STA 2201/442 Assignment 4

1. Based on a random sample of size n from a p-dimensional multivariate normal distribution, derive a formula for the large-sample likelihood ratio test statistic G^2 for the null hypothesis that $\boldsymbol{\Sigma}$ is diagonal (all covariances between variables are zero). You may use material on the slides from the multivariate normal lecture, which will be provided with the quiz if this question is asked. You may also use the fact that the unrestricted MLE is $\widehat{\theta} = (\overline{\mathbf{x}}, \widehat{\boldsymbol{\Sigma}})$, where

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

   Hint: Because zero covariance implies independence for the multivariate normal, the joint density is a product of marginals under H_0.

2. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are given in the file `sat.data`. There is a link on the course web page in case the one in this document does not work.

   Using R, carry out a large-sample likelihood ratio test to determine whether there are any non-zero covariances among the three variables; use $\alpha = 0.05$.

   (a) Calculate the test statistic G^2 and the p-value; also give the degrees of freedom. Your answer is a set of 3 numbers.

   (b) Do you reject H_0? Answer Yes or No.

   (c) Are the three variables all independent of one another? Answer Yes or No.

3. Under carefully controlled conditions, 120 beer drinkers each tasted 6 beers and indicated which one they liked best. Here are the numbers preferring each beer.

	Preferred Beer					
	1	2	3	4	5	6
Frequency	30	24	22	28	9	7

   This question calles for Wald tests, so the first part of your R printout should show the function that calculates the minus log likelihood, the MLE and the asymptotic covariance matrix.

(a) The first question is whether preference for the 6 beers is different in the population from which this sample was taken. Using R, conduct a Wald test to answer this question. Your printout should show the calculation of the test statistic, the degrees of freedom and the p-value. Do you reject H_0 at $\alpha = 0.05$?

(b) It seems that the first 4 beers are lagers and the last two are ales. No one would expect preference for lagers and ales to be the same.[1] So please use R to test whether preference for the 4 lagers is different, and at the same time, whether preference for the 2 ales is different. What is the null hypothesis? Use a Wald test. Your printout should show the calculation of the test statistic, the degrees of freedom and the p-value. Do you reject H_0 at $\alpha = 0.05$?

(c) Is *average* preference for lagers different from average preference for ales? Again, please display the calculation of the test statistic, the degrees of freedom and the p-value. Do you reject H_0 at $\alpha = 0.05$? If yes, which kind of beer is liked more on average?

4. A market researcher wants to know whether people are more likely to own desktop computers or laptops. But of course some own both and some own neither. So she selects a random sample of n adults, and records the data in an array with n rows and two columns of ones and zeros. Column one is for desktops and column two is for laptops, with a "1" indicating that the person owns that kind of machine, and a "0" indicating that the person does not own one.

   It is straightforward to estimate the probability of owning each type of computer, but we want a *test* comparing the two probabilities. There are quite a few reasonable ways to do this, and you should try to think of some. But suppose somebody says "Oh, just do a matched t-test."

   At first this might seem crazy, because the t-test is based on a normal distribution and these data are binary. But maybe if n is large it's not so crazy after all. Think about this carefully, do some calculations, and decide whether you think it's an acceptable way to analyze the data.

   Here is a good way to approach the problem. Start by proposing a believable statistical model – as opposed to the model underlying the t-test, which is unbelievable. The model will include a probability distribution for the data and a set of unknown parameters that determine the probability distribution. State the null hypothesis. Now, assuming that the null hypothesis is true, what happens to the t statistic asymptotically? Does it converge in distribution to something reasonable? Or does it go off to something unreasonable?

[1] Actually, I am making all this up with only a vague idea of what these terms mean.