

AMOD-5210H: Foundations of Modelling

Jasmeet Singh Saini

2023-03-02

After Loading required packages and then reading the excel file.

```
library(readxl)
dataset_excel <- read_excel("ass3data.xlsx")
```

Let's performing data extraction.

```
set.seed(0758054)
index <- sample(1:nrow(dataset_excel), 200)
AMOD5210 <- dataset_excel[index, ]
```

Question 1

Report the frequencies for males and females in your subsample, as well as the mean, median, standard deviation, minimum and maximum values for the variable “age”.

```
table(AMOD5210$Gender)
```

```
##
## Female    Male
##      119      81
```

Hence, the *frequency* for males and females is **81** and **119** respectively.

The mean, median, standard deviation, minimum and maximum values for the variable “age” is given below,

```
#Minimum
min(AMOD5210$Age, na.rm = TRUE)
```

```
## [1] 19
```

```
#Maximum
max(AMOD5210$Age, na.rm = TRUE)
```

```
## [1] 79
```

The *minimum* and *maximum values* for the variable “age” is **19** and **79** respectively.

```
#Standard Deviation
sd(AMOD5210$Age, na.rm = TRUE)
```

```
## [1] 14.19257
```

```
#Mean
mean(AMOD5210$Age, na.rm = TRUE)
```

```
## [1] 46.41
```

```
# Median
median(AMOD5210$Age, na.rm = TRUE)
```

```
## [1] 46
```

The *Standard Deviation*, *Mean* and *Median* values for the variable “age” is **14.19257**, **46.41** and **46** respectively.

```
library(psych)
describe(AMOD5210$Age)
```

```
##      vars    n  mean    sd median trimmed  mad min max range skew kurtosis se
## X1      1 200 46.41 14.19     46   46.26 16.31  19  79    60 0.12   -0.73  1
```

Question 2

Are the continuous variables of “age”, “AG”, and “LTW” in your subsample normally distributed? If not, how would you describe these distributions and what could you do to make them more normal?

To check the continuous variables of “age”, “AG”, and “LTW” in the normally distributed, we will use Shapiro-Wilks test.

For “age”

Let’s test for “age”

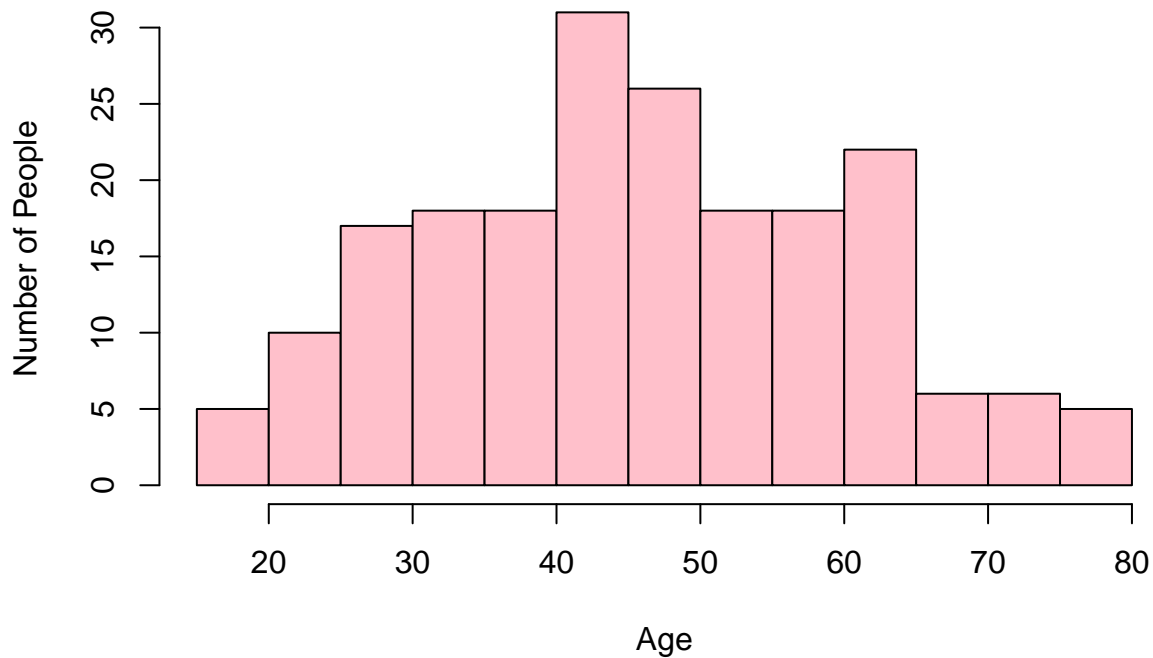
```
shapiro.test(AMOD5210$Age)

##
##  Shapiro-Wilk normality test
##
## data:  AMOD5210$Age
## W = 0.9844, p-value = 0.02596
```

