Problem Set 12.Rmd

2023-12-09

```
library(ggplot2)
```

1.

```
sis <- c(69, 64, 65, 63, 65, 62, 65, 64, 66, 59, 62)
bro <- c(71, 68, 66, 67, 70, 71, 70, 73, 72, 65, 66)
```

(a)

```
corr <- cor(sis, bro)
R_SQ <- corr^2
# The sample coefficient of determination is
R_SQ
```

```
## [1] 0.3114251
```

(b)

```
Test_Corr <- cor.test(sis, bro)
Test_Corr
```

```
##
##  Pearson's product-moment correlation
##
## data:  sis and bro
## t = 2.0175, df = 9, p-value = 0.07442
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06286527  0.86751705
## sample estimates:
##       cor
## 0.5580547
```

The p-value is not less than alpha. So, we fail to reject the null hypothesis that a sister's height alone is sufficient to predict the brother's height

(c)

```
fit <- lm(bro ~ sis)
ConfidenceInterval <- confint(fit, level = 0.9)
ConfidenceInterval
```

```
##                     5 %      95 %
## (Intercept) -3.20446954 65.568106
## sis          0.05401643  1.127802
```

Therefore, the 0.9 level confidence interval for the slope of the population regression line for predicting y from x is 0.054 to 1.127.

2.

(a)

The null hypothesis (H0): There is no significant difference between the average anxiety levels for male and female students The alternate hypothesis (H1): There is a significant difference between the average anxiety levels for male and female students

```
df <- read.table("examanxiety.txt", sep = "\t", header = TRUE)
df
```

```
##      Code Revise Exam Anxiety Gender
## 1       1      4   40  86.298   Male
## 2       2     11   65  88.716 Female
## 3       3     27   80  70.178   Male
## 4       4     53   80  61.312   Male
## 5       5      4   40  89.522   Male
## 6       6     22   70  60.506 Female
## 7       7     16   20  81.462 Female
## 8       8     21   55  75.820 Female
## 9       9     25   50  69.372 Female
## 10     10     18   40  82.268 Female
## 11     11     18   45  79.044   Male
## 12     12     16   85  80.656   Male
## 13     13     13   70  70.178   Male
## 14     14     18   50  75.014 Female
## 15     15     98   95  34.714   Male
## 16     16      1   70  95.164   Male
## 17     17     14   95  75.820   Male
## 18     18     29   95  79.044 Female
## 19     19      4   50  91.134 Female
## 20     20     23   60  64.536   Male
## 21     21     14   80  80.656   Male
## 22     22     12   75  77.432   Male
## 23     23     22   85  65.342 Female
## 24     24     84   90  56.116 Female
## 25     25     23   30  71.790 Female
## 26     26     26   60  81.462 Female
## 27     27     24   75  63.730   Male
## 28     28     72   75  27.460 Female
## 29     29     37   27  73.402 Female
## 30     30     10   20  89.522   Male
## 31     31      3   75  89.522 Female
## 32     32     36   90  75.014 Female
## 33     33     43   60  43.580   Male
## 34     34     19   30  82.268   Male
## 35     35     12   80  79.044   Male
## 36     36      9   10  79.044 Female
## 37     37     72   85  37.132   Male
## 38     38     10    7  81.462   Male
## 39     39     12    5  83.074 Female
## 40     40     30   85  50.834   Male
## 41     41     15   20  82.268   Male
## 42     42      8   45  78.238 Female
## 43     43     34   60  72.596   Male
## 44     44     22   70  74.208 Female
## 45     45     21   50  75.820 Female
## 46     46     27   25  70.984   Male
## 47     47      6   50  97.582   Male
## 48     48     18   40  67.760   Male
## 49     49      8   80  75.014   Male
## 50     50     19   50  73.402 Female
## 51     51      0   35  93.552 Female
## 52     52     52   80  58.894 Female
## 53     53     38   50  53.252 Female
## 54     54     19   49  84.686   Male
```

```
## 55    55     23   75  89.522 Female
## 56    56     11   25  71.790 Female
## 57    57     27   65  82.268   Male
## 58    58     17   80  69.372   Male
## 59    59     13   50  62.118   Male
## 60    60     42   70  68.566 Female
## 61    61      4   40  93.552   Male
## 62    62      8   80  84.686 Female
## 63    63      6   10  82.268   Male
## 64    64     11   20  81.462 Female
## 65    65      7   40  82.268   Male
## 66    66     15   40  91.134   Male
## 67    67      4   70  91.940 Female
## 68    68     28   52  86.298 Female
## 69    69     22   50  72.596   Male
## 70    70     29   60  63.730 Female
## 71    71      2   80  63.730   Male
## 72    72     16   60  71.790 Female
## 73    73     59   65  57.282   Male
## 74    74     10   15  84.686 Female
## 75    75     13   85  84.686   Male
## 76    76      8   20  77.432 Female
## 77    77      5   80  82.268 Female
## 78    78      2  100  10.000   Male
## 79    79     38  100  50.834 Female
## 80    80      4   80  87.910   Male
## 81    81     10   10  83.880   Male
## 82    82      6   70  84.686 Female
## 83    83     68  100  20.206 Female
## 84    84      8   70  87.104   Male
## 85    85      1   70  83.880 Female
## 86    86     14   65  67.760   Male
## 87    87     42   75  95.970 Female
## 88    88     13   85  62.118 Female
## 89    89      1   30  84.686   Male
## 90    90      3    5  92.746   Male
## 91    91      5   10  84.686 Female
## 92    92     12   90  83.074 Female
## 93    93     19   70  73.402   Male
## 94    94      2   20  87.910 Female
## 95    95     19   85  71.790   Male
## 96    96     11   35  86.298   Male
## 97    97     15   30  84.686 Female
## 98    98     23   70  75.820   Male
## 99    99     13   55  70.984 Female
## 100  100     14   75  78.238 Female
## 101  101      1    2  82.268   Male
## 102  102      9   40  79.044   Male
## 103  103     20   50  91.134 Female
```

