Chi-squared tests

Stat 250

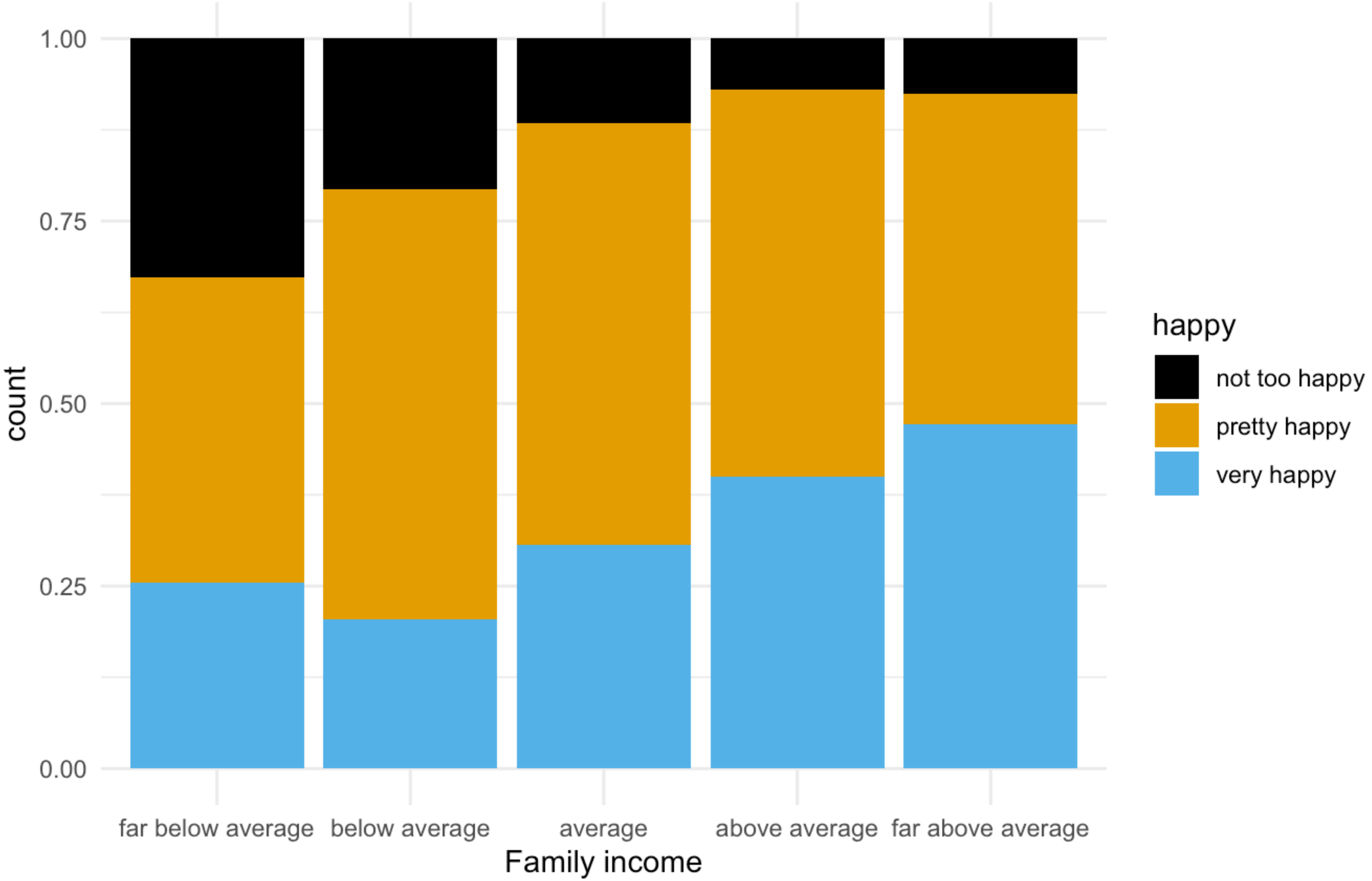
Click here for PDF version

Can money buy you happiness?

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. We’ll consider two questions:

- Compared with American families in general, would you say your family income is far below average, below average, average, above average, or far above average?
- Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

Can money buy you happiness?



