University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Lab 7

Part A:
Read the following article (New York Times - Wednesday, 26 November 2014), "Obama to Introduce Sweeping New Controls on Ozone Emissions":

`http://www.nytimes.com/2014/11/26/us/politics/obama-to-introduce-sweeping-new-controls-on-ozone-emissions.html?emc=edit_th_20141126&nl=todaysheadlines&nlid=33035119&_r=0 .`

President Obama's proposed regulation would lower the current threshold for ozone pollution from 75 parts per billion to a range of 65 to 70 parts per billion. This range is less strict than the 60 parts per billion sought by environmental groups, but the Environmental Protection Agency (E.P.A.) keeps open the possibility that the final rule could be stricter.

Use the California ozone data ($n = 175$):

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/ozone.txt", header=TRUE)
```

Answer the following questions:

a. Compute the sample mean and sample standard deviation of ozone in this data set.

b. Simulations-based inference for the mean:
Here is how you can randomly select with replacement a random sample of $n = 175$ from our data:

```
y <- sample(a$o3, replace=TRUE)
```

Note: Because we sample with replacement an observation from our original data may be selected more than once. Find the sample mean of this new sample: `mean(y)`

What to do: Select 1000 samples each one of size $n = 175$. For each sample compute the sample mean. When you finish you will have 1000 sample means. Use them to construct a histogram. Finally, find a 95% confidence interval for the population mean. *Hint*: You need to write a short for loop. See for example the for loop we used for testing one proportion. Of course here we will sample from the ozone data and each time we will compute the sample mean. But the idea of a for loop is the same!

Figure 1: Emissions from a power plant in Kentucky. Source: *New York Times.*

Part B:
Using the same data we will test the hypothesis that two population means are equal. We will divide the number of ozone monitoring stations into two parts. You can choose north and south, east and west, etc. There are $n = 175$ ozone monitoring stations, so approximately you should have about half of the data set in each group. Perhaps you can use the 36th parallel to divide the data into south and north. Suppose n_1 observations are in group 1 and n_2 observations are in group 2. Compute the sample mean difference $\bar{x}_1 - \bar{x}_2$. We will use this later at the end of the lab!

Under the null hypothesis the two means are equal, so you can treat the two samples as one sample. We then sample without replacement n_1 observations from the data set (and therefore the remaining are the n_2 observations). Compute the sample means and find the difference $\bar{x}_1 - \bar{x}_2$. Repeat 1000 times and construct the histogram of the 1000 sample means differences. To test the hypothesis that the two means are equal we simply count how many of the simulated sample means differences exceed our sample means difference from the original data.