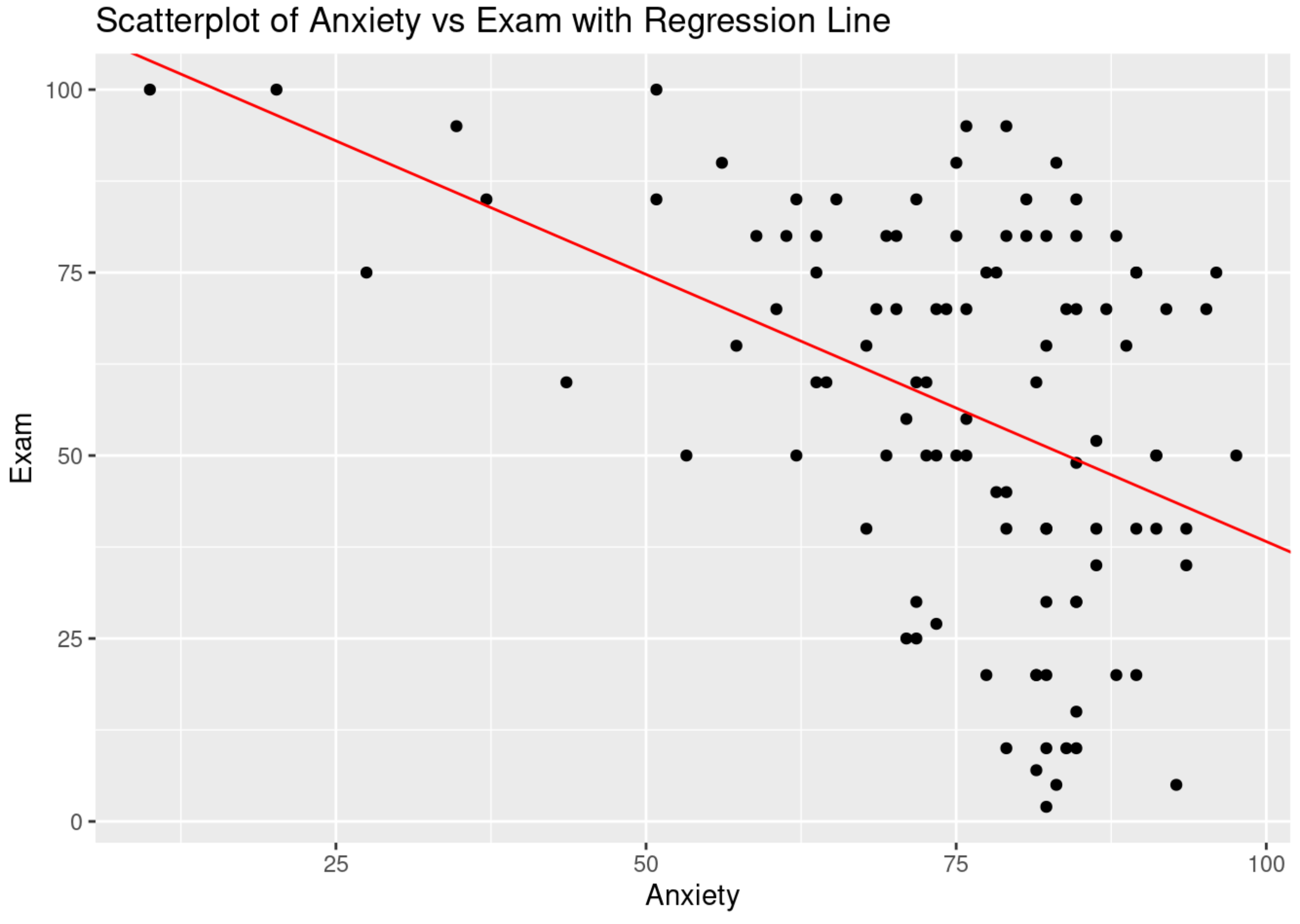
```
male <- df$Anxiety[df$Gender == "Male"]
female <- df$Anxiety[df$Gender == "Female"]
# two-sample t-test
t_test <- t.test(male, female)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  male and female
## t = -0.32961, df = 100.41, p-value = 0.7424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.147444  5.110827
## sample estimates:
## mean of x mean of y
##  74.38373  75.40204
```