Happiness contingency table

How can we explore whether opinion on income and happiness are associated?

happy	far below average	below average	average	above average	far above average	Total
not too happy	50	123	120	33	4	330
pretty happy	64	350	602	253	24	1293
very happy	39	121	319	190	25	694
Total	153	594	1041	476	53	2317

R:

```r
tabyl(happy2018, happy, finrela) |>
  adorn_totals(where = c("row", "col"))
```

Test statistic

H_0 : the variables are independent

What would the contingency table look like under H_0?

happy	far below average	below average	average	above average	far above average
not too happy	21.79111	84.60078	148.2650	67.79456	7.548554
pretty happy	85.38153	331.48123	580.9292	265.63142	29.576608
very happy	45.82736	177.91800	311.8058	142.57402	15.874838

Test statistic

How can we compare what we observe to what would be expected under H_0?

Observed:

happy	far below average	below average	average	above average	far above average	Total
not too happy	50	123	120	33	4	330
pretty happy	64	350	602	253	24	1293
very happy	39	121	319	190	25	694
Total	153	594	1041	476	53	2317

Expected:

happy	far below average	below average	average	above average	far above average	Total
not too happy	21.79111	84.60078	148.2650	67.79456	7.548554	330
pretty happy	85.38153	331.48123	580.9292	265.63142	29.576608	1293

happy	far below average	below average	average	above average	far above average	Total
very happy	45.82736	177.91800	311.8058	142.57402	15.874838	694
Total	153	594	1041	476	53	2317

Permutation test

1. Store the data in a table: one row per observation, one column per variable.
2. Calculate a test statistic for the original data.
3. **Repeat**
    - Randomly permute the rows in one of the columns.
    - Calculate the test statistic for the permuted data.

    Until we have enough samples
4. Calculate the p-value as the fraction of times the random statistics exceed the original statistic.

Permutation test setup

Calculate the observed test statistic

```
# Have mosaic package loaded
observed_table <- tally(~ happy + finrela, data = happy2
observed <- chisq(observed_table)
```

Remove any missing values on variables of interest

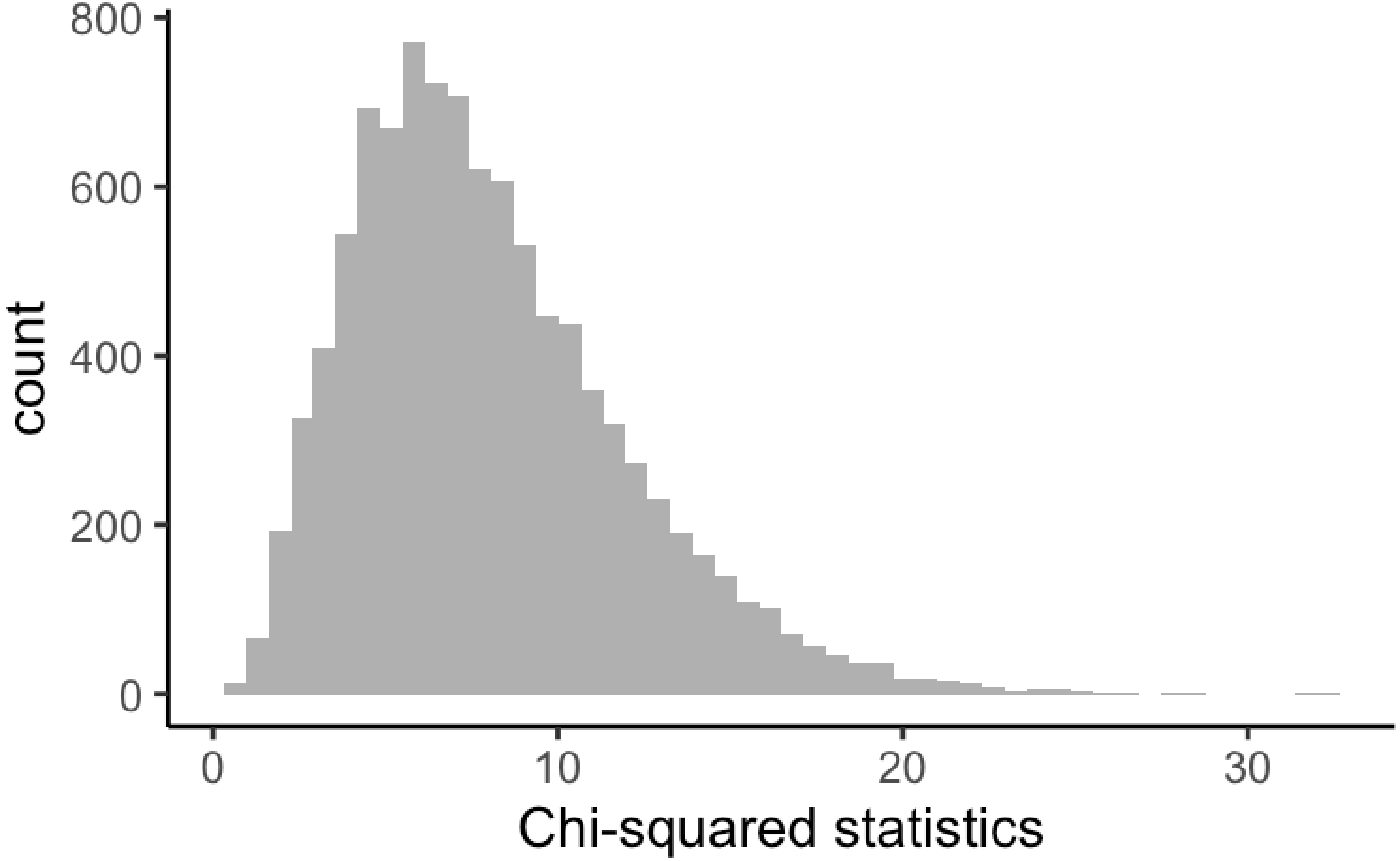
```
library(tidyr) # for drop_na()
happy_complete <- drop_na(happy2018, happy, finrela)

# Extract columns of interest
happy <- happy_complete$happy
finrela <- happy_complete$finrela
```

