MA117 - Worksheet 10
Chi-Square Tests
Week 3, Monday

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.

In each of the following, suppose χ^2 is a random variable with a chi-square distribution with the indicated number of degrees of freedom. Sketch a picture of a relevant region in a chi-square distribution, and then use the R functions `pchisq` and `qchisq` to calculate the indicated quantity.

(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.

Problem 2. A study examined microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China. In this region, woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, deciduous forests make up 39.6%, and "other" makes up the rest. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, 61 as deciduous forests, and the rest as "other." Conduct a hypothesis test to determine if the barking deer of Hainan Island prefer to forage in certain habitats over others.

Problem 3. 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

Conduct a hypothesis test to determine if opinions on this issue are correlated with whether or not an individual has graduated from college.

Problem 4. Here's the FiveThirtyEight data set we looked at before about singular vs plural usages of the word "data."

`https://sagrawalx.github.io/teaching/fa21-b1_ma117/class/datasgpl.csv`

See Worksheet 9 for more information about this data set. Load this data frame into RStudio.

(a) Throw out data from the data frame for people with no response recorded under `Prefer` or `CareAbout`. Hint: If you called your data frame `df`, use a command of the form

```
dfs <- subset(df, ???)
```

with an appropriate condition `???` to create a subseted data frame `dfs`. The subsetted data frame `dfs` should have 1087 observations. After this, you might run the following lines of code to clean some things up:

```
dfs$Prefer <- as.factor(as.character(dfs$Prefer))
dfs$CareAbout <- as.factor(as.character(dfs$CareAbout))
```

(b) Generate a contingency table for `Prefer` and `CareAbout`. (Hint: Use the `table` function with two inputs, namely, the two relevant columns of your subsetted data frame.)

(c) 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?

(d) If you ran a chi-square test to test the hypothesis that `Prefer` and `CareAbout` are uncorrelated, how many degrees of freedom would you have?

(e) The R function `chisq.test` takes as input a contingency table and outputs the p-value of the data under the hypothesis that the two variables are uncorrelated. Calculate the p-value of our survey data under the hypothesis that `Prefer` and `CareAbout` are uncorrelated using a command of the form

```
chisq.test(???)
```

where `???` is the command you used to generate the table in part (b) above.

(f) What conclusion can you draw from the p-value calculation above?

Problem 5. If you just have a contingency table (not a data frame), you can still input your 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 the data in problem 3.

```
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 and check your answer to problem 3.

[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.