SOC 301 Practice Problems for Final Exam

Make sure to use your cheatsheets to solve these problems as you prepare.

The final exam will be Tuesday, December 13th from 8:30 AM to 11:00 AM PST in Walter Annex 101. It will be a "pencil-and-paper" exam, but you can use the `dplyr` and `ggplot2` cheatsheets. You should know precisely where things are on those cheatsheets and I will be harsher on syntax on this exam since so much of it is laid out clearly on those cheatsheets.

To prepare for the exam, you should **at the very minimum** redo all of the quizzes, both of the exams, and the Exam II review. A ZIP file containing blank versions of all of these is available here. I will likely take some problems directly from those assignments for the final exam. In addition, you can find many sample problems below that I may take straight from or modify slightly for the final exam.

1 OKCupid profiles data

Recall the OKCupid profiles data stored in the `profiles` data frame in the `okcupiddata` package. A smaller number of variables was selected from this data and is stored in the `profiles_little` data frame with the first few rows shown below:

```
library(okcupiddata); data(profiles)
profiles_little <- profiles %>% select(age:drugs, height, orientation, sex, status)
head(profiles_little, 16) %>% kable()
```

age	body_type	diet	drinks	drugs	height	orientation	sex	status
22	a little extra	strictly anything	socially	never	75	straight	m	single
35	average	mostly other	often	sometimes	70	straight	m	single
38	thin	anything	socially	NA	68	straight	m	available
23	thin	vegetarian	socially	NA	71	straight	m	single
29	athletic	NA	socially	never	66	straight	m	single
29	average	mostly anything	socially	NA	67	straight	m	single
32	fit	strictly anything	socially	never	65	straight	f	single
31	average	mostly anything	socially	never	65	straight	f	single
24	NA	strictly anything	socially	NA	67	straight	f	single
37	athletic	mostly anything	not at all	never	65	straight	m	single
35	average	mostly anything	socially	NA	70	straight	m	available
28	average	mostly anything	socially	never	72	straight	m	seeing someone
24	NA	NA	often	NA	72	straight	m	single
30	skinny	mostly anything	socially	never	66	straight	f	single
29	thin	mostly anything	socially	never	62	straight	f	single
39	fit	strictly anything	socially	NA	65	straight	f	single

Write down the FULL `dplyr` commands that will produce the following tables for the `profiles_little` data frame (one for each part):

a) the median and mean `age` based on values of `drinks`

b) the top five heights based on `sex` for `single` status

c) the total number of each category in `body_type`

d) the total number in each category for `drinks` combined with `sex`. In other words, lay out all possible combinations of `drinks` and `sex` and give how many are in each combination.

e) choose only the data with `age` between 40 and 50 or never for `drugs` or vegetarian for `diet`

f) pick only the data with strictly vegan for `diet` and socially for `drinks`

2 There is Only One Test

Replicate the "There is Only One Test" diagram from memory. Be sure to know how to apply the diagram to each of the five scenarios we discussed in class:

- One Mean
- One Proportion
- Two Proportions
- Two Means (Independent Samples)
- Two Means (Paired Samples)

3 Inference Question 1

This example involves thinking about county level data on the percentage of Asian American residents by gender: `male`, `female`, or `non-binary`. All that is collected is a random representative sample of 300 US counties. Describe how the process of bootstrapping could be used to create plots and a range of possible values for the percentage of Asian American residents, on average, by county AND gender throughout the entire US.

- Layout what the tidy data set would look like for this sample of 300 counties.
- You should carefully lay out each step of the bootstrapping process being as specific as possible. For example, you should be clear about the size of each sample and how many times you are repeating the process.
- Additionally, you should sketch a plot (free hand) of what the three distributions might look like and how one could use the distributions to help address the problem. (Your numbers may not necessarily be correct, but it's important to get a sense of what the plot might look like.)

4 Inference Question 2

Now suppose we are interested in comparing the mean percentage of hispanic residents to the mean percentage of black residents by county in the southernmost 10 states in the US. Researchers believe that hispanic residents make up a larger mean percentage based on immigration patterns and other factors. Researchers have collected a random sample of 50 counties from these ten states with the percentage of hispanic and black residents in those 50 counties.

