Eric Slud Spring 2009

Handout/Worksheet on Asymptotics for 2×2 Tables

As described in class, we consider 2×2 tables of frequency counts n_{ij}, $i = 1, 2,\ j = 1, 2,$ and corresponding probabilities $p_{ij},\ i = 1, 2,\ j = 1, 2.$ These Tables have row and column totals denoted by similar notations,

$$n_{i\cdot} = \sum_{j=1}^{2} n_{ij} \quad , \qquad p_{i\cdot} = \sum_{j=1}^{2} p_{ij}$$

$$n_{\cdot j} = \sum_{i=1}^{2} n_{ij} \quad , \qquad p_{\cdot j} = \sum_{i=1}^{2} p_{ij}$$

where the table-totals are respectively

$$n \equiv \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \quad , \qquad 1 = \sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij}$$

Thus, the table entries p_{ij} form a doubly indexed probability vector. The frequency-count entries n_{ij} together with their marginal totals are regarded as random **data** gathered according to one of the following 5 sample designs.

Design 1. The table total n is fixed, and

$$(n_{11},\, n_{12},\, n_{21},\, n_{22}) \sim \text{Multinomial}(n,\, ((p_{11},\, p_{12},\, p_{21},\, p_{22}))$$

Design 2. Marginal totals $n_{1\cdot}$ and $n_{2\cdot}$ are fixed, $n = n_{1\cdot} + n_{2\cdot}$, and

$$n_{11} \sim \text{Binom}(n_{1\cdot},\, p_{11}/p_{1\cdot}) \qquad \text{indep. of} \qquad n_{21} \sim \text{Binom}(n_{2\cdot},\, p_{21}/p_{2\cdot})$$

Design 3. Marginal totals $n_{\cdot 1}$ and $n_{\cdot 2}$ are fixed, $n = n_{\cdot 1} + n_{\cdot 2}$, and

$$n_{11} \sim \text{Binom}(n_{\cdot 1},\, p_{11}/p_{\cdot 1}) \qquad \text{indep. of} \qquad n_{12} \sim \text{Binom}(n_{\cdot 2},\, p_{12}/p_{\cdot 2})$$

Design 4. Marginal totals $n_{1\cdot},\ n_{2\cdot},\ n_{\cdot 1},\ n_{\cdot 2}$ are fixed, $n = n_{1\cdot} + n_{2\cdot} = n_{\cdot 1} + n_{\cdot 2}$, and

$$n_{11} \sim \text{Ext.Hypergeom}\Big(\frac{p_{11}\, p_{22}}{p_{12}\, p_{21}},\, n,\, n_{1\cdot},\, n_{\cdot 1}\Big)$$

where the *Extended Hypergeometric* is a probability law on the nonnegative integers with probability mass function defined (for integer parameters $0 \le r,\, m \le n$) at $k \ge \max(0, r+m-n),\ k \le \min(r, m)$ by

$$\text{ExtHyp}(\vartheta,\, n,\, m,\, r)(k) \;=\; \vartheta^k \binom{m}{k} \binom{n-m}{r-k} \Big/ \sum_{j \ge 0} \vartheta^j \binom{m}{j} \binom{n-m}{r-j}$$

Design 5. None of the marginal totals is fixed, but a parameter n_0 is, and all table entries and marginal totals are Poisson-distributed random variables:

$$\text{for } i = 1,\, 2,\ j = 1,\, 2, \qquad n_{ij} \sim \text{Poisson}(n_0\, p_{ij}) \qquad \text{are independent}$$

We now consider the asymptotic probability distributions for large sample sizes of these 5 designs, and we show how they are related. Throughout, the probability vector $\mathbf{p} = (p_{11},\, p_{12},\, p_{21},\, p_{22})$ is fixed, and in those designs (#2,3,4) where some or all of the marginal totals are fixed, we assume respectively, as $n \to \infty$:

In design 2, $\quad n_{1\cdot}/n \to p_{1\cdot}$;
In design 3, $\quad n_{\cdot 1}/n \to p_{\cdot 1}$;
In design 4, $\quad n_{1\cdot}/n \to p_{1\cdot}$ and $n_{\cdot 1}/n \to p_{\cdot 1}$.

Facts to Check as Problems.

(I). (a) In all five designs, a maximal nonsingular (i.e. not linearly degenerate) subvector of $\mathbf{n} = (n_{11},\, n_{12},\, n_{21},\, n_{22})$ follows a natural and canonical exponential family distribution; and

(b) In all five designs, $P(n_{11} = k \,|\, n,\, n_{1\cdot},\, n_{\cdot 1}) = \text{ExtHyp}(p_{11}p_{22}/(p_{12}p_{21}), n\,,\, m,\, r)(k)$ is the same.

(II). In all five models except possibly Design 4, the sufficient vector statistic T found in **(I)(a)** follows a (multivariate) Central Limit Theorem with normalizing factor $1/\sqrt{n}$.

