UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143
PROBABILITY & STATISTICAL DATA ANALYSIS

HUMAN POPULATION IN ASIA

GROUP PROJECT

GROUP 1 SECTION 08

NO.	NAME	MATRIC NO.
1	FARAH IZZAH BINTI MOHD AISHA NUDDIN	A21EC0022
2	KHALISAH NAJAH BINTI NAWAWI	A21EC0038
3	NOR ALYA BINTI MOHAMAD NOR	A21EC0105
4	NURFAZLIANA SURAYA BINTI BAHARUDDIN	A21EC0118

LECTURER'S NAME : DR. SHARIN HAZLIN BINTI HUSPI

SUBMISSION DATE : 3rd JULY 2022

TABLE OF CONTENT

1.0 INTRODUCTION

The population is one of the important factors which helps to balance the environment. If the population is balanced, then all the needs and demands of the people can be easily fulfilled, which helps to preserve the environment of the country. But of course, there will be consequences if the population increases too much and leads to overpopulation. There will be increasing demand for food, water, housing, energy, healthcare, transportation, and other resources as there are more people. And all that consumption raises the probability of major catastrophes like pandemics, ecological destruction, and greater violence. The objective of this project is to discover the statistics and data of Asian Countries' populations.

2.0 DATASET

We picked Asian Countries as our dataset for this project. The dataset contains the data of 151 countries in Asia. The variables that were chosen are shown in Table 1 and Table 2.

Name	Variable	Data Type
Data names	Name of the Asian countries	Nominal
Population	Population of the Asian countries	Ratio
World Share	World Share of Asian countries	Ratio
Infertility	Infertility rate of the Asian countries	Ratio
Life expectancy	Life expectancy of 5 subregions in Asia	Ratio
Density	Density of the Asian countries per people	Ratio
Processed Data		
Growth	Growth of Asian countries' population	Ratio

Table 1: Variables in Country Data

Selected Variable(s)	Test	Description
Population	Hypothesis testing (1 sample test)	**Explanation:** The variables are chosen to determine whether the mean percentage of yearly change based on the asian population is greater than 1.0%. **Possible outcome:** The mean percentage of yearly change based on the asian population is greater than 1.0%.
Fertility rate, Density	Correlation analysis	**Explanation:** The variables are selected to test whether the linear relationship exists between the fertility rate and the density of the country using Pearson's Product-Moment Correlation Coefficient. **Possible outcome:**
Population, World Share	Regression analysis	**Explanation:** The variables are selected to test whether the value of population size depends on the world share percentage, using population as dependent variable x and world share as independent variable y. **Possible outcome:** The population of the countries in Asian, depend on the world share percentage. The higher the world share percentage, the larger the population size.
Growth	ANOVA test	**Explanation:** A random sample is selected from a population to test the equality of the means of world share rate for all subregions by analyzing the sample variances at 5% of significance level. **Possible outcome:** The means of population growth are the same for all continents at 5% of significance level.

Table 2: Summary of Selected Variables and Test with their Respective Description

3.0 DATA ANALYSIS

3.1 Hypothesis Testing - 1 Sample

Based on this project, it is of interest to determine whether the mean percentage of yearly change based on the Asian population is greater than 1.0%. To conduct a Hypothesis testing, we must establish the following :

1) Hypothesis statement

   Let μ = mean percentage of yearly change based on the asian population

   H_0: μ = 1.0

   H_1: μ > 1.0

2) Test statistic

   The type of hypothesis test is test on mean, variance unknown. This test concerns one mean for data that can be assumed to follow a normal distribution and the sample is large (n>30). Thus, the test statistic is :

