SOC 301
Social Statistics
Professor Chester Ismay

Name: KEY
April 11, 2017

# Midterm Exam

- Read carefully through each problem. Take your time and relax.
- Please write your name at the top of each page of the exam.
- You are to work on this exam alone, but you can use the following RStudio Cheatsheets:
  - **Data Transformation with dplyr**
  - **Data Visualization with ggplot2**
- Point allocations for each part of each problem are in the box to the left of the question. Please use this as a way to help you manage your time during the exam.
- One goal of my exams (and this course) is for you to show me that you have gained the ability to think critically and statistically about problems. If you are ever in doubt, carefully write what you are struggling with and your thought processes on the back of the sheet of paper corresponding to that problem.

  Use the back of the sheets as scratchwork. You may receive partial credit if I can understand what you are thinking. Place your final answer on the front page below the problem statement. Remember that this is largely an exam focused on checking for your understanding of writing R code and interpreting statistical output though.

  BLANK ANSWERS I can guarantee will NOT provide you with any credit.
- You can assume that all needed R packages have been installed and that the `library` function has been run on all of these packages.
- Take time after you have completed the exam to carefully review your answers. Make sure that you have answered all parts to all questions asked.

| Question: | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Points: | 12 | 13 | 12 | 4 | 9 | 50 |

1. Recall the `gapminder` data frame in the `gapminder` package. Five rows of the `gapminder` data frame are below.

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Equatorial Guinea | Africa | 2002 | 49.348 | 495627 | 7703.496 |
| Japan | Asia | 1982 | 77.110 | 118454974 | 19384.106 |
| Madagascar | Africa | 1962 | 40.848 | 5703324 | 1643.387 |
| Saudi Arabia | Asia | 1957 | 42.868 | 4419650 | 8157.591 |
| Turkey | Europe | 1962 | 52.098 | 29788695 | 2322.870 |

Also recall the `country_groups` data frame which gives the `region` and `subregion` classification of each country. Three rows of `country_groups` are below.

| name | region | subregion |
|---|---|---|
| Brazil | Americas | South America |
| Ireland | Europe | Northern Europe |
| Lesotho | Africa | Southern Africa |

For each of the following questions, be sure to use `dplyr` and `%>%` whenever possible for full credit.

3 (a) What code is necessary to add the `subregion` values from `country_groups` to `gapminder` in a new data frame with name `gap_full`? (Five lines of `gap_full` are below for reference.)

| country | continent | year | lifeExp | pop | gdpPercap | region | subregion |
|---|---|---|---|---|---|---|---|
| Equatorial Guinea | Africa | 2002 | 49.348 | 495627 | 7703.496 | Africa | Middle Africa |
| Japan | Asia | 1982 | 77.110 | 118454974 | 19384.106 | Asia | Eastern Asia |
| Madagascar | Africa | 1962 | 40.848 | 5703324 | 1643.387 | Africa | Eastern Africa |
| Saudi Arabia | Asia | 1957 | 42.868 | 4419650 | 8157.591 | Asia | Western Asia |
| Turkey | Europe | 1962 | 52.098 | 29788695 | 2322.870 | Asia | Western Asia |

**Solution:**

```
gap_full <- inner_join(x = gapminder, y = country_groups,
                       by = c("country" = "name"))
```

6 (b) What code would be needed to produce the following summarized table for Africa for 2007 using `gap_full`? (Recall that `year` is a numeric variable NOT a character variable.)

```
# A tibble: 5   3
        subregion mean_lifeExp sd_lifeExp
            <chr>        <dbl>      <dbl>
1  Eastern Africa     54.00129   9.909632
2   Middle Africa     51.57544   6.969391
3 Northern Africa     70.20567   5.833684
4 Southern Africa     47.03560   5.662191
5  Western Africa     54.08673   6.732993
```

**Solution:**

```
gap_full %>% filter(year == 2007, continent == "Africa") %>%
  group_by(subregion) %>%
  summarize(mean_lifeExp = mean(lifeExp), sd_lifeExp = sd(lifeExp))
```

3 (c) Which subregion in Africa in 2007 had the least variability in its countries' average life expectancy?