(III). In all five models except possibly Design 4, if

$$\vartheta \;\equiv\; \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}} \quad , \qquad \hat{\vartheta} \;\equiv\; \frac{n_{11}\, n_{22}}{n_{12}\, n_{21}}$$

then

$$\sqrt{n}\,(\log \hat{\vartheta} \;-\; \log \vartheta) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\Big(0,\, \sum_{i,j} 1/p_{ij}\Big)$$

Remarks. The objective of all of this work in (I)–(III) is to deduce the CLT result (II) and asymptotic distribution as in (III) for the Design 4 (Extended Hypergeometric) case, at least when the values $n_{1\cdot}$, $n_{\cdot 1}$ conditioned on are respectively not too far ($o(n^{2/3})$) away from $np_{1\cdot}$ and $np_{\cdot 1}$. We sketch here why that is possible. It depends on a slightly more refined form of the DeMoivre-Laplace Theorem (a so-called *local limit theorem*) given in the Feller (1957, p. 172, equation 2.18) reference saying that the *ratio* of Binom(n, p) to normal-approximating densities

$$P(a - 1/2 < X \leq a + 1/2) \Big/ \Big[\frac{1}{\sqrt{np(1-p)}}\, \phi(\frac{a - np}{\sqrt{np(1-p)}})\Big] \;\to\; 1$$

when n gets large if the 'deviation' ratio $(a - np)/\sqrt{np(1-p)}$ is of smaller order than $n^{-1/6}$. The idea is to use this result in connection with (I)(b) for Design 2, when $n_{1\cdot} = np_{1\cdot} + o(n^{2/3})$ and $n_{\cdot 1} = np_{\cdot 1} + o(n^{2/3})$.

Here are some steps:

(1^o) First, suppose that $n_{1\cdot} = m = m_n$ and $n_{2\cdot} = n - m = n - m_n$ form a sequence of (nonrandom) values where $m_n/n \to p_{1\cdot}$ as $n \to \infty$, and let $q_j = p_{j1}/p_{j\cdot}$, $j = 1, 2$. Then uniformly over all k, r such that $k - mq_1$, $r - k - (n - m)q_2 = o(n^{2/3})$,

$$P(n_{11} = k) \Big/ \Big(\Phi(\frac{k + 1/2 - mq_1}{\sqrt{mq_1(1-q_1)}}) \;-\; \Phi(\frac{k - 1/2 - mq_1}{\sqrt{mq_1(1-q_1)}})\Big) \;\to\; 1$$

$$P(n_{21} = r-k) \Big/ \Big(\Phi(\frac{r - k + 1/2 - (n-m)q_2}{\sqrt{(n-m)q_2(1-q_2)}}) - \Phi(\frac{r - k - 1/2 - (n-m)q_2}{\sqrt{mq_1(1-q_1)}})\Big) \;\to\; 1$$

and it is easy to check that the denominators in these expressions can be replaced by

$$\frac{1}{\sqrt{mq_1(1-q_1)}}\, \phi(\frac{k - mq_1}{\sqrt{mq_1(1-q_1)}}) \;\;, \quad \frac{1}{\sqrt{(n-m)q_2(1-q_2)}}\, \phi(\frac{r - k - (n-m)q_2}{\sqrt{(n-m)q_2(1-q_2)}})$$

(2^o) Second, it follows from (1^o), using independence of n_{11}, n_{21} under Design 2, that uniformly under the same range restrictions

$$P(n_{11}+n_{21}=r) \;=\; \Big\{ \sum_{k:|k-mq_1|,\,|r-k-(n-m)q_2|\le n^{5/8}} \frac{1}{\sqrt{mq_1(1-q_1)}}\,\phi\Big(\frac{k-mq_1}{\sqrt{mq_1(1-q_1)}}\Big)$$

$$\cdot\frac{1}{\sqrt{(n-m)q_2(1-q_2)}}\,\phi\Big(\frac{r-k-(n-m)q_2}{\sqrt{(n-m)q_2(1-q_2)}}\Big)\Big\}\;\cdot\;(1+o(1))$$

$$=\;\frac{1+o(1)}{\sqrt{mq_1(1-q_1)+(n-m)q_2(1-q_2)}}\,\phi\Big(\frac{r\;-\;mq_1\;-\;(n-m)q_2}{\sqrt{mq_1(1-q_1)+(n-m)q_2(1-q_2)}}\Big)$$

(3^o) Results (1^o) and (2^o) already imply that

$$P(n_{11}=k\,|\,n_{\cdot 1}=r) \;=\; P\Big(\big|\sqrt{mq_1(1-q_1)}\,Z_1-k+mq_1\big|\;\le\;1\;\Big|$$

$$\big|\sqrt{mq_1(1-q_1)}\,Z_1+\sqrt{(n-m)q_2(1-q_2)}\,Z_2-r+(mq_1+(n-m)q_2)\big|\;\le\;1\Big)$$

where Z_j are independent $\mathcal{N}(0,1)$ random variables. From this, after defining the ratio

$$\alpha \;=\; \sqrt{mq_1(1-q_1)}\,\Big/\,\Big\{mq_1(1-q_1)\;+\;(n-m)q_2(1-q_2)\Big\}^{1/2}$$

it is easy to check that $(n_{11}-mq_1)/\sqrt{mq_1(1-q_1)}$ given $n_{\cdot 1}=r$ is approximately normal (uniformly in $r-mq_1-(n-m)q_2=O(n^{5/8})$), with

$$\text{mean}\;\;=\alpha\,\frac{r-mq_1-(n-m)q_2}{mq_1(1-q_1)}\qquad\text{and variance}\;\;=\;\;1-\alpha^2$$

(4^o) From this point, it is a straightforward delta-method-exercise [**not** assigned as part of the exercise set on this worksheet] to check that, with r,m,n fixed subject to the requirements above, $\sqrt{n}\,(\hat{\vartheta}-\vartheta)$ has the same asymptotic normal distribution found for the other designs in part (III). Alternatively, you could use the result of (III), say for Design V together with the fact that this last conditional-normal essentially does not depend on any of the values r,m (only on their limiting ratios over n, embodied in probabilities p_{jk}) to deduce the result.