STA 2101/442 Assignment Five[1]

Please bring your R printouts to the quiz. The non-computer questions are just practice for the quiz, and are not to be handed in, though you may use R as a calculator. Bring a real calculator to the quiz.

1. 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 might want to read the data from `cauchy.data`. 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 might want to read the data from `beta.data`. 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 might want to read the data from `mystery.data`.

(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 might want to read the data from `gamma.data`.

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.

2. For the data of Problem 1b, conduct a large-sample likelihood ratio test of $H_0 : \alpha = \beta$, using R. Your printout should display the value of G^2, the degrees of freedom and the p-value. Do you reject H_0 at the 0.05 significance level? If yes, which parameter seems to be higher based on $\widehat{\alpha}$ and $\widehat{\beta}$? *Bring your printout to the quiz.*

3. For a random sample from a Multinomial$(1, \boldsymbol{\theta})$ distribution, show that the likelihood ratio test statistic can be written as

$$G^2 = 2n \sum_{j=1}^{c} \overline{Y}_j \log \left(\frac{\overline{Y}_j}{\widehat{\theta}_j} \right),$$

where $\widehat{\theta}_j$ is the *restricted* maximum likelihood estimate of θ_j. That is, it's the MLE under H_0. You may use without proof the fact that the unrestricted MLE is $\overline{Y}_j$.

4. In yet another marketing study, consumers are given samples of three different laundry detergents, which we will call "A," "B" and "C." The labels that the people see are randomly assigned. Suppose $n = 150$ consumers participate in the product test. Thirty-eight prefer "A", 55 prefer "B,", and 57 prefer "C."

   (a) Test the null hypothesis of equal preference using a large-sample likelihood ratio test. Of course you may use Question 3.

   (b) Test the same null hypothesis using a Wald test. I used numerical maximum likelihood even though I knew the MLE.

   In both cases, use R. Your printout should display the value of the test statistic, the degrees of freedom and the p-value. Do you reject H_0 at the 0.05 significance level? *Bring your printout to the quiz.*

5. 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 (including a useful version of the likelihood function), which will be provided with the quiz if this question is asked. You may also use without proof 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.

6. This question uses the SAT data again: 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; guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Are the three variables all independent of one another? Answer Yes or No. *Bring your printout to the quiz.*