

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”

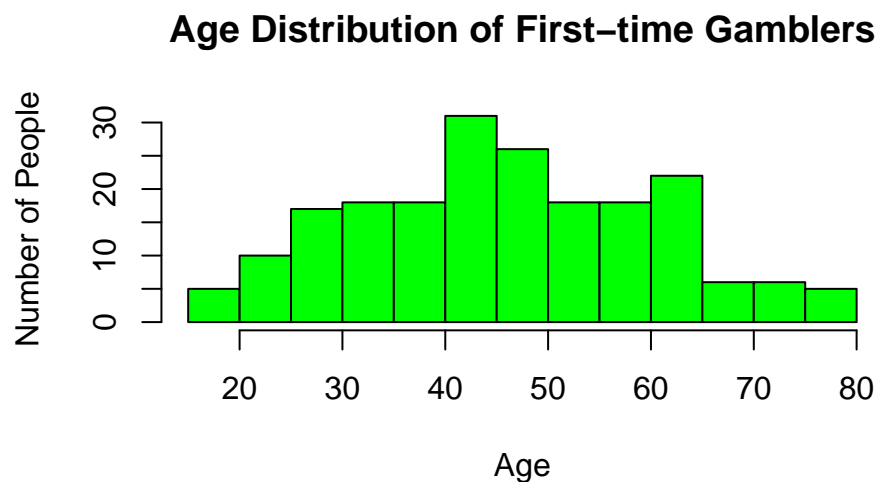
Let’s test for “age”

```
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 = "green")
```



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

For “AG”

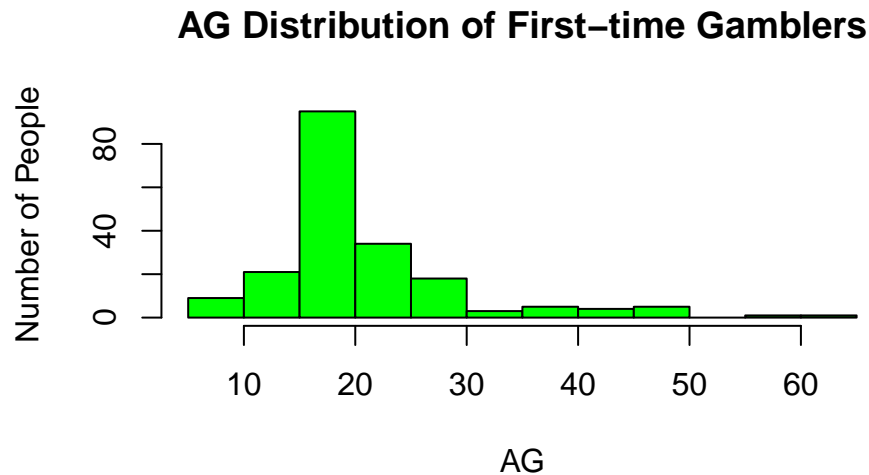
Let’s test for “AG” (continuous variable for age of 1st time gambling for money).

```
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 = "green")
```



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

For “LTW”

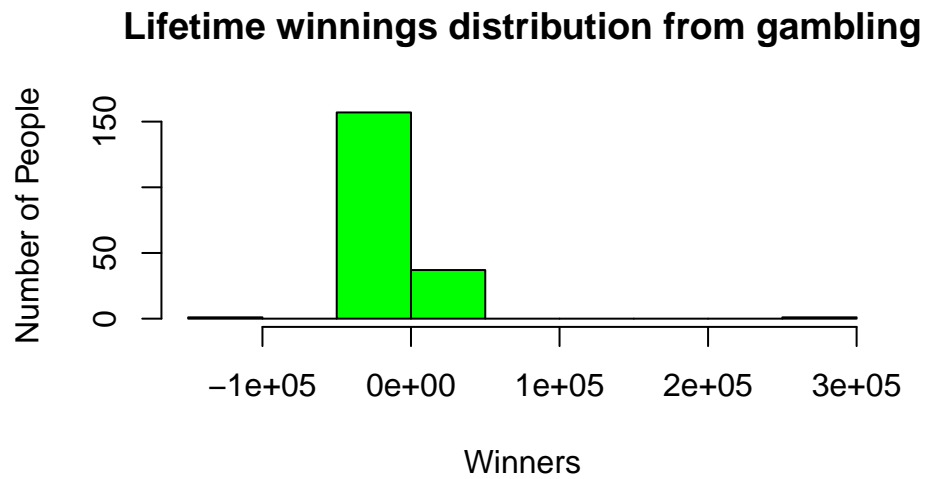
Finally, let’s test for “LTW” (continuous variable for estimated lifetime winnings from gambling)

```
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 = "green")
```



