# STAT 446: Final Exam

For the exam, you may use any course materials or information you find on the internet; however, use citations as appropriate. You may not discuss questions or work together with classmates. You are welcome to contact the instructor with any questions. For complete (and partial credit) please show all work and turn in a reproducible document (PDF or DOC) as well as your source code (.RMD).

## Yelp Dataset

This first part of the exam will be focused on a data set with Yelp restaurant reviews. The dataset contains reviews for a subset of restaurant locations in Arizona, Nevada, and Wisconsin.

```
set.seed(12032019)
Yelp <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/yelp_exam.csv')
```

### Q1. Simple Random Sampling (26 points)

For this question, the goal is to compare average star ratings across a set of chain restaurants.

The dataset `Yelp` contains average ratings across six chain restaurants: Burger King, McDonald's, Panda Express, Pizza Hut, Subway, and Taco Bell.

#### a. (4 points)

Assume the goal is to make inferences about the McDonald's restaurants in the dataset. Define the sampled population and sampling frame, then defend a scope of inference for an appropriate target population for this scenario.

#### b. (8 points)

If researchers are interested in understanding the mean rating across the set of 114 McDonald's restaurants to within .2 stars, estimate how many samples need to be collected. State and defend any assumptions in your calculations. You may use the ten restaurants in `McDonalds_10` as a pilot study.

```
McDonalds_10 <- Yelp %>% filter(name == "McDonald's") %>% sample_n(10)
```

#### c. (4 points)

Regardless of your results in part b, 50 SRS samples have been taken. Define a simple random sample. Then, in this scenario, what is the parameter of interest that you hope to identify andvwhat statistic you will use as your point estimator.

#### d. (6 points)

Use the sample in `McDonalds_50` to construct a point estimate and a confidence interval for the mean rating.

```
McDonalds_50 <- Yelp %>% filter(name == "McDonald's") %>% sample_n(50)
```

#### e. (4 points)

Summarize the results in part d in a way that a food critic could understand.

### Q2 Statified Random Sampling. (24 points)

Based on your dining experiences at the chain restaurants, you believe that there may be differences in ratings across the restaurants. So you will implement a stratified sampling approach to estimate the mean rating across all of the restaurants in the dataset.

#### a. (4 points)

Define a stratified random sample and then describe why a stratified sampling approach can be useful.

#### b. (4 points)

Assume 91 samples will be taken. What information would be necessary to optimally allocate those samples across the 6 strata?

#### c. (6 points)

A stratified sample has been taken for you in `Yelp_Strat`. First calculate the confidence intervals and point estimates of the mean star rating within each of the 6 strata.

```r
Yelp_Strat <- Yelp %>% group_by(name) %>% sample_frac(.25)
```

#### d. (6 points)

Now calculate a confidence interval for the point estimate of the overall mean star rating across the 6 restaurants.

#### e. (4 points)

Summarize the results in part c and part d in a way that a food critic could understand.

## Air BNB Dataset

The last part of the exam uses a dataset that contains rental listings from airbnb.com in Los Angeles, CA.

```
airbnb <- read_csv('http://www.math.montana.edu/ahoegh/teaching/stat446/airbnb_LA.csv')
```

### Q3. Cluster Sampling (22 points)

In order to facilitate efficient data collection, the city will only sample a subset of neighborhoods.

#### a. (4 points)

Define cluster sampling. Then in this situation, describe the primary sampling units and secondary sampling units.

#### b. (2 points)

Given that rentals within a neighborhood are likely similar, would this suggest that cluster sampling or stratified sampling would result in lower variance?

#### c. (6 points)

A single stage cluster sample has been taken for you in the dataset `cluster_sample`. Compute a point estimate and confidence interval for the mean listing price of airbnb units in LA.

```
neighborhoods <- airbnb %>% group_by(neighbourhood) %>% tally()

neighborhoods_sample <- neighborhoods %>% sample_n(50)

cluster_sample <- airbnb %>% filter(neighbourhood %in% neighborhoods_sample$neighbourhood)
```

#### d. (4 points)

Describe how the calculations in part c would change if the sample was selected with probability proportional to the number listings in the neighborhood. You can assume samples were taken *with* replacement.

#### e. (6 points)

Now we will take a SRS of 5000 accomodations `SRS_airbnb`. Use a bootstrap procedure to estimate confidence intervals for the mean price across the four room types: Entire home/apt, Hotel room, Private room, and Shared room.

```
SRS_airbnb <- airbnb %>% sample_n(5000)
```