UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI 2143

PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER 2 (SESSION 2019/2020)

PROJECT 2 : INFERENTIAL STATISTICS

PROJECT TITLE : CHILDREN STATISTICAL IN CERTAIN CASES

NAME : NURSYAHIDATUL ASYIQIN BINTI YUSOF

MATRIC NO : A19EC0140

SECTION : 10

LECTURER'S NAME : DR AZURAH BINTI ABU SAMAH

Table of content

1.0 Introduction

Child marriage is a human right violation. Despite laws against it the practice remains widespread. In statistics around one in every five girls is married before reaching age 18. Therefore, when relating to children, UNICEF is a stand or specific organization for every child. UNICEF works in over 190 countries and territories to save children's lives, to defend their rights, and to help them fulfil their potential, from early childhood through adolescence.

Despite all the dataset UNICEF collected, some of their data was used in this analysis. We want to prove that percentage of child marriage in this world was related to a country. For example this may be due to the environment country itself that affects the life of a population in that country. For this, we investigate and analyze data that we get from UNICEF website which is country and also gender. As above explanation, child marriage had more in effect for woman therefor we need a right analysis to prove those statement.

Beside, we also used other variables to show is the violation that happen nowadays related to child marriage? Therefore, we investigate some of the potential variables that might causes because of child marriage. Thus, this analysis is for finding relationships between these variables using a correlation and regression statistical analysis.

Touching Up Data

Before starting the analysis, there are some facts that need to be clear. Firstly since the dataset variables all are in difference excel or dataset, therefore all the dataset was combined together to ease the analysis. Secondly, to ease the testing only 15 countries were selected out of all the countries listed in the dataset. This is because some countries might be missing values in the sample that need to be addressed.

2.0 Statistical Analysis

2.1 Statistical Analysis : Hypothesis Testing Two Sample Test

2.1.1 Description of case study and data

As we can see increasing in the number of percentage on child marriage keep on being an issue from time to time. Therefore does it equate to higher number of child marriage for female which result also for a higher number of child marrigae for male? Or is it the latter? Is there sufficient evidence that the average child marriage percent for female is greater compared from child marriage percent for male?

The sample for child marriage percent female and child marriage percent male are treated as independent. Even though they have similar countries there are totally different when involved with gender specifically.

2.1.2 Scenario

To do the hypothesis we must perform the two tailed tests. The confidence level that will be used is 95% . The further calculation was done using R :

```
        Welch Two Sample t-test

data:  male and female
t = -4.3305, df = 14.567, p-value = 0.0006336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.683442  -9.049891
sample estimates:
mean of x mean of y
 4.266667 22.133333
```

Table 1

2.1.3 Summary of Analysis using Two Sample Test Testing Hypothesis

H0: The average of child marriage percent for female lower or equal of child marriage percent for male (u1 <= u2).

H1: The average pf child marrigae percent for female is greater than child marriage percents for male (u1>u2).

From R, the hypothetical different t-value is -4.3305 and the p-value is 0.0006336. Since the p-value is less than 0.05, therefore we reject the null hypothesis. There is enough evidence that the mean of child marriage percent for female is greater than child marriage percent for male.

2.1.4 Justification and Conclusion

This shows that the world had a crisis issue in discriminating against the female side. Why does female becoming had more number in child marriage comparing to male? As some people believe that females just no need to have a great education instead, their life more focusing on families matters. This thinking was totally wrong.

2.2 Statistical Analysis : Data analysis using Correlation and Linear Regression

2.2.1 Description of case study and data

The unicef dataset child statistics is used. With the combinations of these dataset it contain the percentage of child marriage, justification of wife beating among adolescence, out of school rates and violent discipline on certain country. 15 countries had been chosen and their percentage according to the variables above was recorded.

We want to observed the percentage of child marriage influnce the percentage of others variables or not.

2.2.2 Scenario

We found, based on scatterplot and correlation coefficient, **that child marriage and justification wife beating among adolescence** are positive but it has weaker correlation and not linear . So apparently we can conclude that there is no relationship between the two variables which is child marriage(x) and wife beating among adolescence(y). We wish to further this by developing an equation (using linear regression method) predicting child marriage based on wife beating among adolescents. The details of the data analysis using linear Regression are described in **Appendix A.**

2.2.3 Summary of Analysis using linear regression result

The scatter graph in **Figure 1 of Appendix A** suggests that there is weak positive linear association between the child marriage and wife beating among adolesccence. Its does not depending on which country has larger percentage on child marriage will have larger pencentage in wife beating among adolescence. The correlation coefficient (0.02080489) is smaller enough and close to 0.