This how a histogram looks like:

```
hist(AMOD5210$Age, xlab = "Age"
     , ylab = "Number of People"
     , main = "Age Distribution of First-time Gamblers"
     , prob = FALSE
     , col = "pink")
```

Age Distribution of First-time Gamblers



By Shapiro-Wilk normality test the p-value < 0.05 . Hence, it is not normally distributed.

For “AG”

Let's test for “AG” (continuous variable for age of 1st time gambling for money).

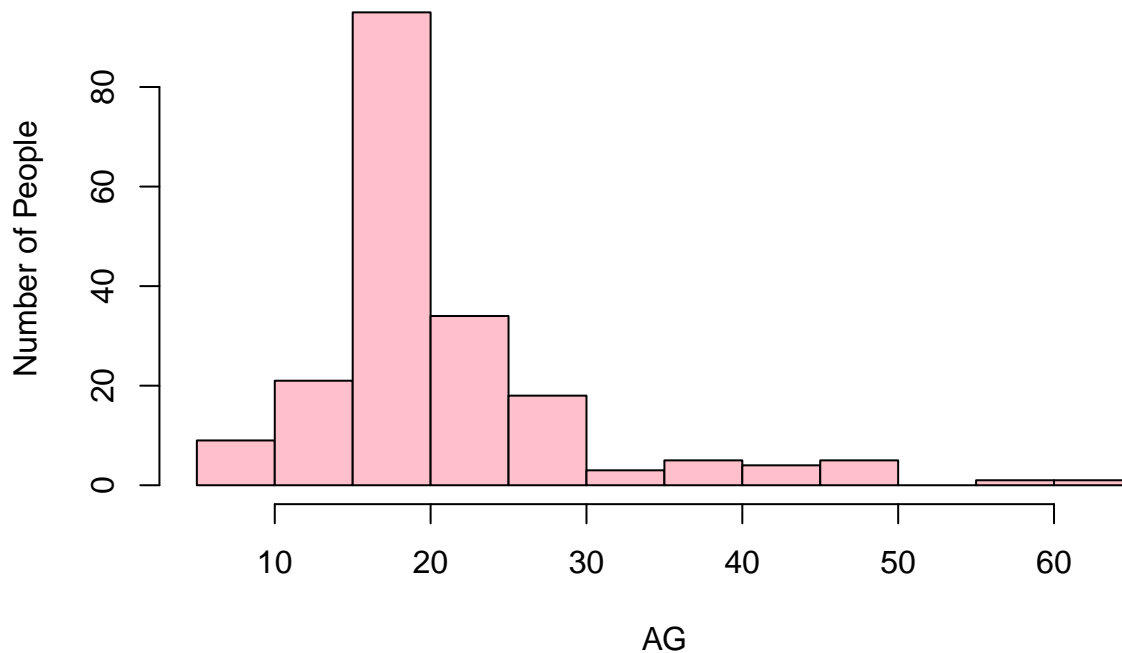
```
shapiro.test(AMOD5210$AG)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  AMOD5210$AG  
## W = 0.81998, p-value = 2.679e-14
```

This how a histogram looks like:

```
hist(AMOD5210$AG, xlab = "AG"  
     , ylab = "Number of People"  
     , main = "AG Distribution of First-time Gamblers"  
     , prob = FALSE  
     , col = "pink")
```

AG Distribution of First-time Gamblers



By Shapiro-Wilk normality test the p-value < 0.05 . Hence, it is not normally distributed.

