STA 2101/442 Assignment 6[1]

There is a lot of R work on this assignment (Questions 3, 5 and 6), so here are some rules about the printouts you bring to the quiz.

- Write nothing on the printouts by hand except possibly your name and student number.
- **Show full input as well as output on your printouts.** We need to see how you got your answers.
- You may indicate the question numbers in comment statements; in fact it may be a good idea.
- Comment statements may indicate the question you are trying to answer or the null hypothesis you are trying to test, but *not* the answer or conclusion.
- The rule is that you may include any reasonable comment you could have made **before seeing the results**.
- Of course comment statements may not have answers to the non-computer questions.
- You may compare numerical answers, but you may not look at anyone else's code or show anyone yours.
- It is acceptable to get help with your computer assignments from someone outside the class, but the help must be limited to general discussion and examples that are not the same as the assignment. *As soon as you get an outside person to actually start working on one of your computer assignments, you have committed an academic offence.*

The non-computer questions on this assignment are practice for the quiz on Friday October 20th, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz. As usual, you may use anything on the formula sheet unless you are directly asked to prove it.

1. Suppose that $\sqrt{n}\left(\mathbf{T}_n - \boldsymbol{\theta}\right) \stackrel{d}{\rightarrow} \mathbf{T} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$. We will say that $\mathbf{T}_n$ is *asymptotically normal* and use well-known properties of the multivariate normal, even though under the surface we are using Slutsky lemmas about convergence in distribution.
   (a) What is the asymptotic mean of $\mathbf{T}_n$?
   (b) What is the asymptotic covariance matrix of of $\mathbf{T}_n$?
   (c) What is the asymptotic distribution of $\mathbf{LT}$?
   (d) If $\boldsymbol{\theta}$ is $k \times 1$ and $\mathbf{L}$ is $r \times k$ with $r \leq k$, what conditions are needed for $\left(\mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^\top\right)^{-1}$ to exist?
   (e) Assuming those conditions are satisfied and that $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ is true, what is the asymptotic distribution of $\left(\mathbf{LT}_n - \mathbf{h}\right)^\top \left(\mathbf{L}\frac{1}{n}\boldsymbol{\Sigma}\mathbf{L}^\top\right)^{-1} \left(\mathbf{LT}_n - \mathbf{h}\right)$?

(f) If $\widehat{\mathbf{\Sigma}}_n \xrightarrow{p} \mathbf{\Sigma}$, give a useable statistic for a large-sample test of $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$.

2. Let $X_1, \ldots, X_n$ be a random sample from a $B(1, \theta)$ distribution.

   (a) Denoting the test statistic of Problem 1 by W_n, write down and simplify the W_n statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

   (b) Your answer is related to a Z-test of this same null hypothesis. Write down the formula for the Z statistic.

3. For the SAT (Scholastic Aptitude Test) data of Assignment Two (available here), suppose you are interested in testing whether mean performance is higher for the Verbal test or the Math test.

   (a) Using R, please calculate the W_n statistic to test this hypothesis. Feel free to use my `Wtest` function. Note that the statistic $\mathbf{T}_n$ is of dimension *two*. Guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Be able to state your conclusion in plain, non-statistical language. If a directional conclusion is possible, state it.

   (b) As a cross-check, carry out a matched t-test. How does it compare to the large-sample distribution-free test?

   Bring your R printout for this question to the quiz.

4. A team of botanists grew fungus in a nutrient solution in test tubes. Each day for seven days, one of their graduate students carefully measured the length of the fungus in each of n tubes. The scientists were interested in lots of things, including whether average growth was linear or not. Denote the expected amount of fungus at day j by μ_j.

   (a) What is the null hypothesis, in symbols?

   (b) Assuming that the scientists wish to make as few assumptions as possible and n is large, the W_n statistic is natural for this problem. What is $\mathbf{T}_n$?

   (c) What is $\mathbf{L}$?

   (d) What is $\mathbf{h}$?

   (e) What is a convenient choice for $\widehat{\mathbf{\Sigma}}_n$? How many rows and columns?

