Statistical Analyses on SleepStudy Dataset

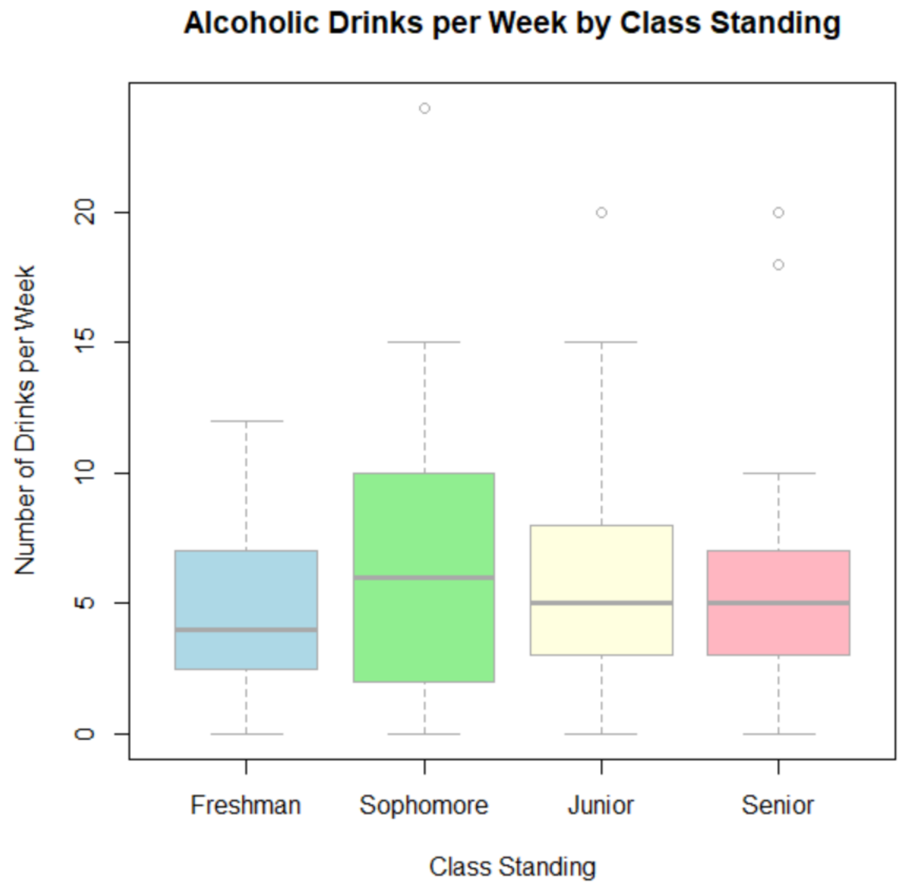
Data Visualization:

1.a. Comparative Boxplot of Alcoholic Drinks per Week by Class Standing

```
data = SleepStudy

# 1.a.
# Convert ClassYear to factor with labels
data$ClassYear = factor(data$ClassYear, levels = 1:4, labels = c("Freshman", "Sophomore", "Junior", "Senior"))

# Construct boxplot
windows()
boxplot(Drinks ~ ClassYear, data = data,
        main = "Alcoholic Drinks per Week by Class Standing",
        xlab = "Class Standing",
        ylab = "Number of Drinks per Week",
        col = c("lightblue", "lightgreen", "lightyellow", "lightpink"),
        border = "darkgray")
```



