Assignment 5: Comparing Distributions

Asa Hayes

24 March, 2021

Instructions

Download the assignment05.Rmd file and open it in RStudio. Complete this assignment by filling in the answers below in the R Markdown Notebook document.

Data

In this assignment, you will apply inferential statistics to the PUF Lands drilling pad measurements that you collected for Assignment #4. The data is contained within the "Pads.csv" file which is posted on Canvas under Assignment #5.

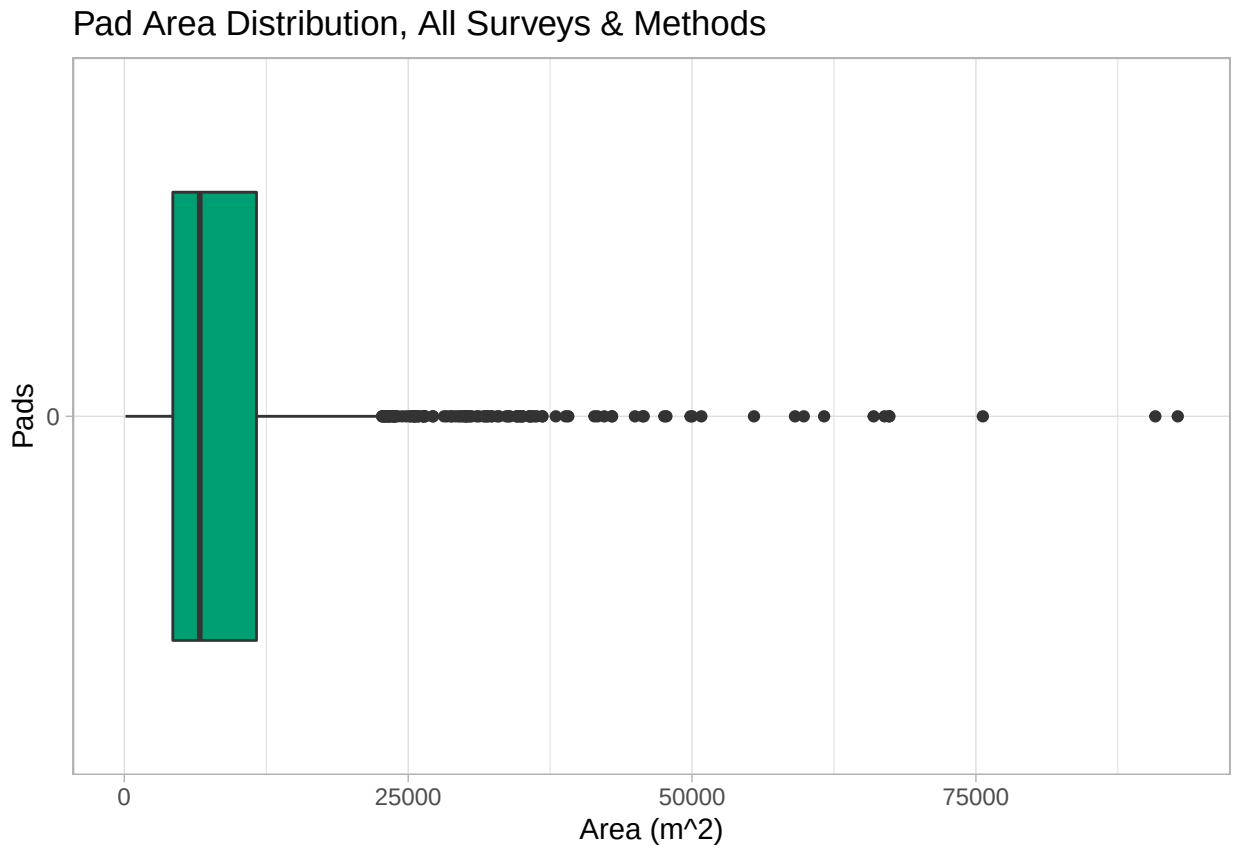
Deliverables

Please submit to Canvas the following items:
1. An HTML (or Word or PDF) file knitted from the .Rmd file
2. A completed .Rmd file

