STA 2101/442 Assignment 5[1]

1. Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x) = \frac{\theta e^{\theta(x-\mu)}}{(1+e^{\theta(x-\mu)})^2}$$

for x real, where $-\infty < \mu < \infty$ and $\theta > 0$. Here are some numerical data:

```
4.82  3.66  4.39  1.66  3.80  4.69  1.73  4.50  9.29  4.05  4.50 -0.64  1.40
4.18  2.70  5.65  5.47  0.55  4.64  1.19  2.28  7.16  4.80  3.19  2.33  2.57
2.31  0.35  2.81  2.35  2.52  3.44  2.71 -1.43  7.61  0.93  2.52  6.86  6.14
4.37  3.79  5.04  4.50  1.92  3.25 -0.06  2.81  3.09  2.95  3.69
```

You can read the data from
http://www.utstat.toronto.edu/~brunner/data/legal/mystery.data.txt.

   (a) Find the maximum likelihood estimates of μ and θ.
   (b) Obtain an approximate 95% confidence interval for θ.
   (c) Test $H_0 : \mu = 0$ at the $\alpha = 0.05$ significance level with a large-sample Z test.

   Do it the easiest way possible.

2. Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$, where $\alpha > 0$ and $\beta > 0$. Here are some numerical data:

```
0.45 0.42 0.38 0.26 0.43 0.24 0.32 0.50 0.44 0.29 0.45 0.29 0.29 0.32 0.30
0.32 0.30 0.38 0.43 0.35 0.32 0.33 0.29 0.20 0.46 0.31 0.35 0.27 0.29 0.46
0.43 0.37 0.32 0.28 0.20 0.26 0.39 0.35 0.35 0.24 0.36 0.28 0.32 0.23 0.25
0.43 0.30 0.43 0.33 0.37
```

You can read the data from
http://www.utstat.toronto.edu/~brunner/data/legal/beta.data.txt.

   (a) Find the maximum likelihood estimates of α and β.
   (b) Test $H_0 : \alpha = \beta$ at the 0.05 significance level with a large-sample likelihood ratio test. Give the p-value.
   (c) Test $H_0 : \alpha = \beta$ at the 0.05 significance level with a Wald test. You may use my `Wtest` function.

   If you are getting a lot of warnings, maybe it's because the numerical search is leaving the parameter space. If so, you can set bounds on the parameters with the `nlminb` function or the `maxLik` package.

3. Looking at the expression for the multivariate normal likelihood on the formula sheet, how can you tell that for *any* fixed positive definite $\boldsymbol{\Sigma}$, the likelihood is greatest when $\boldsymbol{\mu} = \overline{\mathbf{y}}$?

4. 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 the likelihood function on the formula sheet. You may also use without proof the fact that the unrestricted MLE is $\widehat{\theta} = (\overline{\mathbf{y}}, \widehat{\boldsymbol{\Sigma}})$.

   Hint: Because zero covariance implies independence for the multivariate normal, the joint density is a product of marginals under H_0. To be direct, I am suggesting that you *not* use the likelihood function on the formula sheet to calculate the numerator of the likelihood ratio. You'll eventually get the right answer if you insist on doing it that way, but it's a lot more work.

5. The file http://www.utstat.toronto.edu/~brunner/data/illegal/bp.data.txt has diastolic blood pressure, education, cholesterol, number of cigarettes per day and weight in pounds for a sample of middle-aged men. There are missing values; `summary` will tell you what they are.

   Assuming multivariate normality and using R, carry out a large-sample likelihood ratio test to determine whether there are any non-zero covariances among the five variables; guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Are the five variables all independent of one another? Answer Yes or No. For this question, let's agree that we will base the sample covariance matrix only on *complete observations*. That is, there will be no missing values on any variable. Don't forget that $\widehat{\boldsymbol{\Sigma}}$, like $\widehat{\sigma}^2_j$, has n in the denominator and not $n-1$. What is n?

