Coffee Roast Preference Study

From STAT 206 – Statistical Inference II – Winter 2020

Presentation recreated by Lance Elson

Overview / Background

- 3-person project, my role was to design a study and conduct data analysis after the study was conducted
- 2 studies conducted (Mon, Wed) from 3–5 PM, next to a Starbucks inside the Drexel LeBow College of Business (business building reservation table)
- 3 Starbucks coffee roasts (dark, medium, light) given to each participant in a random order
- Water used as a palate cleanser in between taste tests

- Objective: Find relationships between coffee preferences, demographic, and habitual data for marketing applications, using parametric, non-parametric, and linear regression analysis techniques

Data Collected

- Tester - Unique number assigned to each participant (Scrapped in analysis)
- DayofWeek - Which day (Mon, Wed) the test was conducted
- DarkRoastRating - Roast rating on a scale of -10 (hated it) to 10 (loved it)
- MediumRoastRating - Roast rating on a scale of -10 (hated it) to 10 (loved it)
- LightRoastRating - Roast rating on a scale of -10 (hated it) to 10 (loved it)
- FavCoffee - Calculated value observing which roast participant liked the most

Demographic

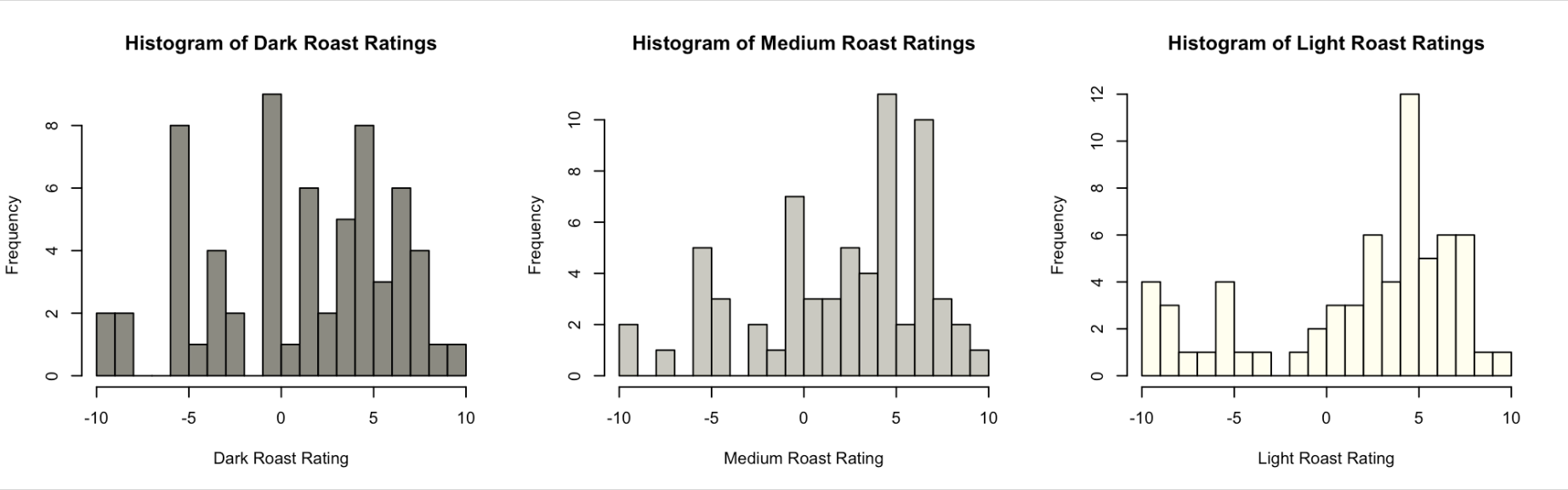
- Major - What college major someone is. Cleaned up for consistency.
- Gender - Participant gender
- Age - Participant age