Questions

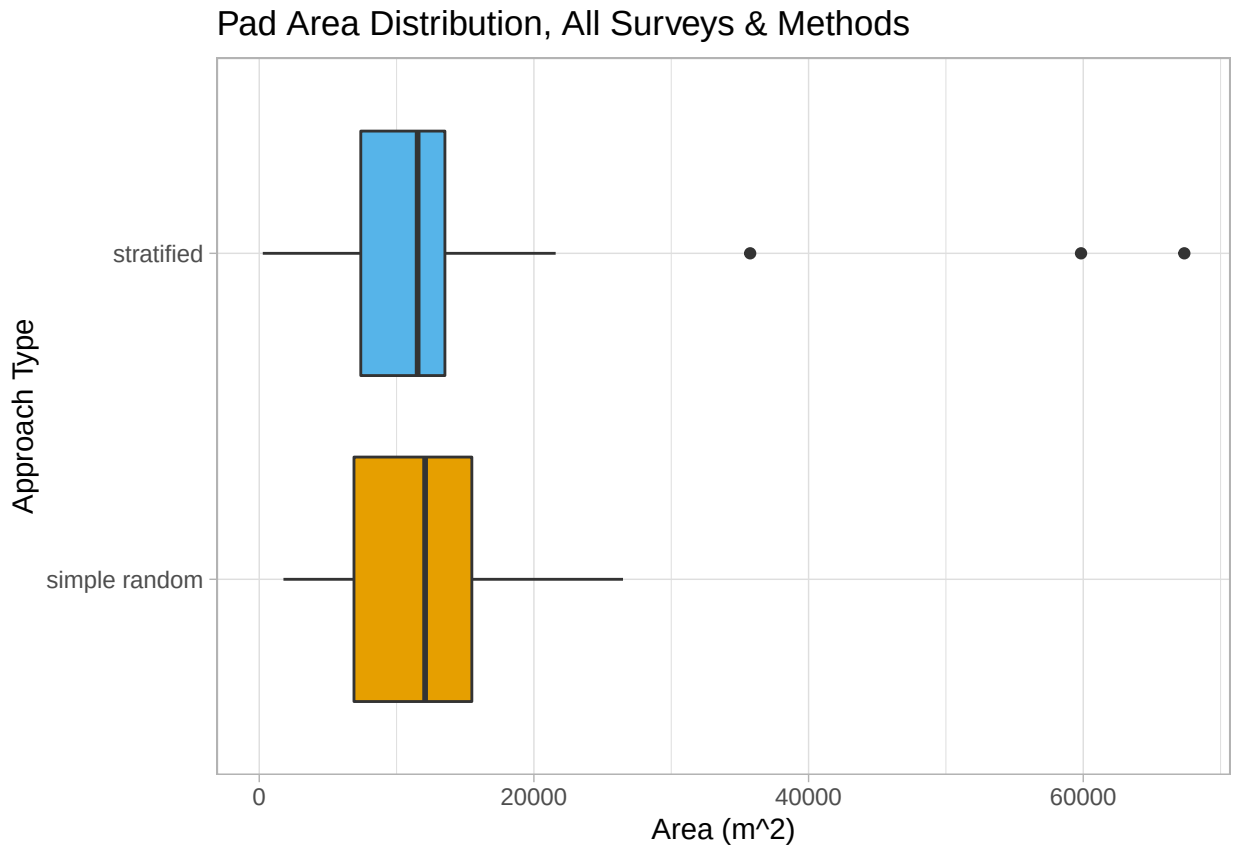
1. Download the Pads.csv file from Canvas, read it into RStudio using the read.csv() function and assign it to the variable "Pads". Harnessing the immense power of R, I compiled and cleaned the class's KML files and created a CSV file containing the following variables for each polygon observation: (1) student; (2) pad_id; (3) survey; (4) approach; (5) lat; (6) lon; (7) area_m2 (the area of the pad). If your measurements are not included in the spreadsheet, or there is a problem with your data, just select another student's observations and use theirs. In a single figure, compare the distribution of all pad areas in the CRIS and Andrews surveys. As usual, all graphics created in this assignment should be done using the ggplot2 package and will be graded based on the highest standards of data visualization

```
path = "C:\\Users\\A\\Desktop\\GEOG-312\\Labs\\Lab5\\Pads.csv"
pads = read.csv(path, header=T)
library(ggplot2)
colorGood <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7", "#999999")

ggplot(data = pads, aes(area_m2, y=factor(0))) +
  geom_boxplot(fill = "#009E73") +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Area (m^2)",
       y = "Pads",
       title = "Pad Area Distribution, All Surveys & Methods")
```