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** (H_0), 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 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 "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** (H_0), that is, there is a no significantly difference between the score of males and females on the variable “LTW”.

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 “Gambled”. 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 heterogeneity of variance. The mean of “Gambled” for male was -611.790(SD = 33558.163) and for female it was -975.278(SD = 3912.717). An independent sample T-test showed that, $t(194) = -0.1151715$, $p > 0.05$, concluding that the mean difference in “Gambled” between male and female in the sample was not statistically significant.

–X–

Inferential statistic for “Gambled” 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 “Gambled” (that is, amount of money gambled in past 12 months).

H_A : There is significantly difference between the score of males and females on the variable “Gambled” (that is, amount of money gambled in past 12 months).

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

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

```
## # A tibble: 2 x 5
##   Gender variable      n mean    sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 Female Gambled    119  68.6  110.
## 2 Male   Gambled     81  84.1  121.
```

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

```
identify_outliers(group_by_gender, Gambled)
```

```
## # A tibble: 15 x 11
##   Gender   ID   Age Income MS      AG   LTW Gambled Onset is.ou~1 is.ex~2
##   <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <lg1> <lg1>
## 1 Female 3252   47  75000 married 33 -1 e4    300 Late TRUE FALSE
## 2 Female 2436   37  25000 single  10 3 e3    400 Early TRUE  TRUE
```

```
## 3 Female 2295 42 45000 divorced 16 -1 e3 250 Early TRUE FALSE
## 4 Female 2715 47 135000 married 19 -1 e4 700 Late TRUE TRUE
## 5 Female 539 41 85000 married 16 6.5e3 425 Early TRUE TRUE
## 6 Female 842 66 55000 married 45 -2 e3 250 Late TRUE FALSE
## 7 Female 1409 64 35000 married 20 1 e3 500 Late TRUE TRUE
## 8 Male 2997 56 110000 married 16 5 e3 300 Early TRUE FALSE
## 9 Male 3455 41 75000 divorced 8 2.5e3 400 Early TRUE TRUE
## 10 Male 1964 51 55000 married 22 -1.5e4 300 Late TRUE FALSE
## 11 Male 2391 40 85000 married 13 -1 e4 300 Early TRUE FALSE
## 12 Male 1890 62 55000 married 30 -1 e3 500 Late TRUE TRUE
## 13 Male 2633 67 35000 married 18 -1.5e5 450 Early TRUE TRUE
## 14 Male 17 48 80000 married 19 -1.5e4 500 Late TRUE TRUE
## 15 Male 426 44 85000 married 20 -3 e3 400 Late TRUE TRUE
## # ... with abbreviated variable names 1: is.outlier, 2: is.extreme
```

Here, we get extreme outliers in the Gambled 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, Gambled)
```

```
## # A tibble: 2 x 4
##   Gender variable statistic      p
##   <chr> <chr>      <dbl> <dbl>
## 1 Female Gambled      0.648 1.77e-15
## 2 Male   Gambled      0.710 2.40e-11
```

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, Gambled ~ 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  198      0.833 0.363
```

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 “Gambled”.

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

```
t_test(Gambled ~ 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 Gambled Female Male    119    81    -0.944    198 0.347
```

As $p > 0.05$, therefore, we have enough evidence to **accept** the **Null Hypothesis** (H_0), that is, there is a no significantly difference between the score of males and females on the variable “Gambled”.

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 “Gambled”. 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 heterogeneity of variance. The mean of “Gambled” for male was 84.123(SD = 120.928) and for female it was 68.588(SD = 109.573). An independent sample T-test showed that, $t(198) = -0.9435949$, $p > 0.05$, concluding that the mean difference in “Gambled” between male and female in the sample was not statistically significant.