$$z = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

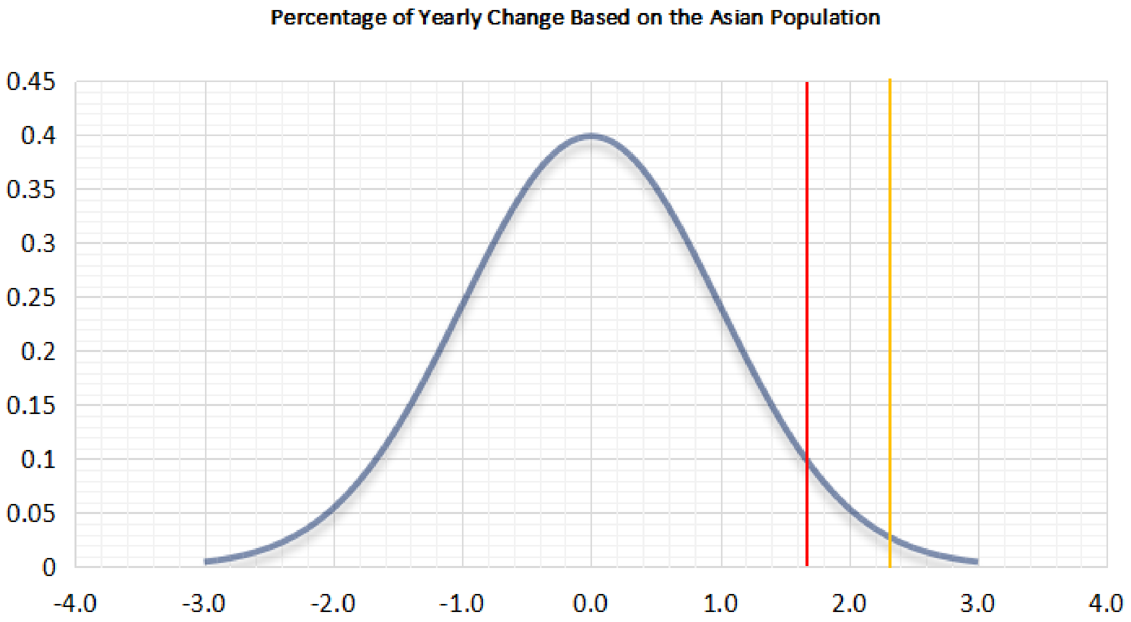
$$z = \frac{1.267843 - 1.0}{\frac{0.8220835}{\sqrt{51}}} = 2.32675$$

3) Significance level, α = 0.05

   $Z_{0.05} = 1.645$

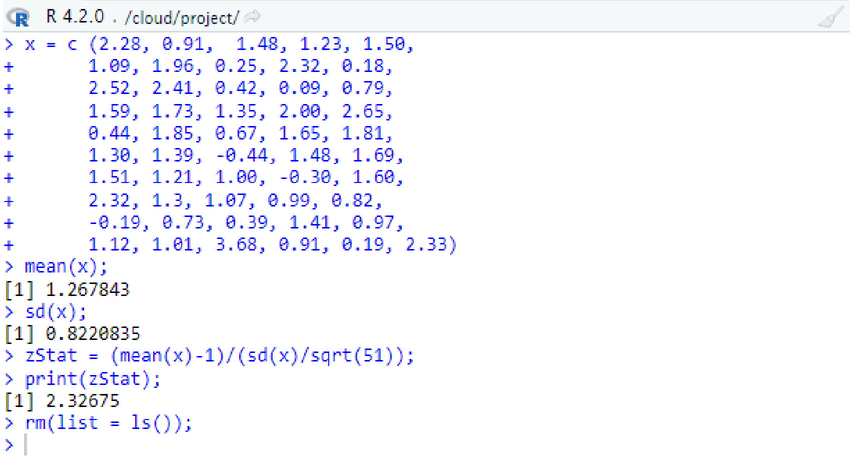
4) Conclusion

   Since the test statistic (Z = 2.32675) > $Z_{0.05}$ = 1.645, H_0 should be rejected. There is sufficient evidence that the mean percentage of yearly change based on the asian population is greater than 1.0%.