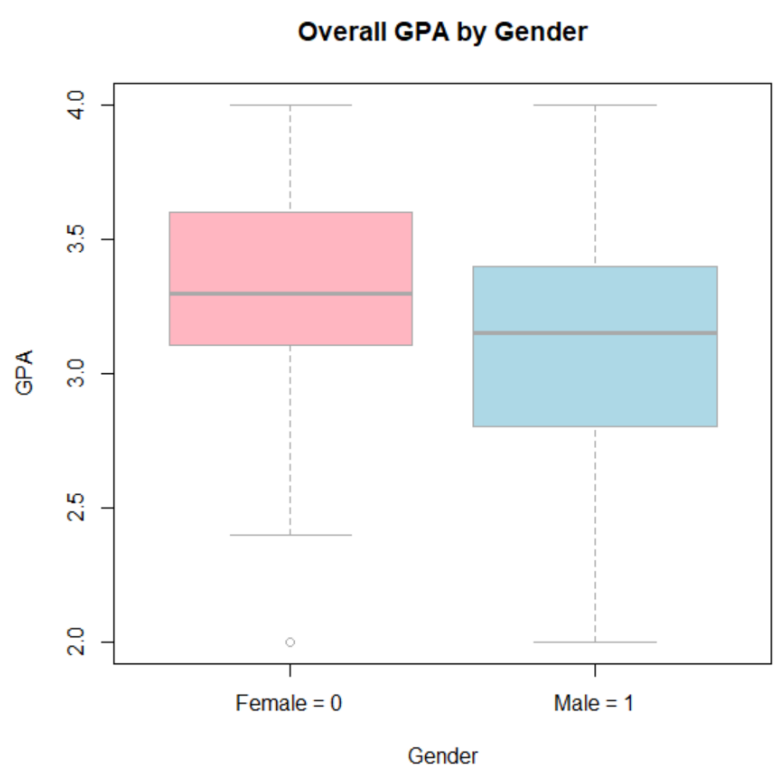
The comparative boxplot suggests:

- Sophomores have the highest median, followed by juniors and seniors, then freshman. This suggests that sophomores drink more *on average* than other classes.
- Sophomores show the widest spread (highest IQR), suggesting high variability in drinking habits. Seniors have the narrowest IQR, suggesting more consistent drinking habits.
- Upperclassmen (sophomores to seniors) have extreme outliers.
- All in all, **all upperclassmen** (sophomores to seniors) **drink more than freshman**.

1.b. Comparative Boxplot of Overall GPA by Gender

```r
# 1.b.
# Convert Gender to character (0=female, 1=male)
data$Gender = as.character(data$Gender)

# Construct boxplot
windows()
boxplot(GPA ~ Gender, data = data,
        main = "Overall GPA by Gender",
        xlab = "Gender",
        ylab = "GPA",
        col = c("lightpink", "lightblue"),
        names = c("Female = 0", "Male = 1"),
        border = "darkgray")
```



The comparative boxplot suggests:

- Female students have a slightly higher median GPA than male students.
- Male students show slightly more variability in performance, but females cluster more towards the end.

- One female student has a very low GPA (~2.0), which is unusual given the otherwise strong performance of the group.
- All in all, **females outperform males** in **GPA** at **every quartile**.

Hypothesis Testing:

2.a. Stress Level by Gender

1) Hypothesis

μ_F = average stress level for female students

μ_M = average stress level for male students

$H_0: \mu_F = \mu_M$ *(no difference in stress)*
$H_1: \mu_F \neq \mu_M$ *(difference in stress)*

2) Preparation

- Alpha = 5%
- n > 30
- σ-unknown
- Independent samples (female vs. male students)
- Data Provided

So, a two-tailed, two-sample t-test is performed.

3) Computation & Comparison

```
# 2
# Subset the data by gender
MaleSleep=subset(data, data$Gender=="1")
FemaleSleep=subset(data, data$Gender=="0")

alpha = 0.05

# 2.a.
# Perform t-test
t.test(FemaleSleep$StressScore, MaleSleep$StressScore, alternative=
"two.sided", conf.level=1-alpha)
```

```
        Welch Two Sample t-test

data:  FemaleSleep$StressScore and MaleSleep$StressScore
t = 2.9552, df = 225.43, p-value = 0.003457
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9773598 4.8892809
sample estimates:
mean of x mean of y 
10.649007  7.715686 
```

4) Interpretation

- p-value (0.003) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis**.
- We have **enough statistical evidence** to conclude that the stress level for female and male students are **different**.

2.b. Weekday Sleep by Gender

1) Hypothesis

μ_F = average sleep on weekdays for female students

μ_M = average sleep on weekdays for male students

H_0: $\mu_F \leq \mu_M$ (females sleep less or equal than males)

H_1: $\mu_F > \mu_M$ (females sleep more than males)

2) Preparation

- Alpha = 5%
- n > 30
- σ-unknown
- Independent samples (female vs. male students)
- Data Provided

So, a right-tailed, two-sample t-test is performed.