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, $X^2(2, N = 200) = 2.6943, p > 0.05$.

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?

Let’s check the statistics of the dataset:

```
library(psych)
describe(AMOD5210, fast = TRUE)

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf
## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

##      vars   n   mean    sd   min   max range    se
## ID         1 200 2029.27 1084.54    17  3935  3918  76.69
## Age        2 200   46.41  14.19    19    79    60   1.00
## Income     3 200 59950.00 36329.69 10000 175000 165000 2568.90
## MS         4 200      NaN     NA    Inf  -Inf  -Inf     NA
## Gender     5 200      NaN     NA    Inf  -Inf  -Inf     NA
## AG         6 196   21.83   9.35     5    65    60   0.67
## LTW        7 196 -825.06 21702.37 -150000 254000 404000 1550.17
## Gambled    8 200   74.88  114.26     0    700   700   8.08
## Onset      9 196      NaN     NA    Inf  -Inf  -Inf     NA
```

We will now check **assumption** on all variables, before performing *correlation test*.

Let’s check for the outliers for all four variables, that is, “LTW”, “AG”, “Age”, “Gambled” :

```
identify_outliers(AMOD5210, LTW)

## # A tibble: 49 x 11
##       ID   Age Income MS      Gender   AG   LTW Gambled Onset is.out~1 is.ex~2
##   <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <lgl> <lgl>
## 1   349   65  45000 married Male    17  5000    10 Early TRUE  FALSE
## 2  1180   61  55000 married Female  45  4000    50 Late  TRUE  FALSE
## 3  2152   60  45000 married Male    16 -10000   200 Early TRUE   TRUE
## 4  3464   34  45000 married Female  18 -5000    80 Early TRUE  FALSE
```

```
## 5 2997 56 110000 married Male 16 5000 300 Early TRUE FALSE
## 6 96 49 55000 married Female 16 5000 60 Early TRUE FALSE
## 7 3252 47 75000 married Female 33 -10000 300 Late TRUE TRUE
## 8 1377 60 85000 married Male 16 -5000 150 Early TRUE FALSE
## 9 2436 37 25000 single Female 10 3000 400 Early TRUE FALSE
## 10 589 29 75000 single Male 19 5000 200 Late TRUE FALSE
## # ... with 39 more rows, and abbreviated variable names 1: is.outlier,
## # 2: is.extreme
```

```
identify_outliers(AMOD5210, AG)
```

```
## # A tibble: 16 x 11
##      ID   Age Income MS      Gender   AG   LTW Gambled Onset is.out~1 is.ex~2
##    <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <lgl> <lgl>
## 1 3597   67 35000 married Male    50    0   100 Late TRUE FALSE
## 2 1180   61 55000 married Female 45 4000   50 Late TRUE FALSE
## 3 899    63 75000 married Male    45 -200   10 Late TRUE FALSE
## 4 2683   72 10000 divorced Female 40 -200    0 Late TRUE FALSE
## 5 2321   76 135000 married Male    50 -10    2 Late TRUE FALSE
## 6 3454   65 35000 married Female 40 50    0 Late TRUE FALSE
## 7 674    52 55000 married Female 48 -3000 60 Late TRUE FALSE
## 8 1816   50 175000 married Female 45 -500 200 Late TRUE FALSE
## 9 2378   78 55000 married Male    60 0   100 Late TRUE TRUE
## 10 2140   59 35000 married Female 50 -1200 200 Late TRUE FALSE
## 11 2118   52 110000 married Female 39 -5000 200 Late TRUE FALSE
## 12 2036   72 35000 divorced Female 65 -300 50 Late TRUE TRUE
## 13 2310   60 10000 divorced Female 40 -2000 75 Late TRUE FALSE
## 14 2623   64 110000 married Female 40 -100 0 Late TRUE FALSE
## 15 3117   77 10000 divorced Female 50 -100 200 Late TRUE FALSE
## 16 842    66 55000 married Female 45 -2000 250 Late TRUE FALSE
## # ... with abbreviated variable names 1: is.outlier, 2: is.extreme
```

```
identify_outliers(AMOD5210, Age)
```

```
## [1] ID      Age      Income    MS      Gender    AG
## [7] LTW      Gambled  Onset    is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

