Inference for multiple proportions

Stat 120

May 13 2022

Tests for Categorical Variable(s)

Chi-square test for association

- Determine if a relationship between two categorical variables is statistically significant
- E.g. Does M&M color distribution depend on type (chocolate vs. peanut)?

Chi-square test for association hypothesis

- Hypotheses look like

$$H_0 : \text{two categorical variables are not associated}$$

$$H_A : \text{two categorical variables are associated}$$

- E.g. Does M&M color distribution depend on type (chocolate vs. peanut)?

$$H_0 : \text{there is no association between M\&M color and type}$$

$$H_A : \text{there is an association between M\&M color and type}$$

Expected Counts and p-value

The expected counts for each combination in a two-way table

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{expected count} - \text{observed count})^2}{\text{expected count}}$$

- For both types of test, large chi-square test stat values support the **alternative** hypothesis so

$$p - value = P(\chi^2 \geq \text{observed } \chi^2)$$

- always a **right-tailed** value

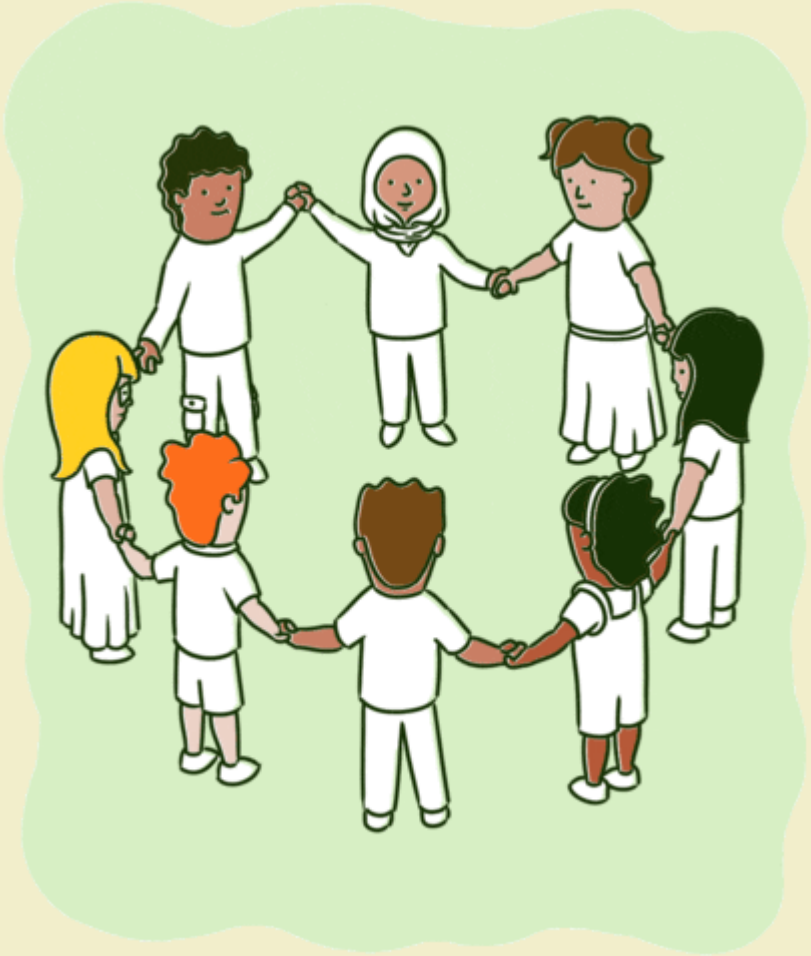
Chi-Square test for association

Options for computing the p-value:

- randomization/permutation: simulate new data consistent with H_0 and recompute the χ^2 test stat
  - **Association:** permute the values of one variable column to break the link that could exist in the data between both variables
- Chi-square distribution (probability model)
  - **Association:** use $(r-1)(c-1)$ where $r =$ number of rows and $c =$ number of columns
  - need n large enough so expected counts are at least 5

Your Turn 1

05:00



- Go over to the in class activity file
- Complete Example 1

Association example

Does political comfort level depend on religion?

- H_0 : There is no association between religion and comfort level
  - implies: the distribution of comfort level is the same for all three religion types
- H_A : There is an association between religion and comfort level
  - implies: the distribution of comfort level is the different for at least one religion type.