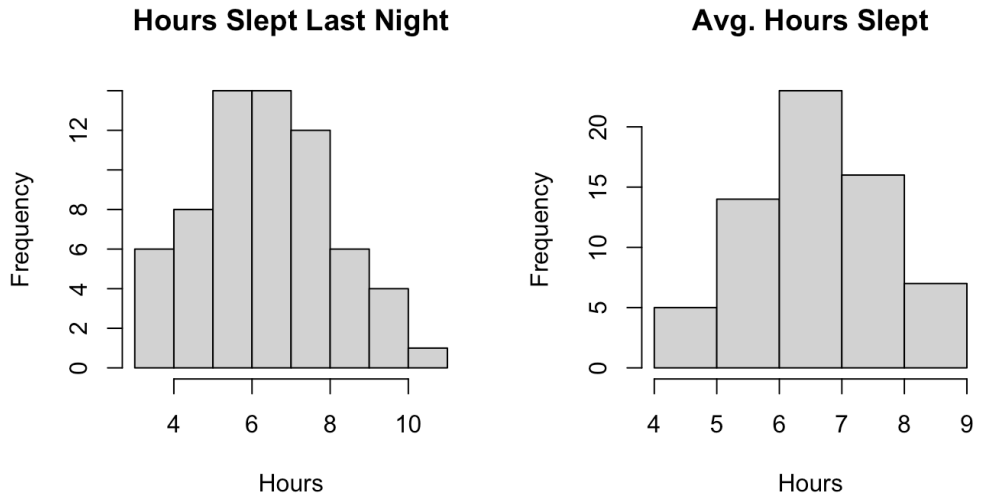
Habitual

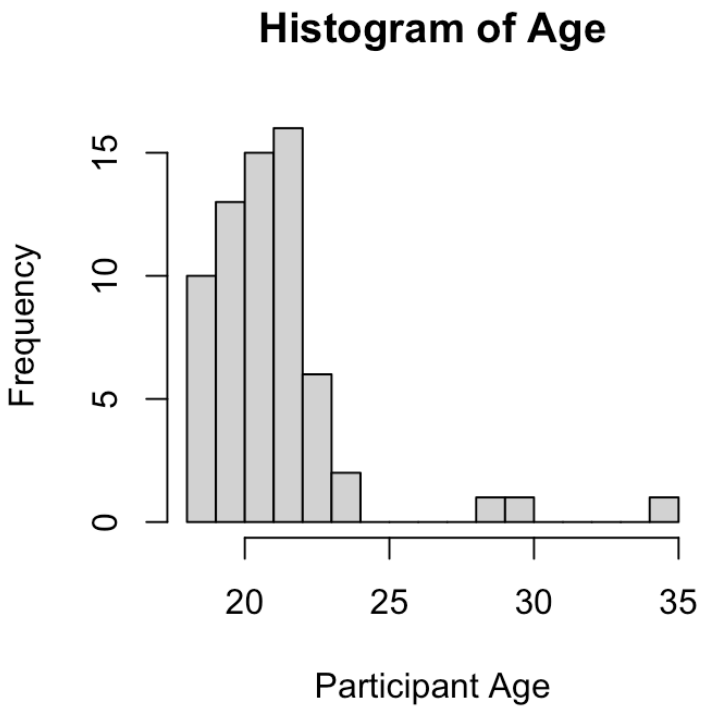
- AgeStartDrinkCoffee - Participant's age at which they started drinking coffee
- Smoker - Whether participant smokes regularly
- HrsSleptPrev - How many hours the participant slept before the day of the test
- SleptAvgFactor - Whether participant slept more or less than mean slept hours
- HrsSleptAvg - How many hours the participant usually sleeps per night
- SleptAvgFactor - Whether participant sleeps more or less than mean sleeping hours
- DaysWeekDrinkCoffee - How many days a week participant drinks coffee
- CoffeeflozPerDay - How much coffee a participant drinks on days they drink coffee
- CaffieneIndex - How a person drinks their coffee on a scale of decaf (0) to extremely caffinated (10)
- OtherCaf - Whether a person regularly drinks other caffeinated beverages besides coffee
- EstWeekCoffeeCafCons - Calculated field using the above metrics determining how much caffeine participant gets from coffee every week

