Hypothesis Testing

By Eric Lin

The Data Set

- The dataset we are working on is a collection of data from various music parties, and is named: party.

- Numerical Variables: Attendance, Rating, Ticket

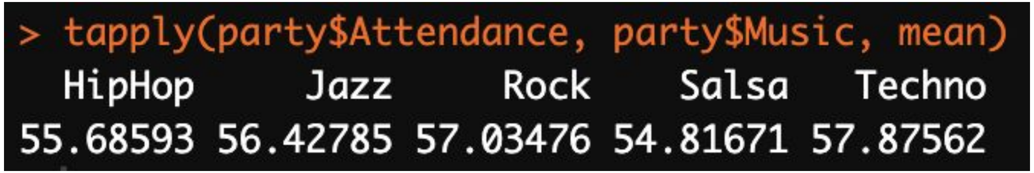
- Categorical Variables: Music, DJ, Day

- In the following tests, we mainly be focusing on the effects of categorical variables on the mean of Attendance.

Part A - Forming multiple hypotheses

Hypothesis I

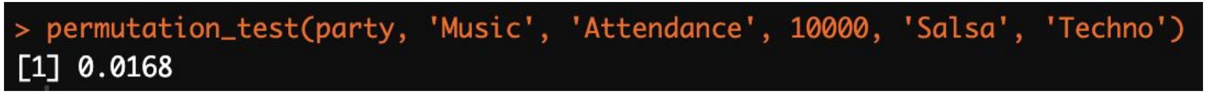
- It is common sense that music likely affects the Attendance of music parties, so we can test that hypothesis.



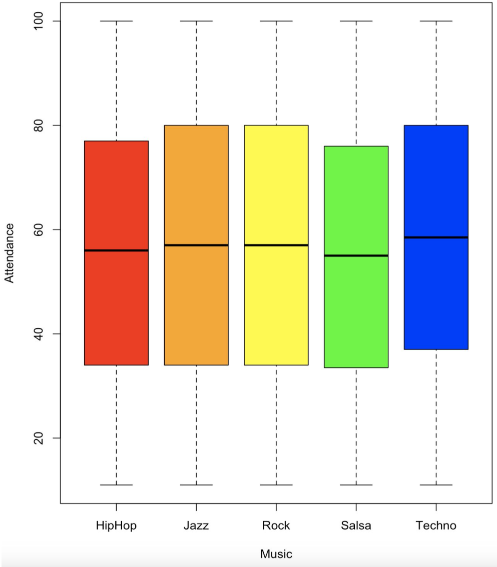
```
> tapply(party$Attendance, party$Music, mean)
  HipHop     Jazz     Rock    Salsa   Techno 
55.68593 56.42785 57.03476 54.81671 57.87562 
```

- We can clearly see that there is a large difference of Attendance mean between Salsa and Techno.

- If we use a permutation test, we can also see there is significance behind our hypothesis.



```
> permutation_test(party, 'Music', 'Attendance', 10000, 'Salsa', 'Techno')
[1] 0.0168
```

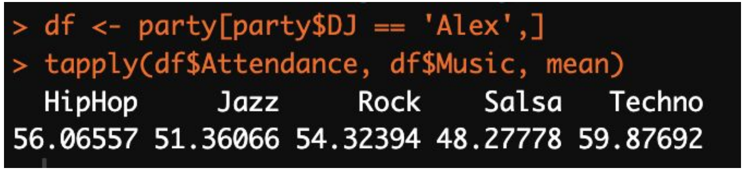


Hypothesis I:

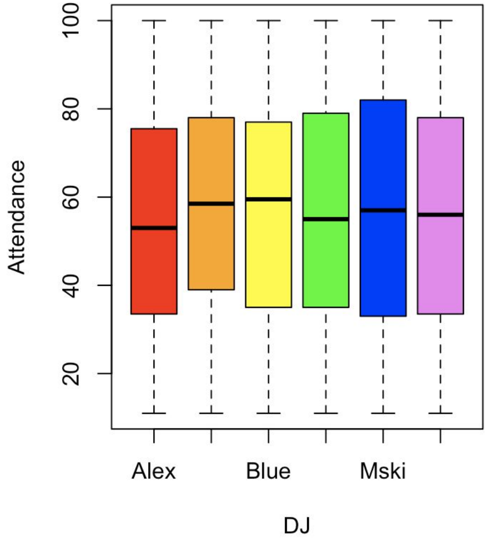
The average Attendance is higher when Music is Techno than when Music is Salsa.

Hypothesis II

- Besides the music, the person who is controlling the music is also highly likely to affect the Attendance of a party.