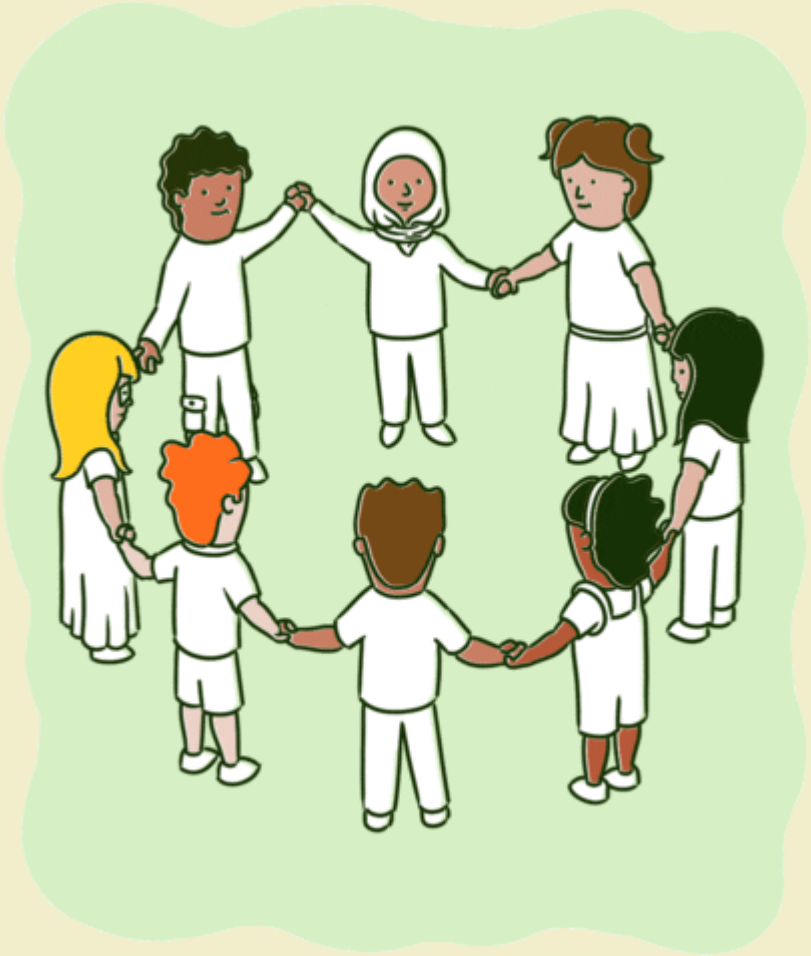
Association example

- EDA for two categorical variables
  - change comfort level names then reorder

```
survey <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Survey.csv")
summary(survey$Question.9)
   Length     Class      Mode
      374 character character
levels(survey$Question.9) <- c("almost always", "rarely", "sometimes")
survey$Question.9 <- fct_relevel(survey$Question.9,
                      "almost always", "sometimes", "rarely")
summary(survey$Question.9)  # always check work!
   almost always comfortable rarely, if ever, comfortable
                         171                           51
       sometimes comfortable                         NA's
                         150                            2
```

Your Turn 2

05:00



- Go over to the in class activity file
- Skim through Example 2 in your group

Association example

- Observed distribution of comfort level given religiousness

```
counts <- table(survey$Question.8, survey$Question.9)
counts
```