5. In a study of the psychology of attention, subjects attempted to solve word problems while listening to distracting background noise. The distracting material was either music, or spoken words related to the problem they were trying to solve. The distracting material was presented at three different levels of loudness. Each subject attempted 10 problems at each combination of loudness and type of distraction, for a total of 60 problems. Order of presentation was randomized. Data for each subject are number correct in each of the six treatment combinations. The data are available at

   http://www.utstat.utoronto.ca/~brunner/data/legal/distract.data.txt.
   See `help(read.table)` if necessary.

   (a) Produce a table showing the sample mean for each of the six treatment conditions.

   (b) Give a large-sample 95% confidence interval for each treatment mean.

   (c) Now test whether the six treatment means (expected values) are equal; as usual, $\alpha = 0.05$. You may use my `Wtest` function. Just to make sure we are doing things the same way, my test statistic value is $W_n = 757.293$. In plain, non-statistical language, what do you conclude?

   (d) Now we will compare *averages* of expected values. Those who have had a course in experimental design will recognize that we are testing differences between marginal means. Test the difference between the average expected test performance for Voice distraction and the average expected test performance for Music distraction. Be able to state a *directional* conclusion in plain, non-statistical language, if a conclusion is justified by the test.

   (e) Now just for Voice distraction, is there any effect of volume? Do the test and state a conclusion in plain language. Don't bother with follow-up tests yet; we'll do that later.

   (f) Now just for Music distraction, is there any effect of volume? Do the test and state a conclusion in plain language. Don't bother with follow-up tests yet; we'll do that later.

   Please bring your R printout for this question to the quiz.

6. For each of the following distributions and associated data sets, obtain the maximum likelihood estimate numerically with `R`. *Bring your printout for each problem to the quiz.*; you may be asked to hand it in. There are links to the data from the course web page in case the ones from this document do not work.

   (a) $f(x) = \frac{1}{\pi[1+(x-\theta)^2]}$ for x real, where $-\infty < \theta < \infty$. Data:

   ```
   -3.77  -3.57   4.10   4.87  -4.18  -4.59  -5.27  -8.33   5.55  -4.35  -0.55
   5.57 -34.78   5.05   2.18   4.12  -3.24   3.78  -3.57   4.86
   ```

   You can read the data from
   http://www.utstat.toronto.edu/~brunner/data/legal/cauchy.data.txt. For this one, try at least two different starting values and *plot the minus log likelihood function!*

(b) $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$, where $\alpha > 0$ and $\beta > 0$. 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. If you are getting a lot of warnings, maybe it's because the numerical search is leaving the parameter space. If so, try `help(nlminb)`.

(c) $f(x) = \frac{\theta e^{\theta(x-\mu)}}{(1+e^{\theta(x-\mu)})^2}$ for x real, where $-\infty < \mu < \infty$ and $\theta > 0$. 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.

(d) $f(x) = \frac{1}{m!} e^{-x} x^m$ for $x > 0$, where the unknown parameter m is a positive integer. *This means your estimate will be an integer.* Data:

```
 8.34  7.65  6.72  3.84  7.12  1.88  5.07  2.69  4.50  5.78  4.88  5.23  6.17
11.76  7.84  5.87  5.23  6.55  8.34  5.35  4.98 13.81  8.62  7.88  6.34  5.16
 6.64  4.35  6.77  5.83  5.85  2.46  8.33  3.74  5.10  3.95  7.84  4.70  6.09
 5.23  1.44  6.11  4.88  7.24  7.89  8.98  1.78  5.46  5.34  4.25
```

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

For each distribution, be able to state (briefly) why differentiating the log likelihood and setting the derivative to zero does not work. For the computer part, bring to the quiz one sheet of printed output for each distribution. The sheets should be separate, because you may hand only one of them in. Each printed page should show the following, *in this order*.

- Definition of the function that computes the likelihood, or log likelihood, or minus log likelihood or whatever.
- How you got the data into R – probably a `scan` statement.
- Listing of the data for the problem.
- The `nlm` or `nlminb` statement and resulting output.
- For the Cauchy example, a plot of the minus log likelihood.

7. 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.