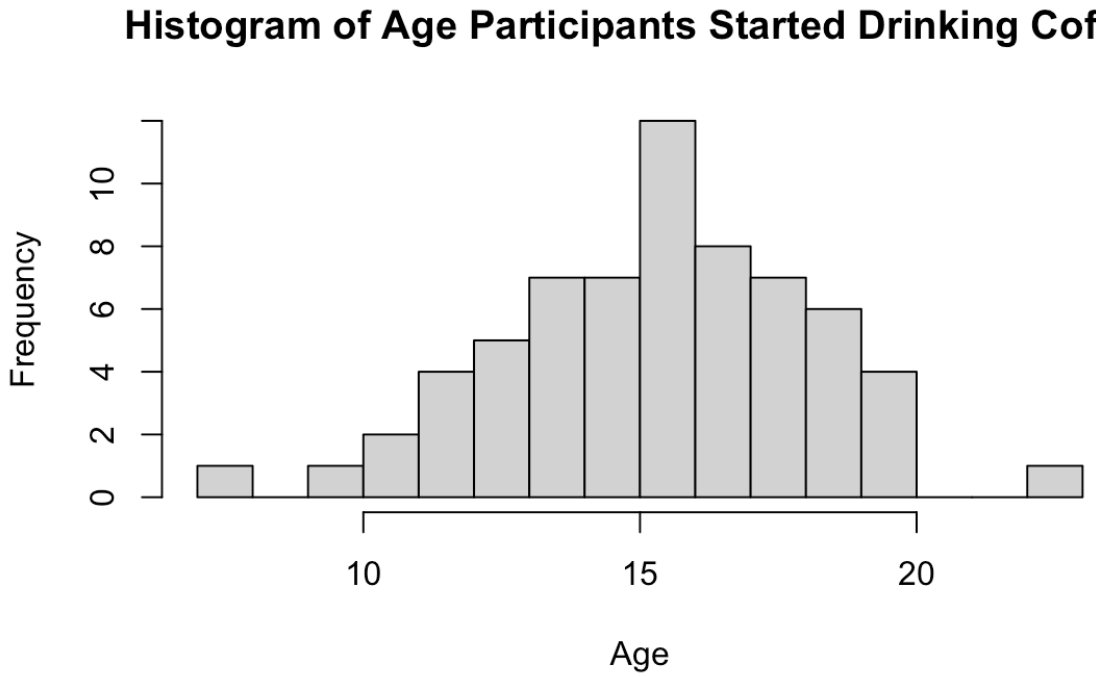
Error Sources

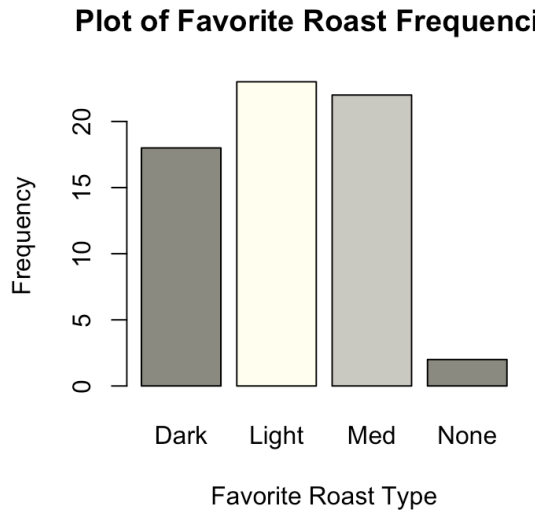
- Test was conducted next to a Starbucks in a college building. Our population, therefore, was skewed towards coffee drinkers.
- Population age was heavily skewed towards college students
- Only 7 out of 65 participants are smokers
- Abnormally distributed data for coffee roast preferences. Generally, people like coffee (giving it rating of 5). Also spike at disliking coffee (-5)




Hours Slept Last Night
Frequency
0
2
4
6
8
12
4
6
8
10
Hours
Avg. Hours Slept
Frequency
0
5
10
15
20
4
5
6
7
8
9
Hours


Histogram of Age
Frequency
0
5
10
15
20
25
30
35
Participant Age


Histogram of Age Participants Started Drinking Coffee
Frequency
0
2
4
6
8
10
10
15
20
Age


Plot of Favorite Roast Frequencies
Frequency
0
5
10
15
20
Dark
Light
Med
None
Favorite Roast Type