2. In a single figure, compare the distribution of pad areas between the two sampling approaches that you used when you did your spatial sample in the previous assignment. If your name is not in the Pads data frame, just use another person's samples. Hint: create a new data frame from Pads, containing just your sample observations.

```
myPads <- subset(pads, student == "hayesasarodman")

ggplot(data = myPads, aes(area_m2, y=approach, fill=approach)) +
  geom_boxplot(show.legend = FALSE) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Area (m^2)",
       y = "Approach Type",
       fill = "Type",
       title = "Pad Area Distribution, All Surveys & Methods")
```



3. Using the t-test, compare the pad areas between your two sampling approaches. Determine if your two samples are statistically different at the *99% confidence interval*. Print the result of the t-test and describe your results in terms of the null hypothesis. Note that if your samples are not included in the Pads data frame, just select a random person's observations to use.

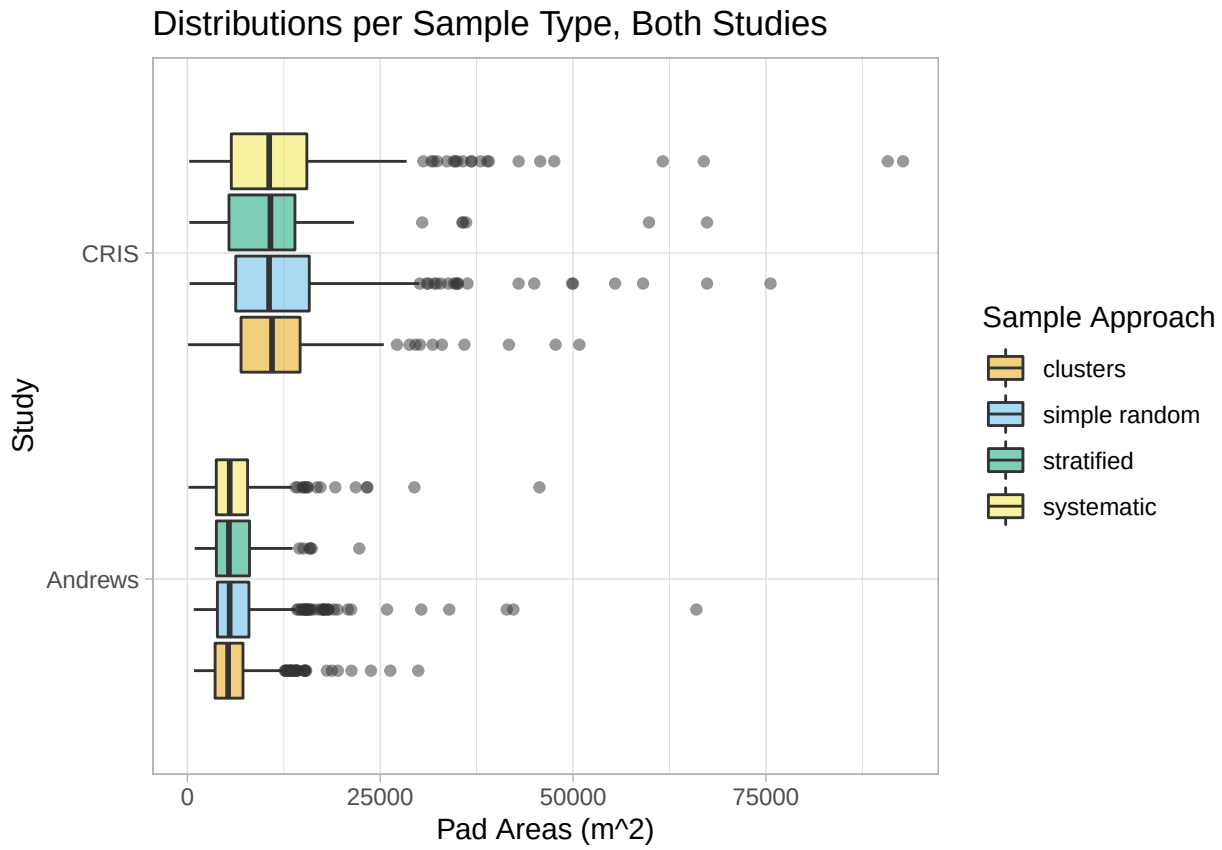
```
t.test(area_m2 ~ approach, data=myPads, conf.level=0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  area_m2 by approach
## t = -0.91717, df = 43.491, p-value = 0.3641
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -9258.244  4555.012
## sample estimates:
## mean in group simple random    mean in group stratified
##                    11282.18                    13633.80
```

