STA 215: Statistical Inference
HW 3 Due Wed Feb 29 2012

1. Consider $k_1 \times k_2$ two-way category counts $X = ((X_{ij}))_{i,j=1}^{k_1,k_2}$, modeled as $X \sim \mathsf{Multinomial}(n, p)$ where $p \in \Delta_{k_1 k_2}$. We will write p in a $k_1 \times k_2$ two-way table form: $p = ((p_{ij}))_{i,j=1}^{k_1,k_2}$. For a level-$\alpha$ testing of independence, i.e., $H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}$, $\forall i, j$, where $p_{i\cdot} = p_{i1} + \cdots + p_{ik_2}$ and $p_{\cdot j} = p_{1j} + \cdots + p_{k_1 j}$, one rejects H_0 if the corresponding Pearson's chi-square statistic $S(x)$ exceeds the $(1-\alpha)$-th quantile of the $\chi^2((k_1-1)(k_2-1))$ distribution. In truth the size of this test is only approximately α, and the quality of this approximation may be poor when n is not much larger than $k_1 k_2$. Fisher proposed an alternative way to carry out a level-α test, known as *Fisher's exact test of independence* which describe below.

   Let $\mathcal{T}_x$ contain all $k_1 \times k_2$ category counts $\tilde{x}$, with total n, such that $\tilde{x}_{i\cdot} = x_{i\cdot}$ and $\tilde{x}_{\cdot j} = x_{\cdot j}$ for every i, j, where $x_{i\cdot} = x_{i1} + \cdots x_{ik_2}$ and $x_{\cdot j} = x_{1j} + \cdots + x_{k_1 j}$ are the row and column totals of the observed count x, and $\tilde{x}_{i\cdot}$ and $\tilde{x}_{\cdot j}$ denote the same for $\tilde{x}$. The set $\mathcal{T}_x$ has finitely many elements, which we denote by $\tilde{x}^{(1)}, \cdots, \tilde{x}^{(M)}$.

   For each $\tilde{x}^{(m)} \in \mathcal{T}_x$, calculate $\tilde{s}^{(m)} = S(\tilde{x}^{(m)})$ and

$$w^{(m)} = \binom{n}{\tilde{x}_{11}^{(m)} \ \tilde{x}_{12}^{(m)} \ \cdots \ \tilde{x}_{k_1 k_2}^{(m)}},$$

   $m = 1, \cdots, M$. Calculate the normalized weights $\tilde{w}^{(m)} = w^{(m)} / [w^{(1)} + \cdots + w^{(M)}]$, $m = 1, \cdots, M$. Find a rearrangement $\{i_1, \cdots, i_M\}$ such that $\tilde{s}^{(i_1)} \le \tilde{s}^{(i_2)} \le \cdots \le \tilde{s}^{(i_M)}$. Find the largest m such that $\tilde{w}^{(i_1)} + \cdots + \tilde{w}^{(i_m)} < 1 - \alpha$. Reject H_0 if $S(x) > s^{(i_m)}$.

   Prove that Fisher's test indeed has size $\le \alpha$.

   [Hint: Despite the long description, the mathematics of checking size is fairly straightforward. BUT, notice that in calculating $P(S(X) > s^{(i_m)})$ both the statistic $S(X)$ and the bound $s^{(i_m)}$ depend on X.]

   [A real hint: Try proving that when p satisfies H_0, the conditional distribution of X given $X_{i\cdot} = x_{i\cdot}$, $X_{\cdot j} = x_{\cdot j}$, $i = 1, \cdots, k_1$, $j = 1, \cdots, k_2$ is the discrete distribution over $\mathcal{T}_x$ with the probability of $\tilde{x}^{(i)}$ being $\tilde{w}^{(i)}$.]

   [And a note: The strategy outlined above is not exactly what is followed in practice, simply because a complete enumeration of $\mathcal{T}_x$ is difficult. Instead, various Monte Carlo approximations are used to sample $\tilde{x}$ from $\mathcal{T}_x$.]