- Layout what the tidy data set would look like for this sample of 50 counties from the southernmost 10 states.
- Describe how each of the elements of the "There is Only One Test" diagram applies to this problem.
- Explain in detail how shuffling with note cards and a calculator (and many hours...) could be used to create a null distribution.

5 R errors

Over the course of this semester you've encountered many errors in running R code. As I suggested at the beginning of the semester, it's good practice to keep track of these errors so that when you run into them you can easily diagnose them. In the problems below, explain why the error following was given.

a)
```
library(nycflights13); library(dplyr); library(ggplot2); data(flights)
flights %>% filter(carrier %in% c("UA", "AA")) %>%
    ggplot(aes(x = dep_delay, y = arr_delay)) +
    geom_point(color = carrier)
```

```
Error in layer(data = data, mapping = mapping, stat = stat, geom = GeomPoint, :  object
'carrier' not found
```

b)
```
library(nycflights13); library(dplyr); library(ggplot2); data(flights)
flights %>%
    ggplot(aes(x = dist)) +
    geom_histogram(fill = purple, bins = 10)
```

```
Error in layer(data = data, mapping = mapping, stat = stat, geom = GeomBar, :  object
'purple' not found
```

c)
```
library(readr)
my_data <- read_csv("my_data.csv")
```

```
Error:  'my_data.csv' does not exist in current working directory ('/home/cismay/final_project').
```

d)
```
library(nycflights13); library(dplyr); library(ggplot2); data(flights)
flights %>% select(month, day, dep_delay) %>%
  ggplot(aes(x = arr_delay)) +
  geom_histogram(color = "white")
```

```
Error in eval(expr, envir, enclos) :  object 'arr_delay' not found.
```

e)
```
library(dplyr); library(ggplot2); library(tidydata); data(raceelection)
raceelection %>%
  ggplot(aes(x = race)) +
  geom_bar()
```

```
Error in library(tidydata) :  there is no package called 'tidydata
```

f)
```
library(nycflights13); library(dplyr); library(ggplot2); data(flights)
flights %>% filter(carrier %in% c("UA", "AA")) %>%
  ggplot(aes(x = carrier, y = distance)) +
  geom_histogram(bins = 20)
```

```
Error:  stat_bin() must not be used with a y aesthetic.
```

