Simonson_HW5

Martin Simonson

February 22, 2019

1.) Training in coordinate geometry

28 9th year students are to receive coordinate geometry training in one of two ways: a *conventional* way and a *modified* way. After training, students were asked to solve a coordinate geometry problem. The **time to compute the problem** was recorded, but five students in the *conventional* group failed to complete the problem in time. Use an appropriate method to test if the modified training reduces the time it takes to solve the coordinate geometry problem. Data are provided in *case0402.csv*. [Note: the normality assumption is automatically violated when some observations are censored, and the sample means are not meaningful estimates of the population mean]

- **Answer:** Applying the randomization test is appropriate since the two groups are independent but the normality assumption is violated. Let μ_1 represent the mean time in seconds to solve a geometry problem for students in the conventional group, and let μ_2 represent the mean time in seconds to solve a geometery problem for students in the modified group.
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
df<-read.csv("Data/case0402.csv", header = T)
#str(df)
df<-subset(df, Censored == 0) #remove 5 censored obs.
aggregate(Time~Treatment, data = df, FUN = mean)
```

```
##      Treatment     Time
## 1 Conventional 182.0000
## 2     Modified 125.2857
```

```
# conventional - traditional
obs.diff<-182-125.2857 # difference of mean times
obs.diff
```

```
## [1] 56.7143
```

```
# permutation test
rand.test(Time ~ Treatment, df, plot = F)
```

```
##
##  Randomization test
##
## data:  Time by Treatment
## mean.diff = 56.714, MC sample size = 10000, p-value = 0.0243
## alternative hypothesis: two.sided
## sample estimates:
## mean in group Conventional     mean in group Modified
##                   182.0000                   125.2857
```

- With a p-value of 0.0261 we have some evidence to reject the null hypothesis that there is no true difference in mean time to complete a geometry problem bewtween traditional and modified training styles. The data indicates that the students who experienced traditional training took longer to complete the geometry problem compared to students who underwent modified training.

2.) Pollen Removal, Revisited

Data found in *bee.csv* Ignore the first column of the proportion of pollen removed, the second column is the duration of visit **(unknown units)** the third column indicates worker (W) or queen (Q) bees. Consider the data on duration of visit… find out whether workers spend more time in flowers than do queens. Use a randomization/Permutation test.

a) Use a permutation test on the raw data to test if mean durations are different between worker and queen queen bees.

- **Answer:** Let μ_1 represent the mean duration of visit for queen bees, and let μ_2 represent the mean duration of visit for worker bees.
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```r
df<-read.csv("bee.csv")
#str(df)
rand.test(Duration ~ Type, df, plot = F)
```

```
##
##  Randomization test
##
## data:  Duration by Type
## mean.diff = 19.662, MC sample size = 10000, p-value = 6e-04
## alternative hypothesis: two.sided
## sample estimates:
## mean in group Q mean in group W
##        35.83333        16.17143
```

- Observed mean difference is 19.6619 units of time that are probably seconds. The p-value after a randomization test is less than 0.001, therefore, we have very strong evidence to reject the null hypothesis that there is no difference in mean visitation times between queen bees and worker bees. We can further conclude that queen bees spend more time visiting flowers than worker bees.

b) Use a permutation test on the log-transformed data to determine if there is a difference in median duration of visit between worker bees and queen bees.

- **Answer:** Let μ_1 represent the median duration of visit for queen bees, and let μ_2 represent the median duration of visit for worker bees.
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```r
rand.test(log(Duration) ~ Type, df, plot = F)
```

```
##
##  Randomization test
##
## data:  log(Duration) by Type
## mean.diff = 0.65117, MC sample size = 10000, p-value = 0.0084
## alternative hypothesis: two.sided
## sample estimates:
## mean in group Q mean in group W
##        3.277111        2.625945
```

