CS57300 Homework Shaobo Fang (00242-60324)	— Homework 2 , Problem 1 —	(February 19, 2016)

Q1:
i. Do you agree with Jimmy or Ronald? Justify using statistical arguments (e.g. equations) with an example or counter-example.
ii. Using the 2012 donation data above, explain in detail how you would verify the Democrats' claims. Give a formal statistical argument and present your conclusions. Also argue whether you can prove the DEMs to be right or wrong.
A:
(i). Given that no data is provided, a solid conclusion can hardly be drawn. That being said, I am more intend to agree with Ronald's at this point, that the the number obtained from an article proves nothing.

1. DEM's argument is that DEM gets smaller donations over the years (a hypothesis) while the evidence Jimmy provided is the number for 2012 only (incomplete data). We can not draw a solid conclusion without testing hypothesises. We will design a hypothesis testing in the following part
2. Media is not necessarily trustworthy. To verify the empirical mean we have conducted the computation ourselves based on the data:

$$empirical\ mean:\ \mu^{(DEM)} = \frac{\sum_{i=1}^{N} x_i^{(DEM)}}{N}$$

   Similarly,

$$empirical\ mean:\ \mu^{(GOP)} = \frac{\sum_{i=1}^{N} x_i^{(GOP)}}{N}$$

   If we do not exclude negative samples:
   $\mu_{DEM} = 1887$ and $\mu_{GOP} = 2064$, which match the numbers from the article.
   However, if we exclude the negative samples in the computation we obtain:
   $\mu_{DEM} = 1980$ and $\mu_{GOP} = 2150$, which are different from the numbers from the article.

(ii). We conduct hypothesis testing from two samples (democratic donation pool and republican donation pool). Denote x_1 is drawn from republican pool, and x_2 is drawn from democratic pool. The procedures to conduct the hypothesis testing would be:

1. Compute the empirical standard error:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

   where

$$s_i^2 = \frac{1}{n_i} \sum_{k=1}^{n_i} (x_k^{(i)} - \bar{x}^{(i)})^2$$

   and $\bar{x}_i$ are the corresponding empirical class mean.
2. The degrees of freedom would then be:

$$DF = \left\lfloor \frac{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^2}{(\sigma_1^2/n_1)^2/(n_1-1) + (\sigma_2^2/n_2)^2/(n_2-1)} \right\rfloor$$

3. Compute test statistic (t-score, also known as Welsh's t)

$$t_d = \frac{(\bar{x}_1 - \bar{x}_2) - d}{SE}$$

4. Our **Null hypothesis** is: $\mu_1 \geq \mu_2$, accordingly: $\mu_1 - \mu_2 \geq 0$. The **Alternative Hypothesis** would than be: $\mu_1 - \mu_2 < 0$

5. The p value can then be computed as:

$$\int_{d \geq 0} P[T_{DF} > t_d] \partial d$$

The empirical variance of the GOP is $\sigma_1^2 = \sigma_{GOP}^2 = 101468387$.
The empirical variance of the DEM is $\sigma_2^2 = \sigma_{DEM}^2 = 26660303$.
Use the equation of empirical standard error:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$SE = 36.0479$$

Then compute the test statistic (t-score) as:

$$t_d = \frac{(\bar{x}_1 - \bar{x}_2) - d}{SE}$$

$$t_d = 4.712$$

We assume the $\bar{x}_i$ follows the Gaussian Distribution. To verify our assumption, we plot the histogram of random 200 sub-samples for each class and make visual observation (seems this is sufficient for now).

As this is a one-tail t-test, we compute the probability to reject the Null Hypothesis as:

$$\int_{d \geq 0} P[T_{DF} > t_d] \partial d$$

Q2: Can you add?
A:
The question asks if I can add.

$$1/3 + 2/3 = 1 \tag{1}$$

The equation shows I can.