Statistical analysis

[Name of the student here]

[Name of the institution here]

# **Paper 1**

H01: There are no negative consequences of alcohol and other drug usage before the school leaving celebrations

Ha1: There are significant negative consequences of alcohol and other drugs usage before the school leaving celebrations

H02: There are no negative consequences of alcohol and other drug usage during the school leaving celebrations

Ha2: There are significant negative consequences of alcohol and other drug usage during the school leaving celebrations

The independent variable is alcohol usage on an average celebration day, other drug use, protective behavior strategy score, gender, accommodation/location, and survey modality. Hangover, emotional outburst, vomiting, heated argument, accident/injury, physical aggression, blackout, inability to payout for things, unprotected sex, and unhappiness towards the sexual situation, sexual risk/problem and any legal risk/problem have been used as the dependent variables. Another variant of this study can be conducted to compare the differences among groups e.g. between genders. The extent to which the independent variables explain the variation in the dependent variable should also be observed. A lower value in this regard will mean that more variables should be added to the analysis.

The dual-frame sampling method was used in this study. The main advantage of this method over the single-frame design is that a large number of people can be added to the dual-frame method. As in this study, one frame included the people who gave online responses and people from the other frame were conducted with the help of face to face interviews. In dual-frame sampling, the number of persons is increased and it provides a better representation of the population under study. There were surveys conducted before and after the celebrations took place. There were some small gifts given to the participants which also kept them interested in the surveys. The face-to-face interviews gave better responses as compared to the online survey but there were not many resources available to undertake the whole research with the interviews. Another advantage of such a technique is to minimize the bias related to the similarity of responses from all the respondents. Another positive aspect of personal interviews in this study is that these include actual happenings as responses.

The first survey included 56% of females which make 303 out of 541. Out of these females, 275 were 17 years old and 28 were 18 years old. 87% of these females were enrolled in an independent school. 52% of the people filled the online surveys. The survey after celebrations were conducted by using 405 people out of which 203 were females. Out of these females,191 were aged 17 years and the rest were aged 18 years or above. 192 of these females were from some independent school. Table 4 in this study shows that 209 of the respondents indulged in a hangover, 139 were involved in an emotional outburst.

As an inferential statistic, a series of Wilcoxon signed-rank tests were used because we had to determine that both the samples have been drawn from the same population. The relationships between the dependent and independent variables were studied by using the logistic regression model. A logistic regression model is very similar to the multiple regression model and all the assumptions of multiple regression apply to it. Since the dependent variables are nominal as all of these cannot have a direct numerical value and there are many independent variables in this study so logistic regression analysis is the most appropriate to use. Males consumed 18.44 and females 13.24 Australian standard drinks on an average day when they celebrated their school leaving. This consumption was higher than their last social event consumption and ecstasy consumption. The number of drinks consumed per hour remained similar across the different contexts. A majority of respondents experienced at least one of the negative outcomes specified in the study.

Odds ratio tells us the chance of occurrence of an event provided that some other event has already taken place. It can also be stated in terms of the association between exposure and an outcome. The odds ratio shows the odds that an outcome will occur given a particular exposure. Another concept related to the odds ratio is the confidence interval. The higher odds ratio will have a huge difference between the lower and upper limit of the confidence interval. The interpretation of this huge difference between boundaries of confidence intervals is that the probabilities of the occurrence of such an event are very high. As depicted by the table, the odds ratio for indulging in unprotected sex against those using the safety strategies is very high. This shows a very high probability of indulging in unprotected sex if the safety strategies are not implemented. The confidence interval limits are very wide in this scenario which shows a very high probability of a person indulging in unprotected sex if the safety strategies are not implemented properly. If we observe in detail, there are high odds ratios in most cases in the analysis. Gender is the only criterion that has odd ratios that are lower than 1 showing that there is less chance in getting involved in certain activities based on gender. The lowest score is between survey modality online and regrets on a sexual encounter.