2.2.3.1 Linear Regression Model

From the output of linear regression, as shown in **Figure 2 of Appendix A.**

- The estimated regression line equation can be written as follow :
  wife-beating among adolescents = 64.97 + 0.049*child marriage
- The intercept (b0) is 64.97.It can be interpreted as the predicted justification wife-beating among adolescent percentage(y) when zero value is child marriage(x).
- The regression beta coefficient for the variable child married(b1), also known as the slope, is 0.049.

2.2.3.2 Summary of Model Assessment

The output of the six components are as follows

```
Residuals:
    Min      1Q  Median      3Q     Max
-61.607 -25.650   2.836  34.950  51.245


Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           64.96686   20.20453   3.215  0.00676 **
data$`Child marriage%`  0.04924    0.65621   0.075  0.94133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 40.27 on 13 degrees of freedom
Multiple R-squared:  0.0004328,       Adjusted R-squared:  -0.07646
F-statistic: 0.005629 on 1 and 13 DF,  p-value: 0.9413
```

Table 2

- Residuals

  The median is 2.836 which is so far from zero. The absolute value of minimum and maximum are far, there are not equal.

- Coefficient

  The statistical hypotheses are :

  Null hypothesis (HO) : **the coefficients are equal to zero. There is no relationship between child marriage and wife-beating among adolescent.**

  Alternative Hypothesis (H1) **: There is a relationship between the variables**

  Since for our variable, t value < p value where 0.075 < 0.94133 so we cannot reject the null hypothesis. There means that there is no significant association between the predictor and the outcomes variables.

- Residuals Standard Error

  For these variables, RSE = 40.27, larger RSE meaning the observed wife-beating among adolescence values deviate from the true regression line by approximately 40.27 units on average.

- R squared

  In our dataset, the R-squared is 0.0004328. This show that child marriage explains about 0.04328% of the variation in wife-beating among adolescents. R squared was nearly to 0. The value of wife-beating among adolescent does not depend on child marriage. (None of variation in y is explained by variation in x)

- F-statistic

  F-statistic equal 0.005629 producing a p-value 0.9413 which is not too significant. The p-value is larger means the results are not significant. Cannot reject the null Hypothesis.

- P-value

  Refer section 2.2.3.2 (*table 2*)

2.2.4 Justification and conclusion

The above analysis implies that child marriage is not a good predictor or factor that gives significant high impact in predicting justification of wife-beating among adolescents . Currently the problem children under 18 are facing today is not just because of one case. It can be from any causes also just as poverty, county type and others.

2.3 Statistical Analysis : Data analysis using Chi Square test of independence

2.1.1 Description of case study and data
The dataset of child marriage is used. It involves variables of country and gender. Data show the percentage of child marriage based on the gender in a certain contry. This data involves around 15 countries in the world and the observation of percentage has been recorded.

2.1.2 Scenario
We found, based on the bar chart, between male and female gender, it has a huge different in child marriage percentage. The detailed graph can be seen in **Figure 1 of Appendix B**.

2.1.3 Summary
It shows that the number of females is more likely to have a higher percentage of child marriage compare male when versus with countries. So we further our study using a chi-square test to showthe truth about our statement. The statistical hypotheses are as follow :

H0 : gender is independent of country
H1 : gender is not independent of country /dependent

As for the chi square test, the p-value was (9.157e-07) significant to the alpha value 0.01 so we can reject the null hypothesis. Detail can be seen in **Appendix B section 3** (computation). Therefore we can say that the variable of gender is dependent on the countries.

32.1.4 Justification and Conclusion
Currently, we can conclude that there is a relationship between country and gender. Its can be show as if a certain country has a low and poor poverty, the total of child marriage might be incresess. Beside, the statistic we get shown that females are more than to get married under age maybe because of certain problems.

3.0 Conclusion

To conclude, from our first analysis which is the relationship between countries and gender, it shows that females will have a higher percentage child marriage compare to male. Thus this analysis is strong to back up the statement above. Child marriage threatens girls's lives and health and it limits their future prospects. Girls pressed into child marriage often become pregnant while still adolescents, increasing the risk of complications in pregnancy or childbirth. These complications are the leading cause of death among older adolescent girls.