For “LTW”

Finally, let's test for “LTW” (continuous variable for estimated lifetime winnings from gambling)

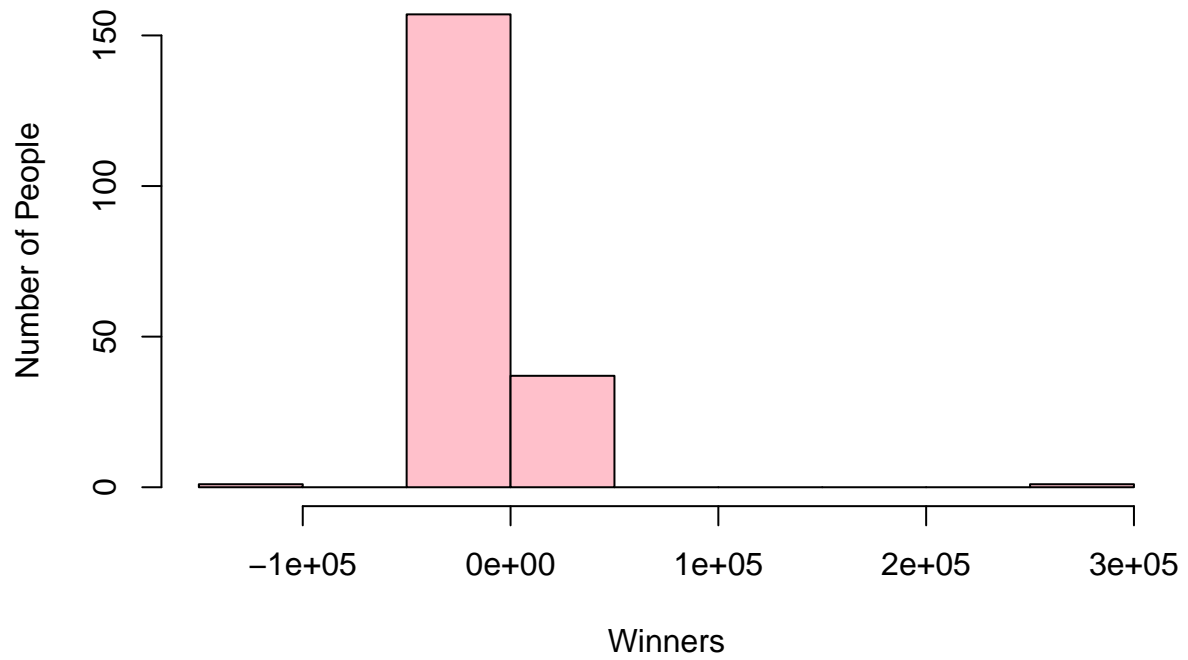
```
shapiro.test(AMOD5210$LTW)
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  AMOD5210$LTW  
## W = 0.21491, p-value < 2.2e-16
```

This how a histogram looks like:

```
hist(AMOD5210$LTW, xlab = "Winners"  
     , ylab = "Number of People"  
     , main = "Lifetime winnings distribution from gambling"  
     , prob = FALSE  
     , col = "pink")
```

Lifetime winnings distribution from gambling



By Shapiro-Wilk normality test the p-value < 0.05 . Hence, it is not normally distributed.

Question 3

Using an appropriate inferential statistic, determine whether males and females scored significantly different on any of the variables “AG”, “LTW”, and “gambled”. Also, evaluate and comment on whether the basic assumptions of your chosen statistic were met.

Inferential statistic for “AG” variable.

Step 1 : Hypothesis & assumptions

Let's define **Null Hypothesis** H_o and **Alternative Hypothesis** H_A to use inferential statistic.

H_o : There is no significant difference between the score of males and females on the variable “AG” (that is, age of first time gambling).

H_A : There is significant difference between the score of males and females on the variable “AG” (that is, age of first time gambling).