Graph 1

The alternative hypothesis is, H_1: $\mu > 1.0$. So, the test is a right-tailed test. Red line in graph 1 represents $Z_{0.05} = 1.645$, meanwhile the test statistic, $Z = 2.32675$ is shown by the yellow line. The test statistic falls within the critical region. So, we should reject the null hypothesis, H_0.



```
> x = c (2.28, 0.91,  1.48, 1.23, 1.50,
+       1.09, 1.96, 0.25, 2.32, 0.18,
+       2.52, 2.41, 0.42, 0.09, 0.79,
+       1.59, 1.73, 1.35, 2.00, 2.65,
+       0.44, 1.85, 0.67, 1.65, 1.81,
+       1.30, 1.39, -0.44, 1.48, 1.69,
+       1.51, 1.21, 1.00, -0.30, 1.60,
+       2.32, 1.3, 1.07, 0.99, 0.82,
+       -0.19, 0.73, 0.39, 1.41, 0.97,
+       1.12, 1.01, 3.68, 0.91, 0.19, 2.33)
> mean(x);
[1] 1.267843
> sd(x);
[1] 0.8220835
> zStat = (mean(x)-1)/(sd(x)/sqrt(51));
> print(zStat);
[1] 2.32675
> rm(list = ls());
> 
```

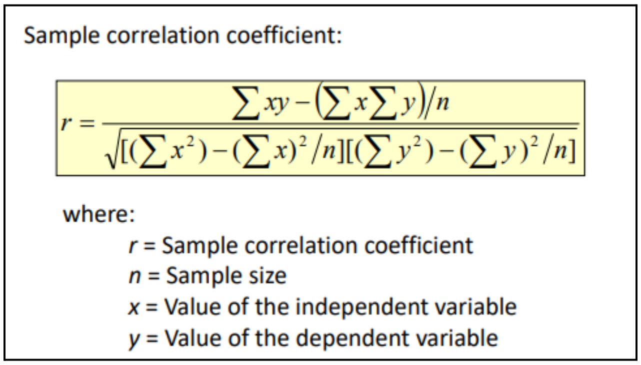
Figure 1

To conclude, there is sufficient evidence that the mean percentage of yearly change based on the asian population is greater than 1.0%. Z-value for this test statistic is obtained using R Programming.

3.2 Correlation Test

In this correlation analysis, we used the variable fertility rate and density of several countries' populations in Asia. We will test whether there is a linear relationship between fertility rate and density using the Pearson's product-moment correlation coefficient. Pearson's technique is to calculate the coefficient of correlation since both data types are ratio type data. Assume the significance level to be 95%, significance level $a = 0.05$.

Calculation of correlation coefficient using Pearson's Technique:



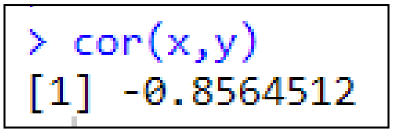
Sample correlation coefficient:

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Pearson's Formula



```
> cor(x,y)
[1] -0.8564512
```

$$r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{[\Sigma x^2 - (\Sigma x)^2/n][\Sigma y^2 - (\Sigma y)^2/n]}}$$

$$r = \frac{3435.5 - [(1574)(118.9)/51}{\sqrt{[50880 - (1574)^2/51][309.65 - (118.9)^2/51]}}$$

$$r = -0.8564512$$

From Rstudio and calculation above, we get the sample correlation coefficient, $r =- 0.8564512$, showing that there is a strong linear association between x,(Density) and y (Fertility Rate).

Hypothesis statement:

H_0: $p = 0$ (no linear correlation)

H_1: $p \neq 0$ (linear correlation exists)

Test statistics:

By using R Studio, we can get the result that t -11.613.

