Tutorial 5

1. The file `student.txt` contains results of a class survey of university students taking Statistics. There were 219 students who took part in the survey. The variables and their corresponding descriptions are given below. Answer all questions from using R, then repeat them in Python.

Variable	Description
id	Identifier code
gender	F = Female, M = Male
workhour	Hours worked per week at a paid job
drivelic	Driving licence (N = No, Y = Yes)
travel	Ever travel outside Asia? (N = No, Y = Yes)

   (a) Give labels to the values in the variables `gender`, `drivelic` and `travel` (for example 'M' can be labeled as 'Male').

   (b) Generate frequency counts for the variables `gender`, `drivelic` and `travel`.

   (c) Create a contingency table of frequency for `gender` and `drivelic`. Is having a driving licence independent of gender? Explain.

   (d) Create a categorical variable `whg` to represent the work hour group using the following rule:
   - The first group includes all the students with 0 work hour. Its label is "None (0 hrs)";
   - The second group includes all the students who work less than 20 hours per week. Its label is "Some (1 - 19 hrs)"; and
   - The third group includes all the students who work 20 or more hours per week. Its label is "Many (20 - 99 hrs)".

   (e) Perform a chi-square test to test the relationship between ever travel outside Asia and work hour group. Are these two variables statistically independent at significance level 0.05? Comments on the standardized residuals of the test and the suitability of the test for these two variables.

2. On the island of Samoa, obesity is a socially desirable trait. Investigators wished to study if obesity causes CVD on Samoan males. Hence they randomly sampled 1201 obese males on the island, and 1431 non-obese males (based on their BMI), and followed them for 20 years to observe if they died from CVD. The data are in the dataset `samoa.csv`. Use R to answer the question (a) - (d) below.

   (a) Form a contingency table for the data given. Which conditional proportion do you think is most informative for these data? Explain.

   (b) Estimate the disease odds ratio from the table. Interpret the value.

   (c) Form a 95% confidence interval for the true odds ratio.

   (d) Is this a retrospective or prospective study? If the latter, provide a point estimate of ralative risk and interpret it.

   (e) In Python, reapeat the steps that you have done with R for the questions above.