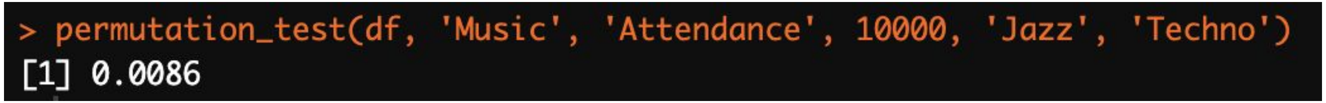


```
> df <- party[party$DJ == 'Alex',]
> tapply(df$Attendance, df$Music, mean)
  HipHop     Jazz     Rock    Salsa   Techno
56.06557 51.36066 54.32394 48.27778 59.87692
```



- We can clearly see that there are differences in Attendance means for different music types of DJ Alex parties, especially between Jazz & Techno.

- If we use a permutation test, we can also see there is significance behind our hypothesis.



```
> permutation_test(df, 'Music', 'Attendance', 10000, 'Jazz', 'Techno')
[1] 0.0086
```

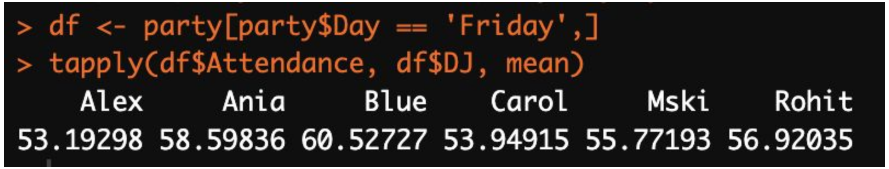
Hypothesis II:

The average Attendance of DJ Alex parties is higher when Music is Techno than when Music is Jazz.

Hypothesis III

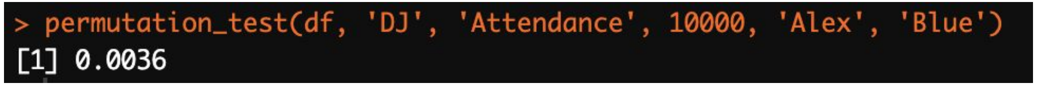
- Like all events, the day is likely to affect the attendance of a party.

- Picking Friday as our day, we see it has the most spread between means for DJs, especially between DJ Blue and DJ Alex.

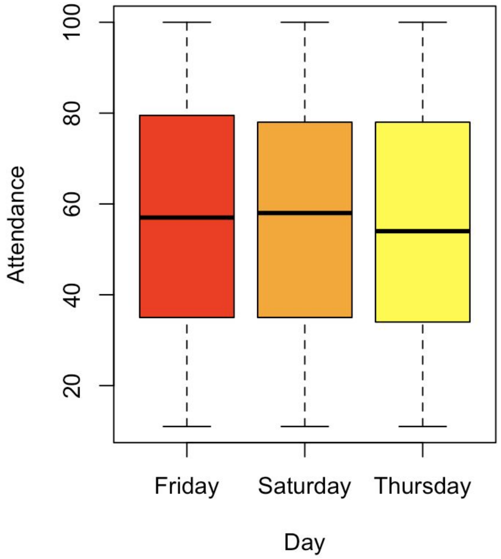


```
> df <- party[party$Day == 'Friday',]
> tapply(df$Attendance, df$DJ, mean)
    Alex     Ania     Blue    Carol     Mski    Rohit 
53.19298 58.59836 60.52727 53.94915 55.77193 56.92035 
```

- If we use a permutation test, we can see there is significance behind our hypothesis.



```
> permutation_test(df, 'DJ', 'Attendance', 10000, 'Alex', 'Blue')
[1] 0.0036
```



Hypothesis III:

The average Attendance of Friday parties is higher when the DJ is Blue than when the DJ is Alex.

Part B - Hypothesis Testing

Hypothesis A

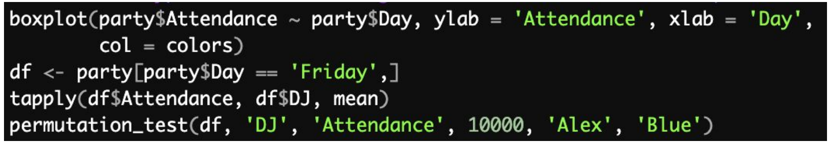
Hypothesis A must have a p-value that is small, but not zero.

Null Hypothesis: The average Attendance of Friday parties is the same when the DJ is Blue and when the DJ is Alex.

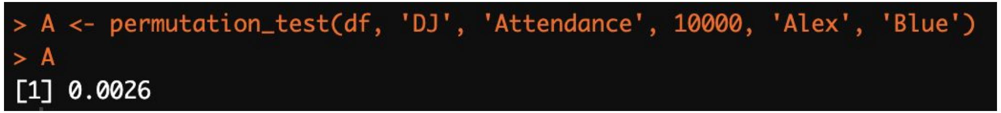
Null Hypothesis: The average Attendance of Friday parties is the higher when the DJ is Blue than when the DJ is Alex.

