Stat 332 – Assignment 3
Prof. Samuel Wong – Winter 2019
Due: Saturday, March 9 at 11:59pm 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. *Cluster sampling variance simulations.* Download the dataset clusters.csv. For this problem, this dataset of 100 units is our "population". You will investigate, computationally, the results of cluster sampling from this population. Units that have the same `cluster` value belong to the same cluster. The response of interest is in the `price` column. Consider two sampling designs:

   - Design A: Draw a SRS of 40 units from the 100 units
   - Design B: Draw a SRS of 4 clusters and measure all 10 units within each of the sampled clusters

   Call $\tilde{\mu}_A$ and $\tilde{\mu}_B$ the estimators of the mean based on Designs A and B, respectively.

   (a) Perform Design A 1000 times. Calculate the estimated population mean for each sample, and report the variance of the 1000 means. [Please don't do this manually. Use a `for` loop or other similar technique!]

   (b) Report the means of the 10 clusters.
   [*Hint:* The R function `tapply(x,y,mean)` will return a vector of means of `x` grouped by each unique value of `y`.]

   (c) Perform Design B 1000 times. Calculate the estimated population mean for each sample, and report the variance of the 1000 means.

   (d) Use your simulation results to estimate the value of $\dfrac{Var(\tilde{\mu}_B)}{Var(\tilde{\mu}_A)}$.

   (e) Calculate the true value of $\dfrac{Var(\tilde{\mu}_B)}{Var(\tilde{\mu}_A)}$ for this population.

2. *Analyzing a cluster sample.* To assess the quality of oranges produced by a Florida citrus grower, suppose an inspector took a sample of 20 crates. Within each crate, 10 oranges were randomly selected and their quality recorded. The data file citrus.csv contains the following variables for each orange:

   - *onum*: a unique identifier for the orange
   - *quality*: the quality score
   - *cnum*: the crate ID from which it was drawn
   - *fpc1*: the total number of crates
   - *fpc2*: the number of oranges in its crate

   (a) Report an estimate and SE of mean orange quality based on this cluster sampling design.

(b) Suppose a fellow inspector ignored that the data came from a cluster design, and treated the data as if it were a SRS of 200 oranges. What would be his SE for mean orange quality? Assume that the SRS FPC ≈ 1.

(c) The fellow inspector remarks: "My analysis is easier and we get a smaller SE." How would you respond, in plain English?

3. The following are the burning times of flares (in minutes) of two different types of torch design. The design engineers are interested in comparing the mean burning times for the two types of design. Sample torches are tested one-at-a-time and the burning times of flares are recorded.

Type 1	65.2	82.3	81.1	67.4	57.3	59.5	66.1	75.0	82.2	70.0
Type 2	67.9	59.8	75.0	72.7	86.3	77.8	62.7	85.6	69.0	82.9

(a) Define the response, factor and its levels, treatments, and experimental units.

(b) To design the experiment, what are the key steps to ensure that it is a completely randomized design?

(c) Is there a significant difference between the two mean burning times? Use $\alpha = 0.05$.

4. Ten judges are randomly selected to rate two brands of beer A and B according to taste (scale: 1 to 10). Each judge is asked to taste both brands of beer and the scores are given below:

Judge	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Beer A	7	3	9	7	6	8	2	6	8	4
Beer B	5	4	7	7	5	6	2	7	6	3

(a) What kind of randomization do you recommend for this experiment? Why?

(b) Determine if there is any significant difference in taste for the two brands of beer. Use $\alpha = 0.05$.

(c) If someone treats this dataset as if it was from a completely randomized design, what would be the conclusion for part (b)?