AMOD-5210H: Foundations of Modelling

Jasmeet Singh Saini - 0758054

2023-03-19

PART 1: EFFECT SIZES

The following questions requires the use of **"healthdata.xlsx"**.

Firstly, loading the required packages and then reading the excel file.

```
health_dataset <- read_excel("health-data.xlsx")
```

Now, let's performing data extraction.

```
set.seed(0758054)
index <- sample(1:nrow(health_dataset),200)
AMOD5210_Part1 <- health_dataset[index, ]
```

Part 1: Question 1

Using an appropriate inferential statistic and effect size, determine whether there is a significant difference between students and non-students on "Health" and "Depress".

For the variable "Health":

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{There is no difference in "health" variable for students and non-students.}$$
$$\text{vs.}$$
$$H_A : \text{There is a difference in "health" variable for students and non-students.}$$

Let's check the head of Health dataset

```
head(AMOD5210_Part1, 3)
```

```
## # A tibble: 3 x 10
##      ID Gender Student Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <dbl> <chr>  <chr>     <dbl>  <dbl>     <dbl>     <dbl>  <dbl>   <dbl> <chr>
## 1    48 Female No           20     19        19        10     12       1 No
## 2   576 Female Yes          21     11        17         8     25      10 No
## 3   525 Female No           21     18        19        19     15       4 No
## # ... with abbreviated variable name 1: Regulation
```

Now, let's grouping with students and checking the summary.

```
grouping_student <- group_by(AMOD5210_Part1, Student)
get_summary_stats(grouping_student, Health, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Student variable     n  mean    sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 No      Health     161  16.1  5.02
## 2 Yes     Health      39  18.5  5.09
```

Now, we need to test some **assumptions** about our data.

```
identify_outliers(grouping_student, Health)
```

```
## # A tibble: 1 x 12
##   Student    ID Gender Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <chr>   <dbl> <chr>    <dbl>  <dbl>     <dbl>     <dbl>  <dbl>   <dbl> <chr>
## 1 No        354 Female      23     17        18        10      3       0 No
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## #   abbreviated variable name 1: Regulation
```

Here, We get no outliers. We will now test for **normality** from the health-data. For that, we will use the **Shapiro-Wilks Test**. If $p > 0.05$, the data is normal.

```
shapiro_test(grouping_student, Health)
```

```
## # A tibble: 2 x 4
##   Student variable statistic     p
##   <chr>   <chr>        <dbl> <dbl>
## 1 No      Health       0.987 0.158
## 2 Yes     Health       0.969 0.347
```

Now, we need to test for **homogeneity of variance**. We can use the **Levene's Test** for this. If $p > 0.05$, variances are homogeneous.

```
levene_test(AMOD5210_Part1, Health ~ Student)
```

```
## 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     0.347 0.557
```

Step 2 : Testing

Now, we will run **Independent t-test**. Since, the homogeneity of variance assumption was not violated, we will set var.equal to TRUE.

```r
t_test(AMOD5210_Part1, Health ~ Student, 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 Health No     Yes      161    39     -2.62   198 0.00959
```

Since, $p < 0.05$, the test shows a significant difference. We have enough evidence to reject the null hypothesis, H_O.

Calculating Cohen's d

Also, the t-Test is significant, we need to calculate **Cohen's d** for our effect size. We need to specify "paired = FALSE" to indicate the groups are independent, and specify "pooled_sd = TRUE" to indicate the variances are equal.

```r
cohens_d(Health ~ Student, data = AMOD5210_Part1, paired = FALSE, pooled_sd = TRUE)
```

```
## Cohen's d |         95% CI
## --------------------------
## -0.47     | [-0.82, -0.11]
##
## - Estimated using pooled SD.
```

Based on Cohen's (1988) conventions we have a small effect.

Step 3 : Conclusion

The current study sought to determine whether or not there is a significant difference between students and non-students on "Health". 200 study participants were randomly sampled from the general public (39 students, 161 non-students). The sample contained no extreme outliers. A Shapiro-Wilks test demonstrated normality by group, and Levene's test demonstrated homogeneity of variance. The mean "Health" variable of non-students in the sample was 16.112 (SD = 5.020) whereas the mean "Health" variable of the students in the sample was 18.462 (SD = 5.088). A Welch's independent t-test showed that the mean difference in "Health" variable between student and non-students in the sample was statistically significant, $t(198) = -2.615876$, $p < 0.05$, $d = -0.47$, with students tending to be more "Healthy" than non-students. According to Cohen's (1988) conventions, this is a small effect.