Finding the critical value:

$a \ = \ 0.05, \ df \ = \ n - 2 = 51 - 2 \ = \ 49$

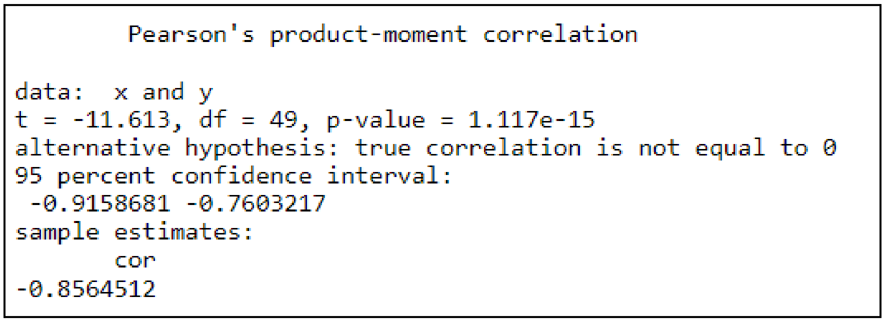
From t distribution table, since it is a two tailed test, there are two critical values:

Lower tail critical value -$t_{a = 0.025, df = 49}$ = -2.0096

Upper tail critical value $t_{a = 0.025, df = 49}$ = 2.0096

FromRstudio, we get the p-value = 1.117e-15.

As this is a two-tailed test, we reject H_0 if t > 2.0096 or t < -2.0096.



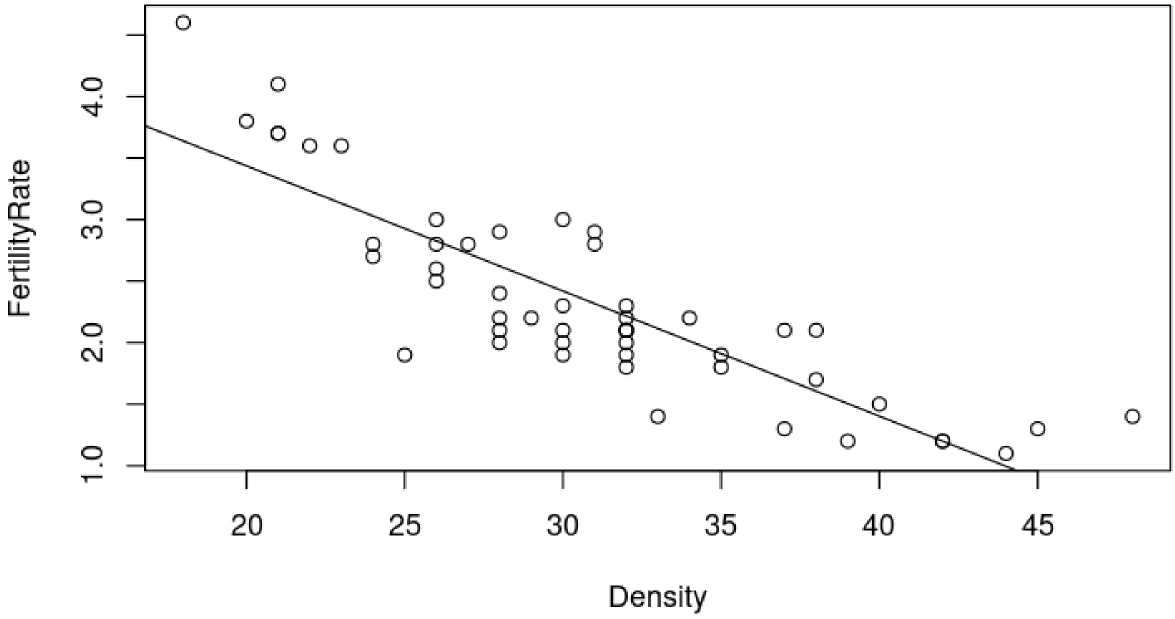
```
        Pearson's product-moment correlation

data:  x and y
t = -11.613, df = 49, p-value = 1.117e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9158681 -0.7603217
sample estimates:
       cor
-0.8564512
```

Figure 2

Decision:

Since the test statistic t = -11.613 < $t_{a = 0.025, df = 49}$ = -2.0096, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between x(Density) and y(Fertility Rate) at $a \ = \ 0.05$.