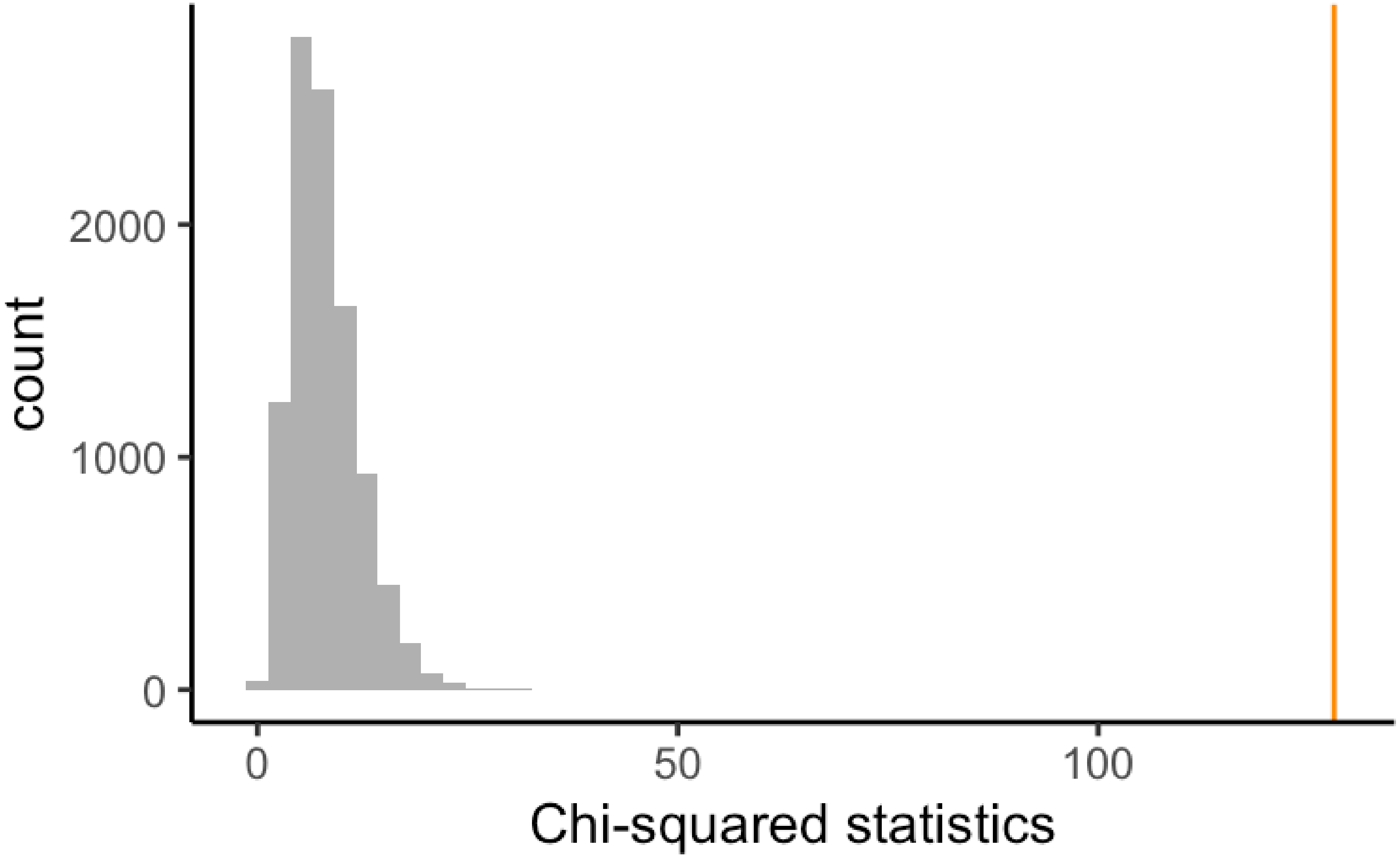
Construct the permutation distribution

```r
set.seed(55057)
N <- 10^4 - 1
result <- numeric(N)
for(i in 1:N) {
  finrela_perm <- sample(finrela)
  perm_table <- tally(~happy + finrela_perm)
  result[i] <- chisq(perm_table)
}
```