The p-value is found to be 0.7424 which means we failed to reject the null hypothesis meaning there is no significant difference. There is no significant difference between the average anxiety levels for male and female students

(b)

```
fit1 <- lm(Exam ~ Anxiety, df)
slope <- coef(fit1)["Anxiety"]
intercept <- coef(fit1)["(Intercept)"]
ggplot(df, aes(x = Anxiety, y = Exam)) +
  geom_point() +
  geom_abline(slope = slope, intercept = intercept, color = "red") +
  labs(x = "Anxiety", y = "Exam") +
  ggtitle("Scatterplot of Anxiety vs Exam with Regression Line")
```



```
fit1
```

```
##
## Call:
## lm(formula = Exam ~ Anxiety, data = df)
##
## Coefficients:
## (Intercept)      Anxiety
##      111.24        -0.73
```

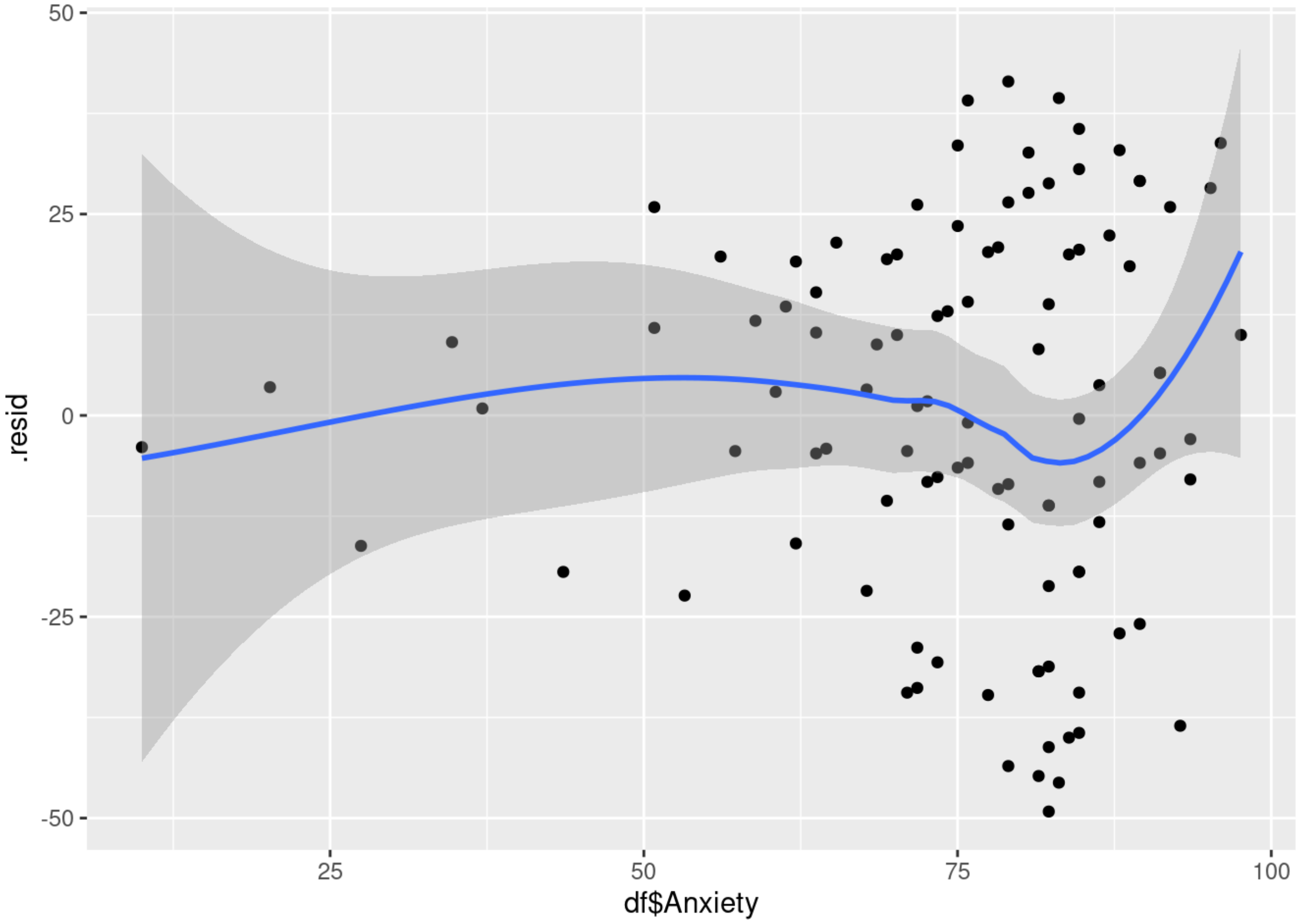
What our regression line hear means that's the data of exam vs anxiety has a pattern following with a negative slope. So with this specific linear model and data, we can infer that the with high anxiety levels, exam scores are generally lower. But, the line doesn't exactly fit it so well because the points are a little scattered. So it's hard to say we can be 100% certain about this fit.