```r
# median difference in duration
exp(0.65117)
```

```
## [1] 1.917783
```

- The observed difference of median values between queen bees and worker bees is 1.917783 time units. With a −value of 0.008 we have strong evidence to reject the null hypothesis and conclude that the difference in median duration times between groups is not due to random chance alone.

3.) Apples

Considerable amounts of fungicides are used in commercial apple orchards. This prevents the development of some fungi that produce unsightly blotches on the skin but have no other effects on the apples. Suppose a researcher in the ISU Plant Pathology department studied whether these blotches could be removed by washing with a bleach solution after harvest. If successful, this would reduce the amount of fungicide needed. She randomly sampled 50 apples from an orchard that did not spray fungicides, so the apples had blotches on them. All apples were very similar in size. She randomly selected 25 apples and washed them individually. She then measured the area covered by fungal blotches on each apple. The average area of fungi for washed apples is 8.41 cm sq. ; the average area of fungi for control apples is 9.37 cm sq. The data are available in *apples.csv*.

a) Identify the treatments, randomization, and if causal conclusions can be justified.

- **Answer:** Treatmens are washed vs. control apples. There are 50 apples (exp. units) and treatments are randomly assigned with 25 apples in each group. Causal conclusions can be drawn because of random sampling and random assignment of treatment.

b) consider a randomization test of the null hypothesis of no difference in mean fungal area between washed and unwashed apples. What is the observed value of the difference between means?

```
df<-read.csv("Data/apples.csv")
rand.test(area ~ treatment, df, plot = F)
```

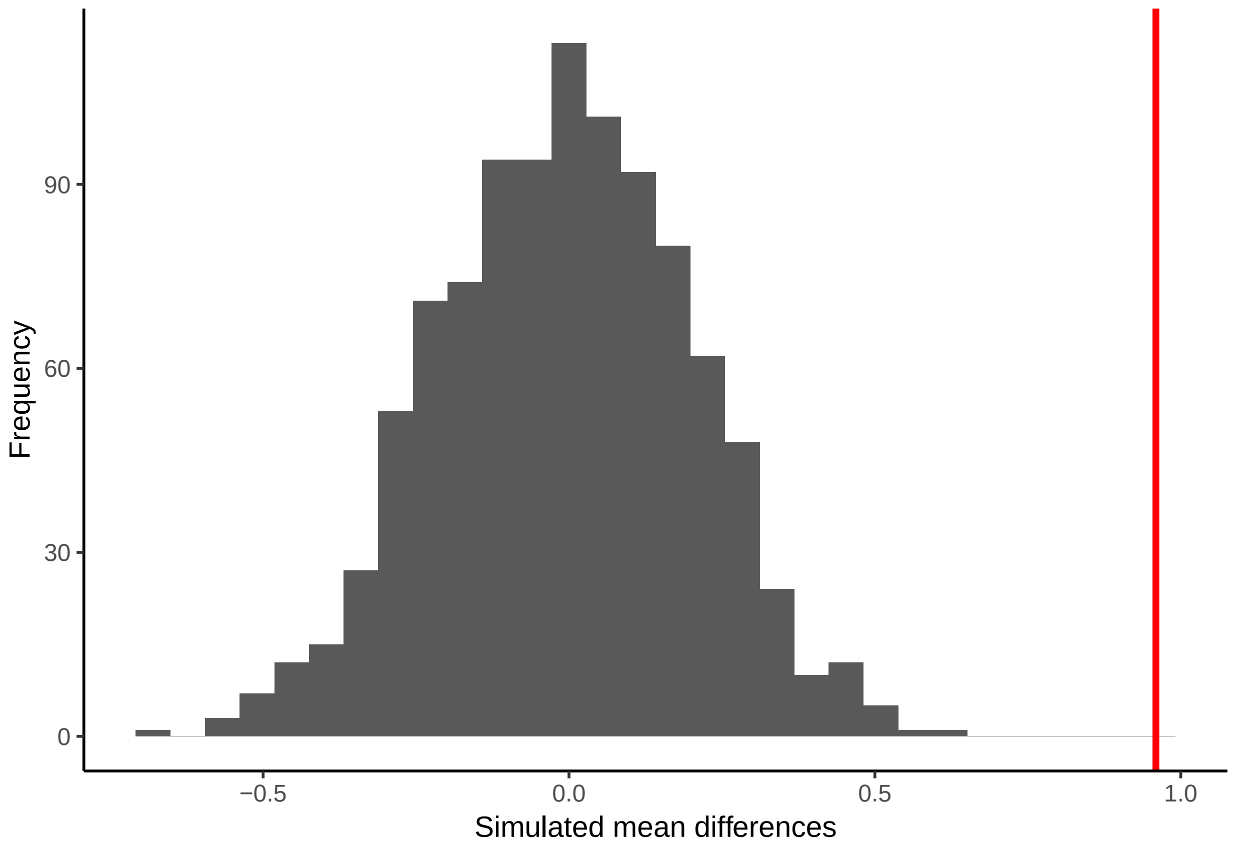
```
##
##  Randomization test
##
## data:  area by treatment
## mean.diff = 0.95932, MC sample size = 10000, p-value = 1e-04
## alternative hypothesis: two.sided
## sample estimates:
## mean in group control  mean in group washed
##              9.371956              8.412640
```

- **Answer:** The observed difference in means in 0.95932 cm^2.

c) Construct a 1,000 iteration randomization of the 50 apples into two groups of 25. Compute the randomization based two sided p-value for testing the null hypothesis of no effect of washign apples. Besides the p-value, show a histogram of the simulated values and mark the location of the observed mean difference.