Graph 2

The dots on the scatter plot graph slope downward, indicating a negative link between density and fertility rate—the lower the density, the greater the fertility rate. This relationship is evident from the scatter plot graph.

The sample coefficient correlation is then determined using Pearson's method as both variables are ratio-type data. We obtained r = -0.8564512, which indicates that there is a reasonably significant negative linear relationship between density and fertility rate.

The association between density and fertility rate is clearly demonstrated by the scatter plot graph's results and correlation test. Due to their inverse linear relationship, the fertility rate declines as density rises.

3.3 Regression Test

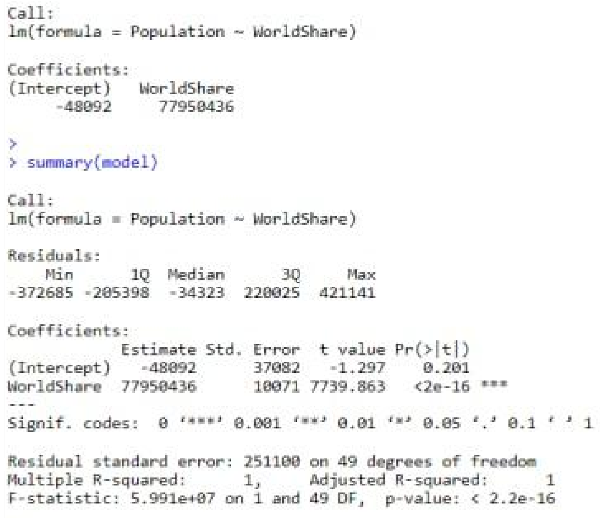
Regression analysis is used to predict the value of a dependent variable based on the value of at least one independent variable. In addition, we can explain the impact of changes in an independent variable on the dependent variable. In this regression test, we wish to examine the relationship between the population and world share of a random sample of 51 Asian countries. The dependent variable, y is the population. Meanwhile world share is the independent variable, x. This regression model is called simple regression since it only involves one single independent variable. The estimated regression model is :

$$\hat{y} = b_0 + b_1 x$$

The formula for b_0 *and* b_1 are :

$$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$



```
Call:
lm(formula = Population ~ WorldShare)

Coefficients:
(Intercept)   WorldShare
     -48092     77950436

>
> summary(model)

Call:
lm(formula = Population ~ WorldShare)

Residuals:
    Min      1Q  Median      3Q     Max
-372685 -205398  -34323  220025  421141

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   -48092      37082   -1.297    0.201
WorldShare  77950436      10071 7739.863   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 251100 on 49 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 5.991e+07 on 1 and 49 DF,  p-value: < 2.2e-16
```

