Stat 332 – Assignment 2
Prof. Samuel Wong – Winter 2019
Due: Friday, February 8 at 10:30am on Crowdmark

General instructions: You may submit your work using one or more of the following ways:

- Type out work, for example using 'Latex', 'R Markdown', or Word.
- Present scans/photographs of handwritten work. If you choose this option ensure your work is legible. Illegible work will receive no credit.

For data analysis problems: When you are using R (which we strongly encourage), you must clearly present your final answers in addition to the commands you used.

1. An activist group wishes to study social and political attitudes of residents of the US. Suppose part of their survey is as follows.

   - Do you think politicians are dishonest, or are they like most other people?
     Yes / No / Don't know
   - Many experts note that national spending on social programs is unsustainable. Do you agree that it is wise to cut spending on social programs?
     Yes / No / Don't know
   - Please tell us which social programs should be cut, and why.

   For each of the survey questions above: (i) briefly discuss a potential problem with the way the question is presented, (ii) suggest how you might improve the question.

2. *Data analysis.* Download the dataset tree.csv. We are interested in the average age of 1210 trees in a region. To get the age of a tree accurately, its rings need to be counted (a tedious task). On the other hand, it is much easier to measure a tree's diameter, a measurement that should be correlated with its age. Suppose that we measured all 1210 tree diameters, and found the average diameter to be 10.3. We then took a SRS of 21 trees, and for these we measured both its diameter and age.

   (a) Ignoring the diameter information for now, estimate the average age of this population based on the SRS. Report an estimate and SE of the estimate.
   (b) Report a scatterplot of the data, with Diameter on x-axis and Age on the y-axis.
   (c) Estimate the mean tree age using ratio estimation. Report an estimate and SE of the estimate.
   (d) Estimate the mean tree age using regression estimation. Report an estimate and SE of the estimate.
   (e) Compare and comment briefly on your three estimates in parts (a), (c), (d).

3. *Variance comparison for ratio and regression estimation of a mean.* Both ratio and regression estimation can be used to account for a known μ_X when estimating μ_Y. An estimate for the variance of $\hat{\mu}_Y + b(\mu_x - \hat{\mu}_x)$ based on a SRS of (x_i, y_i)'s, as shown in class, is

$$\frac{1}{n}\left(1 - \frac{n}{N}\right)(\hat{\sigma}_y^2 + b^2\hat{\sigma}_x^2 - 2b\hat{\sigma}_x\hat{\sigma}_y\hat{\rho})$$

   Now plug in the values of b that we would use for ratio and regression estimation of μ_Y, to show that $\widehat{Var}(\tilde{\mu}_{ratio}) \geq \widehat{Var}(\tilde{\mu}_{reg})$.

4. *Analyzing a stratified sample.* Download the dataset radon.csv. This is data from a radon survey, administered to homes in Minnesota to estimate the prevalence of high indoor radon concentrations. It is believed that high levels of radon are a health risk for lung cancer. The variables are

   - *strata*: The county number of the home
   - *radon*: The radon level (in picocuries/L)
   - *exceed*: Whether that home exceeds the recommended level by the USEPA of 4 pCi/L.
   - N: The number of homes in that county

   (a) Explain why it is sensible for the researchers to stratify by county.

   (b) Look at boxplots of radon levels by county, and comment briefly.

   (c) Estimate the average the radon level of a home in Minnesota. Report the estimate, SE of the estimate, and a 95% CI.

   (d) Estimate the proportion of homes in Minnesota that should be fixed, according to USEPA recommendations. Report the estimate, SE of the estimate, and a 95% CI.