Stat 230 Individual HW 1

Name: Victor Huang

Worked with: Myself

Directions

Homework 1 is due **noon, Tuesday 9/21/21**. Please complete the assignment in this Markdown document, filling in your answers and R code below. Submit a hard copy of the **compiled pdf** either

- in class on Monday 9/20
- in drop-in office hours (Tuesday 9/21)
- in the paper holder outside my CMC 222 office door (hopefully it will be installed by then!)

Make sure to follow the homework guidelines when writing up this assignment.

Tips for using Markdown with homework sets:

- Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
  - Make sure you load your data in the .Rmd and include any needed `library` commands.
- Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
- For your final document, you can change the output type from `html_document` to `word_document` or `pdf_document`. These two to output types are better formatted for printing.
  - on maize: you may need to allow for pop-ups from this site
- If you want to knit to pdf while running Rstudio from your computer (*not* from maize), you will need a LaTeX compiler installed on your computer. This could be MiKTeX, MacTeX (mac), or TinyTex. The latter is installed in R: first install the R package `tinytex`, then run the command `tinytex::install_tinytex()` to install this software.
  - If you are using maize, you don't need to install anything to knit to pdf!

Problem 1

The day 2 handout example compares the average number of farms per county in 1992 in the western and north central regions of the U.S. Your homework problem is to repeat this analysis but compare the southern and north central regions of the U.S.

```
> # data for HW 1
> agstrat <- read.csv("http://people.carleton.edu/~kstclair/data/agstrat.csv")
```

(a) Enter the data in R then use the `filter` command to create a data frame that contains only the southern and north central regions. How many rows are in this new data set? How many counties each in the southern and north central regions?

Answer: As one can see, there are two rows in this new data set with 103 north central counties and 135 southern counties.

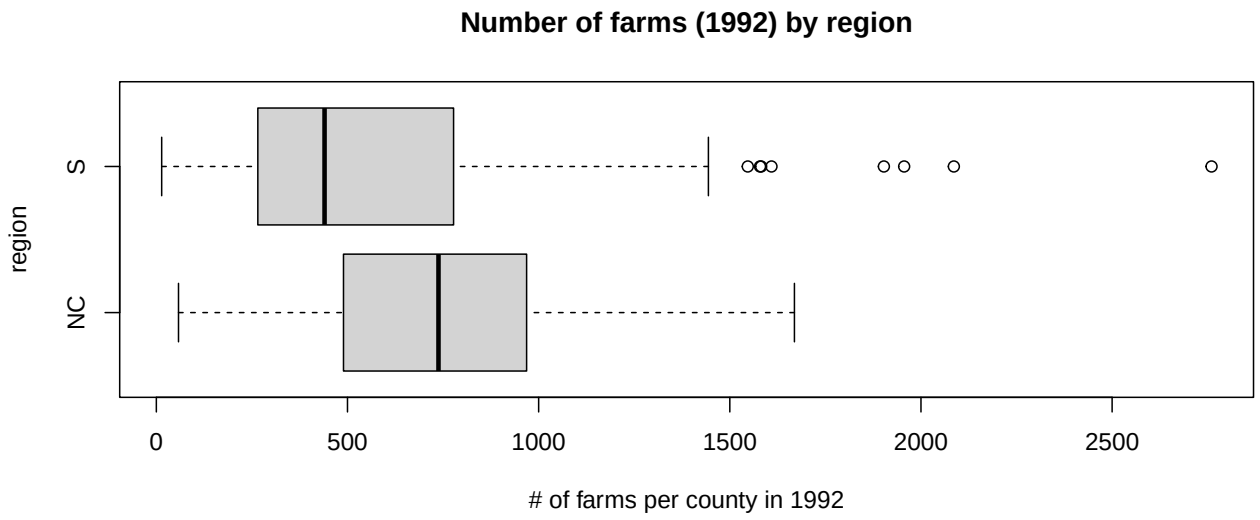
```
> agstrat2 <- filter(agstrat, region %in% c("S", "NC"))
> table(agstrat2$region)

 NC   S
103 135
```