**Solution:** Southern Africa, since its standard deviation is smallest at around 5.66 years.

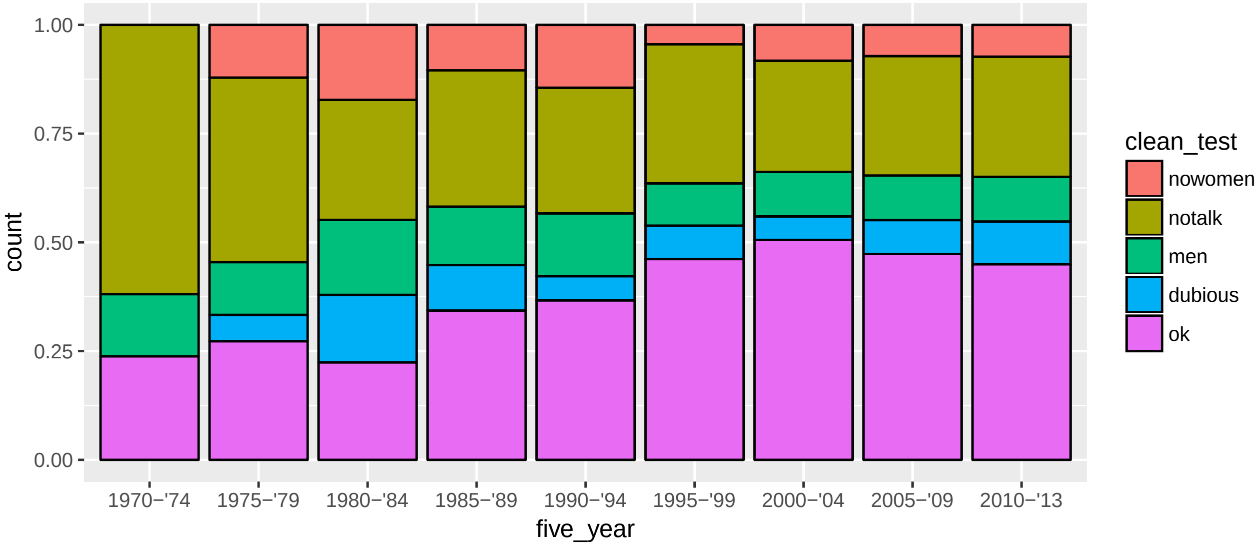
2. Recall the `bechdel` dataset from the `fivethirtyeight` R package. We have modified it here (but kept the same `bechdel` name) as was done on the DataCamp assignments to include five year increments. Six rows of this data frame and some of its columns are below.

| title | year | five_year | clean_test | binary |
|---|---|---|---|---|
| The Skeleton Key | 2005 | 2005-'09 | ok | PASS |
| Astro Boy | 2009 | 2005-'09 | dubious | FAIL |
| Rabbit Hole | 2010 | 2010-'13 | ok | PASS |
| Don't Be Afraid of the Dark | 2010 | 2010-'13 | ok | PASS |
| Final Destination 5 | 2011 | 2010-'13 | men | FAIL |
| 12 Years a Slave | 2013 | 2010-'13 | notalk | FAIL |

3 (a) What is the observational unit in this `bechdel` data frame? Be as specific as possible.

**Solution:** A movie from 1970 to 2013 (that was rated on the Bechdel Test website)

6 (b) What `ggplot2` code is needed to produce the following plot? (Note the color choice for border.)



**Solution:**

```
ggplot(data = bechdel,
       mapping = aes(x = five_year, fill = clean_test)) +
  geom_bar(position = "fill", color = "black")
```

4 (c) Describe how movies have done with respect to passing the Bechdel test over time using this plot.

**Solution:** Focusing on only the `ok` level of the `clean_test` variable, we see that movies have tended to improve in passing the Bechdel test from 25% in 1970-'74 to around 45% in 2010-'13. There has been a decline since 2000-'04 though.