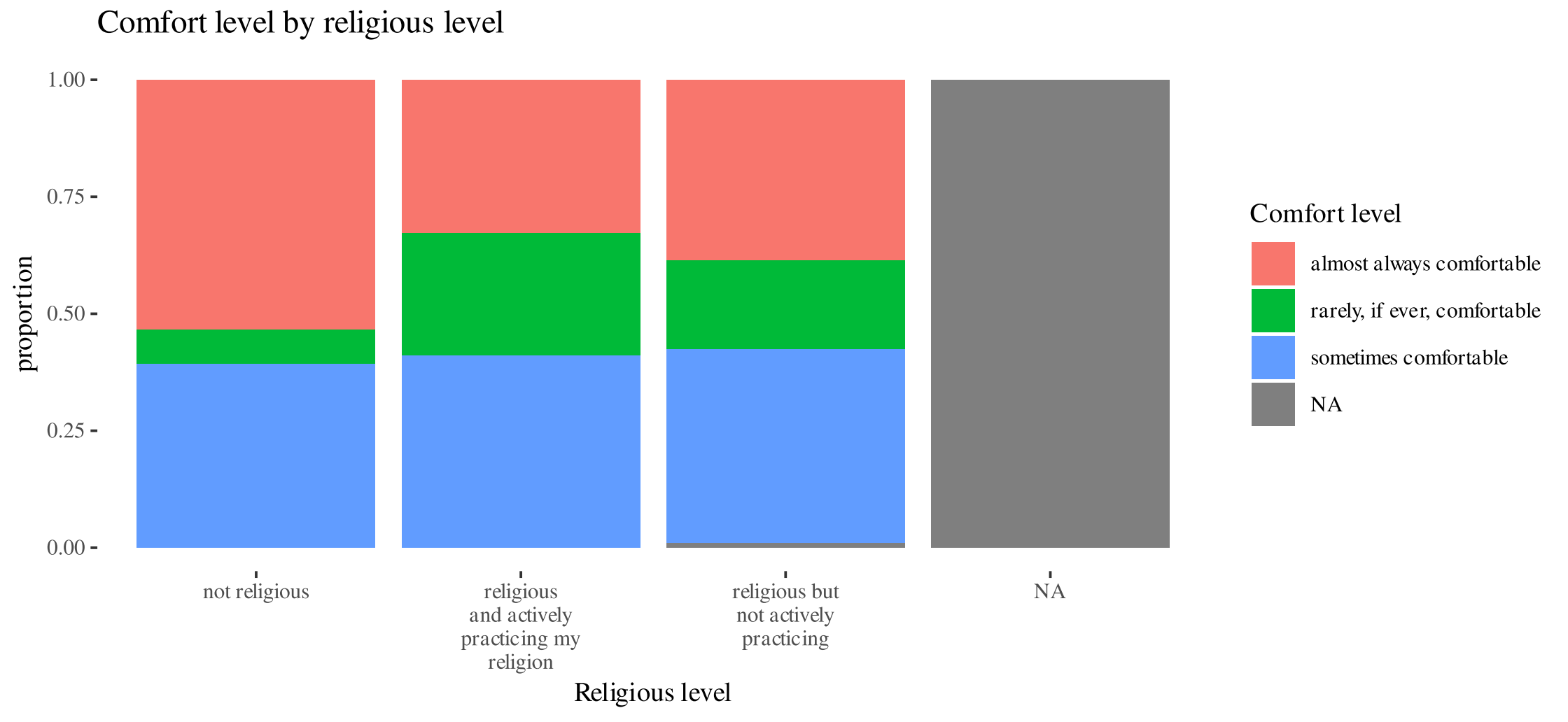
```
                                                 almost always comfortable
  not religious                                                        110
  religious and actively practicing my religion                         20
  religious but not actively practicing                                 41

                                                 rarely, if ever, comfortable
  not religious                                                            15
  religious and actively practicing my religion                            16
  religious but not actively practicing                                    20

                                                 sometimes comfortable
  not religious                                                     81
  religious and actively practicing my religion                     25
  religious but not actively practicing                             44
```

Association example

```
sum(counts)  # number of respondents
[1] 372
prop.table(counts,1)

                                                    almost always comfortable
  not religious                                                     0.53398058
  religious and actively practicing my religion                     0.32786885
  religious but not actively practicing                             0.39047619

                                                    rarely, if ever, comfortable
  not religious                                                      0.07281553
  religious and actively practicing my religion                      0.26229508
  religious but not actively practicing                              0.19047619

                                                    sometimes comfortable
  not religious                                                0.39320388
  religious and actively practicing my religion                0.40983607
  religious but not actively practicing                        0.41904762
```

There is a much higher rate of "almost always comfortable" for the not religious respondents (53.4%) than those that are religious (not active: 32.8%; active: 39%).

