# STAT 210
# Applied Statistics and Data Analysis:
# Problem List 5
# (Due on week 6)

## Exercise 1

The data set `Auto` in the `ISLR` package has information on nine variables for 392 vehicles. For this question, we will be only interested in two of them, `mpg` and `origin`. The first variable, `mpg`, corresponds to fuel consumption in miles per gallon for each vehicle, while `origin` is coded as 1 (American), 2 (European), and 3 (Japanese).

(a) Create a data frame named `q2.df` with the two columns corresponding to these variables. Do a boxplot of `mpg` as a function of `origin`. Comment on what you observe.

(b) Using the information in `mpg`, add a factor `fmpg` to `q2.df` created according to the following rule: if `mpg` is below 20, the value for the factor is `low`; if `mpg` is between 20 and 35, the value is `med`, and if `mpg` is above 35, the value is `high`. One way to do this is using the function `cut`. Also, change the labels in the `origin` factor to `Am`, `Eu`, and `Jap`.

(c) Produce a table of `origin` and `fmpg` and do a mosaic plot. The table should have `origin` as rows and `fmpg` as columns. Comment on what you observe. Produce a second table with proportions calculated relative to the different levels of `origin`. Again, comment on what you observe.

(d) We want to determine whether the fuel consumption categories that we created are homogeneously distributed for the different origins of the vehicles.

- Which test or tests do you know that can be used for this?
- What are the underlying assumptions?
- Are they satisfied in this case?
- Carry out all the tests you mentioned and discuss the results.
- What are your conclusions?

## Exercise 2

The Southern State of the country of Nirvania, malaria is an endemic disease. Studies carried out show that 22% of the population in the state are infected. A new study in the city of Utopia, capital of the state, showed that among 1100 people tested, 198 were infected.

(a) The health authorities want to know if this result is in agreement with previous studies. Which would be your hypotheses in a statistical test? What tests do you know which apply in this situation? Explain why they are adequate and describe their underlying assumptions. Select a test or tests, apply them, and discuss in detail the results.

(b) Utopia is divided by a river into two regions, east and west Utopia. Of the sample used for the survey, 527 subjects were from east Utopia and of those, 89 were infected. Using this information, test whether there is a difference between east and west Utopia. Again, describe clearly the hypotheses you are testing, the reasons for choosing a particular test, the underlying assumptions, and discuss the results.

## Exercise 3

For this question use the data set `Titanic`. Open the help and get familiar with the data included in the set. Observe that this is not a data frame but a table, a type of structure we had not met before. We will focus on two variables, `Class` and `Survived`.

(a) Build a contingency table of `Survived` and `Class`. `Survived` should correspond to the rows of your table. (One way to do this table is to use the function `apply`).

(b) Do a mosaic plot for the table in (a). Differentiate the classes by shades of grey or different colors. Comment on what you observe on this graph.

(c) Add a margin row and column to the table with the corresponding totals.

(d) Build a table with the proportions with respect to the total number of persons in the table. Comment on the results.

(e) Build a table with the proportions with respect to the total number of passengers in each class. Comment on the results.

(f) We want to test whether the distribution of surviving passengers in the different classes is the same. What test would you use for this and why? What conditions need to be satisfied? Discuss whether they are in this example. Carry out this test and comment on your results.

## Exercise 4

In a certain country, the proportion of adults between 18 and 50 years that smoke 'frequently' is 22.5%. The authorities carry out a pilot one-year anti-tobacco campaign in a city and want to evaluate its effect on the proportion of 'frequent' smokers, to decide whether to extend the campaign country-wide.

(i) In a randomly chosen sample of 120 persons in the 18-50 years range, 17 are found to be 'frequent' smokers. Is there evidence of a decrease in the proportion of 'frequent' smokers in the city? Which would be your hypotheses in a statistical test? What tests do you know which apply in this situation? Explain why they are adequate and describe their underlying assumptions. Select a test or tests, apply them and discuss in detail the results.

(ii) In the survey, 65 persons were females and 9 of them were 'frequent' smokers. Look at the proportions for frequent smokers among the male and female populations and test whether there is evidence of a difference between the two. Again, describe clearly the hypotheses you are testing, the reasons for choosing a particular test, the underlying assumptions, and discuss the results.