2. A researchers surveys n college students and counts how many support, how many oppose and how many are undecided about a recently introduced federal policy. Letting X_1, X_2, X_3 denote these counts, she models $X = (X_1, X_2, X_3)$ as $X \sim \mathsf{Multinomial}(n, p)$, $p \in \Delta_3$.

   (a) Give the p-value for testing $H_0 : p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ against $p_0 \ne (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ based on Pearson's chi-square tests for observed counts $X_1 = 140$, $X_2 = 165$, $X_3 = 195$.

   (b) The researcher wants to test whether the actual proportions of supporters and opposers in the entire college are equal. Give a mathematical formulation of this null hypothesis.

   (c) Find the restricted MLE of p under the null hypothesis in part (b) [i..e., maximize the likelihood function only over the null set].

   (d) Give the p-value for testing the null hypothesis in part (b) based on Pearson's chi-square tests, with same observed counts as in part (a).

3. Natality records $X_i = (\mathsf{Low_i}, \mathsf{Cigarettes_i}, \mathsf{Black_i})$, were collected for $i = 1, \cdots, n = 500$ births in the US in the month of June, 1997. The measurements were $\mathsf{Low_i} = 1$ if i-th birth record has birthweight < 2500g, $Y_i = 0$ otherwise; $\mathsf{Black_i}$) = (average) daily number of cigarettes smoked by the mother during pregnancy; $\mathsf{Black_i} = 1$ if the mother is African American and $\mathsf{Black_i} = 0$ otherwise. These data are modeled as

$$\log \frac{P(\mathsf{Low_i} = 1)}{1 - P(\mathsf{Low_i} = 1)} = \beta_1 + \beta_2 \mathsf{Cigarettes}_i + \beta_3 \mathsf{Black}_i + \beta_4 \mathsf{Cigarettes}_i \times \mathsf{Black}_i.$$

For the observed data x (see here), the `glm()` function in `R` produces

$$\hat{\beta}_{\text{MLE}}(x) = \begin{pmatrix} -3.170 \\ 0.079 \\ 1.064 \\ -0.003 \end{pmatrix}, \quad I_x^{-1} = \begin{pmatrix} 0.065 & -0.003 & -0.065 & 0.003 \\ -0.003 & 0.001 & 0.003 & -0.001 \\ -0.065 & 0.003 & 0.192 & -0.011 \\ 0.003 & -0.001 & -0.011 & 0.005 \end{pmatrix}$$

Give a mathematical formalization of the following hypotheses and calculate the corresponding p-values based on the Wald test:

(a) $\mathsf{Cigarettes}$ has no effect on the probability of low birthweight.

(b) $\mathsf{Cigarettes}$ has the same effect on the probability of low birthweight for African American and non-African American mothers.

4. Scalar observations $X_1, \cdots, X_n$ are modeled as $X_i = \mu + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} f \in \mathcal{F}_{sym}$, $\mu \in (-\infty, \infty)$. Here $\mathcal{F}_{sym}$ is the set of all pdfs on $\mathbb{R}$ that are symmetric around zero and have a finite variance. Consider two estimators $\bar{X}$ and X_{med} for μ. It is known that when $X_i = \mu_0 + \epsilon_i$ with $\epsilon_i \overset{\text{IID}}{\sim} f_0 \in \mathcal{F}_{sym}$ one has

$$\sqrt{n}(\bar{X} - \mu_0) \overset{d}{\to} N(0, \sigma^2(f_0)), \quad \sqrt{n}(X_{\text{med}} - \mu_0) \overset{d}{\to} N\left(0, \frac{1}{4 f_0(0)^2}\right),$$

where $\sigma^2(f_0) = \int x^2 f_0(x) dx$ is the variance under f_0. Calculate the asymptotic relative efficiency of X_{med} with respect to $\bar{X}$ for each of the following choices of f:

(a) $f = N(0,1)$, i.e., $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

(b) $f = Lap(0,1)$, i.e., $f(x) = \frac{1}{2} e^{-|x|}$

(c) $f = Logis(0,1)$, i.e., $f(x) = \frac{e^x}{(1+e^x)^2}$