(c)

(i)

```
library(broom)
fit1.aug <- augment(fit1)
ggplot(fit1.aug, aes(df$Anxiety, .resid)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Generally, in such a plot, if the smooth line is close to a horizontal line, we can say there is a linearity. But in our problem we don't see that. So we can't say that there is linearity

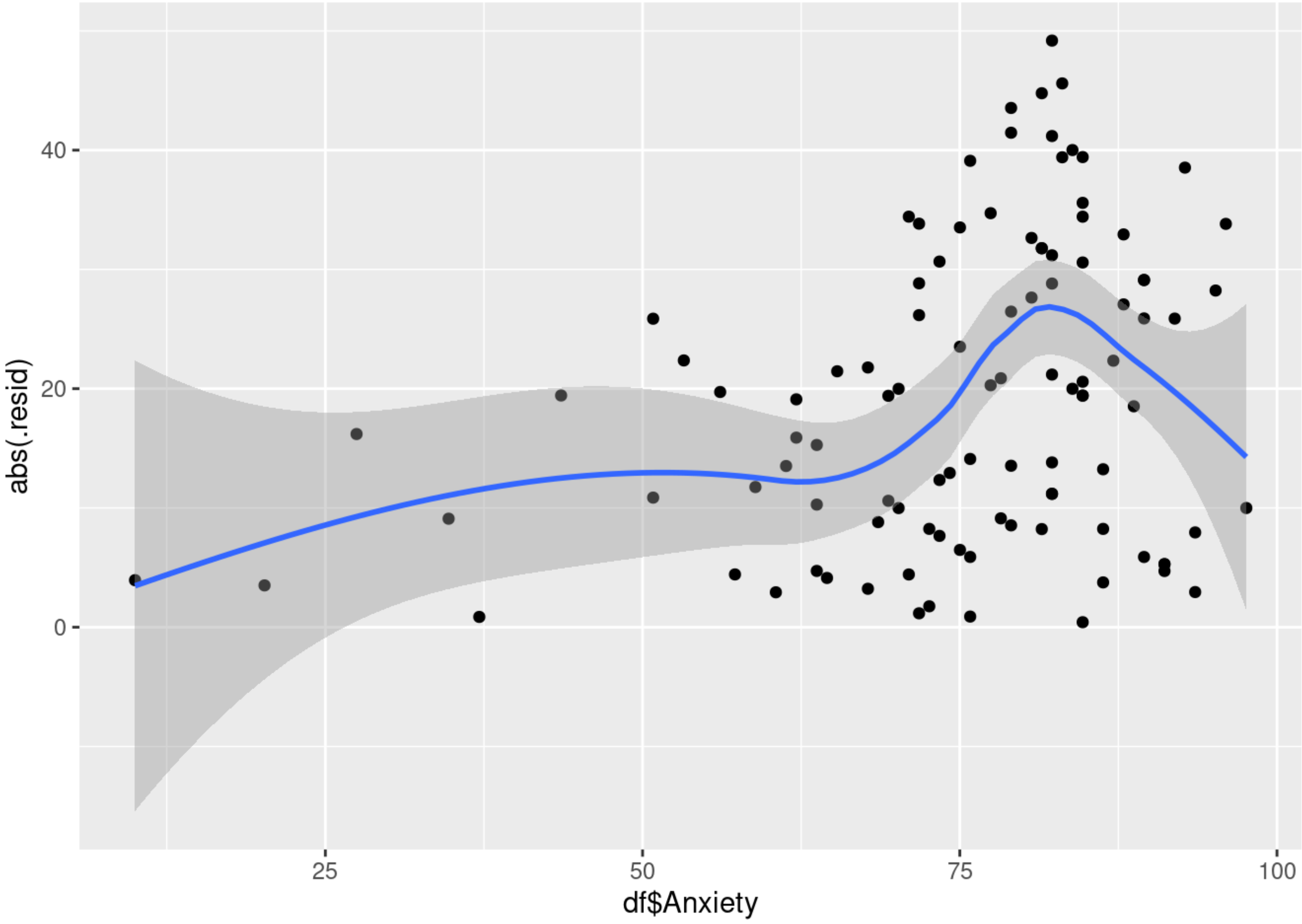
(ii)

For independence in our regression, because our model was fitted using data which is a random sample from a larger population of students, we can say that our data follows an IID distribution. Therefore, we can say that there is independence.

(iii)