Let's organize the data by group and get some descriptive statistics:

```
library(rstatix)

##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##      filter
```

```
group_by_gender <- group_by(AMOD5210, Gender)
get_summary_stats(group_by_gender, AG, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Gender variable      n mean    sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 Female AG         115  23.4  9.55
## 2 Male   AG          81  19.6  8.64
```

Now, we need to test some **assumptions**. Firstly, let's check for extreme outliers.

```
identify_outliers(group_by_gender, AG)
```

```
## # A tibble: 15 x 11
##   Gender  ID  Age Income MS      AG  LTW Gambled Onset is.out~1 is.ex~2
##   <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <lgl>    <lgl>
## 1 Female 1180  61  55000 married 45  4000    50 Late TRUE    FALSE
## 2 Female  674  52  55000 married 48 -3000    60 Late TRUE    FALSE
## 3 Female 1816  50 175000 married 45  -500    200 Late TRUE    FALSE
## 4 Female 2140  59  35000 married 50 -1200    200 Late TRUE    FALSE
## 5 Female 2036  72  35000 divorced 65  -300    50 Late TRUE     TRUE
## 6 Female 3117  77  10000 divorced 50  -100    200 Late TRUE    FALSE
```

```
## 7 Female 842 66 55000 married 45 -2000 250 Late TRUE FALSE
## 8 Male 3597 67 35000 married 50 0 100 Late TRUE TRUE
## 9 Male 899 63 75000 married 45 -200 10 Late TRUE TRUE
## 10 Male 3455 41 75000 divorced 8 2500 400 Early TRUE FALSE
## 11 Male 2321 76 135000 married 50 -10 2 Late TRUE TRUE
## 12 Male 3680 46 75000 married 5 2000 0 Early TRUE FALSE
## 13 Male 2378 78 55000 married 60 0 100 Late TRUE TRUE
## 14 Male 1890 62 55000 married 30 -1000 500 Late TRUE FALSE
## 15 Male 2167 53 45000 married 27 -2300 7 Late TRUE FALSE
## # ... with abbreviated variable names 1: is.outlier, 2: is.extreme
```

Here, we get extreme outliers in the AG variable. But, we will perform the t-test.

Step 2 : Testing

Now, we will test for normality using **Shapiro-Wilks Test**.

```
# To test using Shapiro-Wilks
shapiro_test(group_by_gender, AG)
```

```
## # A tibble: 2 x 4
##   Gender variable statistic      p
##   <chr> <chr>      <dbl> <dbl>
## 1 Female AG          0.851 2.10e- 9
## 2 Male   AG          0.730 6.75e-11
```

Here, $p < 0.05$ for male as well as female. It not a normal distribution and we will use t-test, that is, Levene Test.

Now, to test for homogeneity of variance, we will use Levene Test.

```
# Test for homogeneity of variance(Levene Test)
levene_test(AMOD5210, AG ~ Gender)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic      p
##   <int> <int>      <dbl> <dbl>
## 1     1  194      1.97 0.162
```

Here, $p > 0.05$, Therefore, the variance of male and female is non-homogeneous.

Since, the condition or assumptions is not met during testing. Now, we will test using t-test to get the conclusion between the score of males and females on the variable “AG”.

Now, let’s run t-test for independent.

```
t_test(AG ~ Gender, data = AMOD5210, var.equal = TRUE)
```



```
## # A tibble: 1 x 8
##   .y.   group1 group2    n1    n2 statistic    df      p
## * <chr> <chr> <chr> <int> <int>    <dbl> <dbl>  <dbl>
## 1 AG     Female Male     115    81      2.88   194 0.00442
```

As $p < 0.05$, therefore, we have enough evidence to **reject** the **Null Hypothesis**, that is, there is a significantly difference between the score of males and females on the variable “AG”.

Step 3 : Conclusion

The current study sought to determine whether or not there is a significantly difference between the score of males and females on the variable “AG”. A 200 random samples were taken from a dataset of 3947 observation(81 males, 119 female).The sample contained few extreme outliers. A Shapiro-Wilks test didn’t demonstrated normality. Moreover, Levene’s test demonstrated non-heterogeneity of variance. The mean of “AG” for male was 19.580(SD = 8.637) and for female it was 23.417(SD = 9.550). An independent sample T-test showed that, $t(194) = 2.880084$, $p < 0.05$, concluding that the mean difference in “AG” between male and female in the sample was statistically significant.