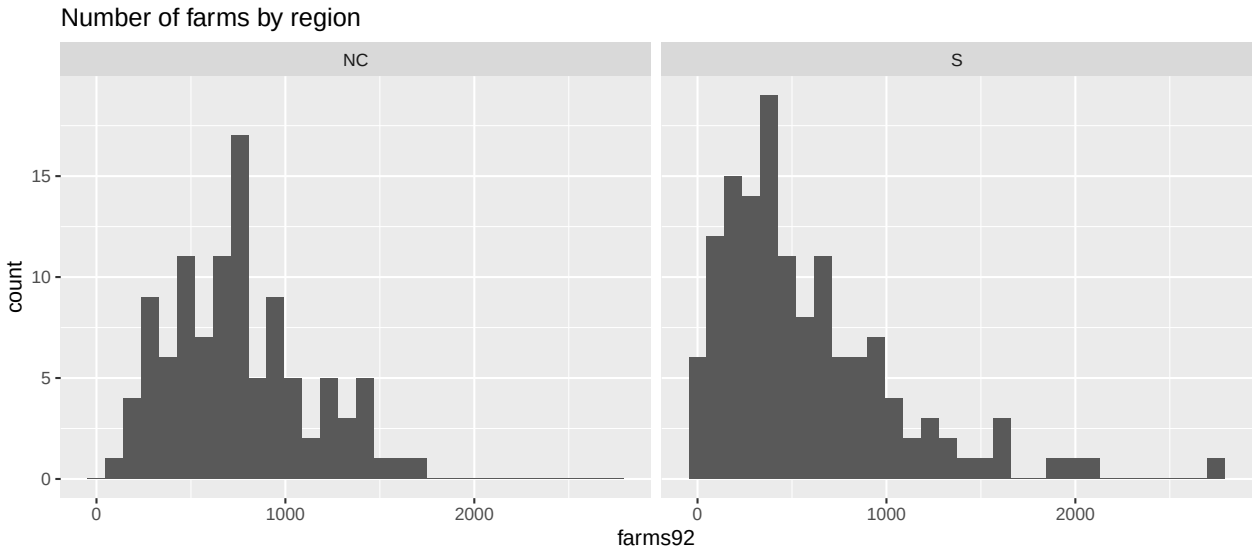
(b) Use basic EDA tools (summary stats and graph(s)) to compare `farms92` in the southern and north central regions. Make three comparative statements using these stats and graph(s). Please refer to the Notes Graph Formatting section to resize your graphs from the large default size.

Answer: Both county's distributions of farms are slightly skewed to the right. Counties in the NC tend to have more farms with a mean of approximately 750 and a median around 738 whereas counties in the S have a mean of 579 and median of 440. Counties in NC do not seem to have many significant outliers while there exists many outliers in counties in S, this implies that when included, counties in the NC have less variance then counties in the S.

```
> boxplot(farms92 ~ region, data = agstrat2, horizontal=TRUE, main="Number of farms (1992) by region",
```




```
> library(ggplot2)
> ggplot(agstrat2, aes(x=farms92)) + geom_histogram() + facet_wrap(~region) + labs(title = "Number of fa
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
> tapply(agstrat2$farms92, agstrat2$region, summary)
$NC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   58.0   489.5   738.0   750.7   968.5  1669.0

$S
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   14.0   265.5   440.0   578.6   777.5  2760.0
```

(c) There is one extreme outlier in the southern region. Identify this county by row number and by county name.

Answer: We see that the maximum value for county's in the S is 2760, looking through the farm92 dataset, we find that value corresponds with HILLSBOROUGH COUNTY on row 129.


```
> tapply(agstrat2$farms92, agstrat2$region, summary)
$NC
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   58.0   489.5   738.0   750.7   968.5  1669.0

$S
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   14.0   265.5   440.0   578.6   777.5  2760.0
```


(d) Check the inference assumptions. Are the distributions within each region symmetric or skewed? Can you still conduct a t test/CI for this data even if one or both of the distributions are skewed?

Answer: The distribution for both regions or skewed to the right as one can tell from the graph. Since both groups have a large enough sample size with NC at 103 and S at 135, we can still continue to conduct a t-test for the data even with the skewd nature of the data.

(e) Run the t test needed to test:

$$H_0 : \mu_S = \mu_{NC} \quad H_A : \mu_S \neq \mu_{NC}$$

where μ_r is the population mean number of farms per county for all counties within region "r". Interpret the t-statistic for this test, then use the p-value to make a conclusion.

Answer: We get that t=3.1975 and that p=0.001575. The test stat tells us that there exists a SE of 3.1975 from the observed average number of farms above the hypothesized difference of 0. Our p-value states that there is a 0.15755% chance of observing a sample difference in means that is at least 3.1975 SEs away from 0. Since we have such a small p-value we can eject the null hypothesis and accept the alternative hypothesis as true and that there is a difference in mean number of farms per county in the S and NC populations.

```
> t.test(farms92 ~ region, data = agstrat2)

    Welch Two Sample t-test

data:  farms92 by region
t = 3.1976, df = 235.99, p-value = 0.001575
alternative hypothesis: true difference in means between group NC and group S is not equal to 0
95 percent confidence interval:
  66.06586 278.12300
sample estimates:
mean in group NC  mean in group S
        750.6796         578.5852
```

(f) Repeat part (e) with the outlier identified in part (c) omitted. Does your conclusion change? Explain why or why not.

Answer: While we did get different t and p values, they aren't enough to suggest a change in conclusion. The p-value is 0.001575, still far too low to be able to accept the null hypothesis. As such, we still reject the null hypothesis and accept the alternative.

```
> t.test(farms92 ~ region, subset = -129, data = agstrat2)

    Welch Two Sample t-test

data:  farms92 by region
t = 3.6576, df = 233.81, p-value = 0.0003145
alternative hypothesis: true difference in means between group NC and group S is not equal to 0
95 percent confidence interval:
  86.9073 289.8400
sample estimates:
mean in group NC  mean in group S
        750.6796         562.3060
```

(g) Use the confidence interval computed in part (f) to compare the mean number of farms per county within each region. Do not use the work "difference" in your interpretation, instead use directional words like "more" or "less".

Answer: We are 95% confidence that the average number of farms per county for all counties in the north central region is between 86.9073 to 289.8400 farms higher than the average for all counties in the southern region.