3) Computation & Comparison

```
# 2.b.
# Perform t-test
t.test(FemaleSleep$WeekdaySleep, MaleSleep$WeekdaySleep, alternative
="greater", conf.level=1-alpha)
```

```
        Welch Two Sample t-test

data:  FemaleSleep$WeekdaySleep and MaleSleep$WeekdaySleep
t = 1.0218, df = 226.08, p-value = 0.154
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.09308275         Inf
sample estimates:
mean of x mean of y
 7.926887  7.775882
```

4) Interpretation

- p-value (0.154) > alpha (0.05)
- At **α=0.05**, we **fail to reject** the **null hypothesis**.
- We can conclude that there is **no significant evidence** that female students sleep more than male students on weekdays.

Simple Linear Regression:

3. Stress Score Predicted by Anxiety Score

1) Hypothesis

H_0: *There is **no relationship** between anxiety score and stress score.*

H_1: *There is **a relationship** between anxiety score and stress score.*

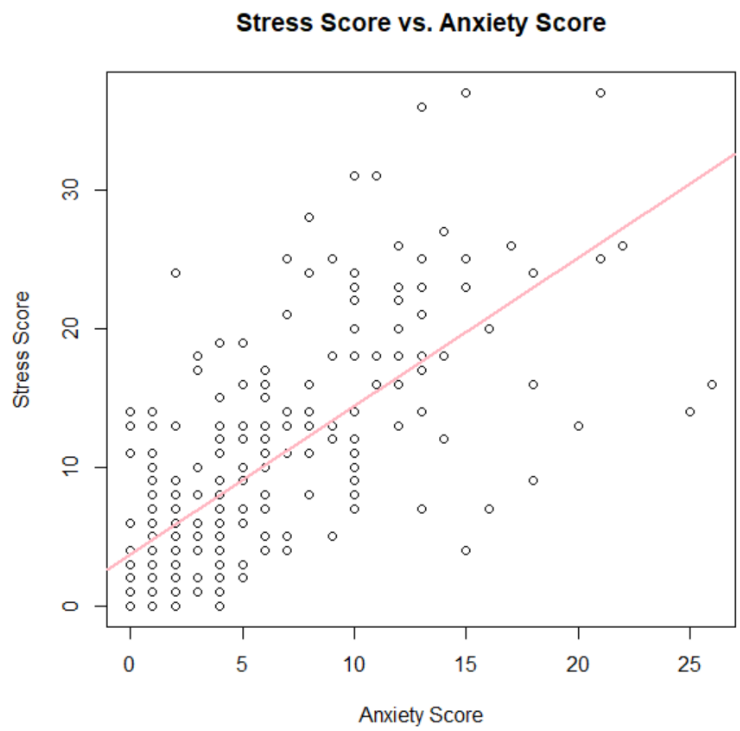
2) Preparation

- Alpha = 5%
- Both StressScore (dependent) and AnxietyScore (independent) are numeric.

```
# 3
# Construct scatter plot
windows()
plot(data$AnxietyScore, data$StressScore, main="Stress Score vs. Anxiety Score", xlab="Anxiety Score", ylab="Stress Score")

# Create model
model_slr1 = lm(StressScore ~ AnxietyScore, data=data)

# Add regression line
abline(model_slr1, col="lightpink", lwd=2)
```



The scatterplot suggests **slightly strong positive linear** relationship between Stress Score and Anxiety Score.

3) Computation & Comparison

```
# Regression analysis
summary(model_slr1)
```

```
Call:
lm(formula = StressScore ~ AnxietyScore, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.4302  -3.6138  -0.8179   2.9984  18.3862 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.72944    0.51637   7.222 6.09e-12 ***
AnxietyScore  1.06803    0.06915  15.445  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.705 on 251 degrees of freedom
Multiple R-squared:  0.4873,    Adjusted R-squared:  0.4852 
F-statistic: 238.5 on 1 and 251 DF,  p-value: < 2.2e-16
```

4) Interpretation

- p-value (<2e-16) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis**.
- We can conclude that anxiety is a **significant predictor** of stress.