The null hypothesis in this case would be that the difference in means between the two samples is 0 (i.e. that the samples are statistically the same), and to prove that we need to obtain a p-value less than the confidence interval (<0.01 for a 99% conf level.) As the p-value is substantially higher than 0.01, we cannot accept the alternate hypothesis and thus fail to reject the null hypothesis. As such, we can conclude that the two samples are statistically similar, even despite the outliers visible on the chart for Q2.

4. Using all the observations in the Pads data frame, plot the distributions of pad areas in Andrews & CRIS, by each of the four sampling approaches. Hint: use facet_wrap() function in ggplot (see R code from ggplot lecture, you should be showing 8 distributions total).

```
ggplot(data = pads, aes(area_m2, y=survey, fill=approach)) +
  geom_boxplot(alpha = 0.5) +
  theme_light() +
  scale_fill_manual(values = colorGood) +
  labs(x = "Pad Areas (m^2)",
       y = "Study",
       fill = "Sample Approach",
       title = "Distributions per Sample Type, Both Studies")
```



5. Subset the Pads data frame to only include observations from the Andrews survey. Set this new data frame to the variable "andrews". Using a parametric test, determine whether there is a statistically significant difference between sample means in all four approaches at the *90% confidence interval* (use all observations in the andrews data frame). Print out the p-value and describe your results in terms of the null hypothesis.

```
andrews <- subset(pads, survey == "Andrews")

andResult <- aov(area_m2 ~ approach, data = andrews)
summary(andResult)
```

```
##               Df    Sum Sq  Mean Sq F value Pr(>F)
## approach       3 1.369e+08 45639166    2.11 0.0971 .
## Residuals   1294 2.799e+10 21627285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(andResult, level=0.90)
```

```
##                               5 %      95 %
## (Intercept)           5792.3385 6628.8018
## approachsimple random  125.4584 1244.5995
## approachstratified    -721.6184  830.9745
## approachsystematic    -597.0003  535.5764
```

6. In a sentence, is it statistically valid to apply a parametric test on the Andrews pad area observations, like we did in the previous question? Why or why not?

7. Determine whether all the pad areas in the Andrews survey are normally distributed at a confidence interval of 95%. Print your results and interpret your them in terms of the null hypothesis.

```
t.test(andrews$area_m2, data=andrews, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  andrews$area_m2
## t = 49.754, df = 1297, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6177.042 6684.155
## sample estimates:
## mean of x
##  6430.599
```

As the p_value is much smaller than 0.05, we can reject the null hypothesis that the areas are normally distributed, and accept the alternate hypothesis that they are not normally distributed.

8. Using a statistically appropriate test, determine whether there is a statistically significant difference between samples in all four approaches in the Andrews survey at the *90% confidence interval* (use all observations in the andrews data frame). Print the test results and describe the results in terms of the null hypothesis. Discuss the reason(s) why the results from Question 4 and this question are the same (or different).

```
#Note: was unable to determine other test besides ANOVA within time limit
```

9. Use the Andrews data and the Kolmogorov-Smirnov test to test whether the pad areas surveyed using the simple random approach are statistically different at the 95% confidence interval than the pad areas surveyed using the clusters approach. Print the D statistic and p-value and describe what each of them is telling you. Then, interpret the results in terms of the null hypothesis.

```
andrews_sim <- subset(andrews, approach == "simple random")
andrews_clu <- subset(andrews, approach == "clusters")
ks.test(x=andrews_sim$area_m2, y=andrews_clu$area_m2, conf.level=0.95)
```

```
## Warning in ks.test(x = andrews_sim$area_m2, y = andrews_clu$area_m2, conf.level
## = 0.95): p-value will be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  andrews_sim$area_m2 and andrews_clu$area_m2
## D = 0.06948, p-value = 0.3269
## alternative hypothesis: two-sided
```

The

10. Print approximately the number of hours this assignment took you to complete.

```
print("2-3")
```

```
## [1] "2-3"
```