3. Recall the `US_births_2000_2014` data frame in the `fivethirtyeight` package that analyzed baby births in the US.

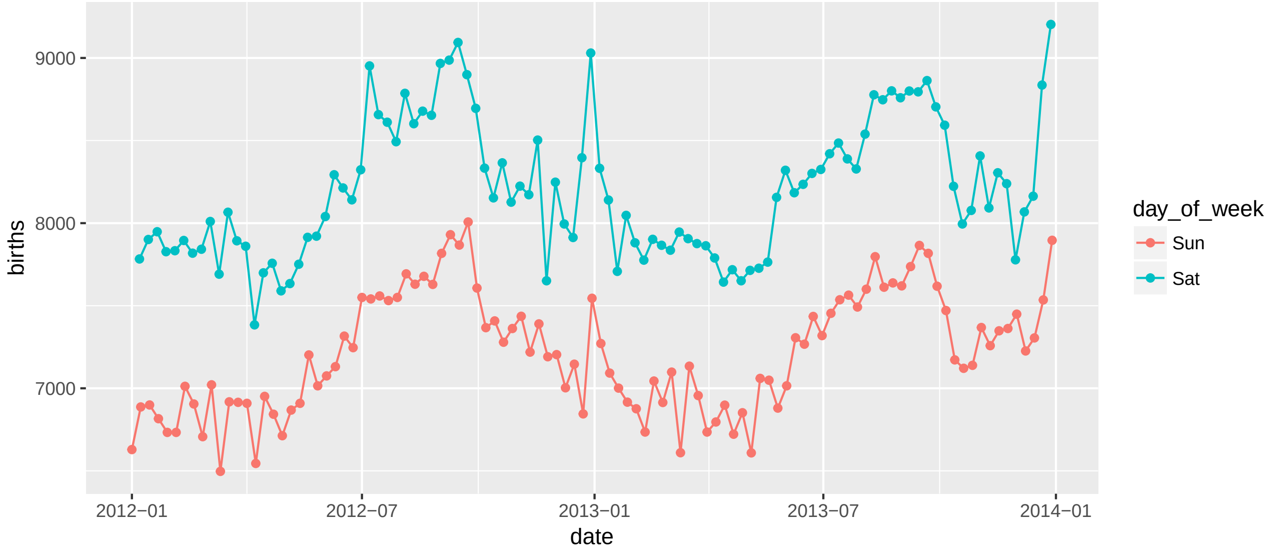
5 (a) Write the `dplyr` code needed to produce a smaller data frame focused on weekend days in 2012 & 2013. Give this data set the name `US_births_12_13`. Six rows of `US_births_12_13` are below.

```
# A tibble: 6  6
   year month date_of_month       date day_of_week births
  <int> <int>         <int>     <date>       <ord>  <int>
1  2012     4            14 2012-04-14         Sat   7699
2  2013     1            27 2013-01-27         Sun   6916
3  2013     5            26 2013-05-26         Sun   6880
4  2013     6            15 2013-06-15         Sat   8235
5  2013     7            14 2013-07-14         Sun   7536
6  2013    12            28 2013-12-28         Sat   9203
```

**Solution:**

```
library(fivethirtyeight)
US_births_12_13 <- US_births_2000_2014 %>% filter(year %in% c(2012, 2013)) %>%
  filter(day_of_week %in% c("Sat", "Sun"))
```

5 (b) What `ggplot2` code is necessary to produce the following plot based on `US_births_12_13`?



**Solution:**

```
ggplot(data = US_births_12_13,
       mapping = aes(x = date, y = births, color = day_of_week)) +
  geom_line() +
  geom_point()
```

2 (c) What date and day of the week corresponds to the highest number of births in 2012 and 2013?

**Solution:** Looking at the plot we see that the highest point occurs near the end of 2013 on a Saturday. We can also look at the rows given in the table above and see that this date must be Saturday, December 28, 2013.

$\boxed{4}$ 4. Match the term with its correct definition.

1. generalizability
2. parameter
3. population
4. sampling
5. sample
6. statistic

(a) The 3. population is the (usually) large pool of observations (instances of observational units) that we are interested in.

(b) The process of selecting observations from a population refers to 4. sampling. There are both random and non-random ways this can be done.

(c) A 2. parameter is a calculation based on one or more variables measured in the population and are almost always denoted symbolically using Greek letters such as $\mu$, $\pi$, $\sigma$, $\rho$, and $\beta$.

(d) A 6. statistic is a calculation based on one or more variables measured in the sample and are usually denoted by lower case Arabic letters with other symbols added sometimes. These include $\bar{x}$, $\hat{p}$, $s$, $r$, and $b$.

(e) The largest group in which it makes sense to make inferences about from the sample collected refers to 1. generalizability. This is directly related to how the sample was selected.

(f) A 5. sample is a smaller collection of observations (instances of observational units) that is selected from the larger pool.

5. We are interested in understanding the percentage of all American adults in favor of decriminalizing marijuana nationally. We aren't able to ask all American adults about their preferences but we are able to survey a random sample of 1000 Oregon adult residents. We find that this sample results in a percentage of 87% in favor of decriminalizing marijuana nationally.

$\boxed{4}$ (a) Give what three rows of a tidy data set might look like for this sample.

**Solution:**

| id | in_favor |
|---|---|
| 0001 | TRUE |
| 0002 | FALSE |
| 0003 | TRUE |

Identify the following in this context and give the value if known from the problem statement:

$\boxed{5}$ (b) i. population

**Solution:** ALL adult Americans

ii. sample

**Solution:** 1000 adult Oregonians

iii. parameter

**Solution:** Percentage of ALL Americans in favor of decriminalizing marijuana

iv. statistic

**Solution:** Percentage of 1000 randomly sampled Oregon adults in favor of decriminalizing marijuana $= 0.87$

v. generalizability

**Solution:** Can only generalize to Oregon adults since only a random sample of Oregon adults was selected and Oregon adults may not be representative of all US adults