Handout/Worksheet on Asymptotics for $2 \times 2$ Tables

As described in class, we consider $2 \times 2$ 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,

$$n_i = \sum_{j=1}^{2} n_{ij}, \quad p_i = \sum_{j=1}^{2} p_{ij}$$

$$n_j = \sum_{i=1}^{2} n_{ij}, \quad p_j = \sum_{i=1}^{2} p_{ij}$$

where the table-totals are respectively

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

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 $n_1$ and $n_2$ are fixed, $n = n_1 + n_2$, and $n_{11} \sim \text{Binom}(n_1, p_{11}/p_1)$ indep. of $n_{21} \sim \text{Binom}(n_2, p_{21}/p_2)$

**Design 3.** Marginal totals $n_1$ and $n_2$ are fixed, $n = n_1 + n_2$, and $n_{11} \sim \text{Binom}(n_1, p_{11}/p_1)$ indep. of $n_{12} \sim \text{Binom}(n_2, p_{12}/p_2)$

**Design 4.** Marginal totals $n_1$, $n_2$, $n_1$, $n_2$ are fixed, $n = n_1 + n_2 = n_1 + n_2$, and

$$n_{11} \sim \text{Ext.Hypergeom}(\frac{p_{11} p_{22}}{p_{12} p_{21}}, n, n_1, n_1)$$
where the Extended Hypergeometric is a probability law on the nonnegative integers with probability mass function defined (for integer parameters $0 \leq r, m \leq n$) at $k \geq \max(0, r + m - n)$, $k \leq \min(r, m)$ by

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

**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:

for $i = 1, 2, j = 1, 2$, $n_{ij} \sim \text{Poisson}(n_0 p_{ij})$ 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 $p = (p_{11}, p_{12}, p_{21}, p_{22})$ 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, $n_{1.}/n \to p_1$;
- In design 3, $n_{1.}/n \to p_1$;
- In design 4, $n_{1.}/n \to p_1$ and $n_{.1}/n \to p_1$.

**Facts to Check as Problems.**

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

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

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

then

$$\sqrt{n} (\log \hat{\vartheta} - \log \vartheta) \overset{D}{\to} \mathcal{N}(0, \sum_{i,j} 1/p_{ij})$$
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 $n_1$, $n_1$ conditioned on are respectively not too far ($o(n^{2/3})$) away from $np_1$ and $np_1$. 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 $\text{Binom}(n, p)$ to normal-approximating densities

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

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 $n_1 = np_1 + o(n^{2/3})$ and $n_1 = np_1 + o(n^{2/3})$.

Here are some steps:

(1) First, suppose that $n_1 = m = m_n$ and $n_2 = n - m = n - m_n$ form a sequence of (nonrandom) values where $m_n/n \to p_1$, as $n \to \infty$, and let $q_j = p_j/p_1$, $j = 1, 2$. Then uniformly over all $k, r$ such that $k - mq_1, r - k - (n - m)q_2 = o(n^{2/3})$,

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

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

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

$$\frac{1}{\sqrt{mq_1(1 - q_1)}} \phi\left( \frac{k - mq_1}{\sqrt{mq_1(1 - q_1)}} \right), \quad \frac{1}{\sqrt{(n - m)q_2(1 - q_2)}} \phi\left( \frac{r - k - (n - m)q_2}{\sqrt{(n - m)q_2(1 - q_2)}} \right)$$
Second, it follows from (1\textsuperscript{o}), using independence of \( n_{11}, n_{21} \) under Design 2, that uniformly under the same range restrictions

\[
P(n_{11} + n_{21} = r) = \sum_{k:|k-mq_1,|r-k-(n-m)q_2|\leq n^{5/8}} \frac{1}{\sqrt{mq_1(1-q_1)}} \phi\left( \frac{k-mq_1}{\sqrt{mq_1(1-q_1)}} \right) \cdot \frac{1}{\sqrt{(n-m)q_2(1-q_2)}} \phi\left( \frac{r-k-(n-m)q_2}{\sqrt{(n-m)q_2(1-q_2)}} \right)
\]

\[
= \frac{1 + o(1)}{\sqrt{mq_1(1-q_1) + (n-m)q_2(1-q_2)}} \phi\left( \frac{r-mq_1-(n-m)q_2}{\sqrt{mq_1(1-q_1) + (n-m)q_2(1-q_2)}} \right)
\]

(3\textsuperscript{o}) Results (1\textsuperscript{o}) and (2\textsuperscript{o}) already imply that

\[
P(n_{11} = k \mid n_1 = r) = P\left( \left| \sqrt{mq_1(1-q_1)} Z_1 - k + mq_1 \right| \leq 1 \right|
\]

\[
|\sqrt{mq_1(1-q_1)} Z_1 + \sqrt{(n-m)q_2(1-q_2)} Z_2 - r+(mq_1+(n-m)q_2)| \leq 1 \}
\]

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

\[
\alpha = \frac{\sqrt{mq_1(1-q_1)}}{\sqrt{mq_1(1-q_1) + (n-m)q_2(1-q_2)}}^{1/2}
\]

it is easy to check that \( (n_{11} - mq_1)/\sqrt{mq_1(1-q_1)} \) given \( n_1 = r \) is approximately normal (uniformly in \( r - mq_1 - (n-m)q_2 = O(n^{5/8}) \)), with

mean \( = \alpha \frac{r-mq_1-(n-m)q_2}{mq_1(1-q_1)} \)

and variance \( = 1 - \alpha^2 \)

(4\textsuperscript{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.