Beside, since for our second analysis there is no relationhisp between percentage child marriage had effects on the investiage varibale, we can said that child marrigae in not the only cause for those violent happen towards adolescene. As sure, from time to time, there is increasing on violence and cases happen to children all around the world. Therefore, we must protect children all around the world to make sure their future is safe.

Appendix A

Explanation of the analysis provided in correlation and Linear Regression

1. Inspect the data

```
        > Dataset_Project2 <- read_excel("Dataset Project2.xlsx")
        > View(Dataset_Project2)
        > data<-Dataset_Project2
        > head(data)

        ## A tibble: 6 x 4
        Country        `Child  marriage%`  `wife-beating  among  adolescents%`  `Violent
discipline%`
        <chr>           <dbl>                 <dbl>                    <dbl>
        1 Bangladesh       63                    29                       82
        2 Cambodia         23                    72                       26
        3 Kenya            26                    82                       88
        4 Ghana            23                    60                       94
        5 Indonesia        21                    93                       24
        6 Kazakhstan        7                     8                       53
```

We want to predict the relationship between children's problems based on certain countries.

2. Visualization

A scatter plot displaying the child marriage (in percentage) versus wife-beating among adolescents (in percentage) was created. Note the value here is in percentage. Smooth line in the scatter plot graph also was added. Here is the code :

```
> library(ggplot2)
> Graph1 = ggplot(data, aes(x=`Child marriage%`,y=`wife-beating among adolescents%`))
> Graph1+
+ geom_point(size = 3, shape = 21, color = "#002344", fill = "#FECB00")+stat_smooth()
```

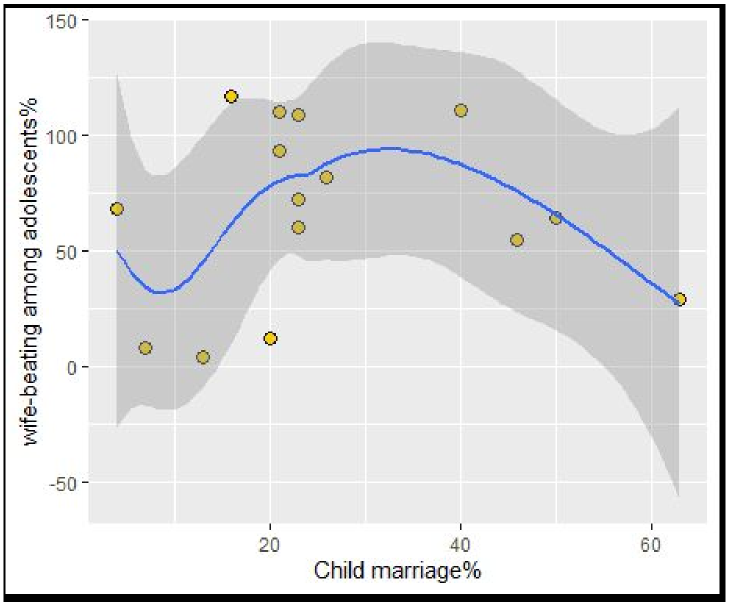

Figure 1 - Scatter graph of child marriage versus wife-beating among adolescents

The graph above suggests a changing parabola or changing trend between the child marriage and wife-beating among adolescents. It is a changing trend because not just an increasing or decreasing trend.

Therefore, we compute the correlation coefficient between the two variables using the R function cor() :

```
> cor(data$`wife-beating among adolescents%`,data$`Child marriage%`)
## [1] 0.02080489
```

In this graph, the correlation coefficient is positive but it has weaker correlation and not linear . So apparently we can conclude that there is no relationship between the two variables which is x and y.

3. Computation

So apparently, the simple linear regression tries to find the best line to predict the wife beating among adolescents on the basic of child marriage. **So in here, we want to determine is there any relationship between the child marriage and wife beating among adolescents and is there any affect between independent variable child marriage(x) with dependent vairable justification wife beating among adolescents (y).**

The population regression model :

Justification wife-beating among adolescents = B0 + B1 x child marriage

```
> Data = lm(data$`wife-beating among adolescents%`~data$`Child marriage%`,
Dataset_Project2)
> Data

Call:
lm(formula = data$`wife-beating among adolescents%` ~ data$`Child marriage%`,
data = Dataset_Project2)

Coefficients:
(Intercept)  data$`Child marriage%`
  64.96686              0.04924
```

The results show the intercept and the beta coefficient for the wife-beating among adolescent variables.