```
identify_outliers(AMOD5210, Gambled)
```

```
## # A tibble: 15 x 11
##      ID   Age Income MS      Gender   AG   LTW Gambled Onset is.ou~1 is.ex~2
##    <dbl> <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <lgl> <lgl>
## 1 2997   56 110000 married Male    16 5   e3   300 Early TRUE FALSE
## 2 3252   47 75000 married Female 33 -1   e4   300 Late TRUE FALSE
## 3 3455   41 75000 divorced Male    8 2.5e3 400 Early TRUE TRUE
## 4 2436   37 25000 single Female 10 3   e3   400 Early TRUE TRUE
## 5 1964   51 55000 married Male    22 -1.5e4 300 Late TRUE FALSE
## 6 2295   42 45000 divorced Female 16 -1   e3   250 Early TRUE FALSE
## 7 2391   40 85000 married Male    13 -1   e4   300 Early TRUE FALSE
## 8 1890   62 55000 married Male    30 -1   e3   500 Late TRUE TRUE
## 9 2633   67 35000 married Male    18 -1.5e5 450 Early TRUE TRUE
```



```
## 10 2715 47 135000 married Female 19 -1 e4 700 Late TRUE TRUE
## 11 539 41 85000 married Female 16 6.5e3 425 Early TRUE TRUE
## 12 17 48 80000 married Male 19 -1.5e4 500 Late TRUE TRUE
## 13 426 44 85000 married Male 20 -3 e3 400 Late TRUE TRUE
## 14 842 66 55000 married Female 45 -2 e3 250 Late TRUE FALSE
## 15 1409 64 35000 married Female 20 1 e3 500 Late TRUE TRUE
## # ... with abbreviated variable names 1: is.outlier, 2: is.extreme
```

All four variables have few outliers. Now, we will follow the next steps, that is, Shapiro-Wilks Test.

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

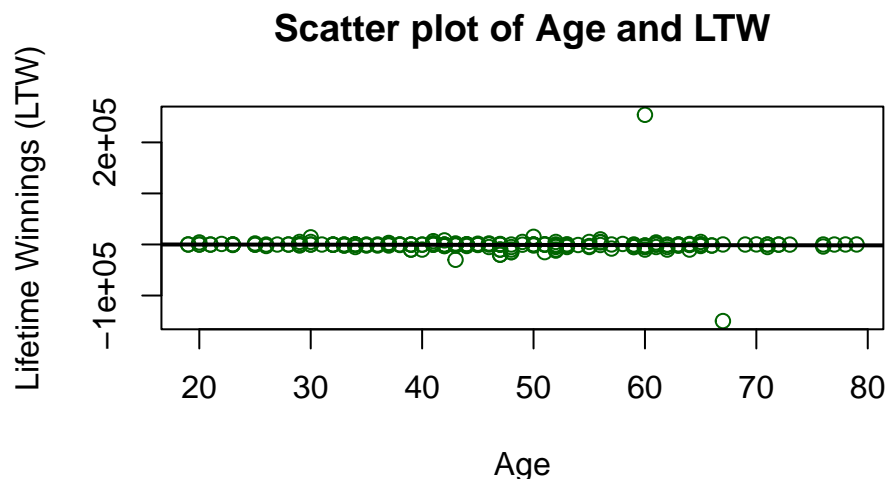
```
# To test using Shapiro-Wilks
shapiro_test(AMOD5210, vars = c("Age", "LTW", "Gambled", "AG"))
```