I think that the sample taken is representative of all the school students across the country. The reason for this statement is that the students were not only given the questionnaires in person but also online surveys were conducted. The questionnaires in person were given to the student from a single geographical area while the online surveys can collect data nationwide. Using dual-frame sampling also ensured that the samples taken were representative of the population. Since the celebrations at the end of schools continue to be popular, there is a need to extend the study to schools all across the country so that solution to youth alcohol and drugs consumption can be found.

# **Paper 2**

The study explored the factors associated with diabetes for the people who reported their disease by themselves. The study was conducted from the year 2001 to 2008. The major aim of the study can be presented in the form of a hypothesis statement as follows:

H0: The self-reporting of diabetes in patients has not improved over the period from 2001 to 2008.

H1: The self-reporting of diabetes has improved over the period 2001-2008.

To study the patients in greater detail, they have been divided into various age groups and also into male and female respondents. Although this study does not involve the comparison across groups, a variant of this study can do the group-wise comparisons. In that case, there were many more hypotheses to be formed and analyzed.

The demographic variables included in the study are age, sex, and income. As far as age is concerned, there are eight distinct groups formed. The first group was below fifteen years of age and then groups were formed with 10 years gap in each class. The last group comprised of the people above seventy-five years of age. As far as the representation is concerned, the age group from 35-44 years had the most number of people as depicted by the highest percentage. The least representative group was the people who were aged more than seventy-five years. For the years 2005 and 2008, the group having the highest respondents has changed to aged 45-54. There was no change to the least represented group throughout the study. The representation of certain groups was decreased throughout the study but the group aged 45-54 showed a consistent increase in the number of people taking part in the study. As far as gender is concerned, there has been a consistency in the representation of two genders in the study group. There are slightly more women participating in this study as compared to men. There are certain groups formed based on the income of households included in the study. Four distinct groups are starting from people having an income of less than 10000 Hong Kong dollars. The last group for income is greater the 50000 Hong Kong dollars. The highest number of people fall in the income bracket of 10000-24999 representing 42.4% of the total subjects. The least representative group is the one that has income higher than 50000 Hong Kong dollars. This ranking of income groups prevails all through the study period.

A binary logistic regression model was used in the study. This method is used when a categorical variable which is usually dichotomous has to be predicted using a set of predictor variables. Another analysis that can be used is the discriminant function which is undertaken when the independent variables are continuous and nicely distributed. Logit regression is used in a situation where all the independent variables are categorical and logistic regression is used when the independent variables are a mix of categorical and continuous variables and if they are not nicely distributed. This study uses logistic regression which is already popular with those studies involving the presence or absence of some disease in some population. As in this study, logistic regression involves that the subjects are analyzed based on their set of scores on several independent variables. This study uses the dichotomous variable of presence or absence of diabetes in the subjects. There is no relationship between the independent variables showing good quality of research. The presence of such a relationship is referred to as multicollinearity. The R squared figure for the regression analysis was calculated at 19.8% which showed that the independent variables have explained 19.8% variation in the dependent variable. This figure is low and should be improved by adding more relevant variables to this study.

When the authors split the data, analysis based on age and gender, the increase in the prevalence of reporting in male respondents was lower than the females. The males showed an increase of 27.8 and 47.9% in 2005 and 2008 whereas females showed an increase of 31.8 and 69.3 % respectively. The lowest income groups showed the maximum increase in prevalence when age and gender variable were separately studied. As far as the individual groups are concerned, there were no significant differences among the respondents based on gender and age groups. Gender is found to have no association with the prevalence of diabetes but the association of disease and older age were more identified in the males as compared to females. Overall the self-report of Diabetes increased by 50% during the period studied.

As per the odd ratios, the males who are aged more than 65 years are the most like to have the disease which is represented by the AOR figure of 141.1.

The first limitation is that the self-reported prevalence of diabetes has not been reported in the past literature anywhere and using this method can be misleading. Self-reported diabetes has not been validated as opposed to that which is reported through a proper medical examination. Another limitation is that there are certainly other variables that affect diabetes and are not considered in this study. One variable of interest can be the person being over or underthe weight. As mentioned earlier, the explained variation for regression model as depicted by the R squared value of 19.8% is very low. This value can be improved by adding other relevant variables to the model. Another limitation is that the population in China is huge and this sample is not representative of the whole population.