```
rand.test(area ~ treatment, df, plot = T, nsim = 1000)
```

```
## Loading required package: ggplot2
```



```
##
##  Randomization test
##
## data:  area by treatment
## mean.diff = 0.95932, MC sample size = 1000, p-value = 0.001
## alternative hypothesis: two.sided
## sample estimates:
## mean in group control  mean in group washed
##              9.371956              8.412640
```

- **Answer:** We have a p-value of 0.001 and therefore have very strong evidence to reject the null hypothesis and conclude that the mean difference of fungi growth on apples is larger for unwashed (control) apples.

d) If you calculated a 95% CI based on the randomization test, would it include 0?

- **Answer:** No, the 95% confidence interval would not contain zero since the randomization test results indicate that there is a significant difference between mean fungi growth between groups.

e) Write a conclusion about the effect of washing on the area covered by fungal blotches.

- **Answer:** We have determined that the data suggests strong evidence to reject the null hypothesis that there is no difference in mean fungal areas on apples between washed and unwashed apples. Therefore, we evaluate the fungal blotches and determine that the observed difference in area covered (mean of unwashed minus mean of washed) was 0.95932 cm^2, indicating that the unwashed apples had a larger area of fungal growth compared to washed apples.

4) Marijuana

Chapter 4, problem 32 (Therapeutic Marijuana). Nausea and vomiting are frequent side effects of cancer chemotherapy, which can contribute to the decreased ability of patients to undergo long-term chemotherapy schedules. The effectiveness of THC (the active ingredient of marijuana) in preventing these side effects was compared with the standard drug Compazine. Of the 46 patients who tried both drugs (but were not told which was which), 21 expressed no preference, while 20 preferred THC and 5 preferred Compazine. Perform (by hand) a sign test, show all your work and clearly explain your conclusions.

Since we only care about non-zero differences, we can eliminate the 21 patients who expressed no preference. We are left with 20 observations of people who preferred THC and 5 that preferred Compazine.

- S = 20
- Z = (20-12.5) / sqrt(25/4) = 3

The p-value is between 0.002 and 0.01 so we have strong evidence against the null hypothesis. We have strong evidence to reject the null hypothesis that there is no difference between patients for preference of THC or Compazine, and conclude that preference for THC is greater than preference for Compazine.

5.) Darwin cross/self-fertilization

Chapter 4, problem 28 in both editions (Darwin cross/self fertilization data). Data are in *ex0428.csv*. Ignore the book's questions, and answer the following:

a) Draw a histogram or boxplot of the differenced data. Is there any indication from the pot that using paired t-tools may be inappropriate? Explain why or why not.

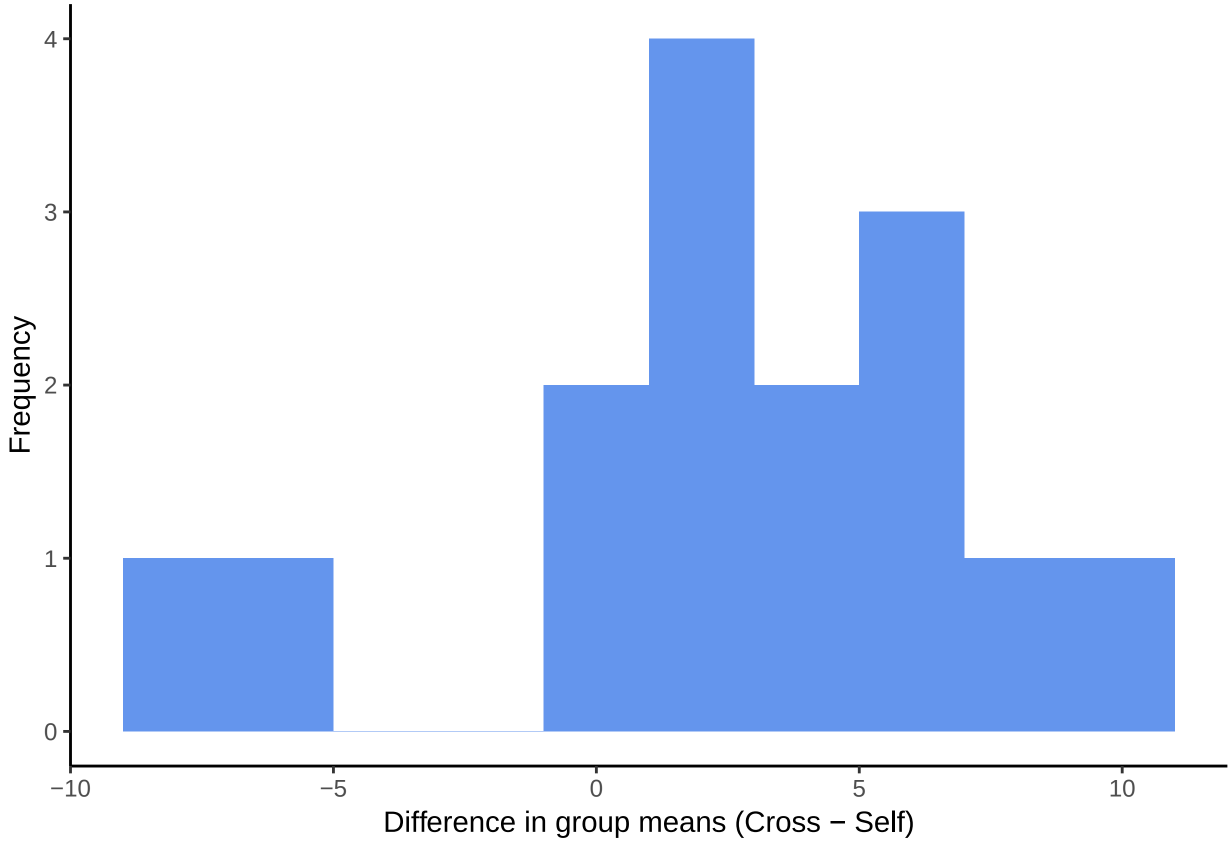
```
df<-read.csv("Data/ex0428.csv", header = T)
str(df)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ Cross: num  23.5 12 21 22 19.1 ...
##  $ Self : num  17.4 20.4 20 20 18.4 ...
```