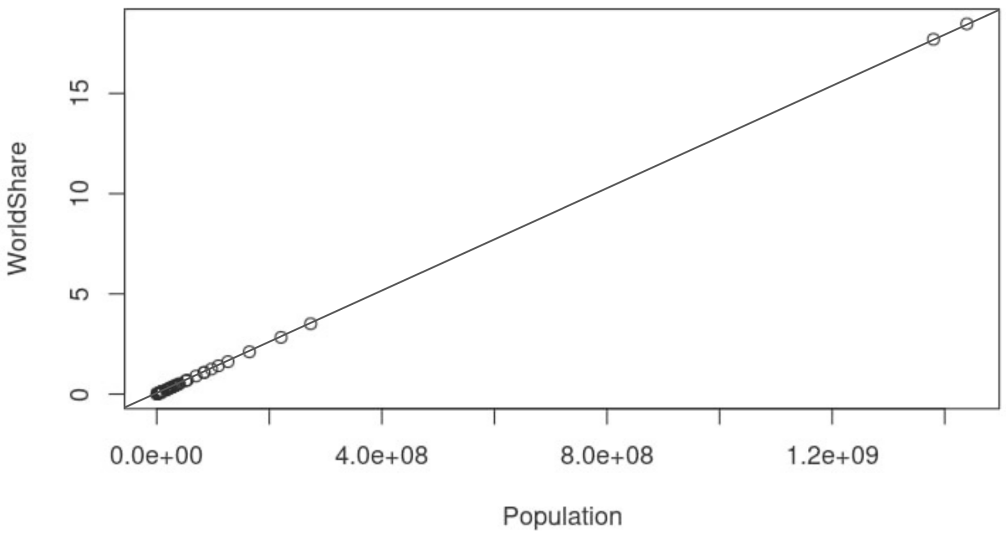
Figure 3

By referring to R studio as well as the estimated regression model, b_0= -48092 indicates that, for population within the range of world share observed, -48092 is the portion of population not explained by the world share. b_1= 77950436 tells us that the average size of population increases by 77950436 on average, for each one unit change in world share.

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

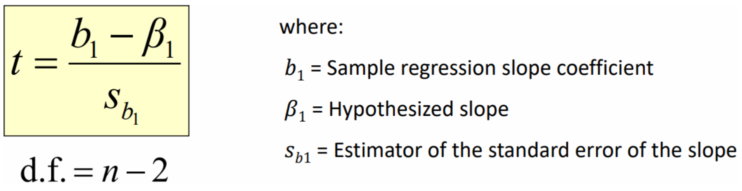
$$R^2 = \frac{3.7783E+18}{3.7783E+18} = 1$$

The coefficient of determination, R^2 is the portion of the total variation in the dependent variable that is explained by variation in the independent variable and the formula is as above. R^2 = 1 in this test shows that 100% of the population is explained by variation in world share.



Graph 3

Graph 3 shows the scatter plot and regression line of the population against world share. Based on the plot, we can conclude that both variables, population and world share, have a positive linear relationship. On top of that, we can conclude that the population of the countries in Asian, depend on the world share percentage.



$$\boxed{t = \frac{b_1 - \beta_1}{s_{b_1}}}$$

$$\text{d.f.} = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Figure 4

Figure 4 shows the formula for the regression t-test. The purpose of this test is to indicate if there is a linear relationship between world share and population. To conduct a regression t-testing, we must establish the following :

1) Hypothesis statement

   H0: $\beta_1 = 0$ (No linear relationship)

   H1: $\beta_1 \neq 0$ (Linear relationship does exist)

2) Sample regression slope coefficient, b_1

   Based on the estimated regression equation, $\hat{y} =- 48092 + 77950436\,x$. The value of b_1 is 77950436.

3) Estimator of the standard error of the slope, S_{b1}

$$S_{b1} = \frac{\sqrt{\frac{\Sigma(y_i - \hat{y})^2}{n-2}}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}$$

$$S_{b1} = 10071.29364$$

4) Test statistic

$$t = \frac{b_1 - \beta_1}{S_{b1}}$$

$$t = 7739.863$$

5) Conclusion

   Since the test statistic ($t = 7739.863$) > $t_{0.025} = 1.960$, H0 should be rejected. There is sufficient evidence that world share affects the population of Asian countries.

3.4 ANOVA Test

In this data analysis, ANOVA test is used to test the equality of growth population means by analyzing the sample variances. Since Asia has five subregions, the population is split into five and the samples are randomly selected to get an equal sample size. The significance level used to test the null hypothesis is α = 0.05.