For the variable "Depress":

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{There is no difference in "depress" variable for students and non-students.}$$
$$\text{vs.}$$
$$H_A : \text{There is a difference in "depress" variable for students and non-students.}$$

```r
get_summary_stats(grouping_student, Depress, type="mean_sd")
```

```
## # A tibble: 2 x 5
##   Student variable     n  mean    sd
##   <chr>   <fct>    <dbl> <dbl> <dbl>
## 1 No      Depress    161  4.83  4.44
## 2 Yes     Depress     39  6.97  5.06
```

Now, we need to test some **assumptions** about our data.

```
identify_outliers(grouping_student, Depress)
```

```
## # A tibble: 7 x 12
##   Student    ID Gender Honesty Leader Persevere Regulat~1 Health Depress Dstatus
##   <chr>   <dbl> <chr>    <dbl>  <dbl>     <dbl>     <dbl>  <dbl>   <dbl> <chr>
## 1 No        249 Female      13     22        10         9     21      20 Yes
## 2 No         96 Male        20     16        17        10     24      21 Yes
## 3 No         92 Female      19     18        21        12     18      17 Yes
## 4 Yes       760 Female      21     15        17        18     29      17 Yes
## 5 Yes       498 Female      24     16        10         6     21      18 Yes
## 6 Yes       557 Female      21     18        15        17     20      18 Yes
## 7 Yes       186 Female      19     13        12        15     26      18 Yes
## # ... with 2 more variables: is.outlier <lgl>, is.extreme <lgl>, and
## #   abbreviated variable name 1: Regulation
```

Here, We get no outliers. We will now test for **normality** from the health-data. For that, we will use the **Shapiro-Wilks Test**. If $p > 0.05$, the data is normal.

```
shapiro_test(grouping_student, Depress)
```

```
## # A tibble: 2 x 4
##   Student variable statistic             p
##   <chr>   <chr>        <dbl>         <dbl>
## 1 No      Depress      0.894 0.00000000241
## 2 Yes     Depress      0.884 0.000807
```

It can be seen that our data is not normal. We are still going to proceed with the test.

Now, we need to test for **homogeneity of variance**. We can use the **Levene's Test** for this. If $p > 0.05$, variances are homogeneous.

```
levene_test(AMOD5210_Part1, Depress ~ Student)
```

```
## 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     0.158 0.691
```

Step 2 : Testing

Now, we will run **Independent t-test**. Since, the homogeneity of variance assumption was not violated, we will set var.equal to TRUE.

```
t_test(AMOD5210_Part1, Depress ~ Student, 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 Depress No     Yes      161    39     -2.63   198 0.00908
```

Since, $p < 0.05$, the test shows a significant difference. We have enough evidence to reject the null hypothesis, H_O.

Calculating Cohen's d

Also, the t-Test is significant, we need to calculate **Cohen's d** for our effect size. We need to specify "paired = FALSE" to indicate the groups are independent, and specify "pooled_sd = TRUE" to indicate the variances are equal.

```
cohens_d(Depress ~ Student, data = AMOD5210_Part1, paired = FALSE, pooled_sd = TRUE)
```

```
## Cohen's d |         95% CI
## --------------------------
## -0.47     | [-0.82, -0.12]
##
## - Estimated using pooled SD.
```

Based on Cohen's (1988) conventions we have a small effect.

Step 3 : Conclusion

The current study sought to determine whether or not there is a significant difference between students and non-students on variable "Depress". 200 study participants were randomly sampled from the general public (39 students, 161 non-students). The sample contained no extreme outliers. A Shapiro-Wilks test demonstrated normality by group, and Levene's test demonstrated homogeneity of variance. The mean "Depress" variable of non-students in the sample was 4.826 (SD = 4.445) whereas the mean "Depress" variable of the students in the sample was 6.974 (SD = 5.055). A Welch's independent t-test showed that the mean difference in "Depress" variable between student and non-students in the sample was statistically significant, $t(198) = -2.634914$, $p < 0.05$, $d = -0.47$, with students tending to be more "Depress" than non-students. According to Cohen's (1988) conventions, this is a small effect.