4. Stress Score Predicted by Depression Score

1) Hypothesis

H_0**:** *There is* ***no relationship*** *between depression score and stress score.*

H_1**:** *There is* ***a relationship*** *between despression score and stress score.*

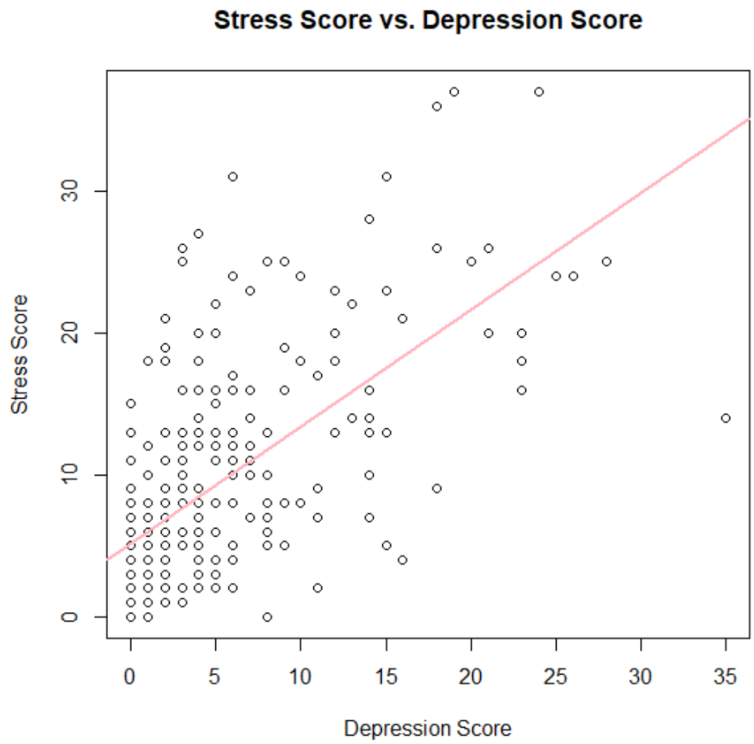
2) Preparation

- Alpha = 5%
- Both StressScore (dependent) and DepressionScore (independent) are numeric.

```
# 4
# Construct scatter plot
windows()
plot(data$DepressionScore, data$StressScore, main="Stress Score vs. Depression Score", xlab="Depression Score", ylab="Stress Score")

# Create model
model_slr2 = lm(StressScore ~ DepressionScore, data=data)

# Add regression line
abline(model_slr2, col="lightpink", lwd=2)
```



The scatterplot suggests **moderate positive linear** relationship between Stress Score and Depression Score.

3) Computation & Comparison

```r
# Regression analysis
summary(model_slr2)
```

```
Call:
lm(formula = StressScore ~ DepressionScore, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.026  -4.179  -1.652   3.513  20.876 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       5.1794     0.5195   9.969   <2e-16 ***
DepressionScore   0.8242     0.0655  12.584   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.239 on 251 degrees of freedom
Multiple R-squared:  0.3868,	Adjusted R-squared:  0.3844 
F-statistic: 158.4 on 1 and 251 DF,  p-value: < 2.2e-16
```

4) Interpretation

- p-value (<2e-16) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis**.
- We can conclude that depression is a **significant predictor** of stress.

Multiple Linear Regression:

5. Stress Score Predicted by Happiness, Weekday Sleep and Weekend Sleep

1) Hypothesis

H_0: $\beta_{Happiness} = \beta_{WeekdaySleep} = \beta_{WeekendSleep} = 0$ *(None of the predictors significantly predict StressScore.)*

H_1: At least one $\beta \neq 0$ *(At least one predictor significantly predicts StressScore.)*

2) Preparation

- Alpha = 5%
- F-test statistic

3) Computation & Comparison

```r
# 5
# Create model
model_mlr1 = lm(StressScore ~ Happiness + WeekdaySleep + WeekendSleep, data = data)

# Regression analysis
summary(model_mlr1)
```

```
Call:
lm(formula = StressScore ~ Happiness + WeekdaySleep + WeekendSleep,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.396  -5.208  -1.722   4.382  27.014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.62902    4.38780   5.157 5.11e-07 ***
Happiness    -0.52271    0.08557  -6.109 3.84e-09 ***
WeekdaySleep -0.50494    0.40413  -1.249    0.213
WeekendSleep  0.54247    0.34519   1.571    0.117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.411 on 249 degrees of freedom
Multiple R-squared:  0.1417,	Adjusted R-squared:  0.1314
F-statistic:  13.7 on 3 and 249 DF,  p-value: 2.665e-08
```