```
## # A tibble: 4 x 3
##   variable statistic      p
##   <chr>      <dbl>    <dbl>
## 1 AG          0.820 2.68e-14
## 2 Age         0.984 2.60e- 2
## 3 Gambled     0.679 2.68e-19
## 4 LTW         0.215 5.28e-28
```

All the four variables in Shapiro-Wilks test has $p < 0.05$. It seem that all the four variables are not Normal Distribution. We will now use Spearman's Rho Correlation method, that is, non-parametric correlation test.

Now, let's check the linearity in all the four variables, that is, "LTW", "AG", "Age", "Gambled":

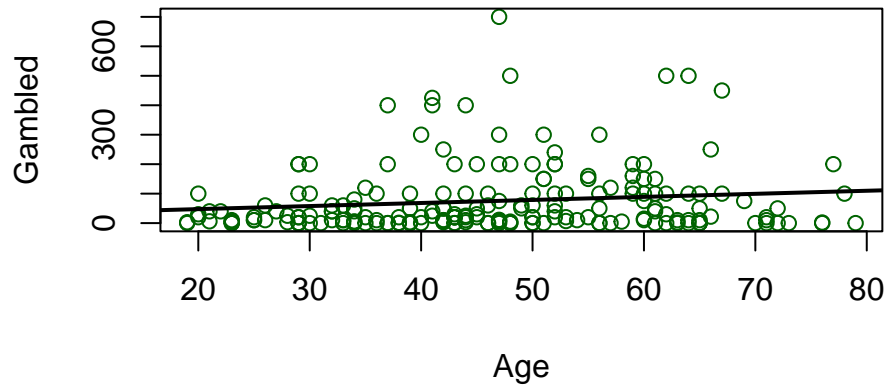
```
# For "Age" and "LTW"
plot(AMOD5210$Age, AMOD5210$LTW, col = "darkgreen",
     xlab = "Age", ylab = "Lifetime Winnings (LTW)", main = "Scatter plot of Age and LTW")
abline(lm(AMOD5210$LTW ~ AMOD5210$Age, data = AMOD5210), col = "black", lw = 2)
```



The scatter plot for "Age" and "LTW" shows that, as the persons age *increases*, the Lifetime Winnings (LTW) *does not changes* too much. Therefore, the regression line, which was seen in the above plot is almost *constant* and does not show **linearity**.

```
# For Age and Gambled
plot(AMOD5210$Age, AMOD5210$Gambled, col = "darkgreen",
     xlab = "Age", ylab = "Gambled" , main = "Scatter plot of Age and Gambled")
abline(lm(AMOD5210$Gambled ~ AMOD5210$Age, data = AMOD5210), col = "black", lw = 2)
```

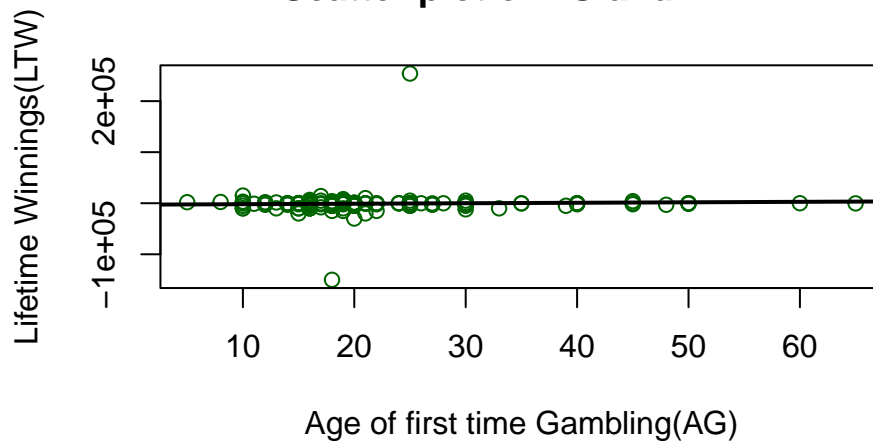
Scatter plot of Age and Gambled



Here, the scatter plot for “Age” and “Gambled” shows that, as the persons age *increases*, the Gambled *changes* very little. Therefore, the regression line, which was seen in the above plot is almost *constant* and does not show **linearity**.

