Homework 5

```
library(dplyr)
library(ggplot2)
```

Due Tuesday August 2 at 10am

Assumptions in inference

There are several websites around that find interesting correlations in seemingly unrelated things. For example, the site Spurious Correlations shows a high correlation between the divorce rate in Maine and per capita consumption of margarine in the U.S.

```
# from http://tylervigen.com/view_correlation?id=1703
divorce = c(5, 4.7, 4.6, 4.4, 4.3, 4.1, 4.2, 4.2, 4.2, 4.1)
margarine = c(8.2, 7, 6.5, 5.3, 5.2, 4, 4.6, 4.5, 4.2, 3.7)
```

1. Make a scatterplot of `divorce` vs. `margarine`. Do the variables appear related?
2. Compute the correlation between `divorce` and `margarine` and use the bootstrap to estimate the SD of your estimate. Report the usual 2 SE confidence interval. Is 0 in this confidence interval? Hence what is the conclusion about whether the divorce rate and the margarine consumption are related, based on this data analysis?
3. Can you think of a mechanism by which such spurious correlations might easily arise?
4. In order to trust the answer in 2., what assumption(s) are you making when you use the bootstrap to estimate the SD? (This is an advanced question for an introductory course, but we discussed the essential point in class.) Do you think these assumption(s) hold here?

Bootstrapping with two parameters, using the bootstrap for testing

Most of our tests about parameters were about a 1-dimensional parameter. There are many times where we will want to ask questions about more than one parameter.

The site Real Clear Politics showed several polls taken near the end of October 2016 that suggested the presidential race was tightening. Let's take two of them. First, an NBC News / Survey Monkey poll of 40816 likely voters shows support for Trump at 44% and support for Clinton at 51% (with the rest supporting Other). A second poll, the ABC / Washington post poll of 1128 likely voters shows support for Clinton at 48% and support for Trump at 47% (with the rest supporting Other).

5. Given a poll based on a simple random sample of size `N` with choices `[Trump, Clinton, Other]`, write a function that uses the bootstrap to estimate the SD of the difference in apparent support between Trump and Clinton, i.e. $\hat{p}_C - \hat{p}_T$. (We use the notation p_C = proportion of voters in the population supporting Clinton, $\hat{p}_C$ = proportion of voters in the sample supporting Clinton, likewise for `Trump` and `Other`.) Use this function to form a confidence interval for $p_C - p_T$, the true difference in support for Clinton and Trump, using the NBC / Survey Monkey poll. Repeat using the ABC / Washington Post poll. (In using this function, we are making the assumption that the polls were simple random samples.

In practice, this is often not the case – polling firms use complicated weighting schemes to make their final estimate. This makes our bootstrap calculations somewhat dubious here.)

6. Use the bootstrap to test the hypothesis $H_0 : p_C = p_T$ using each of the two polls. (Hint: you'll want to bootstrap under the null hypothesis. What is our best estimate of the common value of p if $p_C = p_T$? Note this will be different for the two different polls.)
7. Combine your bootstrap samples of $\hat{p}_C - \hat{p}_T$ from the two polls from the first part into a single data frame. Make a 2D density plot using geom_density2d in ggplot. Where is the density plot centered?