4. Interpretation

- The independent variables is the child marriage. The dependent variable is justification wife beating among adolescents.
- The *estimated regression line/model equation* can be written as follow : **wife-beating among adolescents = 64.97 + 0.049*child marriage.**
- The intercept(b0) is 64.97. It can be interpreted as the predicted justification wife-beating among adolescent percentage(y) when zero value is child marriage(x).
- The regression beta coefficient for the variable child marriage (b1), also known as the slope, is 0.049. It can estimate change in the average value of Y as a result of a one unit change in x.

5. Regression line

```
> library(ggplot2)
> Graph1 = ggplot(data, aes(x=`Child marriage%`,y=`wife-beating among
adolescents%`))
> Graph1+
+ geom_point(size = 3, shape = 21, color = "#002344", fill =
"#FECB00")+stat_smooth()
> Graph1+
+ geom_point(size = 3, shape = 21, color = "#002344", fill =
"#FECB00")+stat_smooth(method = lm)+
+ labs(x="child marriage in %", y="wife-beating among adolescents in
%")+theme_bw()
```

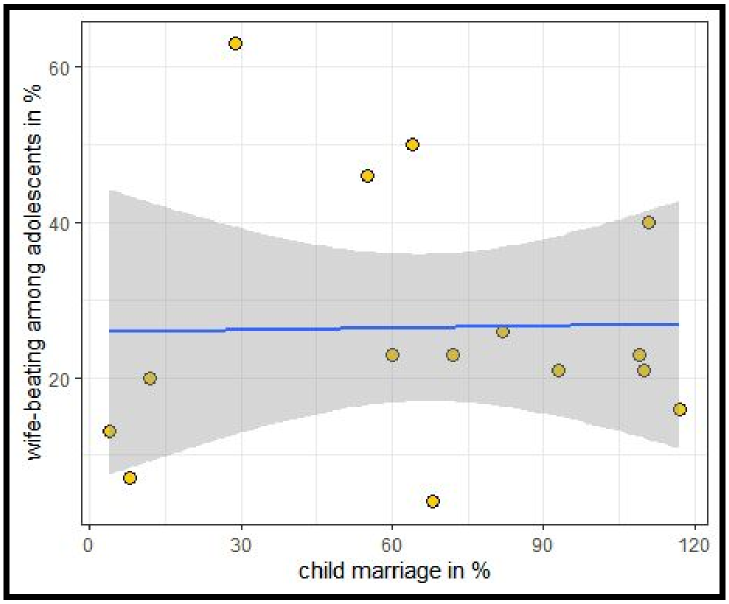

Figure 2 - Regression Line of the case study

6. Model Assessment

```
> summary(Data)


## Call:
## lm(formula = data$`wife-beating among adolescents%` ~ data$`Child marriage%`,
## data = Dataset_Project2)


## Residuals:
##    Min      1Q  Median      3Q     Max
## -61.607 -25.650   2.836  34.950  51.245


## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           64.96686   20.20453   3.215  0.00676 **
## data$`Child marriage%`  0.04924    0.65621   0.075  0.94133
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Residual standard error: 40.27 on 13 degrees of freedom
## Multiple R-squared:  0.0004328,   Adjusted R-squared:  -0.07646
## F-statistic: 0.005629 on 1 and 13 DF,  p-value: 0.9413
```

T-statistic and p-values

The statistical hypotheses are as follow :

Null hypothesis (HO) : the coefficients are equal to zero. There is no relationship between child marriage and wife-beating among adolescent.

Alternative Hypothesis (H1) : There is a relationship between the variables

Since for our variable, t value < p value where 0.075 < 0.94133 so we cannot reject the null hypothesis. There means that there is no significant association between the predictor and the outcomes variables.

Standard errors and confidence intervals

```
> confint(Data)
##                              2.5 %        97.5 %
##     (Intercept)              21.317630    108.616086
##     data$`Child marriage%`   -1.368426    1.466896
```

Model accuracy

Residual standard error (RSE)

For these variables, RSE = 40.27, meaning the observed wife-beating among adolescents values deviate from the true regression line by approximately 40.27 units on average.