The null hypothesis goes as, H_0: μ1 = μ2 = μ3 = μ4 = μ5 . While for the alternative hypothesis, H_1: at least one mean is different. The number of samples (n), mean of samples ($\bar{x}$) and standard deviation of samples (s) for each sample are calculated as below:

Samples	***n***	**x̄**	***s***
Southern Asia	5	1.636	0.606
Eastern Asia	5	0.236	0.411
South-Eastern Asia	5	1.120	0.196
Western Asia	5	1.824	0.760
Central Asia	5	1.640	0.417

Table 3: The n, $\bar{x}$ and s for each sample

The mean between samples that we have calculated is $\bar{\bar{x}}$ = 1.291, the standard deviation between samples is $s_{\bar{\bar{x}}}$= 0.6456, variance between samples is $ns^2_{\bar{x}}$ = 0.4167 and variance within samples is s^2_p= 0.2652. To obtain $F_{\text{test statistic}}$ value, we used the formula below:

$$F = \frac{variance\ between\ samples}{variance\ within\ samples} = \frac{ns^2_{\bar{x}}}{s^2_p}$$

By using the above formula, the value of we obtained for $F_{\text{test statistic}}$ = 1.5713. Then the critical value of F with α = 0.05 is obtained from the F-distribution table which goes by $F_{\text{critical value}}$ = 3.01.

From the result, since the $F_{\text{test statistic}} < F_{\text{critical value}}$ (1.5713 < 3.01), we fail to reject the null hypothesis, H_0 as there is sufficient evidence to claim that different continents have the same means of the growth population.

4.0 CONCLUSION

From this project, we did 4 tests on the dataset of Asian countries. Hypothesis Testing, Correlation Test, Regression Test and ANOVA Test. In the Hypothesis Testing, we wanted to determine whether the mean percentage of yearly change based on the asian population is greater than 1.0%. We let H_0: μ = 1.0, and H_1: μ > 1.0. After calculating throughout the test, we conclude that H_0 should be rejected because there is no sufficient evidence to support the statement. Next we did the Correlation test where we test whether the is a linear relationship between fertility rate and density using the significance level, $\alpha = 0.05$. with H_0: $p = 0$ (no linear correlation) and H_1: $p \neq 0$ (linear correlation exists) as the hypothesis statements, we did the test statistic by using the R studio. In result, we reject the null hypothesis. There is sufficient evidence to conclude that there is a linear relationship between x(Density) and y(Fertility Rate)

Following after is the Regression Test. We examined the relationship between the population and world share of a random sample of 51 Asian countries. We had a dependent variable, y and independent variable, x. In result, we conclude that The R^2 value which is nearly in this test shows that 99.366% of the population is explained by variation in world share. The last test we did was the ANOVA test. We wanted to test the equality of growth population means by analyzing the sample variances. From this test we conclude that different continents have the same means of population growth.

Last but not least, from this project, we learned how to relate the datasets with one another to find certain relations by doing several analyses. We also learned and used RStudio which was a huge help in plotting graphs. We learned how to apply this to a real life situation which would be very useful in our job scope. We understand the importance of doing tests to conclude any data with enough evidence.

5.0 APPENDICES

5.1 Raw Dataset

Dataset	Population	World Share	Fertility Rate	Density (P/Km²)	Growth
China	1,439,323,776	18.47 %	1.7	38	0.39 %
India	1,380,004,385	17.70 %	2.2	28	0.99 %
Indonesia	273,523,615	3.51 %	2.3	30	1.07 %
Pakistan	220,892,340	2.83 %	3.6	23	2.00 %
Bangladesh	164,689,383	2.11 %	2.1	28	1.01 %
Japan	126,476,461	1.62 %	1.4	48	-0.30 %
Philippines	109,581,078	1.41 %	2.6	26	1.35 %
Vietnam	97,338,579	1.25 %	2.1	32	0.91 %
Turkey	84,339,067	1.08 %	2.1	32	1.09 %
Iran	83,992,949	1.08 %	2.2	32	1.30 %
Thailand	69,799,978	0.90 %	1.5	40	0.25 %
Myanmar	54,409,800	0.70 %	2.2	29	0.67 %
South Korea	51,269,185	0.66 %	1.1	44	0.09 %
Iraq	40,222,493	0.52 %	3.7	21	2.32 %
Afghanistan	38,928,346	0.50 %	4.6	18	2.33 %