Analysis I

Linear Regression - Variables vs. How People Rate Roasts

- Via linear regression, none of these variables have a significant relationship with roast rating

```
Call:
lm(formula = Rating ~ Age + AgeStartDrinkCoffee + DayofWeek +
    Smoker + Gender + CoffeeType + HrsSleptAvg + OtherCaf, data = blockingdata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6439  -2.9917   0.9946   3.6323   8.8980

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.09631    4.35977  -0.022   0.9824
Age                  0.15995    0.14283   1.120   0.2642
AgeStartDrinkCoffee -0.12370    0.13807  -0.896   0.3714
DayofWeekWednesday  -1.41549    0.78861  -1.795   0.0743 .
SmokerYes            1.20674    1.26232   0.956   0.3403
GenderMale           0.81683    0.79187   1.032   0.3036
CoffeeTypeLight      0.64615    0.89290   0.724   0.4702
CoffeeTypeMed        1.00000    0.89290   1.120   0.2642
HrsSleptAvg          0.12637    0.34303   0.368   0.7130
OtherCafYes         -1.14819    0.75956  -1.512   0.1323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.09 on 185 degrees of freedom
Multiple R-squared:  0.05339,   Adjusted R-squared:  0.00734
F-statistic: 1.159 on 9 and 185 DF,  p-value: 0.3236
```

Analysis II

Linear Regression - Variables vs. Average Sleep

- Via linear regression, people who drink coffee later in life tend to sleep more.
- Smokers tend to sleep more than nonsmokers
- People who drink other caffeinated beverages apart from coffee tend to sleep less

```
Call:
lm(formula = HrsSleptAvg ~ Age + AgeStartDrinkCoffee + DayofWeek +
    CoffeeType + Smoker + Gender + OtherCaf, data = blockingdata)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26179 -0.69303 -0.04792  0.80311  2.23143

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.069e+00  7.745e-01   9.128   <2e-16 ***
Age                 -4.771e-02  3.033e-02  -1.573   0.1174
AgeStartDrinkCoffee  7.061e-02  2.906e-02   2.430   0.0160 *
DayofWeekWednesday  -2.757e-02  1.686e-01  -0.164   0.8702
CoffeeTypeLight     -5.496e-15  1.909e-01   0.000   1.0000
CoffeeTypeMed       -5.459e-15  1.909e-01   0.000   1.0000
SmokerYes            5.905e-01  2.663e-01   2.217   0.0278 *
GenderMale           8.893e-02  1.691e-01   0.526   0.5996
OtherCafYes         -3.828e-01  1.599e-01  -2.394   0.0177 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.088 on 186 degrees of freedom
Multiple R-squared:    0.1,     Adjusted R-squared:  0.06132
F-statistic: 2.584 on 8 and 186 DF,  p-value: 0.01066
```

Analysis III

Linear Regression - Variables vs. Est. Weekly Caffeine Consumption from Coffee

- Via linear regression, people who started drinking coffee later in life regularly consume less caffeine
- Smokers regularly consume more caffeine

```
Call:
lm(formula = EstWeekCoffeeCafCons ~ Age + AgeStartDrinkCoffee +
    CoffeeType + Smoker + Gender + OtherCaf + HrsSleptAvg, data = blockingdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1929.42  -743.97   -76.82   568.70  3057.91

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.428e+03  8.881e+02   4.986 1.41e-06 ***
Age                  1.122e+01  2.861e+01   0.392   0.6955
AgeStartDrinkCoffee -1.924e+02  2.812e+01  -6.840 1.11e-10 ***
CoffeeTypeLight     -1.125e-12  1.819e+02   0.000   1.0000
CoffeeTypeMed       -1.123e-12  1.819e+02   0.000   1.0000
SmokerYes            5.346e+02  2.488e+02   2.149   0.0329 *
GenderMale           3.481e+01  1.573e+02   0.221   0.8251
OtherCafYes         -2.999e+02  1.527e+02  -1.963   0.0511 .
HrsSleptAvg         -1.604e+01  6.987e+01  -0.230   0.8186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1037 on 186 degrees of freedom
Multiple R-squared:  0.2553,    Adjusted R-squared:  0.2233
F-statistic:  7.97 on 8 and 186 DF,  p-value: 3.252e-09
```

Study Question I

Do people prefer certain roasts over others?