–X–

Inferential statistic for “LTW” variable.

Step 1 : Hypothesis & assumptions

Let’s define **Null Hypothesis** H_o and **Alternative Hypothesis** H_A to use inferential statistic.

H_o : There is no significantly difference between the score of males and females on the variable “LTW” (that is, lifetime winnings from gambling).

H_A : There is significantly difference between the score of males and females on the variable “LTW” (that is, lifetime winnings from gambling).

Let’s organize the data by group and get some descriptive statistics:

```
library(rstatix)
get_summary_stats(group_by_gender,LTW,type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Gender variable      n mean    sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 Female LTW      115 -975. 3913.
## 2 Male   LTW       81 -612. 33558.
```

Now, we need to test some **assumptions**. Firstly, let’s check for extreme outliers.

```
identify_outliers(group_by_gender, LTW)
```

```
## # A tibble: 49 x 11
##   Gender  ID  Age Income MS      AG    LTW Gambled Onset is.out~1 is.ex~2
##   <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl>    <dbl> <chr> <lg1>    <lg1>
## 1 Female 3227  42 10000 married 27  -3000    50 Late  TRUE    FALSE
## 2 Female 1180  61 55000 married 45   4000    50 Late  TRUE    TRUE
```

```
## 3 Female 3464 34 45000 married 18 -5000 80 Early TRUE TRUE
## 4 Female 96 49 55000 married 16 5000 60 Early TRUE TRUE
## 5 Female 1176 63 55000 married 30 -2500 2 Late TRUE FALSE
## 6 Female 3682 43 65000 married 18 2000 0 Early TRUE FALSE
## 7 Female 3252 47 75000 married 33 -10000 300 Late TRUE TRUE
## 8 Female 2442 55 175000 married 30 -3000 150 Late TRUE FALSE
## 9 Female 2436 37 25000 single 10 3000 400 Early TRUE TRUE
## 10 Female 1522 37 55000 married 19 2000 0 Late TRUE FALSE
## # ... with 39 more rows, and abbreviated variable names 1: is.outlier,
## # 2: is.extreme
```

Here, we get extreme outliers in the LTW variable. But, we will perform the t-test.

Step 2 : Testing

Now, we will test for normality using **Shapiro-Wilks Test**.

```
# To test using Shapiro-Wilks
shapiro_test(group_by_gender, LTW)
```

```
## # A tibble: 2 x 4
##   Gender variable statistic      p
##   <chr> <chr>      <dbl> <dbl>
## 1 Female LTW      0.709 8.69e-14
## 2 Male   LTW      0.290 6.48e-18
```

Here, $p < 0.05$ for male as well as female. It is not a normal distribution and we will use t-test, that is, Levene Test.

Now, to test for homogeneity of variance, we will use Levene Test.

```
# Test for homogeneity of variance(Levene Test)
levene_test(AMOD5210, LTW ~ Gender)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic      p
##   <int> <int>      <dbl> <dbl>
## 1     1  194      4.51 0.0350
```

Here, $p < 0.05$, Therefore, the variance of male and female is homogeneous.

Since, the condition or assumptions is not met during testing. Now, we will test using t-test to get the conclusion between the score of males and females on the variable "LTW".

Now, let's run t-test for independent.

```
t_test(LTW ~ Gender, data = AMOD5210, var.equal = TRUE)
```

```
## # A tibble: 1 x 8
##   .y.   group1 group2    n1    n2 statistic    df    p
## * <chr> <chr> <chr> <int> <int>    <dbl> <dbl> <dbl>
## 1 LTW   Female Male     115    81    -0.115   194 0.908
```

As $p > 0.05$, therefore, we have enough evidence to **accept** the **Null Hypothesis**, that is, there is a no significantly difference between the score of males and females on the variable “LTW”.

Step 3 : Conclusion

abc

Question 4

Using an appropriate inferential statistic, determine whether marital status is significantly dependent on reporting an early or late onset of gambling (“Onset”)?