```
ggplot(fit1.aug, aes(df$Anxiety, abs(.resid))) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



For homoskedasticity, as long as the spread of the residuals changes as we go from left to right, we can say that there is equal variance of errors. Here it does look like the spread of the residuals changes as we go from left to right. So it does look like the data has homoskedasticity.

(iv)

3.

(a)

There are three assumptions to be made for the analysis of variance F-test: The observations are independent, All of the populations are normal, and Homoscedasticity.

Observations are indenpendent: All the rats are put into 4 cages by splitting them up randomly. So we can assume that the observations are independent.

All the populations are normal: When we look at the graphs, they are close to a straight line. So we can say that they are approximately normal.

Homoscedasticity: Generally with real data it’s hard to get standard deviations of all the sample very close to each other. Although Fruit has a higher standard deviation than the rest while the other are closer to each other, we can still say that there is homoscedasticity.

So the data here is approximately consistent with homoscedasticity.

(b)

```
M_fruit <-  83.5
SD_fruit <- 16.9
M_carb <- 92.3
SD_carb <- 14.6
M_meat <- 88.6
SD_meat <- 14.2
M_mixed <- 99.4
SD_mixed <- 14.1
total <- 140
sample <- 35

average <- mean(M_fruit, M_carb, M_meat, M_mixed)
std_dev <- c(SD_fruit, SD_carb, SD_meat, SD_mixed)
cal <- sample * (M_fruit - average)^2 + sample * (M_carb - average)^2 + sample * (M_m
eat - average)^2 +sample * (M_mixed - average)^2
BDF <- 3
B_MS <- cal/BDF
SS <- (sample-1) * SD_fruit^2 + (sample-1) * SD_carb^2 + (sample-1) * SD_meat^2 + (sa
mple-1) * SD_mixed^2
DF <- 136
within_MS <- SS/DF
F_value <- B_MS/within_MS
```

Therefore, we get a F value of 6.96, between mean square of 1566.25, within mean square of 224.805, Total 35272.23, Total DF 139, and a p-value of 0.00021

(c)

Well the p-value being very low we have to say that there is a difference in the means of weight gained the 4 different groups. All the means could be different or even 3 are same but 1 is different. But we could always still make use of more data at the end of the day.