- H_0 = All people like the same coffee roast equally
- H_A = There is a difference in the means of the ratings between at least one pair of the roasts

Using both a parametric and non-parametric test, we fail to reject the null hypothesis that all people like each roast of coffee equally

```
             Df Sum Sq Mean Sq F value   Pr(>F)
CoffeeType    2   33.4   16.71   0.967    0.383
Tester       64 2818.6   44.04   2.549 3.59e-06 ***
Residuals   128 2211.9   17.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Kruskal-Wallis rank sum test

data:  Rating by CoffeeType
Kruskal-Wallis chi-squared = 1.6807, df = 2, p-value = 0.4316
```

ANOVA with 2 blocked factors: Participant and the day of the week when the data was recorded

Kruskal test, considering abnormal roast rating distributions

Study Question II

Is there a relationship between gender and coffee preference?

- H_0 = Each gender likes the same coffee roast equally for all roasts
- H_A = There is a difference in the means of the roast preferences between genders for at least one coffee roast

Using a parametric test, we nearly have a relationship between gender and coffee roast preference. This can be further studied moving forward.

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
Gender             1   59.9   59.86   3.454   0.0654 .
CoffeeType         2   33.4   16.71   0.964   0.3840
Tester            63 2758.8   43.79   2.527 5.13e-06 ***
Gender:CoffeeType  2   28.4   14.18   0.818   0.4436
Residuals        126 2183.6   17.33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Kruskal-Wallis rank sum test