```
df$diff<-df$Cross - df$Self
summary(df$diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.380   1.375   3.000   2.617   5.625   9.380
```

```
ggplot(data=df,aes(x=diff))+
       geom_histogram(fill="cornflowerblue", binwidth = 2)+
       ylab("Frequency")+
       xlab("Difference in group means (Cross - Self)")+
       theme_classic()
```



- **Answer:** The histogram indicates that there is a non-normal distribution of values after taking the difference of Cross-pollinated and Self-pollinated. The violation of that assumption indicates a paired t-test may be inappropriate.

b) Regardless of your answer in part a), conduct a Wilcoxon signed-rank test to test the hypothesis of no treatment effect. Write down the null and the alternative hypotheses. Report a p-value, and write a conclusion within the context of the data.

- **Answer:**
- H_0 : mean heights from cross fertilized plants are comparable to mean heights from self-fertilized plants.
- H_A : mean heights from cross fertilized plants are not comparable to mean heights from self-fertilized plants.

```
wilcox.test(df$Cross,df$Self,paired=T)
```

```
##
##  Wilcoxon signed rank test
##
## data:  df$Cross and df$Self
## V = 96, p-value = 0.04126
## alternative hypothesis: true location shift is not equal to 0
```

```
mean(df$diff)
```

```
## [1] 2.616667
```

The p-value is 0.04126 so we have some evidence against the null hypothsis, suggesting that the mean heights of cross-fertilized plants are greater than the mean heights of self-fertilized plants.