4) Interpretation

- Overall p-value (2.665e-08) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis** and there is **at least one predictor** that is **statistically significant**.
- **Happiness**: Significant => p-value (3.84e-09) < alpha
- **WeekdaySleep**: Not significant => p-value (0.213) > alpha
- **WeekendSleep**: Not significant => p-value (0.117) > alpha
- **Happiness significantly predicts stress** while sleep variables do not.

6. Stress Score Predicted by Anxiety and Depression

1) Hypothesis

H_0: $\beta_{AnxietyScore} = \beta_{DepressionScore} = 0$ *(Neither AnxietyScore nor DepressionScore significantly predict StressScore.)*

H_1: At least one $\beta \neq 0$ *(At least one predictor significantly predicts StressScore.)*

2) Preparation

- Alpha = 5%
- F-test statistic

3) Computation & Comparison

```
# 6
# Create model
model_mlr2 = lm(StressScore ~ AnxietyScore + DepressionScore, data = data)

# Regression analysis
summary(model_mlr2)
```

```
Call:
lm(formula = StressScore ~ AnxietyScore + DepressionScore, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.7552  -3.0351  -0.6377   2.7937  16.8150 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      2.65676    0.48065   5.527 8.16e-08 ***
AnxietyScore     0.79279    0.07073  11.208  < 2e-16 ***
DepressionScore  0.49045    0.06126   8.006 4.48e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.1 on 250 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5886 
F-statistic: 181.3 on 2 and 250 DF,  p-value: < 2.2e-16
```

4) Interpretation

- Overall p-value (2.2e-16) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis** and there is **at least one predictor** that is **statistically significant**.
- **AnxietyScore**: Significant => p-value (<2e-16) < alpha
- **DepressionScore**: Significant => p-value (4.48e-14) > alpha
- **Both anxiety** and **depression significantly predicts stress** in the model, with anxiety showing a stronger effect.

Chi Square for Categorical Variables:

7. Association between Gender and Alcohol Use

1) Hypothesis

***H_0*: *Gender and alcohol use are* not associated.**

***H_1*: *Gender and alcohol use are* associated.**

2) Preparation

- Alpha = 5%
- Both Gender and AlcoholUse are categorical.
- Check expected counts ≥ 5 in all cells (for Chi-square validity)

3) Computation & Comparison

```
> # 7
> # Create contingency table
> table1 = table(
+   "Gender" = ifelse(data$Gender == 0, "0=Female", "1=Male"),
+   "AlcoholUse" = data$AlcoholUse
+ )
> table1
          AlcoholUse
Gender      Abstain Heavy Light Moderate
  0=Female       20     5    60       66
  1=Male         14    11    23       54
>
> # Perform chi-square test
> chisq.test(table1)

        Pearson's Chi-squared test

data:  table1
X-squared = 11.961, df = 3, p-value = 0.007517
```

The two-way table suggests:

- Alcohol use patterns vary significantly by gender, with females more likely in the "Light" and "Moderate" groups, while males are slightly more likely to be in the "Heavy" group.

4) Interpretation

- Test Statistic: $\chi^2 = 11.96$
- p-value (0.007517) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis** and can conclude that there is a **significant association** between gender and alcohol use.

8. Association between Gender and All-nighters

1) Hypothesis

H_0: *Gender and all-nighters are **not associated**.*
H_1: *Gender and all-nighters are **associated**.*

2) Preparation

- Alpha = 5%
- Both Gender and All-nighters are categorical.
- Check expected counts ≥ 5 in all cells (for Chi-square validity)

3) Computation & Comparison

```
> # 8
> # Create contingency table
> table2 = table(
+   "Gender" = ifelse(data$Gender == 0, "0=Female", "1=Male"),
+   "AllNighter" = ifelse(data$AllNighter == 0, "0=No", "1=Yes")
+ )
> table2
          AllNighter
Gender     0=No 1=Yes
  0=Female  139    12
  1=Male     80    22
>
> # Perform chi-square test
> chisq.test(table2)

	Pearson's Chi-squared test with Yates' continuity correction

data:  table2
X-squared = 8.5746, df = 1, p-value = 0.003409
```

The two-way table suggests:

- There is a **difference** in all-nighter behavior by gender. **Males** show a **higher** frequency of pulling all-nighters than females.

