Geography 411 – Spring 2026
Homework #2 - ANOVA
Due February 20th

Use the "precipSample.csv" dataset to test the hypothesis that average annual precipitation in Buffalo is equal for the three time periods 1940-1967, 1968-1996, and 1997-2025. This dataset consists of a random sample of 30 years from the period 1940-2025. You will use R to complete ANOVA, Kruskal-Wallis, and the median tests. Additionally, you will perform the Levene and Kolmogorov-Smirnov diagnostic tests but do not have to interpret them in the homework. Then you will repeat the analyses for San Diego. You should hand in a document that includes all of the plots and statistical output from the analyses. The six-step hypothesis testing procedure should be included for ANOVA, Kruskal-Wallis, and the median tests. There are some questions regarding the results to be answered after all the hypothesis tests are completed (on page 7).

Please work together in the lab to complete the computer portion of the assignment, but interpret results and turn in assignments individually. The R commands necessary to carry out the operations are provided in `courier` font. The commands are also provided in an R script file along with some comments on their use. Use the `help()` command at any time to find out more about a command and the possible options.

1. Download the precipitation sample data file, "precipSample.csv," and the R script file from the class UBLearns website to your working folder. This could be a local drive on your home computer, a USB drive, or a folder on UB Box. It is probably a good idea to make a separate directory for this exercise.

2. Set your working directory in RStudio.

3. Open the script file with the *Open file…* option from the *File* menu. This file contains all of the commands that you need to complete the assignment.

4. Load the precipitation sample data file into R. The `read.csv()` command is used to open comma delimited data files. The data set is read and saved as a data frame called precip. A data frame is the R term for a data table. Typing the data frame name will print the contents of the data frame in the R console.

```
precipSample <- read.csv("precipSample.csv")
precipSample
```

5. Before performing our difference of means tests, we need to create the grouping or factor variable for our three time periods. The next section of code creates a new variable called "Period" and assigns the proper label (Period1, Period2, Period3) to each sample depending on the year.

```
precipSample$Period[(precipSample$Year <= 1967)] <-
  "Period1"
precipSample$Period[(precipSample$Year > 1967 &
  precipSample$Year < 1997)] <- "Period2"
precipSample$Period[(precipSample$Year >= 1997)] <-
  "Period3"
precipSample
```

6. The table command will tabulate a frequency table for a variable. The next set of commands save some quantities that we can use later for calculating the ANOVA statistic. The quantities n1, n2, and n3 are the sizes of each of the samples. These three should add up to the overall sample size n.

```
table(precipSample$Period)
n1 <- sum(precipSample$Period == "Period1")
n2 <- sum(precipSample$Period == "Period2")
n3 <- sum(precipSample$Period == "Period3")
n <- length(precipSample$Buffalo)
```

We should also set a variable k equal to the number of categories.

```
k <- 3
```

7. The sample means can be calculated in a similar manner for each of the three samples.

```
mean1 <- mean(precipSample$Buffalo[precipSample$Period ==
  "Period1"])
mean2 <- mean(precipSample$Buffalo[precipSample$Period ==
  "Period2"])
mean3 <- mean(precipSample$Buffalo[precipSample$Period ==
  "Period3"])
```

8. You can simply type an R object name to list is in the console. Type mean1 in the console to see the sample mean for the first time period. If you want to add some text to label quantities, you can use the print and paste commands.

```
print("sample means")
print(paste("Overall",mean(precipSample$Buffalo)))
print(paste("Period1",mean1))
print(paste("Period2",mean2))
print(paste("Period3",mean3))
```

9. A good early step in any analysis is to plot some useful graphs. The box plot is a good tool for examining the similarities and differences between distributions.

```
boxplot(Buffalo ~ Period, data = precipSample, ylab =
   "annual precipitation (in.)")
```

10. Start recording the graphic window history so that you can compare plots. You might also want to save the plots or copy and paste them to a document as you go.

11. A stacked histogram plot can be useful as well. This takes quite a bit of code to complete a histogram and include a vertical line at the mean for each of the three time periods. The layout command partitions the plot area. The default settings of the histogram function must be changed to produce aligned histograms. The abline command is used to draw a vertical line on each histogram at the mean precipitation value for each time period.

```
layout(matrix(c(1,2,3)))
minMax                                                <-
   c(min(precipSample$Buffalo),max(precipSample$Buffalo)+
   1)
BuffaloPrecipBreaks <- seq(22,54,2)
hist(precipSample$Buffalo[precipSample$Period          ==
   "Period1"], breaks = BuffaloPrecipBreaks, xlim =
   minMax, ylim = c(0,5), main = "Period 1", xlab =
   "Annual Precipitation (in.)")
abline(v = mean1, lwd = 2)
hist(precipSample$Buffalo[precipSample$Period          ==
   "Period2"], breaks = BuffaloPrecipBreaks, xlim =
   minMax, ylim = c(0,5), main = "Period 2", xlab =
   "Annual Precipitation (in.)")
abline(v = mean2, lwd = 2)
hist(precipSample$Buffalo[precipSample$Period          ==
   "Period3"], breaks = BuffaloPrecipBreaks, xlim =
   minMax, ylim = c(0,5), main = "Period 3", xlab =
   "Annual Precipitation (in.)")
abline(v = mean3, lwd = 2)
layout(matrix(c(1)))
```

12. The next section of code in the script file is included for those interested in seeing how to calculate the intermediate values needed for ANOVA analysis. I will not provide a detailed description of this section. The loops in this section correspond to the summation symbols in the formulas we use to calculate the sum of squares.