Step 1: Hypothesis

Let's define **Null Hypothesis** H_o and **Alternative Hypothesis** H_A to use inferential statistic.

H_o : Marital Status is NOT significantly dependent on reporting an early or late onset of gambling.

H_A : Marital Status is significantly dependent on reporting an early or late onset of gambling.

Step 2: Testing

To conduct the Test of Independence, that is, **Chi-Squared Test**, we need to build the table of frequency for Onset and MS:

```
frequency_table <- table(AMOD5210$Onset, AMOD5210$MS)
frequency_table
```

```
##
##      divorced married single
## Early      9      64     13
## Late     20      78     12
```

For Chi-Squared Test, we know

```
chisq.test(x = frequency_table, correct = FALSE)
```

```
##
## Pearson's Chi-squared test
##
## data: frequency_table
## X-squared = 2.6943, df = 2, p-value = 0.26
```

Since $p > 0.05$, we have enough evidence to accept **Null Hypothesis** H_o , that is, Marital status is not dependent on the Onset of gambling.

We know there is an effect, but we don't know where that effect is since we have a 2 x 3 contingency table. We need to perform a post-hoc test to know where the effect is

```
# First, let's install and load a useful package
#install.packages("chisq.posthoc.test")
library(chisq.posthoc.test)
# Now, let's run a chi-square post-hoc test
chisq.posthoc.test(frequency_table)
```

```
## Dimension      Value divorced married single
## 1 Early Residuals -1.509900  0.5457321  0.8761884
## 2 Early p values  0.786414  1.0000000  1.0000000
## 3 Late Residuals  1.509900 -0.5457321 -0.8761884
## 4 Late p values  0.786414  1.0000000  1.0000000
```

Step 3: Conclusion

The present research seeks to determine whether marital status is significantly dependent on reporting an early or late onset of gambling. A 200 sample of (81 Males, 119 Female) were taken and then divided based on marital status: Single ($N = 26$), Married ($N = 144$) and Divorced ($N = 30$). A Chi-square Test of Independence revealed that the marital status is independent on reporting an early or late onset of gambling, $X^2(2, N = 200) = 2.6943, p > 0.05$. A post-hoc test of proportions concluded that singles reported an early onset of gambling than married or divorced.

Question 5

What are the correlations (reported to 3 decimals) for the following pairs of variables: “age” and “LTW”; “age” and “gambled”; and “AG” and “LTW”. Report the p-values for each correlation. For each of the relevant correlations, what is the slope and intercept when “LTW” is the Y variable (i.e., dependent variable)? One of the key assumptions when interpreting a correlation is that the x and y variables are linearly related. Do you think this assumption is met for each of the 3 correlations?

Answer

Question 6

Using an appropriate inferential statistic, determine whether an individual's income level differs across married, single, and divorced individuals ("MS"). Also, evaluate and comment on whether the basic assumptions of your chosen statistic were met.

To test whether an individual's income level differs across married, single, and divorced individuals ("MS"), we will test using **Independent ANOVA Test**.

Step 1: Hypothesis and Assumptions

Let's define **Null Hypothesis** H_o and **Alternative Hypothesis** H_A to use inferential statistic.

H_o : An individual's income level doesn't differs across married, single, and divorced individuals ("MS").

H_A : An individual's income level differs across married, single, and divorced individuals ("MS").

Let's organize the data by group and get some descriptive statistics.

```
# install.packages("datarium")
# install.packages("rstatix")
library(rstatix)
library(datarium)
ms_group <- group_by(AMOD5210, MS)
get_summary_stats(ms_group, Income, type = "mean_sd")
```

```
## # A tibble: 3 x 5
##   MS      variable      n  mean    sd
##   <chr>    <fct>    <dbl> <dbl> <dbl>
## 1 divorced Income      30 35333. 20634.
## 2 married  Income     144 68958. 35389.
## 3 single   Income      26 38462. 34257.
```

Let's test some **assumptions**.

