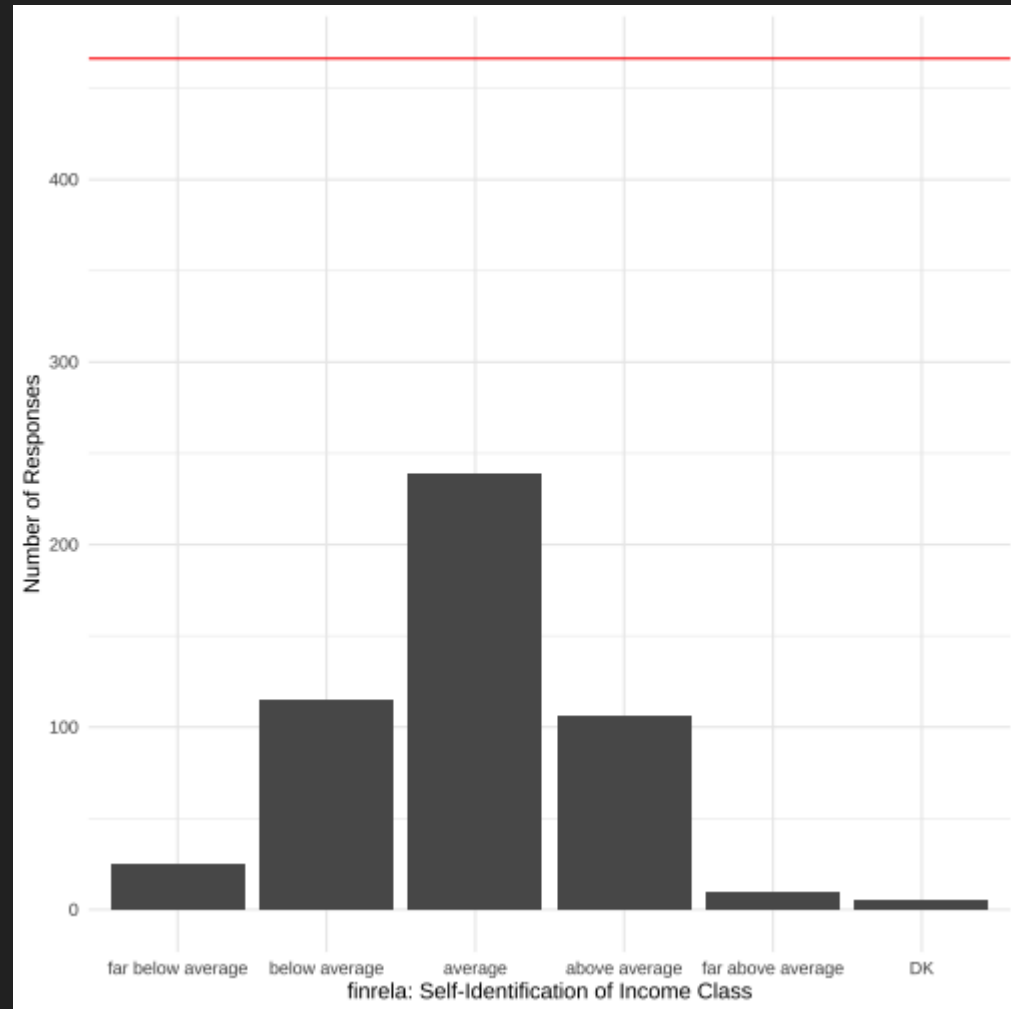
Calculate the p -value using the true χ^2 distribution



```
##      X-squared
## 1.082094e-05

## # A tibble: 1 × 3
##   statistic chisq_df  p_value
##   <dbl>     <int>    <dbl>
## 1      30.7         5 0.0000108
```

```
gss %>%
  ggplot2::ggplot() +
  ggplot2::aes(x = finrela) +
  ggplot2::geom_bar() +
  ggplot2::geom_hline(yintercept = 466.3,
                      col = "red") +
  ggplot2::labs(x = "finrela: Self-Identifi",
               y = "Number of Responses")
  theme_minimal()
```



It seems like a uniform distribution may not be the most appropriate description of the data--many more people describe their income as average than than any of the other options. Lets now test whether this difference in distributions is statistically significant.



```

gss %>%
  specify(response = finrela) %>%
  hypothesize(null = "point",
    p = c("far below average" = 1
          "below average" = 1/6,
          "average" = 1/6,
          "above average" = 1/6,
          "far above average" = 1
          "DK" = 1/6)) %>%
  calculate(stat = "Chisq")
## Response: finrela (factor)
## Null Hypothesis: point
## # A tibble: 1 × 1
##   stat
##   <dbl>
## 1  488.

