Math 338 Exam 2 Lab Solutions

Question 1

Let's get the data into RStudio first.

```
## Read the data set into RStudio
library(readr)
florida <- read_csv("~/Math 338 Summer 2017/florida_mcd_bk.csv")
head(florida)  # view the first few entries in the data set
```

```
## # A tibble: 6 × 3
##   restaurant            city violations
##        <chr>           <chr>      <dbl>
## 1 McDonald's     TALLAHASSEE          5
## 2 McDonald's     TARPON SPGS          8
## 3 McDonald's        SARASOTA          1
## 4 McDonald's            LUTZ          8
## 5 McDonald's     HAINES CITY          3
## 6 McDonald's DEERFIELD BEACH          8
```

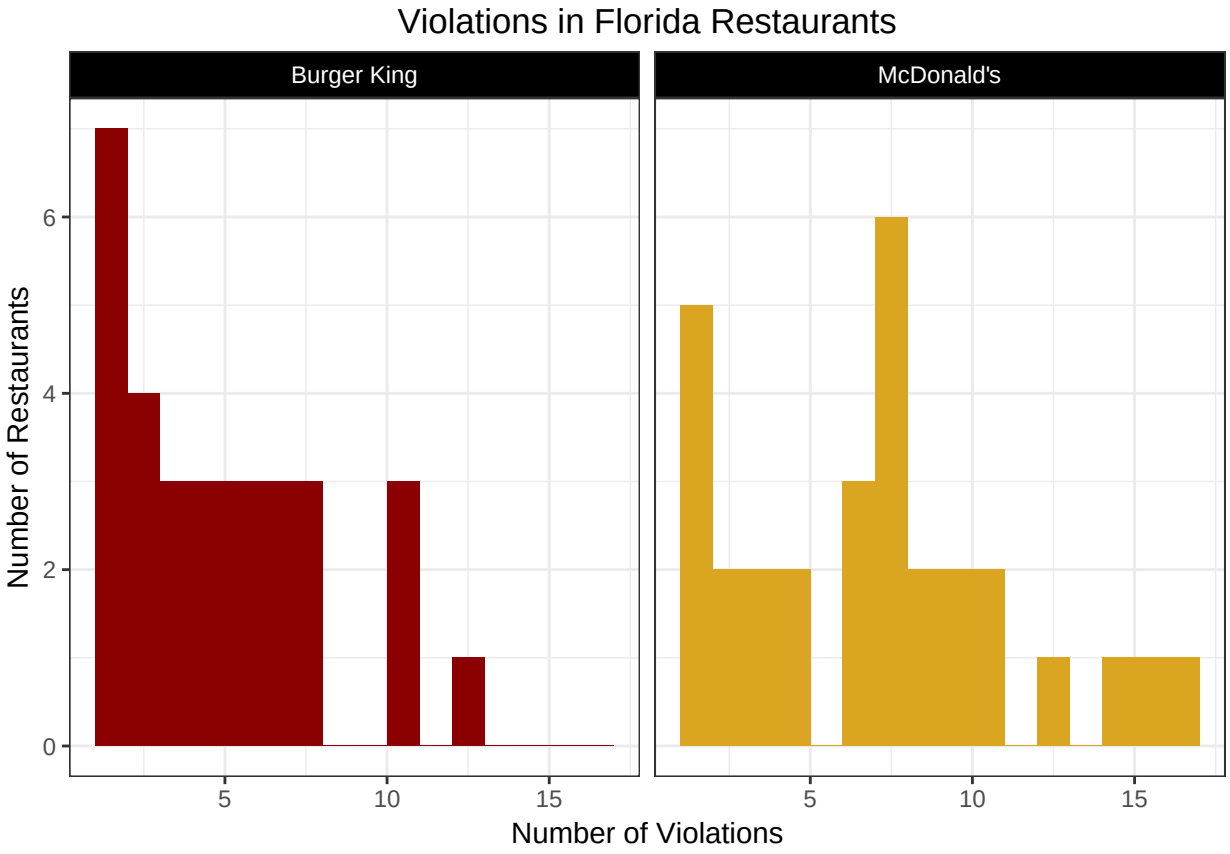
Part A

Since we have two independent samples, we should use a **two independent samples t-test**.

Part B

```
library(ggplot2)
restaurant_plot <- ggplot(data = florida, mapping = aes(x = violations,
    fill = restaurant))
restaurant_histogram <- restaurant_plot + geom_histogram(binwidth = 1,
    center = 0.5)
restaurant_histogram_labeled <- restaurant_histogram +
    facet_grid(. ~ restaurant) + labs(x = "Number of Violations",
    y = "Number of Restaurants", title = "Violations in Florida Restaurants")

# I wrote some extra code to make the histograms
# look nicer. I don't expect you to actually have
# done any of this, but I wanted to show off some
# more ggplot2 capabilities.
restaurant_histogram_final <- restaurant_histogram_labeled +
    theme_bw() + scale_fill_manual(values = c("darkred",
    "goldenrod")) + theme(legend.position = "none",
    plot.title = element_text(hjust = 0.5), strip.text = element_text(color = "white"),
    strip.background = element_rect(fill = "black"))
print(restaurant_histogram_final)
```



Part C

We are told that we have a simple random sample and that the cases are all independent. Therefore, we need only check whether we have a sufficiently large sample. The histogram for Burger King is quite clearly skewed to the right, and the histogram for McDonald's is also probably skewed to the right. However, our rule of thumb states that if the combined sample size is at least 40, we can use two-sample t procedures even if the histograms show quite a bit of skew. Therefore, **we are okay to use two-sample t procedures here**.

Part D

Before running our hypothesis test, we should define H_0, H_a, and α:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

where μ_1 is the population mean number of violations at Burger King and μ_2 is the population mean number of violations at McDonald's. By default, we will use a significance level of $\alpha = 0.05$.

```
mcd_bk_test <- t.test(violations ~ restaurant, data = florida)
print(mcd_bk_test)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  violations by restaurant
## t = -2.1853, df = 54.657, p-value = 0.03317
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.1539112 -0.1794221
## sample estimates:
## mean in group Burger King  mean in group McDonald's
##                  5.266667                  7.433333
```

We can find specific parts of the output:

```
mcd_bk_test$statistic  # gives us the value for the test statistic
```

```
##         t
## -2.185292
```

```
mcd_bk_test$p.value  # gives us the p-value
```

```
## [1] 0.03317182
```

Since our observed p-value of 0.0332 is less than our significance level of 0.05, we reject the null hypothesis. We conclude that **Burger King and McDonald's restaurants in Florida do have a different population mean number of violations**.

Question 2

Part A

First we have to filter out only the McDonald's restaurants:

```
library(dplyr)
mcd_only <- florida %>% filter(restaurant == "McDonald's")
```

Now we can use the t.test function to produce our confidence interval:

```
mcd_ci <- t.test(mcd_only$violations, conf.level = 0.99)
print(mcd_ci)
```

```
##
##  One Sample t-test
##
## data:  mcd_only$violations
## t = 9.4936, df = 29, p-value = 2.117e-10
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  5.275125 9.591541
## sample estimates:
## mean of x
##  7.433333
```

The 99% confidence interval output by RStudio is (5.28, 9.59). **We are 99% confident that the population mean number of violations at McDonald's in Florida is between 5.28 and 9.59**.

Part B

If we ran created a 99% confidence interval 1000 times, we would expect **99%** of the resulting confidence intervals to contain the true population mean number of violations.