data:  Rating by interaction(Gender, CoffeeType)
Kruskal-Wallis chi-squared = 4.1357, df = 5, p-value = 0.53
```

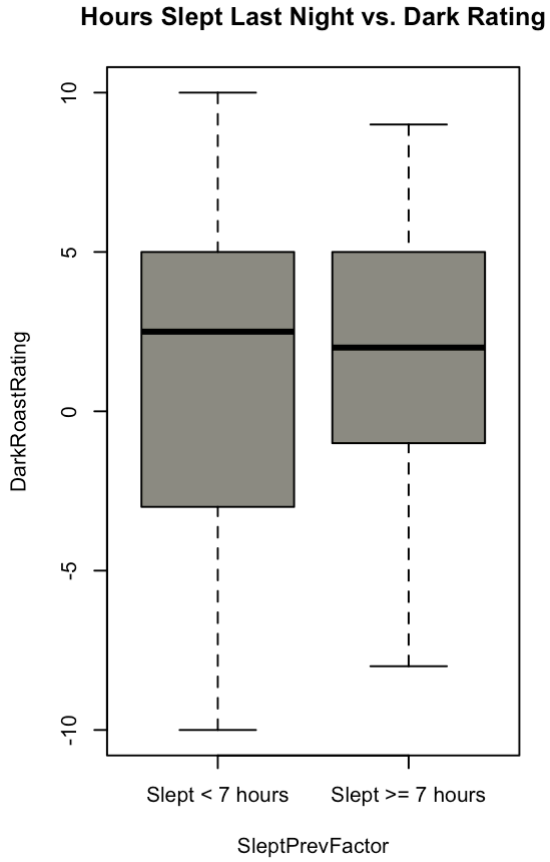
2-way ANOVA with interaction and 2 blocked factors: Rating vs. Gender vs. Coffee Roast, blocked Participant and the day of the week when the data was recorded

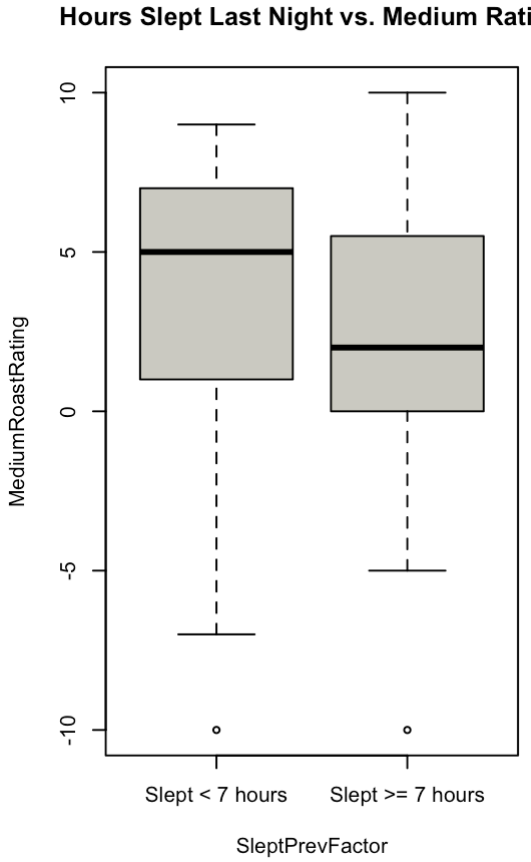
Kruskal test, considering abnormal roast rating distributions

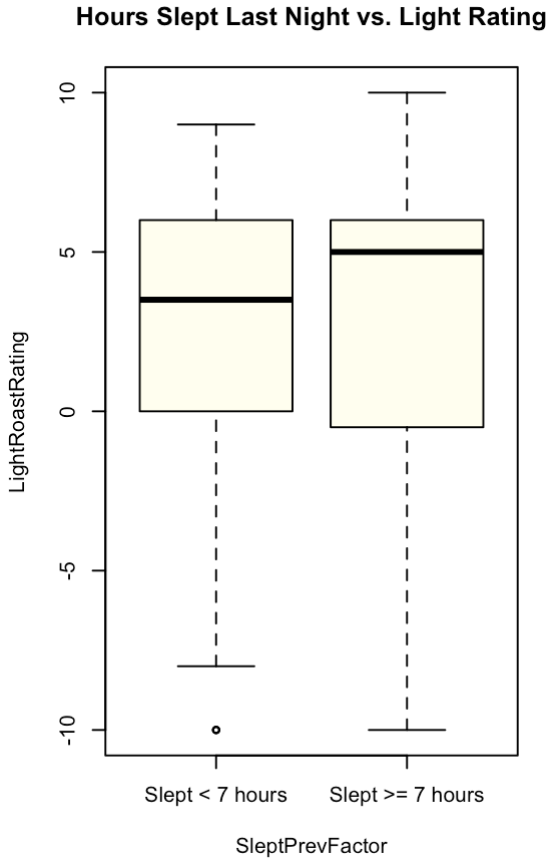
Study Question III

Is there a relationship between sleeping habits and coffee roast preference?

- H_0 = People who slept greater than or equal to 7 hours before the test have the same coffee preferences as people who slept less than 7 hours
- H_A = There is a difference in the means of coffee preferences between people who slept more than or equal to 7 hours before the test, and people who slept less than 7 hours before the test







Study Question III

Is there a relationship between sleeping habits and coffee roast preference?

- Using both a parametric and non-parametric test, we can conclude that there is not sufficient evidence to say that coffee tastes relate to how many hours people slept the previous night

```
                            Df Sum Sq Mean Sq F value  Pr(>F)
SleptPrevFactor              1    0.0    0.00   0.000   0.998
CoffeeType                   2   33.4   16.71   0.963   0.385
Tester                      63 2818.6   44.74   2.577 3.3e-06 ***
SleptPrevFactor:CoffeeType   2   24.2   12.08   0.696   0.501
Residuals                  126 2187.7   17.36
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Kruskal-Wallis rank sum test

data:  Rating by interaction(SleptPrevFactor, CoffeeType)
Kruskal-Wallis chi-squared = 3.1709, df = 5, p-value = 0.6737
```

2-way ANOVA with interaction and 2 blocked factors: Rating vs. Slept more/equal/less than 7 hours of sleep the night before the test vs. Coffee Roast, blocked Participant and the day of the week when the data was recorded

Kruskal test, considering abnormal roast rating distributions

Key Conclusions

- Statistically significant relationship between regularly smoking and regularly consuming more caffeine
  - Lifestyle and habitual similarities between smokers and coffee drinkers can be used in target marketing
- Statistically significant relationship between drinking coffee earlier and life and regularly consuming more caffeine
  - Marketing towards a younger demographic, like students and young professionals can encourage customer loyalty once they are engaged in coffee consumption
- Nearly significant relationship between gender and coffee roast preference
  - Can be explored using further study