```

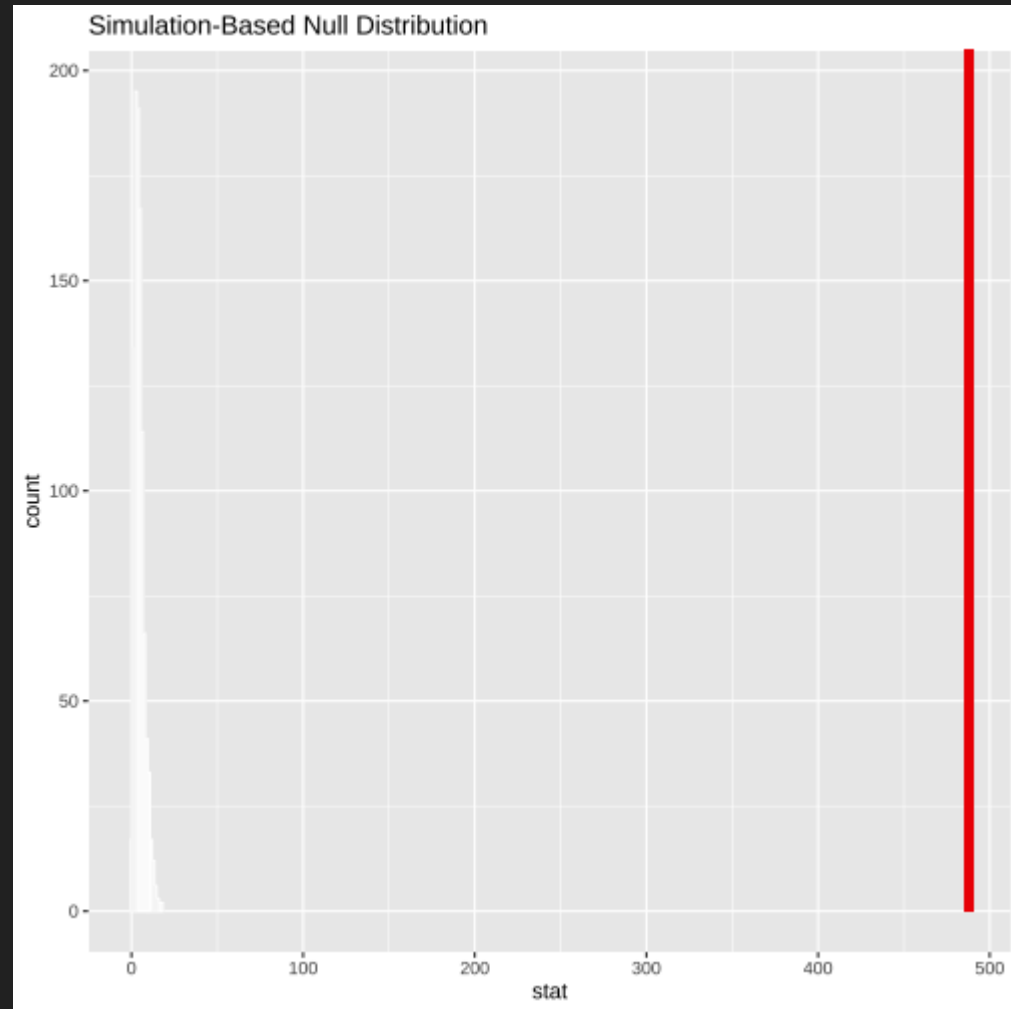




```
gss %>%  
  specify(response = finrela) %>%  
  hypothesize(null = "point",  
    p = c("far below average" = 1  
      "below average" = 1/6,  
      "average" = 1/6,  
      "above average" = 1/6,  
      "far above average" = 1  
      "DK" = 1/6)) %>%  
  generate(reps = 1000, type = "draw") %>%  
  calculate(stat = "Chisq")  
## Response: finrela (factor)  
## Null Hypothesis: point  
## # A tibble: 1,000 × 2  
##   replicate  stat  
##   <fct>      <dbl>  
## 1 1          1.86  
## 2 2          3.16  
## 3 3         10.1  
## 4 4          0.88  
## 5 5         15.9  
## 6 6          3.47  
## 7 7          2.15  
## 8 8          4.48  
## 9 9          6.11  
## 10 10         11.8  
## # ... with 990 more rows
```



```
null_dist_gof %>%  
  visualize() +  
  shade_p_value(observed_gof_statistic,  
                direction = "greater")
```



```
null_dist_gof %>%
```

```
  get_p_value(observed_gof_statistic,  
              direction = "greater")
```

```
## Warning: Please be cautious in reporting a p-value of 0. The  
## approximation based on the number of `reps` chosen in the  
## `get_p_value()` for more information.
```

```
## # A tibble: 1 × 1  
##   p_value  
##   <dbl>  
## 1      0
```



If each self-identified income class was equally likely to occur, our approximation of the probability that we would see a distribution like the one we did is approximately 0



Calculate the p -value using the true χ^2 distribution

```
## [1] 3.131231e-103
```



Thats it!