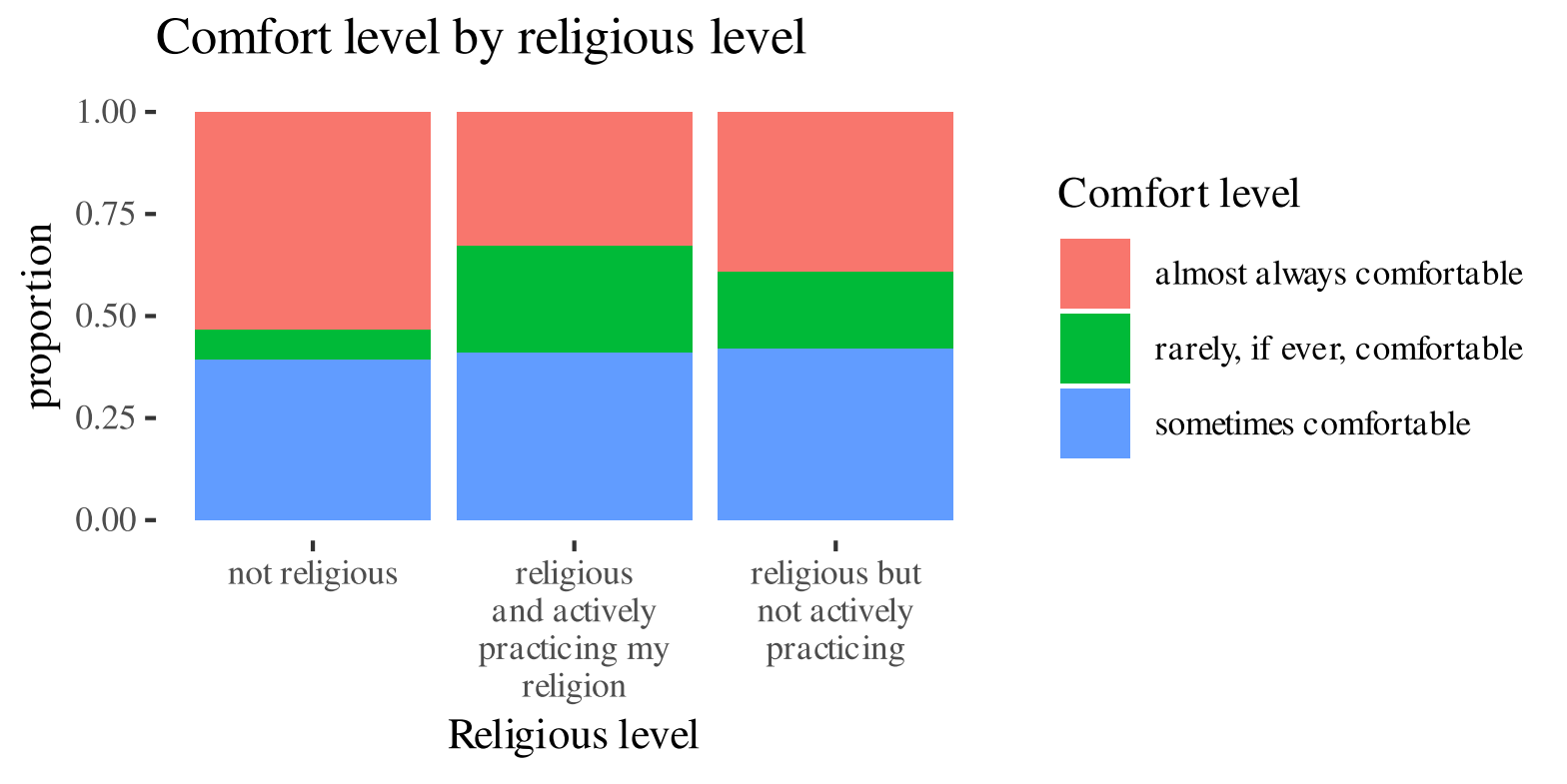
2b. Association example



Association example

- Both variables have missing value(s). Remove using `drop_na` then redo plot.

```
library(tidyr)
survey_ex2 <- drop_na(survey, Question.8, Question.9) # removes missing values
```



Association example

- Expected counts assuming no association (null)?
  - expected number of respondents who are "not religious" and "almost always comfortable"?
  - is **not** 1/9 of all respondents!
- There are 206 "not religious" respondents (row total)
- The overall rate (ignoring religion) of "almost always comfortable" is $\frac{171}{372}$, or about 46%.
- If religion isn't related to comfort level, the expected number is about

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n} = 206 \times \frac{171}{372} = 94.694$$