13. R provides quantile functions for distributions that can be used instead of the tables at the back of the book. Note that the default for these functions is to give values corresponding to the lower part or left tail of the function. The critical value for our ANOVA test is calculated with the command below. I have changed the *lower.tail* parameter to *FALSE* so that the function returns the critical *F* value for $\alpha=0.05$ and degrees of freedom are *k-1* and *n-k*. You might check to be sure that this value matches Table A.5 in the textbook.

```
Fcrit <- qf(0.05, (k-1), (n-k), lower.tail = FALSE)
Fcrit
```

14. The ANOVA statistic is calculated in two steps. The first command computes a linear model of precipitation as a function of time period. I will explain this in depth during the second half of the course. The second step calculates the ANOVA table for the model.

```
modelBuffalo <- lm(Buffalo ~ as.factor(Period), data =
   precipSample)
anova(modelBuffalo)
```

15. The Levene test is used to evaluate the assumption of homoskedasticity. We start by calculating the absolute deviation of each observation from its time period mean.

```
precipSample$BuffaloAbsoluteDeviations <-
   abs(resid(modelBuffalo))
precipSample
```

You can create a boxplot of these deviations as a visual check of the variation in each time period.

```
boxplot(precipSample$BuffaloAbsoluteDeviations ~ Period,
   data = precipSample)
```

I have combined the two commands in step 14, to compute the Levene test statistic.

```
anova(lm(precipSample$BuffaloAbsoluteDeviations ~
   as.factor(precipSample$Period)))
```

16. The exercise script contains code to repeat the stacked histogram from step 11 with appropriate normal curves drawn over relative histograms. This would be a simple plot to assess the normality assumption. The cumulative histogram approach and the corresponding Kolmogorov-Smirnov test is a more robust approach.

17. The Kolmogorov-Smirnov (K-S) test is used to evaluate the assumption of normality. The test statistic is the maximum difference between the empirical CDF and the normal CDF. The first three lines of code create the empirical cdf and plot it with the normal cdf. The *ks.test* command calculates the test statistic. The code for the other time periods is provided in the exercise script.

```
#Period 1
period1ECDF                                                <-
   ecdf(precipSample$Buffalo[precipSample$Period           ==
   "Period1"])
plot(period1ECDF)
curve(pnorm(x,
   mean=mean(precipSample$Buffalo[precipSample$Period   ==
   "Period1"]),
   sd=sd(precipSample$Buffalo[precipSample$Period          ==
   "Period1"])), add=TRUE)
ks.test(precipSample$Buffalo[precipSample$Period           ==
   "Period1"],                                        "pnorm",
   mean(precipSample$Buffalo[precipSample$Period           ==
   "Period1"]),
   sd(precipSample$Buffalo[precipSample$Period             ==
   "Period1"]))
```

18. The non-parametric Kruskal-Wallis test is executed with the kruskal.test command.

```
kruskalWallis <- kruskal.test(Buffalo ~ as.factor(Period),
   data = precipSample)
kruskalWallis
```

The critical value can be calculated as in Step 13, but now we use the chi-square distribution (Table A.6).

```
qchisq(0.05, k-1, lower.tail = FALSE)
```

19. Calculation of the non-parametric median test requires several steps. The first step is to calculate the median.

```
medianPrecipBuffalo <- median(precipSample$Buffalo)
medianPrecipBuffalo
```

A box plot with the median superimposed can be used to visualize the test. Under the null hypothesis, the three distributions are evenly distributed above and below the median.

```
boxplot(Buffalo ~ Period, data = precipSample, ylab =
   "annual precipitation (in.)")
abline(h = medianPrecipBuffalo, lwd = 2)
```

A new variable is created that indicates whether each observation is above or below the median.

```
precipSample$BuffaloMedian <- ifelse(precipSample$Buffalo
   > medianPrecipBuffalo, "Greater than median", "Less
   than or equal")
precipSample
```

Now we can create our table that we use for the median test.

```
medianTable      <-      table(precipSample$BuffaloMedian,
   precipSample$Period)
medianTable
```

Finally we calculated our chi-square test statistic. Remember that this is a special case of the chi-square goodness of fit test.

```
chisq <- chisq.test(medianTable)
chisq$expected
chisq
```

You may receive the following warning, but please ignore it for now

```
.
Warning message:
In  chisq.test(medianTable)  :  Chi-squared  approximation
   may be incorrect
```

20. Remember to repeat these analyses for the San Diego precipitation data. The full code is provided in the script file.

Discussion Questions (Word document)

a. Discuss some of the reasons for differences and the ordering of the significance values (p-values) for the ANOVA, Kruskal-Wallis, and median tests for each city.
b. If you were going to report one of these tests in a publication, which should you choose for each dataset? Why?
c. Are these data evidence of a change in climate?

Part II

Portfolio Development - GitHub

At the top of your R file, add:

 # Geography 411 – Homework 2
 # Name: Your Name
 # Description: Analysis of variance

i. Create a new repository in GitHub

Fill in:

- Repository name:
  Analysis of variance
- Description:
  Write out your description
- Set to Public
- Add a README file

ii. Upload files to GitHub
iii. Commit changes
iv. In your ReadMe file, write out your summary using this structure (be concise):
   a. Title
   b. Overview
   c. Methods
   d. Key findings
   e. Tools
   f. Reflection

Click **Commit changes** when finished.