4) Interpretation

- Test Statistic: $\chi^2 = 8.57$
- p-value (0.0034) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis** and can conclude that there is a **significant association** between gender and all-nighters.

ANOVA:

9. GPA Differences by Class Standing

1) Hypothesis

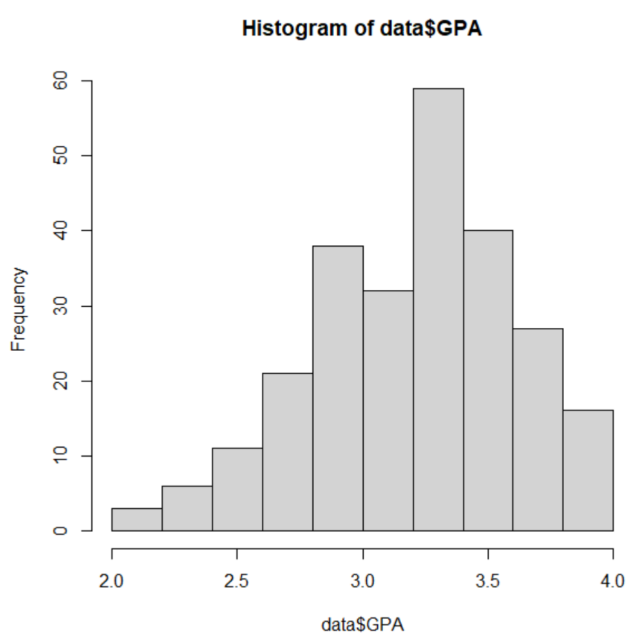
H_0: $\mu_{Freshman} = \mu_{Sophomore} = \mu_{Junior} = \mu_{Senior}$ *(There is **no significant difference** in mean GPA among class standings.)*

***H_1: At least one** class standing has a **significantly different** mean GPA from the others.*

2) Preparation

- Variables
  - Independent: ClassYear (categorical, 4 levels)
  - Dependent: GPA (continuous)
- ANOVA Conditions
  - **Normality**: n > 30

```
# Check Normality
# Check Numerical Variable
windows();
hist(data$GPA)
```



The **histogram** of GPA suggests a reasonably normal distribution. Since our sample size is 40 (**n > 30**), we assume normality by the **Central Limit Theorem (CLT).**

- **Independence:** Assume random sampling
- **Homogeneity/ Equality of Variances**

```
> sd_by_gpa = tapply(data$GPA, data$ClassYear, sd)
> sd_by_gpa
        1         2         3         4 
0.3661250 0.3736898 0.4198895 0.3639701 
```

```
> sds = c(0.3661250, 0.3736898, 0.4198895, 0.3639701)
> vector <- sds
> result <- FALSE
> 
> # Loop through each element in the vector
> for (i in 1:length(vector)) {
+   for (j in 1:length(vector)) {
+     if (i != j && vector[i] > 2 * vector[j]) {
+       result <- TRUE
+       break
+     }
+   }
+   if (result) break
+ }
> 
> # Print the result: We want to see FALSE
> print(result)
[1] FALSE
```

To check the assumption of homogeneity of variances, we verified that no group's standard deviation exceeded **twice** that of another group. Since this condition was met (FALSE result from the test), we assume **equal variances** and proceed with ANOVA.