2d. Association example

- Chi-square contribution for "not religious" and "almost always comfortable" cell?
- The contribution to the chi-square test stat from this category is 2.47.

$$\frac{(110 - 94.694)^2}{94.694} = 2.474$$

2e. Association example

- Use `chisq.test` to finish the test stat calculation (who wants to add 9 cell contributions!)

```
ComfortReligion <- chisq.test(survey_ex2$Question.8, survey_ex2$Question.9)
ComfortReligion
```

```
    Pearson's Chi-squared test

data:  survey_ex2$Question.8 and survey_ex2$Question.9
X-squared = 21.362, df = 4, p-value = 0.0002684
```

- The test stat value is 21.362.
- There are 3 categories for each variable, so the degrees of freedom will be $df = (3-1)(3-1) = 4$.

2f. Association example

- **Interpret:** If there is no association between comfort level and religiousness, then we would see a chi-square test stat of 21.362, or one even larger, only about 0.03% of the time.
- Conclusion?
- We have strong evidence that there is an association between political comfort level and religiousness ($\chi^2 = 21.362$, df = 4, p-value = 0.00026).

2g. Association example

- Are the expected counts above 5?

```
ComfortReligion$expected
                                                    survey_ex2$Question.9
survey_ex2$Question.8                                almost always comfortable
  not religious                                                    94.69355
  religious and actively practicing my religion                    28.04032
  religious but not actively practicing                            48.26613
                                                    survey_ex2$Question.9
survey_ex2$Question.8                                rarely, if ever, comfortable
  not religious                                                      28.241935
  religious and actively practicing my religion                       8.362903
  religious but not actively practicing                              14.395161
                                                    survey_ex2$Question.9
survey_ex2$Question.8                                sometimes comfortable
  not religious                                                    83.06452
  religious and actively practicing my religion                    24.59677
  religious but not actively practicing                            42.33871
```

2h. Association example

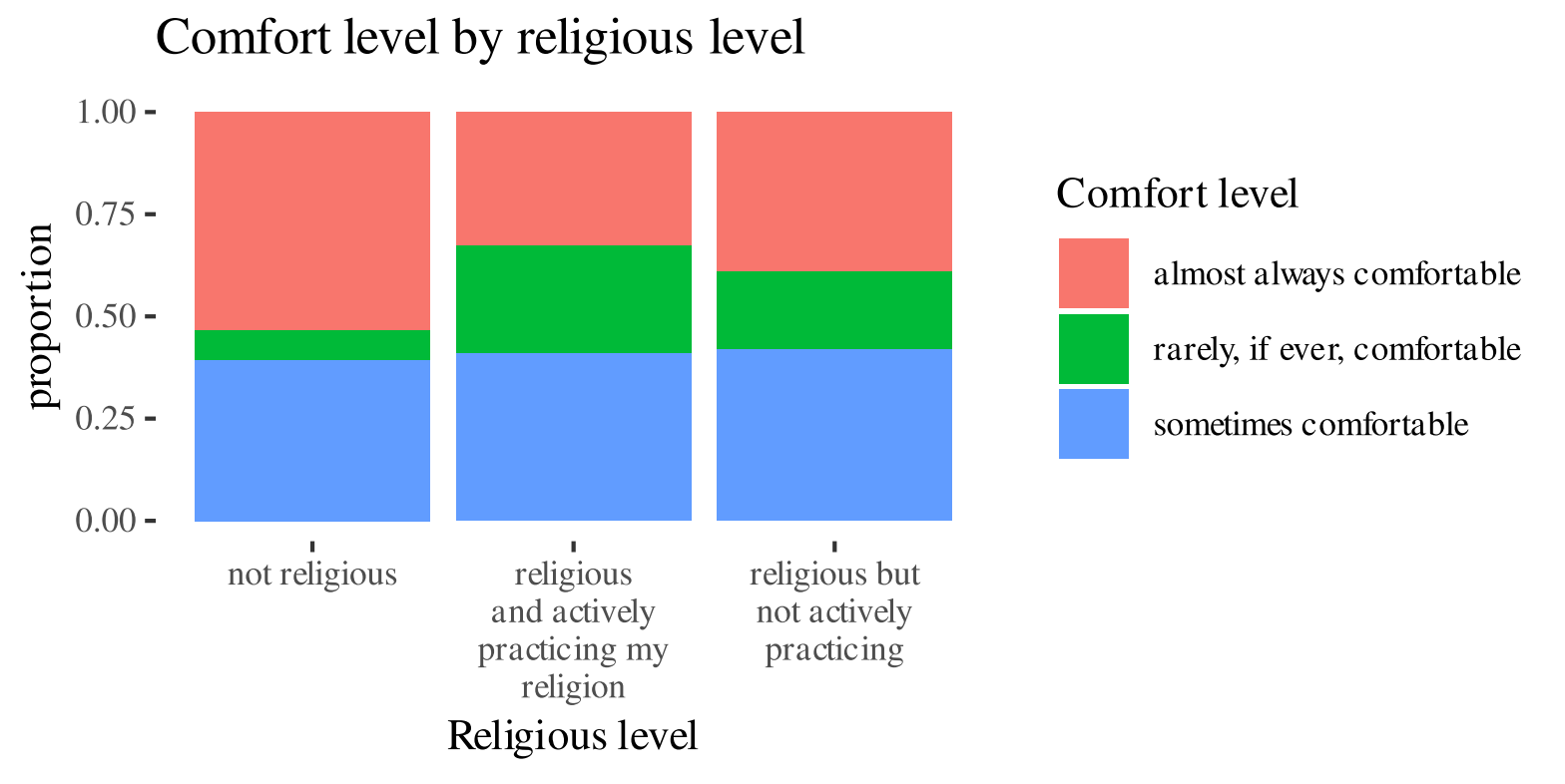
- If we get a red warning when running `chisq.test`, it usually means the sample size conditions aren't met to use the chi-square model.
- Instead run a randomization test with

```
chisq.test(survey_ex2$Question.8, survey_ex2$Question.9,
           simulate.p.value = TRUE)
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  survey_ex2$Question.8 and survey_ex2$Question.9
X-squared = 21.362, df = NA, p-value = 0.0004998
```

