MA117 - Worksheet 7
Categorical Variables
Week 2, Thursday

Problem 1. The R function `pchisq` takes as input two numbers `q` and `df` and outputs the percentile of `q` in the chi-squared distribution with `df` degrees of freedom. Conversely, `qchisq` takes as input two numbers `p` and `df` and outputs the observation at percentile `p` in the chi-squared distribution with `df` degrees of freedom. Use these functions to calculate the following. Also sketch a picture.

(a) df $= 5$. Calculate $P(\chi^2 \leq 10)$.
(b) df $= 10$. Calculate $P(\chi^2 \leq 10)$.
(c) df $= 15$. Calculate $P(\chi^2 \leq 10)$.
(d) df $= 15$. Calculate $P(\chi^2 \geq 10)$.
(e) df $= 15$. Calculate $P(5 \leq \chi^2 \leq 10)$.
(f) df $= 15$. Calculate the observation that is larger than exactly 80% of observations.
(g) df $= 15$. Calculate the observation that is less than exactly 80% of observations.

Background for problems 2–6. FiveThirtyEight conducted a survey about singular versus plural usage of the word "data." Specifically, the survey asked respondents the following questions:

A. How would you write the following sentence?
   (a) Some experts say it's important to drink milk, but the data is inconclusive.
   (b) Some experts say it's important to drink milk, but the data are inconclusive.
B. When faced with using the word "data," have you ever spent time considering if the word was a singular or plural noun?
C. How much, if at all, do you care about the debate over the use of the word "data" as a singluar or plural noun?

Problem 2. Answer the above questions A–C for yourself! ☺

The results of FiveThirtyEight's survey are recorded in the following `csv` file:

`https://sagrawalx.github.io/teaching/data/datasgpl.csv`

Responses are encoded as follows:

A. The `Prefer` column records respondents' responses to question A (`Singular` indicates sentence (a) and `Plural` indicates sentence (b)).
B. The `ThoughtAbout` column records respondents' responses to question B (`Yes` or `No`).
C. The `CareAbout` column records respondents' responses to this question C (`Not at all`, `Not much`, `Some`, or `A lot`).

The data also records some demographic information. Note that some respondents did not respond to some of the questions, so you may want to throw out blank responses below.

Problem 3. (a) Calculate and interpret a 98% confidence interval for the proportion of people who prefer `Singular` usages of the word "data."
(b) Calculate and interpret a p-value for the data under the hypothesis that half of all people prefer `Singular` usages of the word "data."

Problem 4. Let p_1 denote the proportion of people who prefer the `Singular` usage of the word "data" among people who have thought about this issue, and p_2 the proportion who prefer the `Singular` usage among people who haven't thought about this issue.

(a) Calculate and interpret a 99% confidence interval for $p_1 - p_2$.

(b) Calculate and interpret a p-value for the data under the hypothesis that $p_1 = p_2$.

Recall. If you named your data frame `df`, typing

```
count(df, Prefer, ThoughtAbout)
```

will give you some relevant counts.

Problem 5. (a) If `Prefer` and `CareAbout` were uncorrelated, how many people would you expect there to be who both care `A lot` about singular versus plural usages of the word "data" and also prefer the `Singular` usage? Is the observed number more or less than the expected number?

(b) In base R, you can run a chi-square test to calculate a p-value for the data under the hypothesis that `Prefer` and `CareAbout` are uncorrelated using

```
chisq.test(table(df$Prefer, df$CareAbout))
```

Alternatively, to use tidier and more modern syntax, you can install the `tidymodels` package using `install.packages("tidymodels")`, equip it using `library(tidymodels)`, and then run

```
chisq_test(df, Prefer ~ CareAbout)
```

Run one of these commands and interpret the p-value.

Problem 6. How would you use this data to determine if `Prefer` is correlated with `Gender`?

Problem 7. A survey asked 827 randomly sampled registered voters in California "Do you support or oppose drilling for oil and natural gas off the Coast of California, or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.

	College Grad	Not College Grad	Total
Support	154	132	286
Oppose	180	126	306
Don't Know	104	131	235
Total	438	389	827

You can input this contingency table directly into R so that you can get R to run a chi-square test![1] Here's some code that'll create a table called `drill` out of this data.

```
drill <- matrix(c(154, 132, 180, 126, 104, 131), ncol=2, byrow=TRUE)
colnames(drill) <- c("College Grad", "Not College Grad")
rownames(drill) <- c("Support", "Oppose", "Don't Know")
drill <- as.table(drill)
```

Run `chisq.test` on this table to calculate a p-value. Then interpret this p-value.

[1] You can also use `chisq.test` to run chi-square tests of the "single categorical variable" type. Reading through the help file at `?chisq.test` will show you more about how to do this.