Part 1: Question 2

Using an appropriate inferential statistic and effect size, determine whether there is a significant difference in the proportion of men and women diagnosed with or without depression.

We are going to use a $\chi 2$ Test of Independence for this question.

Step 1 : Hypothesis & Assumptions

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

H_0 : There is no difference in the proportion of men and women diagnosed with or without depression.

vs.

H_A : There is a significant difference in the proportion of men and women diagnosed with or without depression.

Below is the frequency table based on the continuous variable Gender and Dstatus:

```
frequency_table <- table(AMOD5210_Part1$Gender, AMOD5210_Part1$Dstatus)
frequency_table
```

```
##
##           No Yes
##   Female 137  14
##   Male    47   2
```

Step 2 : Testing

Let us now perform the test, $\chi 2$ **Test of Independence**.

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

```
## Warning in chisq.test(x = frequency_table, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  frequency_table
## X-squared = 1.3539, df = 1, p-value = 0.2446
```

```
chisq.posthoc.test(frequency_table)
```

```
## Warning in chisq.test(x, ...): Chi-squared approximation may be incorrect
```

```
##   Dimension     Value        No       Yes
## 1    Female Residuals -1.163565  1.163565
## 2    Female  p values  0.978401  0.978401
## 3      Male Residuals  1.163565 -1.163565
## 4      Male  p values  0.978401  0.978401
```

Since $p > 0.05$ we fail to reject the null hypothesis, H_O.

Now, need to calculate the odds ratio as our effect size.

```r
oddsratio(frequency_table)
```

```
## Odds ratio |       95% CI
## -------------------------
## 0.42       | [0.09, 1.90]
```

Based on Cohens (1988) conventions, we have less than the small category.

Step 3 : Conclusion

The present research seeks to determine whether there is a significant difference in the proportion of men and women diagnosed with or without depression. 200 people (49 Male, 151 Female) reported if they diagnosed with (16) or without depression (184). A Chi-square Test of Independence revealed that there is no difference in the proportion of men and women diagnosed with or without depression, Chi Squared(1, N = 200) = 1.3539, $p > 0.001$, OR = 0.42. According to Cohen's (1988) conventions, this effect was small.

Part 1: Question 3

Researchers are interested to determine whether character strengths are significant predictors of depression symptoms.

a) Using the Pearson's r correlation and r^2, determine whether there are significant correlations between depression symptoms and the four character strengths variables ("Honesty", "Leader", "Persevere", "Regulation"). Report the r and pvalues for each correlation. Also, report r^2 for each correlation.

b) Using multiple linear regression, determine whether the four character strengths variables are significant predictors of depression symptoms. Report the slopes and p-values for each character strength and identify which character strengths were significant predictors. Also report and interpret the multiple R^2 for the overall model.

Part 1: Question 4

Using an appropriate inferential statistic and effect size(s), determine whether participants had significantly different scores across the four character strengths ("honesty", "Leader", "Persevere", "Regulation").

PART 2: DIAGNOSTIC EFFICIENCY STATISTICS

The following questions requires the use of **"diagnostic-data.xlsx"**.

Firstly, loading the required packages and then reading the excel file.

```
library(readxl)
diagnostic_dataset <- read_excel("diagnostic-data.xlsx")
```

Now, let's performing data extraction.

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

```
## # A tibble: 200 x 3
##          ID Diagnosis Test
##       <dbl> <chr>     <chr>
##  1  776715 No        No
##  2 1246562 Yes       Yes
##  3   76389 Yes       Yes
##  4  846832 No        No
##  5 1080233 Yes       No
##  6  704097 No        No
##  7 1344449 No        Yes
##  8  608157 Yes       Yes
##  9 1167439 Yes       No
## 10 1299161 Yes       Yes
## # ... with 190 more rows
```

Part 2: Question 1

Create a 2 x 2 contingency table for the variables Diagnosis and Test. The contingency table you create should include frequencies within each cell, and each row and column of the table should be meaningfully labelled.

Part 2: Question 2

Report the following diagnostic efficiency statistics: a) sensitivity, b) specificity, c) positive prediction value, d) negative prediction value, e) overall correct classification, and f) Kappa

Part 2: Question 3

Based on the diagnostic efficiency statistics reported in Question 2, does the new test accurately diagnose individuals with breast cancer? Explain your answer.