The method and hypothesis is taken from Part A, Hypothesis III.

Essentially, find the day with most spread in DJ Attendance means, then do a permutation test.



```
boxplot(party$Attendance ~ party$Day, ylab = 'Attendance', xlab = 'Day',
        col = colors)
df <- party[party$Day == 'Friday',]
tapply(df$Attendance, df$DJ, mean)
permutation_test(df, 'DJ', 'Attendance', 10000, 'Alex', 'Blue')
```



```
> A <- permutation_test(df, 'DJ', 'Attendance', 10000, 'Alex', 'Blue')
> A
[1] 0.0026
```

Because our p-value < 0.05, we can reject the Null Hypothesis.

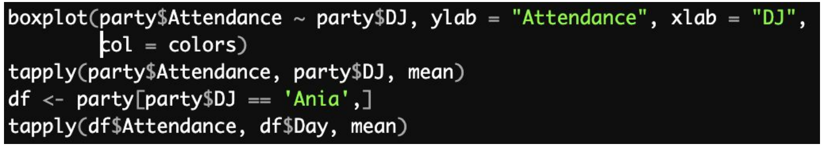
Hypothesis B

Hypothesis B must have a p-value that is close to, but smaller than the critical value of 0.05.

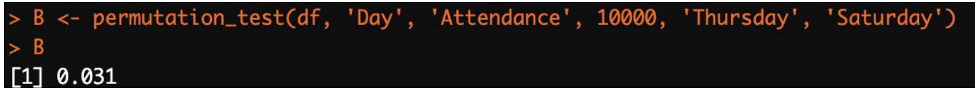
Null Hypothesis: The average Attendance of DJ Ania parties is the same when the Day is Saturday and when the day is Thursday.

Null Hypothesis: The average Attendance of DJ Ania parties is the higher when the Day is Saturday than when the day is Thursday.

Essentially, find the DJ with most spread in Day Attendance means, then do a permutation test.



```
boxplot(party$Attendance ~ party$DJ, ylab = "Attendance", xlab = "DJ",
        col = colors)
tapply(party$Attendance, party$DJ, mean)
df <- party[party$DJ == 'Ania',]
tapply(df$Attendance, df$Day, mean)
```



```
> B <- permutation_test(df, 'Day', 'Attendance', 10000, 'Thursday', 'Saturday')
> B
[1] 0.031
```

Because our p-value < 0.05, we can reject the Null Hypothesis.

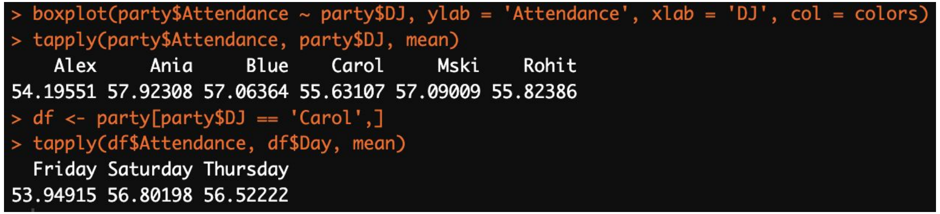
Hypothesis C

Hypothesis C must have a p-value greater than the critical value of 0.05.

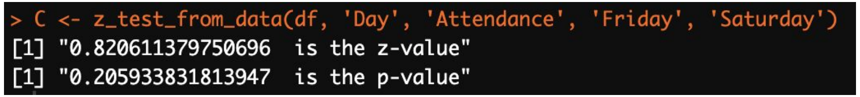
Null Hypothesis: The average Attendance of DJ Carol parties is the <u>same</u> when the Day is Saturday and when the day is Thursday.

Null Hypothesis: The average Attendance of DJ Carol parties is the <u>higher</u> when the Day is Saturday than when the day is Thursday.

Essentially, find the DJ with least spread in Day Attendance means, meaning the difference in means between Days is very small, then do a z-test.



```
> boxplot(party$Attendance ~ party$DJ, ylab = 'Attendance', xlab = 'DJ', col = colors)
> tapply(party$Attendance, party$DJ, mean)
    Alex     Ania     Blue    Carol     Mski    Rohit 
54.19551 57.92308 57.06364 55.63107 57.09009 55.82386 
> df <- party[party$DJ == 'Carol',]
> tapply(df$Attendance, df$Day, mean)
  Friday Saturday Thursday 
53.94915 56.80198 56.52222 
```



```
> C <- z_test_from_data(df, 'Day', 'Attendance', 'Friday', 'Saturday')
[1] "0.820611379750696  is the z-value"
[1] "0.205933831813947  is the p-value"
```

Because our p-value > 0.05, we can fail to reject the Null Hypothesis.

Thank You