Percentage error :

```
> sigma(Data)*100/mean(data$`wife-beating among adolescents%`)
## [1] 60.76682.
```

R-squared and Adjusted R-squared

## Multiple R-squared: 0.0004328,   Adjusted R-squared: -0.07646.

In our dataset, the R-squared is 0.0004328. This show that child marriage explains about 0.04328% of the variation in wife-beating among adolescents. The value near zero, thus we can conclude that the value of y does not depend on x.

F-statistic and P-value

## F-statistic: 0.005629 on 1 and 13 DF, p-value: 0.9413.

F-statistic equal 0.005629 producing a p-value 0.9413 which is not too significant. The p-value is larger means the results are not significant. Cannot reject the null Hypothesis.

Appendix B

Explanation on analysis provided in Chi Square Test & one way contingency table

1. Inspect the data

```
> data<-matrix(c(4,59,4,19,3,23,2,21,5,16,7,0,
+ 5,16,10,40,3,43,5,18,3,17,6,34,
+ 4,9,1,15,2,2),ncol = 2,byrow = T)
> colnames(data)<-c("Male","Female")
> rownames(data)<-c("Bangladesh","Cambodia","Kenya","Ghana","Indonesia",
+ "Kazakhstan","Myanmar","Nepal","Nigeria","Pakistan",
+ "Philippines","Uganda","Ukraine","Timor-Leste","Maldives")
> data
## Male Female
## Bangladesh 4 59
## Cambodia 4 19
## Kenya 3 23
## Ghana 2 21
## Indonesia 5 16
## Kazakhstan 7 0
## Myanmar 5 16
## Nepal 10 40
## Nigeria 3 43
## Pakistan 5 18
## Philippians 3 17
## Uganda 6 34
## Ukraine 4 9
## Timor-Leste 1 15
## Maldives 2 2
```

We want to predict if a relationship exist between country and gender based on child marriage percentage.

2. Visualization

Create a two boxplot of gender which is male and female, displaying percentage of child married versus country.

```
> barplot(data)
> barplot(data, beside = T, main = "child marriage in a country by gender")
```

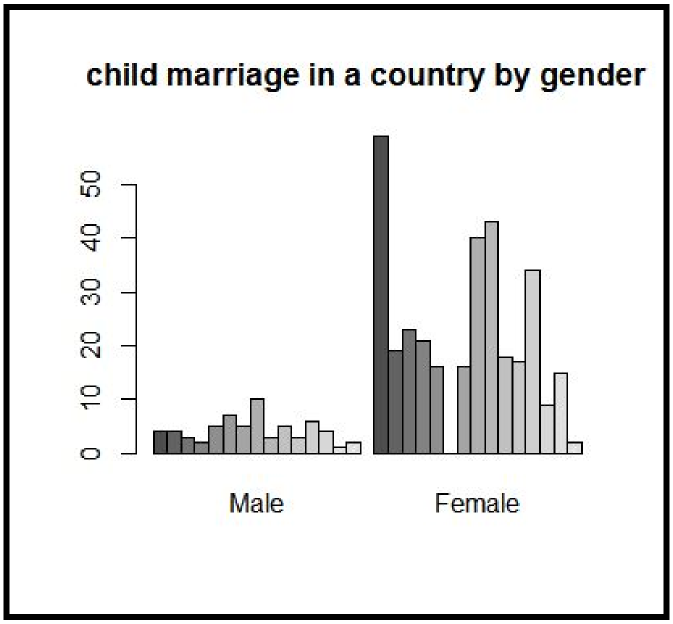

Figure 1 - country versus child marriage percentage

3. Computation

The test statistic was calculated using chi square test based on a two way contingency table.

```
> chisq.test(data, correct = T)
Pearson's Chi-squared test
data: data
X-squared = 54.858, df = 14, p-value = 9.157e-07
```

4. Interpretation

From the output above:

In this case, we may be interested in seeing if there is an association between gender and country variables based on the graph we get. It shows that the number of females is more likely to have a higher percentage of child marriage compare male when versus with country. So we further our study using a chi-square test to show the truth about our statement.

The statistical hypotheses are as follow :

H0 : gender is independent of country

H1 : gender is not independent of country /dependent

As for the chi square test, the p-value was (9.157e-07) significant to the alpha value 0.01 so we can reject the null hypothesis. Therefore we can say that the variable of gender is dependent on the country.