DS 100: Principles and Techniques of Data Science **Date: May 3, 2019**

Exam Review

Name:

Sampling

1. A political scientist is interested in answering a question about a country composed of three states A, B and C. These states have exactly 100, 200, and 300 voting adults respectively. Within each state, assume that there are 10 towns and the population of voting adults in each state is split among its 10 towns uniformly. So for example, state A will have 10 towns, each with $\frac{100}{10} = 10$ voting adults. In addition, assume that each town has an equal number of male and female voting adults. So for example, state A will have 5 male and 5 female voting adults in each town.

   To answer this question, a political survey is administered by randomly sampling 3, 8, and 12 voting adults from each town without replacement in each state, respectively.

   (a) Which sampling plan was used in the survey?

   (a) cluster sampling (b) stratified sampling (c) quota sampling (d) census

   (b) How many strata are there in total?

   (c) What is the probability that the sample generated from the sampling strategy from part (a) will comprise of all females?

EDA & Visualization

2. For each of the following scenarios, determine which plot type is *most* appropriate to reveal the distribution of and/or the relationships between the following variable(s). For each scenario, select only one plot type. Some plot types may be used multiple times.

   A. histogram B. pie chart C. bar plot D. line plot

   E. side-by-side boxplots F. scatter plot G. stacked bar plot H. overlaid line plots

   (a) Sale price and number of bedrooms for houses sold in Berkeley in 2010.

   (b) Sale price and date of sale for houses sold in Berkeley between 1995 and 2015.

   (c) Infant birth weight (grams) for babies born at Alta Bates hospital in 2016.

(d) Mother's education-level (highest degree held) for students admitted to UC Berkeley in 2016.

(e) SAT score and HS GPA of students admitted to UC Berkeley in 2016.

(f) The percentage of female student admitted to UC Berkeley each year from 1950 to 2000.

(g) SAT score for males and females of students admitted to UCB from 1950 to 2000

Estimation

3. Suppose that we try to predict a donkey's weight, y_i from its sex alone. There are 3 different sexes of donkeys, so the sex variable has values: gelding, stallion, and female). Consider the following model consisting of dummy variables:

$$y_i = \beta_F D_{F,i} + \beta_G D_{G,i} + \beta_S D_{S,i}$$

where the dummy variable $D_{F,i} = 1$ if the i^{th} donkey is female and $D_{F,i} = 0$ otherwise. The dummy variables D_G and D_S are dummies for geldings and stallions, respectively.

Prove that if we using the following loss function:

$$L(\beta_F, \beta_G, \beta_S) = \sum_{i=1}^{n} \left(y_i - (\beta_F D_{F,i} + \beta_G D_{G,i} + \beta_S D_{S,i})\right)^2$$

then the loss minimizing value $\hat{\beta}_F = \bar{y}_F$ where $\bar{y}_F$ is the average weight of the female donkeys.

Optimization

4. Fix the following buggy Python implementation of gradient descent:

```
def grad_descent(X, Y, theta0, grad_func, max_iter = 1000, alpha):
    """X: A 2D array, the feature matrix.
    Y: A 1D array, the response vector.
    theta0: A 1D array, the initial parameter vector.
    grad_func: Maps a parameter vector, a feature matrix, and a
        response vector to the gradient of some loss function at the
        given parameter value.
    alpha: Learning rate at each step. The return value is a 1D
        array."""
    theta = theta0
    for t in range(1, max_iter+1):
        grad = grad_func(theta, X, Y)
        theta = theta0 + 1/alpha * grad
    return grad
```

5. Suppose you are given a dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$ is a one dimensional feature and $y_i \in \mathbb{R}$ is a real-valued response. You use $E[Y|X] = \theta(x_i)$ to model the data with β as the model parameter. You choose to use the following regularized empirical risk:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \theta(x_i))^2 + \lambda\beta^2$$

   (a) This regularized empirical risk is best described as:

      (a) mean absolute error with L^2 regularization.
      (b) mean squared error with L^2 regularization.
      (c) mean squared error with L^1 regularization.
      (d) empirical Huber risk with λ regularization.

   (b) Suppose you choose the model $E[Y|X] = \beta x_i^3$. Using the above objective function, derive the risk minimizing estimate for β.

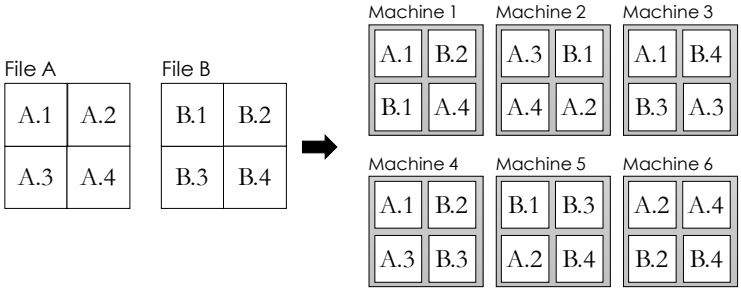
Inference

6. **True or False.** Determine whether the following statements are true or false.
   (a) Suppose we have 100 samples drawn independently from a population. If we construct a 95% confidence interval for each sample, we expect 95 of them to include the **sample** mean.
   (b) We often prefer a pseudo-random number generator because our simulations results can be exactly reproduced by controlling the seed.

7. Suppose we have a Pandas Series called **`thePop`** which contains a census of **25000 subjects**. We also have a simple random sample of **400 individuals** saved in the Series **`theSample`**. We are interested in studying the behavior of the bootstrap procedure on the simple random sample. Fill in the blanks in the code below to construct **10000 bootstrapped estimates** for the **median**.

```
boot_stats = [

        ____________________________

        .sample(n = _______________, replace = _______________)

        .____________________________()


        for j in range(____________________)
    ]
```

8. Consider the following layout of the files `A` and `B` onto a distributed file-system of 6 machines.



Assume that all blocks have the same file size and computation takes the same amount of time.

(a) (1 point) If we were to lose machines $M1$, $M2$, and $M3$ which of the following file or files would we lose (select all that apply).
A. File `A` B. File `B` C. We would still be able to load both files.

(b) (1 point) If each of the six machines fail with probability p, what is the probability that we will lose block $B.1$ of file `B`.?

A. $3p$ B. p^3 C. $(1-p)^3$ D. $1-p^3$