STA 2101/442 Assignment Four[1]

Please bring your R printouts for questions 8 and 9 to the quiz. The other 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. Ten friends have a party right after graduating from university. At the time, none of them has ever been married. The party includes a visit by a fortune teller, who says "Five years from now, 3 of you will still be unmarried, 3 of you will be married for the first time, 2 will be divorced, one will be married for the second time, and one will be widowed."

   How many ways are there for this to happen? The answer is a number. Show your work.

2. A fair die is tossed 8 times. What is the probability of observing the numbers 3 and 4 twice each, and the others once each? The answer is a number.

3. A box contains 5 red, 3 white and two blue marbles. A sample of six marbles is drawn with replacement. Find the probability that

   (a) 3 are red, 2 are white and one is blue

   (b) 2 are red, 3 are white and 1 is blue

   (c) 2 of each colour appears.

   All the answers are numbers.

4. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a $M\left(1, (\theta_1, \ldots, \theta_c)\right)$ distribution. Show why the likelihood function is written $L(\boldsymbol{\theta}) = \theta_1^{n_1}\theta_2^{n_2} \cdots \theta_c^{n_c}$.

5. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a $M\left(1, (\theta_1, \theta_2, \theta_3)\right)$ distribution. Find the maximum likelihood estimator of $(\theta_1, \theta_2, \theta_3)$. Show *all* your work.

6. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_1, \sigma_1^2)$ distribution, and $Y_1, \ldots, Y_n$ (same n) be a random sample from a $N(\mu_2, \sigma_2^2)$ distribution, independent of the first one. Find the asymptotic distribution of $\frac{\overline{X}_n}{\overline{Y}_n}$. What happens when $\mu_2 = 0$?

7. In a political poll, a random sample of registered voters indicate which party they generally like most: Conservative, NDP or Liberal (other preferences were indicated by a small number of respondents; they are excluded from this analysis). A multinomial model seems reasonable for these data, with n_1, n_2 and n_3 denoting the number who chose Conservative, NDP and Liberal respectively. Of course $n = n_1 + n_2 + n_3$.

   The *odds* of an event is the probability of the event divided by one minus the probability. Take the natural log and you have the *log odds*, a quantity that has a prominent role in categorical data analysis.

[1]Copyright information is at the end of the last page.

(a) Give consistent estimators of the log odds of supporting the Conservatives and the log odds of supporting the NDP. How do you know the estimators are consistent?

(b) Find the approximate large sample *joint* distribution of the two log odds estimators. Show your work. The covariance matrix has a fairly nice form.

(c) Express your answer to the last part by saying "They're approximately bivariate normal (what else?) with expected value ..."

(d) Suppose that in a random sample of 200 voters, 91 chose the Conservatives, 71 the NDP and 38 the liberals. Give

i. A point estimate of odds (not log odds) of choosing the NDP. The answer is one number.

ii. a 95% confidence interval for the log odds of choosing the NDP. The answer is a pair of numbers.

iii. Using your answer to the last part (the accepted way to do it), give a 95% confidence interval for the odds of choosing the NDP. The answer is a pair of numbers.

8. 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.

9. For the data of Problem 8b, 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 $\widehat{\beta}$?