3) ANOVA Test

A **one-way ANOVA** is conducted to compare the mean GPA among the class years.

```
> # Perform ANOVA
> one.way1 = aov(data$GPA ~ data$ClassYear)
> summary(one.way1)
                 Df Sum Sq Mean Sq F value   Pr(>F)    
data$ClassYear    3   5.13  1.7109   11.82 2.91e-07 ***
Residuals       249  36.06  0.1448                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Test Interpretation

- p-value (2.91e-07) < alpha (0.05)
- At **α=0.05**, we **reject** the **null hypothesis** and can conclude that there is **at least one** class standing has a **significantly different** mean GPA from the others.

4) Post Hoc Analysis (Tukey’s HSD Test)

Since the ANOVA test showed significance, we perform **Tukey’s HSD** to determine **which groups differ**.

```
# Convert ClassYear to factor with labels
data$ClassYear = factor(data$ClassYear, levels = 1:4, labels = c
("Freshman", "Sophomore", "Junior", "Senior"))
```

```
> tukey_result = TukeyHSD(one.way1)
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ ClassYear, data = data)

$ClassYear
                          diff         lwr        upr     p adj
Sophomore-Freshman -0.40029339 -0.57581330 -0.2247735 0.0000001
Junior-Freshman    -0.31398345 -0.51032301 -0.1176439 0.0002808
Senior-Freshman    -0.29629339 -0.49021368 -0.1023731 0.0005822
Junior-Sophomore    0.08630994 -0.08142649  0.2540464 0.5441416
Senior-Sophomore    0.10400000 -0.06089804  0.2688980 0.3629506
Senior-Junior       0.01769006 -0.16921460  0.2045947 0.9948334
```

Tukey's HSD Test Interpretation

- 3 out of 6 groups - **Sophomore vs. Freshman** ($p = 0.0000001$), **Junior vs. Freshman** ($p = 0.00028$), **Senior vs. Freshman** ($p = 0.00058$) - show a significant difference.

5) Conclusion

- **Freshmen** GPAs **significantly lower** than Sophomores (−0.40), Juniors (−0.31), and Seniors (−0.30)
- **Upperclassmen** (Sophomore–Senior) show **no significant differences**.

10. GPA Differences by Alcohol Use

1) Hypothesis

H_0: $\mu_{Abstain} = \mu_{Light} = \mu_{Moderate} = \mu_{Heavy}$ *(There is **no significant difference** in mean GPA across alcohol use groups.)*

H_1: At least one *alcohol use group has a **significantly different** mean GPA from the others.*

2) Preparation

- Variables
  - Independent: AlcoholUse (categorical, 4 levels)
  - Dependent: GPA (continuous)
- ANOVA Conditions
  - **Normality**: $n > 30$ (Checked normality for GPA in question 9)
  - **Independence:** Assume random sampling
  - **Homogeneity/ Equality of Variances**

```
> sd_by_gpa = tapply(data$GPA, data$AlcoholUse, sd)
> sd_by_gpa
  Abstain     Heavy     Light  Moderate
0.4793122 0.4367589 0.3849260 0.3888629
```

```
> # Standard Deviation Condition for Equality of Variances
> # Is one std dev >= three times another gp Std Dev?
>
> sds = c(0.4793122, 0.4367589, 0.3849260, 0.3888629)
> vector <- sds
> result <- FALSE
>
> # Loop through each element in the vector
> for (i in 1:length(vector)) {
+   for (j in 1:length(vector)) {
+     if (i != j && vector[i] > 2 * vector[j]) {
+       result <- TRUE
+       break
+     }
+   }
+   if (result) break
+ }
>
> # Print the result: We want to see FALSE
> print(result)
[1] FALSE
```

To check the assumption of homogeneity of variances, we verified that no group's standard deviation exceeded **twice** that of another group. Since this condition was met (FALSE result from the test), we assume **equal variances** and proceed with ANOVA.

3) ANOVA Test

A **one-way ANOVA** is conducted to compare the mean GPA among the alcohol use groups.

```
> # Perform ANOVA
> one.way2 = aov(data$GPA ~ data$AlcoholUse)
> summary(one.way2)
                 Df Sum Sq Mean Sq F value Pr(>F)
data$AlcoholUse   3   0.60  0.2004    1.23  0.299
Residuals       249  40.59  0.1630
```

ANOVA Test Interpretation

- p-value (0.299) > alpha (0.05)
- At **α=0.05**, we **fail to reject** the **null hypothesis** and can conclude that there is **no significant difference** in mean GPA across alcohol use groups.