Hypothesis Testing 2

By Eric Lin

The Data Set

- The data set we are using contains a list of details about purchased cars and is called Cars.

- Quantitative Variables include:
  - Buyer’s Age

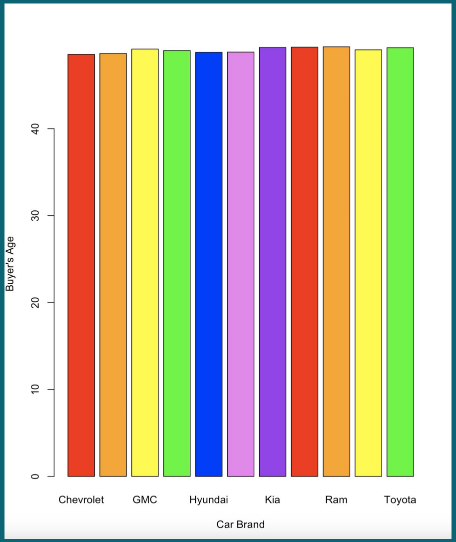
- Categorical Variables include:
  - Car Dealership
  - Season
  - Car Brand
  - Buyer’s Gender

- https://raw.githubusercontent.com/dev7796/data101_tutorial/main/files/dataset/Cars2022.csv

Finding the highest mean Age of Buyer’s

- Using tapply(), we can observe that Ram has the highest mean in terms of Buyer’s Age.

```
> results <- tapply(Cars$Buyer_Age, Cars$Car, mean)
> barplot(results, ylab = 'Buyer\'s Age', xlab = 'Car Brand', col = colors)
> results
Chevrolet      Ford       GMC     Honda   Hyundai      Jeep       Kia    Nissan       Ram    Subaru    Toyota
 48.55071  48.65279  49.15433  48.98918  48.78090  48.80551  49.34194  49.37610  49.41869  49.06511  49.31284
```



- Is this statistically significant enough to be True or is it just random? We can use multiple hypothesis testing to find out for sure.

Bonferroni Correction

- For multiple hypothesis testing, we need to correct the significance level (alpha) using Bonferroni Correction.

- 1.) Find the number of hypothesis tests we will be performing. We find this to be 10.

```
m_count <- length(unique(Cars$Car)) - 1
```

- 2.) Calculate the new significance level using m_count.

```
sig_lvl <- 0.05 / m_count
cat('Significance Level = ', sig_lvl * 100, '% or ', sig_lvl)
```

```
Significance Level =  0.5 % or  0.005
```

Multiple Hypothesis Tests

- Null Hypothesis:

  The average Buyer's Age is the <u>same</u> when the Car is a Ram vs when the Car is a X, where X is any other car in the dataset that is not Ram.

- Alternate Hypothesis:

  The average Buyer's Age is <u>higher</u> when the Car is a Ram vs when the Car is a X, where X is any other acr in the dataset that is not Ram.

Running Multiple Hypothesis Tests

- Run Hypothesis Tests with Car = Ram against all the other possible Cars.

- We find that we fail to reject our Null Hypothesis (H0) in 8/10 of our hypothesis tests.

- Therefore, we cannot reject our Null Hypothesis and it is likely that our alternate hypothesis of Car = Ram having a higher mean Buyer's Age than the rest may be random.

```
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Chevrolet', 'Ram')
# p-value = 0.0010, Reject H0
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Ford', 'Ram')
# p-value = 0.0021, Reject H0
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'GMC', 'Ram')
# p-value = 0.1759, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Honda', 'Ram')
# p-value = 0.0655, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Hyundai', 'Ram')
# p-value = 0.0134, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Jeep', 'Ram')
# p-value = 0.0151, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Kia', 'Ram')
# p-value = 0.3950, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Nissan', 'Ram')
# p-value = 0.4349, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Subaru', 'Ram')
# p-value = 0.1038, Fail to Reject
permutation_test(Cars, 'Car', 'Buyer_Age', 10000, 'Toyota', 'Ram')
# p-value = 0.3517, Fail to Reject
```