4.

```
exp <- read.table("GameEmpathy.txt", sep = " " , header = TRUE)
exp
```

```
##        sex game.type identify  empathy
## 1   female   neutral 3.333333 5.285714
## 2   female   neutral 1.833333 5.571429
## 3     male   neutral 1.000000 4.714286
## 4   female   neutral 1.000000 5.571429
## 5   female   neutral 3.333333 3.142857
## 6   female   neutral 1.000000 5.571429
## 7     male   neutral 5.333333 3.000000
## 8   female   neutral 2.666667 5.285714
## 9   female   neutral 5.666667 5.000000
## 10  female   neutral 3.333333 3.857143
## 11  female   neutral 4.000000 6.000000
## 12    male   neutral 5.833333 3.714286
## 13  female   neutral 1.000000 5.000000
## 14  female   neutral 1.000000 5.142857
## 15  female   neutral 3.666667 4.714286
## 16  female   neutral 2.166667 4.571429
## 17    male       GTA 5.000000 5.142857
## 18  female       GTA 3.666667 6.428571
## 19    male       GTA 6.333333 3.857143
## 20    male       GTA 4.833333 3.714286
## 21    male       GTA 4.666667 6.285714
## 22    male       GTA 4.666667 5.571429
## 23  female       GTA 2.500000 4.285714
## 24    male  HalfLife 6.666667 3.285714
## 25    male  HalfLife 4.000000 5.571429
## 26    male  HalfLife 6.333333 5.428571
## 27  female  HalfLife 3.166667 3.571429
## 28  female  HalfLife 6.333333 5.857143
## 29    male  HalfLife 3.666667 6.000000
## 30    male  HalfLife 5.333333 2.571429
## 31    male  HalfLife 4.000000 4.571429
## 32    male  HalfLife 5.833333 6.000000
## 33  female  HalfLife 5.333333 5.428571
## 34  female  HalfLife 4.833333 4.428571
## 35  female  HalfLife 5.333333 4.428571
## 36    male  HalfLife 5.000000 5.428571
## 37  female  HalfLife 2.500000 7.000000
## 38    male  HalfLife 5.166667 6.166667
## 39  female  HalfLife 4.666667 5.142857
## 40    male  HalfLife 7.000000 6.000000
## 41  female  HalfLife 3.833333 4.571429
## 42    male   neutral 3.500000 3.714286
## 43    male   neutral 5.833333 7.000000
## 44    male   neutral 6.333333 6.857143
## 45    male   neutral 5.500000 4.857143
## 46    male   neutral 4.500000 4.285714
## 47  female   neutral 4.000000 5.857143
## 48    male   neutral 4.166667 4.428571
## 49    male   neutral 4.500000 5.285714
## 50    male   neutral 5.500000 5.000000
## 51    male   neutral 4.333333 5.714286
## 52    male   neutral 6.166667 6.428571
## 53    male   neutral 4.666667 4.142857
## 54    male   neutral 5.333333 5.142857
```

```
## 55    male   neutral 2.500000 5.428571
## 56  female   neutral 3.000000 4.285714
## 57  female   neutral 4.000000 4.833333
## 58    male   neutral 5.666667 6.285714
## 59    male   neutral 1.500000 6.428571
## 60  female   neutral 5.000000 6.285714
## 61  female   neutral 5.833333 6.428571
## 62  female   neutral 2.500000 5.142857
## 63  female   neutral 1.666667 3.714286
## 64    male   neutral 5.833333 3.142857
## 65  female   neutral 2.500000 5.285714
## 66  female   neutral 5.166667 4.857143
## 67  female  HalfLife 5.666667 5.000000
## 68    male  HalfLife 5.500000 6.285714
## 69  female  HalfLife 4.333333 5.571429
## 70  female  HalfLife 4.000000 6.857143
## 71    male  HalfLife 5.500000 5.571429
## 72  female  HalfLife 1.333333 4.714286
## 73    male  HalfLife 3.333333 5.714286
## 74    male  HalfLife 6.166667 5.142857
## 75  female  HalfLife 3.833333 5.428571
## 