Permutation distribution



p-value

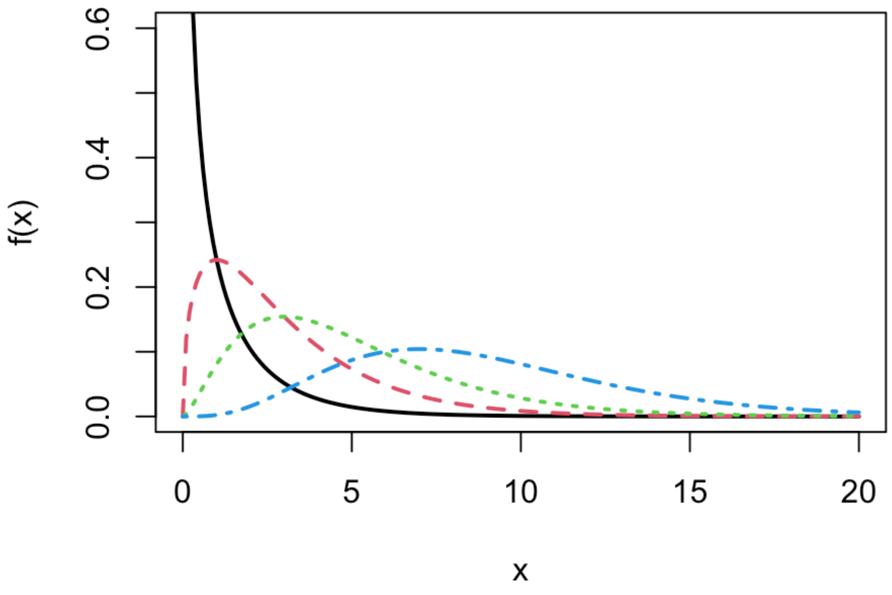


```
(sum(result >= observed) + 1) / (N + 1)
## [1] 1e-04
```