Firstly, we will also look for extreme outlier.

```
identify_outliers(ms_group, Income)
```

```
## # A tibble: 7 x 11
##   MS      ID  Age Income Gender  AG  LTW Gambled Onset is.out~1 is.ex~2
##   <chr>    <dbl> <dbl>  <dbl> <chr>  <dbl> <dbl>    <dbl> <chr>  <lgl>    <lgl>
## 1 divorced 3408   43  85000 Male    20 -30000    20 Late  TRUE    FALSE
## 2 divorced 3479   42  85000 Female  19  8000     10 Late  TRUE    FALSE
## 3 married  2442   55 175000 Female  30 -3000    150 Late  TRUE    FALSE
## 4 married  3068   48 175000 Female  24  -100      5 Late  TRUE    FALSE
## 5 married  1816   50 175000 Female  45  -500    200 Late  TRUE    FALSE
## 6 married  1136   71 175000 Male    20 -5000     20 Late  TRUE    FALSE
## 7 married  3183   44 175000 Female  20  -100     10 Late  TRUE    FALSE
## # ... with abbreviated variable names 1: is.outlier, 2: is.extreme
```

Since, we have two outlier. But, We also need to test the normality assumption using Shapiro-Wilks Test.

```
shapiro_test(ms_group, Income)
```

```
## # A tibble: 3 x 4
##   MS      variable statistic      p
##   <chr>   <chr>      <dbl>    <dbl>
## 1 divorced Income      0.890 0.00479
## 2 married  Income      0.929 0.00000141
## 3 single   Income      0.802 0.000189
```

As $p < 0.05$, the data is not normally distributed for any of the Marital status, that is, divorced, married, single.

Step 2: Testing

Finally, we need to test for homogeneity of variance.

```
levene_test(AMOD5210, Income ~ MS)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##   df1 df2 statistic      p
##   <int> <int>      <dbl> <dbl>
## 1     2   197      3.22 0.0421
```

As $p < 0.05$, thus the distribution shows that homogeneity in variance across marital status is not followed.

Now, let's test two-way independent ANOVA test and view the ANOVA summary table:

```
Ind.ANOVA <- aov(Income ~ MS, ms_group)
Anova(Ind.ANOVA, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: Income
##              Sum Sq Df F value    Pr(>F)
## (Intercept) 3.7453e+10  1  33.419 2.872e-08 ***
## MS          4.1871e+10  2  18.680 3.726e-08 ***
## Residuals   2.2078e+11 197
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As $p < 0.05$, there is enough evidence to reject **reject Null Hypothesis** (H_o). Thus, the income level differs across different categories of Marital Status (MS).

Also, to determine the difference in income levels across categories of MS, we will use Post-hoc test.


```
t_test(AMOD5210, Income ~ MS, var.equal = TRUE, p.adjust.method = "bonferroni")
```

```
## # A tibble: 3 x 10
##   .y.    group1    group2     n1     n2 statistic    df          p    p.adj p.adj~1
## * <chr> <chr>    <chr>  <int> <int>    <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 Income divorced married    30    144    -5.02    172 0.00000127 3.81e-6 ****
## 2 Income divorced single     30     26    -0.420    54 0.676      1 e+0 ns
## 3 Income married  single    144     26     4.06    168 0.0000742 2.23e-4 ***
## # ... with abbreviated variable name 1: p.adj.signif
```

Step 3: Conclusion

The current study determines whether or not the income level differs across married, single, and divorced individuals ("MS"). A 200 random samples taken from the dataset and examined (39 Divorced, 141 Married, 20 Single). The sample contained 2 outliers. A Shapiro-wilks test demonstrated that the distribution across married, single, and divorced individuals ("MS") was not normal. The mean income for divorced was 35333.33 (SD = 20633.64), the mean income for married was 68958.33 (SD = 35389.32), the mean income for single was 38461.54 (SD = 34256.95). The two-way Independent test showed that the Income level differs across married, single, and divorced individuals ("MS"), $F(2, 197) = 18.680$, $p < 0.05$. Bonferroni-corrected pairwise comparisons showed that the divorced cases of MS had more income level than other two, that is, single and married, while married has significantly greater income level than single.