April 9, 2003

Sample Problems for Stat 770 Test

The following four problems are intended to be similar, in coverage and difficulty, to those which I might ask on the in-class test next Monday. (But there will be 2 or 3 problems on the test.)

(1). Consider a 3×3 table with row-levels A, B, C and column-levels 1, 2, 3, with cell-counts following a multinomial distribution with unknown cell-probabilities $\pi_{i,j}$ and table-total $n = 1000$, observed as follows:

$j =$ / $i =$	1	2	3	$Totals$
A	90	90	20	200
B	100	40	160	300
C	210	170	120	500
$Tot.$	400	300	300	1000

(a) Construct hypothesis tests (at the .05 level, using $\chi^2_{1,.05} = 3.84, \chi^2_{2,.05} = 5.99, \chi^2_{3,.05} = 7.81$) of the three successively more restrictive nested hypotheses (each time against the general alternative)

$$H_b: \ \pi_{1j} = \pi_{1+}\,\pi_{+j} \quad \forall \ j$$

$$H_a: \ H_b \cap \{\, \pi_{i1} = \pi_{+1}\,\pi_{i+} \,\} \quad \forall \ i$$

$$H_0: \ \pi_{ij} = \pi_{i+}\,\pi_{+j} \quad \forall \ i,\ j$$

(b) If all cell counts are divided by 3, how do the results change ?

(c) *(Extra)* If H_0 were actually valid, then what is the approximate probability that at least one of these three tests would reject ?

(2). Suppose that a series of 5 independent 2×2 tables are observed, each with the same row-categories (E =*Exposed* and N =*not Exposed*) and column-categories (D =*Disease* and H =*Healthy*), and that the k'th table corresponds to a value $X_k = k$ of a measured risk-factor for the disease. If the observed numbers Y_{ijk} can be assumed *Poisson*-distributed with parameter $\exp(a + bI_{[i=E]} + cX_k\, I_{[j=D]})$, then formulate the constant-relative-risk hypothesis

$$H_0: \ E(Y_{iDk})/E(Y_{iHk}) \qquad \text{is constant} \qquad \forall \ i,\ k$$

as a hypothesis-test for a GLM parameter, and explain as explicitly as you can how you would test it at level .05.

(3). A logistic regression fitted in Splus to a set of 60 data-triples (n_i, Y_i, X_i), with $5 \leq n_i \leq 10$ and $n_i Y_i \sim Binom(n_i, plogis(a + bX_i))$, yields the following output:

```
Coefficients:
(Intercept)           X
    -1.3412      0.3898
Degrees of Freedom: 59 Total (i.e. Null);  58 Residual
Null Deviance:      77.32
Residual Deviance: 63.82
```

and standardized coefficients

```
             Estimate Std. Error   z value
(Intercept) -1.3411556  0.1627916 -8.238484
 X           0.3897926  0.1063123  3.666486
```

NB. The true coefficient values in this *simulated* dataset were: -1, 0.2.

(a) Using the sufficient statistic values $\sum_{i=1}^{60} n_i Y_i = 132$ and $\sum_{i=1}^{60} n_i = 458$, find the maximized log-likelihood for the model with $b = 0$, and use the given deviance values to find the maximized log-likelihood for the fitted lgistic regression.

(b) Use the output information to find the 95% confidence intervals for e^a, e^b. Does the given output determine a confidence interval for e^{a+b} ?

(c) The deviance and Pearson residuals have brief summary statistics respectively given by

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.25100 -0.72460 -0.10640 -0.05093  0.62530  2.72200
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.87700 -0.68330 -0.10580  0.01181  0.64330  2.86800
```

What conclusions does this suggest about the fitted logistic model ? (What might you look at to get a clearer picture ?)

(4). Question about NR and Fisher Scoring and IRLS (in logistic regression setting.)