Saudi Arabia	34,813,871	0.45 %	2.3	32	1.59 %
Uzbekistan	33,469,203	0.43 %	2.4	28	1.48 %
Malaysia	32,365,999	0.42 %	2.0	30	1.30 %
Yemen	29,825,964	0.38 %	3.8	20	2.28 %
Nepal	29,136,808	0.37 %	1.9	25	1.85 %
North Korea	25,778,816	0.33 %	1.9	35	0.44 %
Taiwan	23,816,775	0.31 %	1.2	42	0.18 %
Sri Lanka	21,413,249	0.27 %	2.2	34	0.42 %
Kazakhstan	18,776,707	0.24 %	2.8	31	1.21 %
Syria	17,500,658	0.22 %	2.8	26	2.52 %
Cambodia	16,718,965	0.21 %	2.5	26	1.41 %
Jordan	10,203,134	0.13 %	2.8	24	1.00 %
Azerbaijan	10,139,177	0.13 %	2.1	32	0.91 %
United Arab Emirates	9,890,402	0.13 %	1.4	33	1.23 %
Tajikistan	9,537,645	0.12 %	3.6	22	2.32 %
Israel	8,655,535	0.11 %	3.0	30	1.60 %
Hong Kong	7,496,981	0.10 %	1.3	45	0.82 %
Laos	7,275,560	0.09 %	2.7	24	1.48 %
Lebanon	6,825,445	0.09 %	2.1	30	-0.44 %

Kyrgyzstan	6,524,195	0.08 %	3.0	26	1.69 %
Turkmenistan	6,031,200	0.08 %	2.8	27	1.50 %
Singapore	5,850,342	0.08 %	1.2	42	0.79 %
State of Palestine	5,101,414	0.07 %	3.7	21	2.41 %
Oman	5,106,626	0.07 %	2.9	31	2.65 %
Kuwait	4,270,571	0.05 %	2.1	37	1.51 %
Georgia	3,989,167	0.05 %	2.1	38	-0.19 %
Mongolia	3,278,290	0.04 %	2.9	28	1.65 %
Armenia	2,963,243	0.04 %	1.8	35	0.19 %
Qatar	2,881,053	0.04 %	1.9	32	1.73 %
Bahrain	1,701,575	0.02 %	2.0	32	3.68 %
Timor-Leste	1,318,445	0.02 %	4.1	21	1.96 %
Cyprus	1,207,359	0.02 %	1.3	37	0.73 %
Bhutan	771,608	0.01 %	2.0	28	1.12 %
Macao	649,335	0.01 %	1.2	39	1.39 %
Maldives	540,544	0.01 %	1.9	30	1.81 %
Brunei	437,479	0.01 %	1.8	32	0.97 %

5.2 Processed Dataset

Subregions	Countries	Growth
Southern Asia	India	0.99 %
Southern Asia	Pakistan	2.00 %
Southern Asia	Bangladesh	1.01 %
Southern Asia	Afghanistan	2.33 %
Southern Asia	Nepal	1.85 %
Eastern Asia	China	0.39 %
Eastern Asia	Japan	-0.30 %
Eastern Asia	South Korea	0.09 %
Eastern Asia	Taiwan	0.18 %
Eastern Asia	Hong Kong	0.82 %
South-Eastern Asia	Indonesia	1.07 %
South-Eastern Asia	Malaysia	1.30 %
South-Eastern Asia	Philippines	1.35 %
South-Eastern Asia	Vietnam	0.91 %
South-Eastern Asia	Brunei Darussalam	0.97 %
Western Asia	Turkey	1.09 %

Western Asia	Iraq	2.32 %
Western Asia	Yemen	2.28 %
Western Asia	Syria	2.52 %
Western Asia	Azerbaijan	0.91 %
Central Asia	Uzbekistan	1.48 %
Central Asia	Kazakhstan	1.21 %
Central Asia	Tajikistan	2.32 %
Central Asia	Kyrgyzstan	1.69 %
Central Asia	Turkmenistan	1.50 %