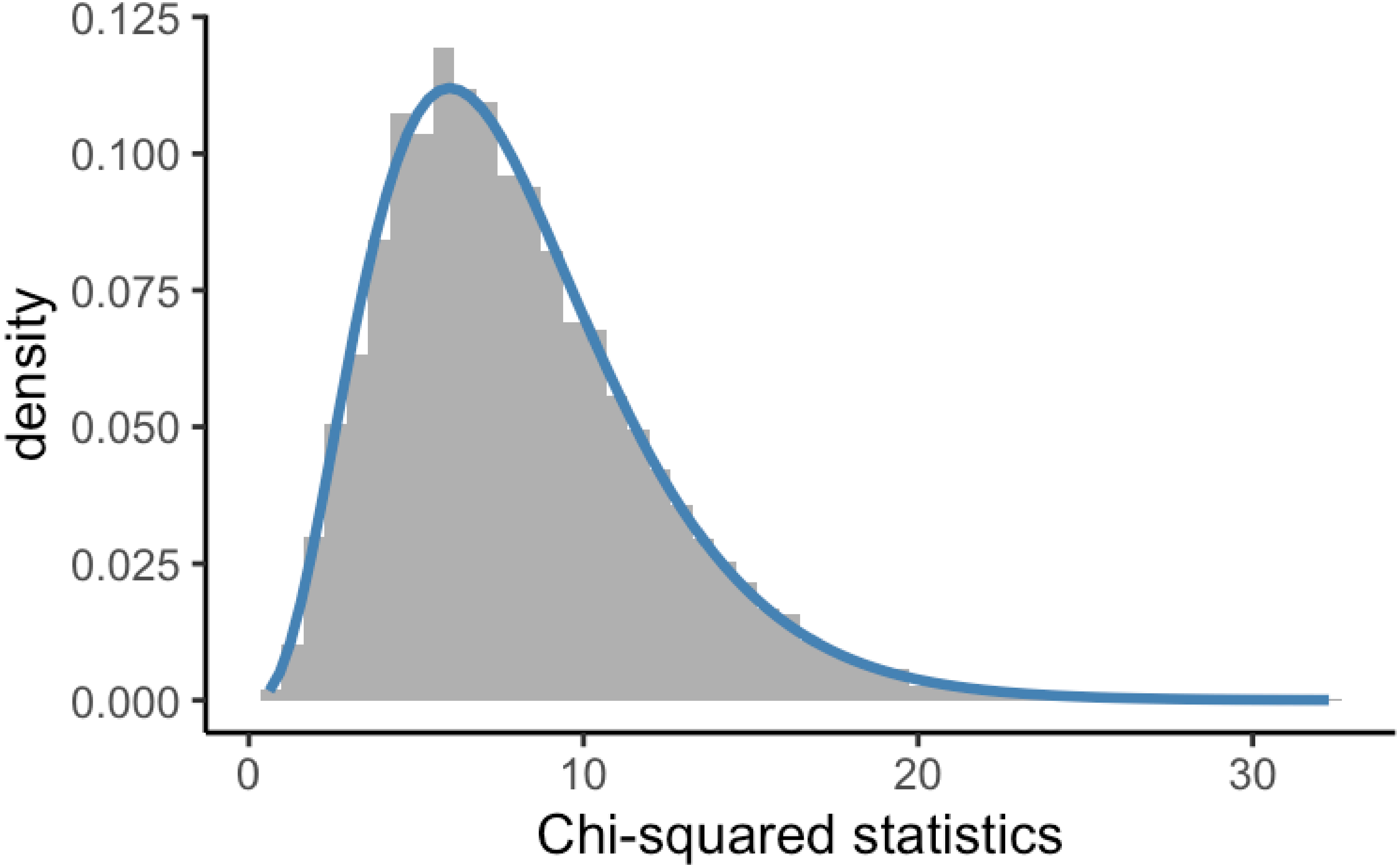
Chi-squared distribution

A random variable follows a χ^2_m distribution if it has PDF

$$f(x|m) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} e^{-x/2}, \; x > 0.$$



Chi-squared reference distribution



Simulation vs. model-based results

Chi-squared test

```
1 - pchisq(observed, df = (3 - 1) * (5 - 1))
```

```
X.squared
        0
```

Permutation test

```
(sum(result >= observed) + 1) / (N + 1)
```

```
[1] 0.0001
```

Caution

The χ^2 distribution provides a reasonable approximation of the null distribution **as long as the sample size is "large enough"**

Common guidelines:

- "Cochran's rule:" All of the cells have **expected counts** > 5
- All expected counts are at least 1 and no more than 20% of cells have **expected counts** < 5

Use a permutation test if the expected counts aren't large enough

Your turn

Work through the example on climate change action by generation with your neighbors.

R code for carrying out chi-squared tests is included on the worksheet