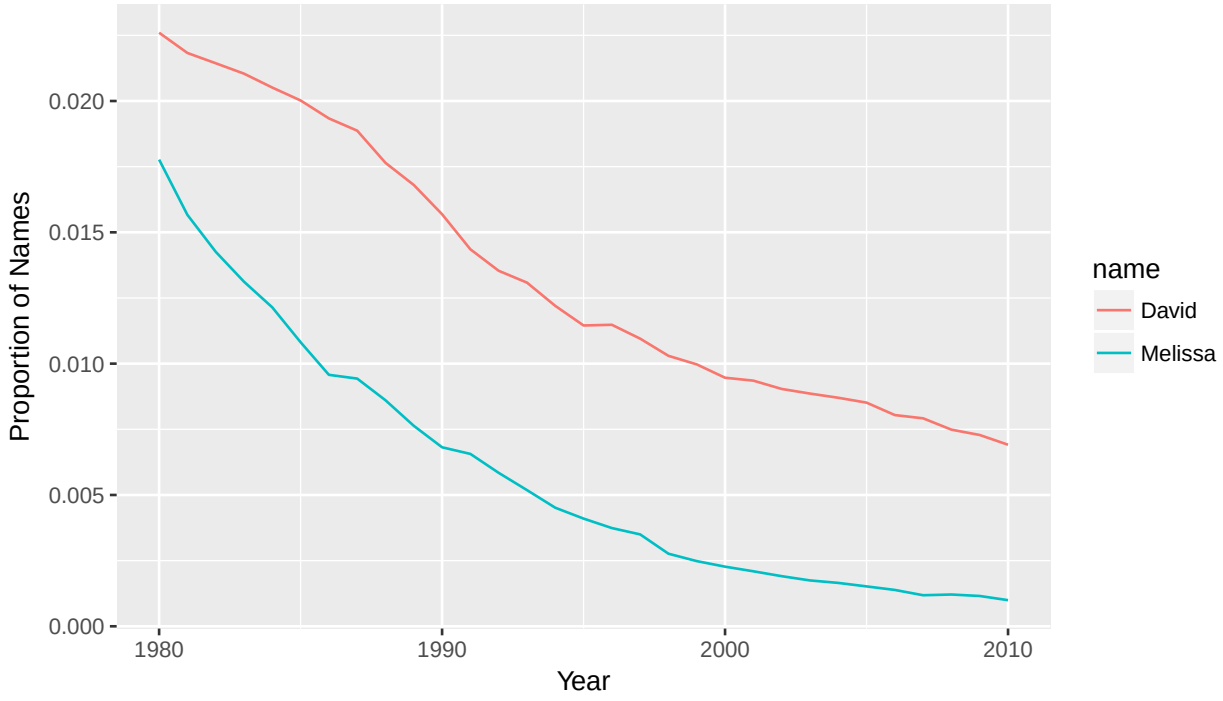
6 Producing a plot for babynames

Give the `dplyr` and `ggplot2` code needed to produce the following plot from the `babynames` data frame in the `babynames` package. Recall that this data can be loaded via the code below.

Note that this selects all males with names David as well as all females with name Melissa. It also focused on only the years 1980 to 2010 inclusive. I've also included the first few rows of the subsetted data for reference. Notice that you'll need to set the labels on the axes as well using `ggplot2`.

```
library(babynames)
data(babynames)
```

```
## # A tibble: 10  5
##     year   sex    name     n  prop
##    <dbl> <chr>   <chr> <int> <dbl>
## 1   1980     F Melissa 31631 0.018
## 2   1980     M   David 41913 0.023
## 3   1981     F Melissa 28006 0.016
## 4   1981     M   David 40643 0.022
## 5   1982     F Melissa 25860 0.014
## 6   1982     M   David 40441 0.021
## 7   1983     F Melissa 23472 0.013
## 8   1983     M   David 39192 0.021
## 9   1984     F Melissa 21886 0.012
## 10  1984     M   David 38471 0.021
```



7 Inference Question 3

Refer to the example provided at http://ismayc.github.io/teaching/sample_problems/two-means-indep.html. Explain in detail what each of the chunks of code below produces. Your explanation should include a line-by-line description of what each line of code is doing. Also describe what point(s) in the "There is Only One Test" graphic the chunk relates to.

a)
```
cleSac <- read.delim("cleSac.txt") %>%
  rename(metro_area = Metropolitan_area_Detailed,
    income = Total_personal_income) %>%
  na.omit()
```

b)
```
inc_summ <- cleSac %>% group_by(metro_area) %>%
  summarize(sample_size = n(),
    mean = mean(income))
```

c)
```
xbar_cle <- inc_summ$mean[1]; xbar_sac <- inc_summ$mean[2]
obs_diff <- xbar_sac - xbar_cle
```

d)
```
set.seed(2016)
many_shuffles <- do(10000) *
  (cleSac %>%
     mutate(income = shuffle(income)) %>%
     group_by(metro_area) %>%
     summarize(mean_inc = mean(income))
   )
```

e)
```
null_distn <- many_shuffles %>%
  group_by(.index) %>% summarize(diffmean = diff(mean_inc))
```

f)
```
null_distn %>% ggplot(aes(x = diffmean)) +
  geom_histogram(bins = 30, color = "white") +
  geom_vline(color = "red", xintercept = obs_diff) +
  geom_vline(color = "red", xintercept = -obs_diff)
```

g)
```
null_distn %>%
  filter( (diffmean >= obs_diff) | (diffmean <= -obs_diff) ) %>%
  nrow() / nrow(null_distn)
```