2i. Association example

- Describe the association!
  - which groups have the most different comfort levels?



2i. Association example

- 95% CI for the difference in the true proportions of "rarely comfortable" people in the not religious and actively religious groups.

$$p = \text{proportion rarely comfortable}$$

  - 95% CI for $p_{not.relig} - p_{active}$

```
table(survey_ex2$Question.8)
```

```
                                  not religious
                                            206
religious and actively practicing my religion
                                             61
        religious but not actively practicing
                                            105
```

$$n_{not.relig} = 206 \quad n_{active} = 61$$

2i. Association example

```
prop.table(counts,1)
```

```
                                                    almost always comfortable
  not religious                                                     0.53398058
  religious and actively practicing my religion                     0.32786885
  religious but not actively practicing                             0.39047619

                                                    rarely, if ever, comfortable
  not religious                                                      0.07281553
  religious and actively practicing my religion                      0.26229508
  religious but not actively practicing                              0.19047619

                                                    sometimes comfortable
  not religious                                               0.39320388
  religious and actively practicing my religion               0.40983607
  religious but not actively practicing                       0.41904762
```

2i. Association example

```
counts
```

	almost always comfortable
not religious	110
religious and actively practicing my religion	20
religious but not actively practicing	41

	rarely, if ever, comfortable
not religious	15
religious and actively practicing my religion	16
religious but not actively practicing	20

	sometimes comfortable
not religious	81
religious and actively practicing my religion	25
religious but not actively practicing	44

$$\hat{p}_{not.rel} = \frac{15}{206} = 0.0728$$

$$\hat{p}_{active} = \frac{16}{61} = 0.2623$$

2i. Association example

- 95% CI for $p_{not.relig} - p_{active}$

$$(0.0728 - 0.2623) \pm 1.96\sqrt{\frac{0.0728(1-0.0728)}{206} + \frac{0.02623(1-0.02623)}{61}}$$
$$= (-0.2430475, -0.1359525)$$

```
(0.0728 - 0.2623) + c(-1,1)*1.96*sqrt(0.0728*(1-0.0728)/206 + 0.02623*(1-0.02623)/61)
```

```
[1] -0.2430475 -0.1359525
```

- I am 95% confident that the percentage of all non-religious students who are rarely comfortable is between 13.6 and 24.3 percentage points lower than the actively religious students.