76  female  HalfLife 7.000000 6.428571
## 77    male  HalfLife 5.000000 5.000000
## 78  female  HalfLife 4.166667 5.285714
## 79    male  HalfLife 2.833333 4.000000
## 80  female  HalfLife 4.833333 6.285714
## 81  female  HalfLife 4.166667 5.428571
## 82    male  HalfLife 4.333333 5.571429
## 83  female  HalfLife 4.833333 5.285714
## 84  female  HalfLife 5.666667 5.142857
## 85  female  HalfLife 4.000000 5.714286
## 86  female  HalfLife 5.166667 6.285714
## 87    male  HalfLife 5.666667 7.000000
## 88  female  HalfLife 7.000000 5.285714
## 89  female  HalfLife 5.166667 6.428571
## 90    male  HalfLife 3.666667 3.857143
## 91    male  HalfLife 3.833333 5.285714
## 92    male  HalfLife 5.333333 5.857143
## 93    male  HalfLife 4.500000 5.571429
## 94  female  HalfLife 6.166667 4.571429
## 95    male  HalfLife 4.166667 4.000000
## 96  female  HalfLife 4.333333 4.428571
## 97  female       GTA 4.666667 5.857143
## 98    male       GTA 5.000000 3.428571
## 99    male       GTA 5.166667 5.428571
## 100   male       GTA 5.166667 2.285714
## 101   male       GTA 4.500000 5.714286
## 102   male       GTA 4.166667 4.571429
## 103   male       GTA 4.333333 5.000000
## 104   male       GTA 3.166667 5.428571
## 105   male       GTA 2.000000 5.857143
## 106   male       GTA 4.500000 4.428571
## 107 female   neutral 3.833333 5.428571
## 108 female   neutral 5.666667 4.428571
## 109   male   neutral 5.000000 4.714286
## 110 female   neutral 5.166667 6.571429
```

```
## 111   male   neutral 3.666667 4.285714
## 112 female   neutral 3.500000 5.714286
## 113 female       GTA 2.833333 6.428571
## 114 female       GTA 1.333333 6.571429
## 115 female       GTA 3.833333 4.714286
## 116 female       GTA 4.000000 5.714286
## 117 female       GTA 6.000000 4.714286
## 118 female   neutral 3.000000 4.428571
## 119 female   neutral 3.333333 5.142857
## 120 female   neutral 4.833333 5.000000
## 121 female  HalfLife 6.333333 5.571429
## 122 female       GTA 2.666667 4.571429
## 123 female       GTA 3.666667 5.714286
## 124 female       GTA 5.333333 6.285714
## 125 female       GTA 3.500000 2.571429
## 126 female       GTA 4.833333 6.428571
## 127 female       GTA 3.333333 5.857143
## 128 female       GTA 4.666667 4.428571
## 129 female       GTA 2.000000 5.285714
## 130 female       GTA 3.666667 3.571429
## 131 female       GTA 3.666667 5.142857
## 132 female       GTA 5.500000 5.571429
## 133 female       GTA 3.666667 4.142857
## 134 female       GTA 4.500000 5.142857
## 135 female       GTA 1.000000 6.714286
## 136 female       GTA 4.333333 5.857143
## 137   male       GTA 4.833333 5.714286
## 138   male       GTA 4.666667 4.285714
## 139   male       GTA 3.666667 5.571429
## 140 female       GTA 4.166667 5.857143
## 141 female       GTA 4.333333 4.000000
## 142   male       GTA 2.500000 3.000000
## 143   male       GTA 4.833333 5.142857
## 144 female  HalfLife 4.333333 6.000000
## 145   male  HalfLife 4.500000 5.571429
## 146   male  HalfLife 3.500000 4.857143
## 147 female  HalfLife 3.500000 6.142857
## 148 female  HalfLife 6.166667 3.714286
## 149   male       GTA 3.666667 5.285714
## 150 female  HalfLife 3.000000 4.857143
## 151   male       GTA 4.500000 6.142857
## 152   male       GTA 6.500000 1.714286
## 153 female       GTA 5.500000 6.000000
```

