Hello everyone, in my part, I will show you some analysis about our project work, and make some conclusions.

Here is the analysis about ratios, frequencies, and figures for qualitative variables. And for these two qualitative variables, “Continent” variable has five groups, and we can get from frequency table that the ratio between the frequency of African countries and European countries is about one, the ratio between the frequency of Asian countries and total observations is about 23%, and so on. For “degree of development of a country”, the number of developing countries nearly twice than the developed countries. The bar plot shows that all European and Oceanian countries are developed countries, and there are only few developed countries in Asia and America.

Here we analyze confidence interval, because in real lives, we can not get exact results, and there is an error interval between sample and the whole population. First, we can get the mean value of a variable, and then calculate the margin of error by using 5% confidence level. From the two values we just calculated, we can get the upper and lower limit bound, in this example, the GDP growth rate is between 0.76% and 0.9%. As we see, the mean value is in this interval, so we can get conclusion that we are 95% confident that the average GDP growth rate is between 0.76% and 0.9%. Similarly, we can analyze confidence interval for other variables.

Here is the cross table about two qualitative variables, which measure the correlation between them. Chi-square and Fisher’s test are used, and we can see the two p-values are less than 5% and are very small. So we need to reject the null hypothesis that two qualitative variables are independent of each other. This situation is normal in our daily life as well, because we know European countries are tend to relatively developed, and on the contrary, African countries are tend to less developed. So the Continent and degree of development of a country are not independent with each other.

The ANOVA table can be analyzed here to show whether there is a significant difference among each group. Here we show the GDP growth rate in developed countries’ group and in developing countries’ group, and the happiness score in each of continents. And the ANOVA tables tell us in both cases, p-values are smaller than 5%, which means we need to reject the hypothesis, and there are significant differences between these groups.

Here we analyze the covariance and correlation between two quantitative variables. For covariance, if there is a positive sign, then it indicates a positive relationship, verse vice. For correlation, the value is between -1 and 1, if the value closes to 1, then it means a strong positive relationship, otherwise, negative relationship. If we consider the happiness score as a dependent variable, from the covariance table and correlation heatmap, GDP growth rate, satisfaction scores towards the whole society, and income should be positively correlated with it. And for income inequality, it is negatively correlated with happiness score. And this phenomenon is normal in current society. If people live in a country with high GDP growth rate, have high income, then his happiness will be high. And if someone lives in a country with high income inequality, then his happiness will be low.

Now we will do some hypothesis test. The first one is mean test. We set the mean value of income is eighteen thousand dollars, which is a number quoted from BBC news. And when we do the hypothesis test with R, we can find the p-value is less than 5% and approaches zero. So we need to reject the null hypothesis. And the actual mean value of the sample is about 6300 dollars, which can prove the decision we made.

The second hypothesis test is about proportion test. Assume in a theoretical and ideal case, the expected proportion of developing countries’ group is 50%. So we can set the null hypothesis is the ideal proportion of developing countries should be 50%. From the results, we can see that the p-value is less than 5%, which indicates we need to reject the null hypothesis. And the actual proportion of developing countries is about 66%, which can prove the decision we made.

The last hypothesis test is about the two sample variance test. Here we create dummy variables, which means if a country is developed, then it can be recorded as 1, otherwise 0. Before the t test, we need to check whether the variances from the sample are equal or not. The p-value is relatively and greater than 5%, so we can not reject the null hypothesis that two variances are equal, and we can use equal variance version of t test in R.

Here comes to the conclusion part, in our project work, we use world happiness report dataset with 110 observations, and after detecting outlier, we have 109 observations with 2 qualitative and 4 quantitative variables. Then we do analysis of two qualitative variables including ratios, frequencies, and cross tables, and analysis of one qualitative and one quantitative variable including ANOVA tables, as well as the analysis of quantitative variables including covariance and correlation. And find the positive relationship between happiness score with GDP, income and satisfaction scores. Finally, we do three tests about the mean value of income variable, the proportion of developing countries, and whether the variances in the sample are the same or not.

All right, that is all of our presentation. Thanks for your attention.