Homework 4 for STA 841

(Due in class November 5)

Problem 1

(To be done individually.) Download the data file `cancermap.dat` and the information file `cancermap.txt` from Blackboard. This is a 66×66 pixel map of Eastern US, with the value for each pixel being the cancer incidence rate at that location. For simplicity, the rates are represented as two classses high/low.

1. Plot this incidence map.

2. It is often believed that cancer rates have a spatial correlation. One of the simplest model to capture such relation is the following autologistic model

$$\text{logit}P(Y) = \alpha + \beta(Y_E + Y_S + Y_W + Y_N),$$

   where the binary rate Y (= high or low) at a pixel, depends on the rate of its four neighboring pixels (East, South, West and North) through a logistic model. Fit this model to the data, and interpret the results. (Because not every pixel have all four neighbors, you are going to have to make some choices to handle such "missing data". Justify your choices.)

3. In the data, there are a total of 2,293 pixels that correspond to Eastern US, and among these, 619 pixels have high cancer rate. If there is actually no sptial correlation in cancer rate, one shouldn't expect the 619 pixels to "cluster". One way to simulate under this null hypothesis is through permuting the high/low class label. Describe how this can be done. Based on this procedure, carry out a permutation test for the null hypothesis $H_0 : \beta = 0$.

4. The wind blows more from West to East than other directions. So maybe pixels are more similar to their neighbors on the East and West than they are to the North and South. Fit an autologistic model that allows for the effect described here. Comment on the practical significance of the coefficient values.

Problem 2

(To be done individually.) The following data were collected in a study of the relation between parental socioeconomic status (SES) and the mental health of children. Any systematic variation is of interest. Analyze the data and write a brief report (1 page at most).

	Mental health status			
Parents' SES	Well	Mild	Moderate	Impaired
A(high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F(low)	21	71	54	71

Problem 3

(To be done individually.) From McCullagh and Nelder (1989). Pick-6 is the weekly Illinois lottery in which the winning ticket comprises six unordered numbers in the range 1-54. The winning numbers are chosen by a physical randomizing device in which 54 numbered ping-pong balls are mixed by a draught of air in a closed transparent container. Six balls carrying the winning numbers are permitted to escape, one at a time, through a hole in the top of the apparatus. The frequency of occurrence of each the 54 numbers in a 12-month period is shown in the following table. (The data set `pick6.dat` can be downloaded from Blackboard.) By fitting appropriate models, test

	Units										
Tens	0	1	2	3	4	5	6	7	8	9	Total
0+		9	8	5	6	9	4	7	7	6	61
10+	5	5	6	3	5	4	5	5	3	2	43
20+	6	3	7	10	9	10	9	7	1	7	69
30+	11	6	2	9	10	4	8	4	9	6	69
40+	5	10	3	7	8	3	4	4	1	4	49
50+	3	5	4	4	5						21
Total	30	38	30	38	43	30	30	27	21	25	312

Source: Chicago Tribune, 14 Nov 1988.

the following hypotheses, all of which refer to the uniformity of the numbers generated by the randomizing device.

1. that the variation of the frequencies is consistent with the hypothesis of uniformity.
2. that the variation of the column totals is consistent with the hypothesis of uniformity.
3. that the variation of the row totals is consistent with the hypothesis of uniformity.
4. that the frequency of occurrence of the numbers 45-54 is the same as that of the remaining numbers.

You are now given the additional information that for the first 24 weeks only the numbers 1-44 were used: thereafter all 54 numbers were used. Test hypotheses (1.) and (2.) above making due

allowance for this change of regime after 24 weeks.
Note: For simplicity you can ignore the fact that the numbers are drawn without replacement.

Problem 4

(To be done individually.) Download the data file `fiji.dat` from Blackboard. This is a study on the relation between the number of children a woman (in Fiji of Indian race) has and three covariate variables—abode type(1: Urban, 2: Rural), education level (1: none, 2: lower elementary, 3: upper elementary, 4: secondary or higher), and years since first marriage (1: <5, 2: 5-9, 3: 10-14, 4: 15-19, 5: 20-24, 6: 25+). Each row corresponds to a particular covariate combination. The first column gives the mean number of children born per woman (rounded to two decimal places) and the second column gives the number of women with the corresponding covariate values.

1. Fit an appropriate model describing how the number of children varies with marital age, mother's abode and education. Give a brief synopsis of the arguments justifying your formulation and choice of model, including checks for model adequacy.

2. Explain the meaning of all parameters in your model. Comment on the major factors affecting the number of children.

3. Construct a 95% confidence interval for the mean number of children born to an urban woman with upper elementary education after ten years of marriage.

4. Estimate the lifetime average number of children born to rural women with secondary education. Give 90% confidence limits.