6. Here is a useful variation on Problem 4. Suppose n independent and identically data vectors $\mathbf{d}_i = \begin{pmatrix} \mathbf{x}_i \\ \hline \mathbf{y}_i \end{pmatrix}$ are multivariate normal. The notation means that $\mathbf{d}_i$ is $\mathbf{x}_i$ stacked on top of $\mathbf{y}_i$. For example, $\mathbf{x}_i$ could be physical measurements and $\mathbf{y}_i$ could be psychological measurements. Derive a likelihood ratio test to determine whether $\mathbf{x}_i$ and $\mathbf{y}_i$ are independent. Your answer is a formula for G^2 and a formula for the degrees of freedom. Part of the job here is to make up good, simple notation.

7. Dead pixels are a big problem in manufacturing computer and cell phone screens. The physics of the manufacturing process dictates that dead pixels happen according to a spatial Poisson process, so that the numbers of dead pixels in cell phone screens are independent Poisson random variables with parameter λ, the expected number of dead pixels. Naturally, λ depends on details of how the screens are manufactured.

   In an effort to reduce the expected number of dead pixels, six assembly lines were set up, each with a different version of the manufacturing process. A random sample of 50 phones was taken from each assembly line and sent to the lab for testing. Mysteriously, three phones from one assembly line disappeared in transit, and 15 phones from another assembly line disappeared. Sample sizes and sample mean numbers of dead pixels appear in the table below.

	Manufacturing Process					
	1	2	3	4	5	6
ybar	10.68	9.87234	9.56	8.52	10.48571	9.98
n	50	47	50	50	35	50

(a) The first task is to carry out a large sample likelihood ratio test to see whether the expected numbers of dead pixels are different for the six manufacturing processes. Using R, calculate the test statistic and the p-value. Also report the degrees of freedom.

You are being asked for a computation, but *most of the task is thinking and working things out on paper*. I got away with only five lines of code: One line to enter the means, one line to enter the sample sizes, one line to compute G^2, one line to compute the p-value, and one other line. Here are some little questions to get you started.

i. Is this a between-cases design or a within-cases design?

ii. Denote the parameter vector by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$. What is p?

iii. What is the null hypothesis?

iv. What is the distribution of a sum of independent Poisson random variables?

v. What is the distribution of $n_j \overline{Y}_j$?

vi. What is the likelihood function? Write it down and simplify.

vii. What is the unrestricted MLE $\widehat{\boldsymbol{\lambda}}$? It's a vector. Work it out if you need to.

viii. What is the restricted MLE $\widehat{\boldsymbol{\lambda}}_0$? It's a vector. Work it out if you need to.

ix. Now you are ready to write the test statistic. There are a lot of cancellations. Keep simplifying!

x. Now use R to compute the test statistic and p-value. For comparison, my p-value is 0.01169133.

(b) Clearly we need to follow up this result to see where it came from, but we'll do Wald tests because they are a little easier. As preparation (and to get some exercise), carry out a Wald test of the overall null hypothesis you just tested above.

You'll need an estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\lambda}}$, and in this case expressing it in closed form is easier than obtaining and inverting the Hessian. So please do it the easy way. The questions below serve as a guide.

i. The asymptotic variance of $\overline{Y}_j$ is just its variance. What is $Var(\overline{Y}_j)$?

ii. The *estimated* asymptotic variance of $\overline{Y}_j$ is the most natural thing you can imagine. What is it?

iii. So what's the estimated asymptotic variance-covariance matrix of the random vector $\widehat{\boldsymbol{\lambda}}$?

Now you can carry out the Wald test. Do it with R, obtaining the test statistic, the degrees of freedom and the p-value.

(c) Finally carry out all pairwise comparisons with a Bonferroni correction, protecting the entire family of tests against Type I error at the *joint* $\alpha = 0.05$ significance level. In plain language, what do you conclude? Remember, with pairwise comparisons you can always draw directional conclusions. I realize I have not told you about Bonferroni corrections, but the article in the Wikipedia will tell you all you need to know.

8. The F distribution is defined as follows. Let $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ be independent, then the random variable $F = \frac{W_1/\nu_1}{W_2/\nu_2}$ is said to have an F distribution with ν_1 and ν_2 degrees of freedom. The formula sheet gives a statistic F^* for testing $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$. Using facts from the formula sheet (several of which you proved last week), show that F^* really does have an F distribution under the null hypothesis.