Here the null hypothesis (H0) will be that the mean empathies across the three games that people played is same (or) U0 = U1 = U2. The alternate hypothesis (H1) will be that the at least one mean is different.

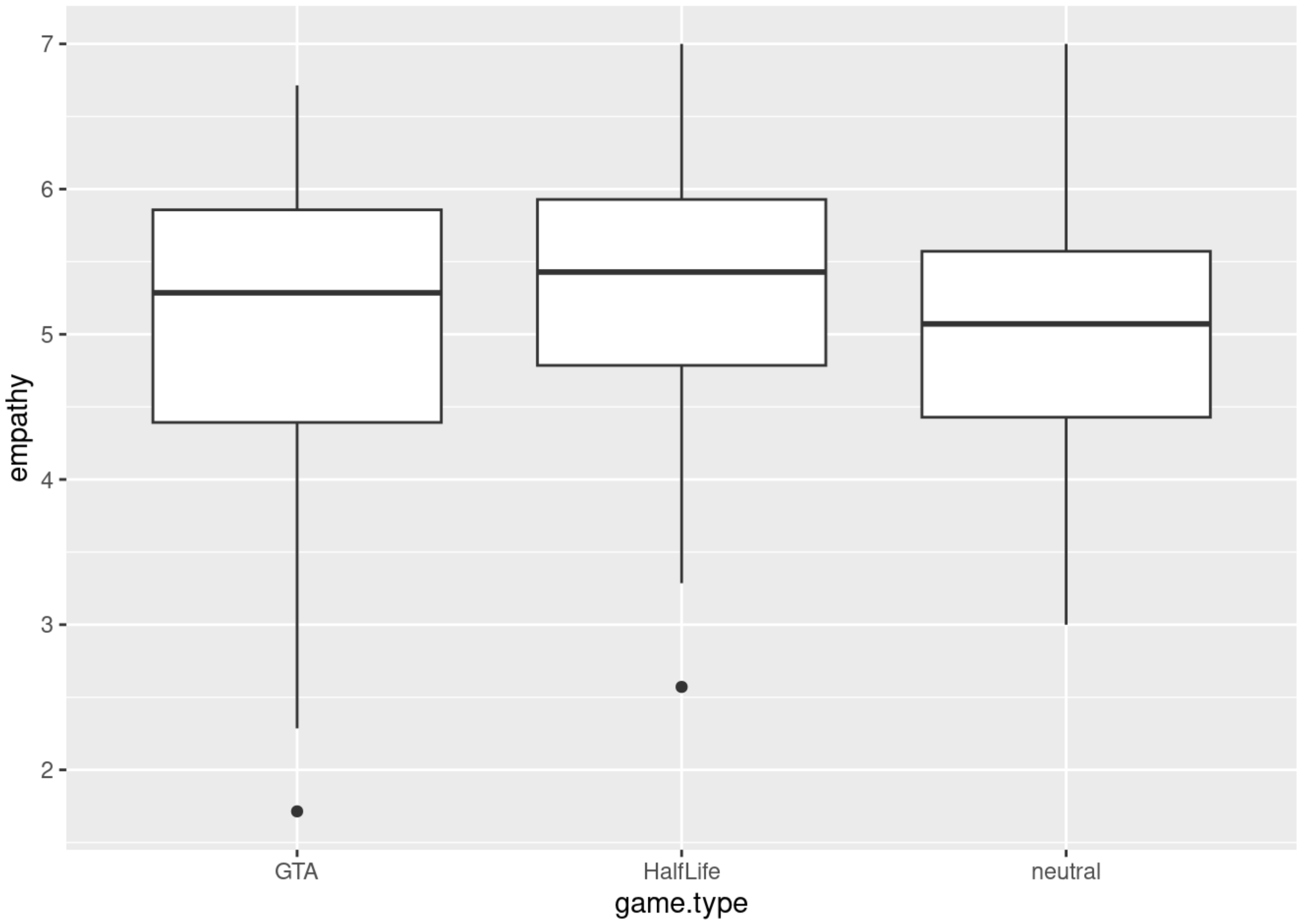
(a)

```
library(ggplot2)

neutral <- exp$empathy[exp$game.type == "neutral"]
HL <- exp$empathy[exp$game.type == "HalfLife"]
gta <- exp$empathy[exp$game.type == "GTA"]

neutral_iden <- exp$identify[exp$game.type == "neutral"]
HL_iden <- exp$identify[exp$game.type == "HalfLife"]
gta_iden <- exp$identify[exp$game.type == "GTA"]


ggplot(exp, aes(x = game.type, y = empathy)) + geom_boxplot()
```



Using just the graph it's hard to tell that these samples from populations with the same mean. We'll need to try a different method.

Let's calculate the anova:

```
anova <- aov(empathy ~ game.type, data = exp)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## game.type     2   2.25   1.125   1.092  0.338
## Residuals   150 154.47   1.030
```

The p-value isn't small, so we cannot reject the null hypothesis which means the empathies across all the three games that people played is the same.

(b)

(i) (ii) (iii)

Null hypothesis (H0): There is no relationship between identification and empathy for student playing one of the particular video game Alternate hypothesis (H1): There is a relationship

I should be writing the Null and Alternate Hypothesis individually for every sub-question but for brevity I'm writing it once.

```
# We have already subset the empathy and identity data video game wise
# Now let's find the correlation between them
neutral_corr <- cor.test(neutral_iden, neutral)
HL_corr <- cor.test(HL_iden, HL)
gta_corr <- cor.test(gta_iden, gta)

# Now lets find their p-values and adjust for multiple testing
p_val <- c(neutral_corr$p.value, HL_corr$p.value, gta_corr$p.value)
p_adjust <- p.adjust(p_val, method = "bonferroni")
p_adjust
```

```
## [1] 1.0000000 1.0000000 0.1835403
```

The adjusted p-values for neutral and Half-Life games indicate that their null hypothesis couldn't be rejected. Although the p-value for gta is comparatively lower, the p-value is not low enough to reject the null. It is quite close to 0.05, so potentially with a more data or analysis, we may find clear relationship. But, for now, we have still failed to reject the null in statistical context.