```
# For AG and LTW
plot(AMOD5210$AG, AMOD5210$LTW, col = "darkgreen",
     xlab = "Age of first time Gambling(AG)", ylab = "Lifetime Winnings(LTW)" , main = "Scatter plot of AG and LTW")
abline(lm(AMOD5210$LTW ~ AMOD5210$AG, data = AMOD5210), col = "black", lw = 2)
```

Scatter plot of AG and LTW



Here, the scatter plot for “AG” and “LTW” shows that, as Age of first time gambling(AG) *increases*, the Lifetime Winnings(LTW) *does not changes* too much. Therefore, the regression line, which was seen in the above plot is almost *constant* and does not show **linearity**.

Hence, none of them the correlation shows the condition of Linearity.

- “Age” and “LTW” Now lets perform correlation test

```
correlation_age_ltw <- cor.test(AMOD5210$Age, AMOD5210$LTW, method = "spearman")
```

```
## Warning in cor.test.default(AMOD5210$Age, AMOD5210$LTW, method = "spearman"):  
## Cannot compute exact p-value with ties
```

```
correlation_age_ltw
```

```
##  
## Spearman's rank correlation rho  
##  
## data: AMOD5210$Age and AMOD5210$LTW  
## S = 1418468, p-value = 0.0686  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho  
## -0.1303528
```

```
# Reporting to 3 decimals points  
round(correlation_age_ltw$estimate,3)
```

```
## rho  
## -0.13
```

Therefore, the *Spearman's rank correlation rho* between “Age” and “LTW” is **-0.13**.

```
lm(AMOD5210$Age ~ AMOD5210$LTW,data = AMOD5210)
```

```
##  
## Call:  
## lm(formula = AMOD5210$Age ~ AMOD5210$LTW, data = AMOD5210)  
##  
## Coefficients:  
## (Intercept) AMOD5210$LTW  
## 4.641e+01 -1.146e-05
```

“LTW” is the dependent variable in the correlation test, therefore, Slope is $-1.146e - 05$ and Y-Intercept is $4.641e + 01$.

- “Age” and “Gambled”

```
correlation_age_gambled <- cor.test(AMOD5210$Age, AMOD5210$Gambled, method = "spearman")
```

```
## Warning in cor.test.default(AMOD5210$Age, AMOD5210$Gambled, method =  
## "spearman"): Cannot compute exact p-value with ties
```

```
correlation_age_gambled
```

```
##  
## Spearman's rank correlation rho  
##  
## data: AMOD5210$Age and AMOD5210$Gambled  
## S = 1184794, p-value = 0.1164  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho  
## 0.1113825
```

```
# Reporting to 3 decimals points  
round(correlation_age_gambled$estimate,3)
```

```
## rho  
## 0.111
```

Therefore, the *Spearman's rank correlation rho* between “Age” and “Gambled” is **0.111**.

- “AG” and “LTW”

```
correlation_ag_ltw <- cor.test(AMOD5210$AG, AMOD5210$LTW, method = "spearman")
```

```
## Warning in cor.test.default(AMOD5210$AG, AMOD5210$LTW, method = "spearman"):  
## Cannot compute exact p-value with ties
```

```
correlation_ag_ltw
```

```
##  
## Spearman's rank correlation rho  
##  
## data: AMOD5210$AG and AMOD5210$LTW  
## S = 1285645, p-value = 0.7331  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho  
## -0.02450817
```

```
# Reporting to 3 decimals points  
round(correlation_ag_ltw$estimate,3)
```

```
## rho  
## -0.025
```

Therefore, the *Spearman's rank correlation rho* between “AG” and “LTW” is **-0.025**.

```
lm(AMOD5210$AG ~ AMOD5210$LTW, data = AMOD5210)
```

```
##  
## Call:  
## lm(formula = AMOD5210$AG ~ AMOD5210$LTW, data = AMOD5210)  
##  
## Coefficients:  
## (Intercept)  AMOD5210$LTW  
##    2.185e+01    1.725e-05
```

“LTW” is the dependent variable in the correlation test, therefore, Slope is $1.725e - 05$ and Y-Intercept is $2.185e + 01$.

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 are going to proceed with the normality test, 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 $p < 0.05$, the distribution shows that non-homogeneity in variance across marital status and thus it is not followed.

Now, lets 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** *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 = FALSE, 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    -7.03   70.1    1.1 e-9  3.3 e-9 ****
## 2 Income divorced single     30     26    -0.406  39.8    6.87e-1  1    e+0 ns
## 3 Income married  single    144     26     4.16   35.3    1.95e-4  5.85e-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 normally distributed and Levene’s test shows the non-homogeneity in variance. 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). An 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 single category had significantly higher income level than both the divorced and married categories, while the married category had significantly higher income level than the divorced category.