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

For all tests in this list use a significance level of $\alpha = 0.02$.

Exercise 1

The data set `PL6-25_q1.csv` has information on 11 socioeconomic variables obtained in a survey of 534 subjects in 1985.

(a) Check whether there are any values missing in the data set. Create a new data frame called `df1` that has no missing values and includes only the variables `gender`, `occupation`, and `married`.

(b) Build a contingency table for `occupation` and `gender`. The rows of the table should correspond to `gender`. Add the totals by row and column. Represent this table as a mosaic plot and discuss the result.

(c) Using the table you built in (b), create a table for the same variables representing the proportions by `occupation`. Discuss the results.

(d) You have to test whether the distribution of occupations across the two genders 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.

(e) Build a contingency table for `occupation` and `married`. The rows of the table should correspond to `married`. Add the totals by row and column. Represent this table as a mosaic plot and discuss the result.

(f) Using the table you built in (e), create a table for the same variables representing the proportions by `occupation`. Discuss the results.

(g) You have to test whether the distribution of occupations across the two values for married 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 2

For all tests in this question, state explicitly the hypotheses, describe the assumptions behind the test, and explain why they are justified.

(a) Health authorities have determined that the proportion of people who smoke tobacco in the country of Arcadina is 23%. The capital city Arcadia carries out a six-month campaign against smoking, and a survey at the end of the campaign shows that out of 1200 persons interviewed, 242 smoke. Is there evidence that the campaign was effective? How would you test whether this? State clearly the hypothesis you are testing and describe the assumptions for the test or tests you propose. Why do you think that the assumptions are satisfied in this case? Carry out the test or tests and comment on the result(s).

(b) In the sample, 698 of those interviewed were males, and 165 said they were smokers. Is there evidence of a difference in the proportion of smokers between males and females?

(c) A similar survey was carried out simultaneously in the city of Avalon, where there was no anti-tobacco campaign, and out of 1150 persons interviewed, 283 were smokers. What test would you use to compare the proportion of smokers in the two cities? Carry out this test and discuss your results.

Exercise 3

For this question, we use the data set `data_q3.csv`. Read the data and store it in a data frame named `q3df`. We will only use two variables in this data set, `blood_type` and `Sugar_in_blood`. The first is an integer-valued variable with values between 1 and 4. The second is a numerical variable measuring the sugar level in blood in milligrams per deciliter (mg/dL). The blood type is coded as follows:

Code	Blood type
1	O
2	A
3	B
4	AB

(a) Create a new factor called `blood_factor` in `q3df` using the information in `blood_type` but using the letter code in the table.

(b) Do boxplots of `Sugar_in_blood` as a function of `blood_factor` and comment on the graph.

(c) We want to divide the data in `Sugar_in_blood` into three groups. Up to 87 mg/dL is `low`, above 87 and up to 100 is `normal`, while above 100 is `high`. Create a new ordered factor in `q3df` called `sugar_level` with this information and levels `low < normal < high`.

(d) Create a contingency table for `blood_factor` and `sugar_level`. Plot this information in a mosaic plot and discuss the results.

(e) We want to test if the proportions of the different values of `sugar_level` are the same for all values of `blood_factor`. What test would be adequate for this? State clearly which assumptions are needed and verify that they are satisfied. Carry out this test and discuss the results.

(f) Assume now that you only have two blood types, `A` and `B`, and that there are only two sugar levels, `normal` and `high`. You want to test if the proportion of `normal` and `high` sugar levels is the same for the different blood types. In this context, describe the null and alternative hypotheses and explicitly identify type I and type II errors. Describe the test statistic and the (asymptotic) sampling distribution.

Exercise 4

(a) A random sample of 200 recent blood donors at a certain blood bank shows that 68 were type A blood. Does this suggest that the actual percentage of type A donations differs from 42%, the percentage of the population having type A blood? What tests do you know that apply in this situation? Explain why they are adequate and describe their underlying assumptions. Carry out a test of the appropriate hypotheses using a significance level of .01. Would your conclusion have been different if a significance level of .05 had been used?

(b) A sample obtained at a different blood bank shows that out of 175 donors, 64 were of type A. Is there evidence to suggest that this proportion differs from that of part (a) of this question? Describe clearly the hypotheses you are testing, the reasons for choosing a particular test, and the underlying assumptions, and discuss the results.