A: 2.4

3. Testing in the General case (J or K greater than 2.)
   a. Score statistic in this case is Pearson Statistic
      i. Calculate expected values $E_{kj} = X_{j+}X_{+k}/X_{++}$
      ii. As in one-dimensional case, $T = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1}(X_{kj} - E_{kj})^2/E_{kj}$.
      iii. $T \sim \chi^2_{(K-1)(J-1)}$ (approximately) for all models, under null:
         - Independent Poisson
         - $J \times K$ category multinomial
         - Separate multinomials
            ▷ either by row or by column
         - Generalization of hypergeometric (to be shown below).
      iv. Same requirement of expectations > 5 .
      v. Under hypothesis for $\theta_{ij} \neq 1$,
         - $E_{00}E_{k\ell}/(E_{k0}E_{0\ell}) = \theta^{\circ}_{k\ell}$, and $E_{i+} = X_{i+}$ $E_{+j} = X_{+j}$.
         - No closed-form solution.
      vi. Note that T and refererence distribution do not depend on which variable you make rows, and which you make columns.
4. $J = K = 2$

A: 2.2.1

a. Square of proportion differences coincids with the Pearson statistic.

i. $Z^2 = T$ for Z the standard normal theory test statistic and $T = \sum_{j,k}(X_{jk} - E_{kj})^2/E_{kj}$.

ii. since

$$\begin{aligned} Z^2 &= [(\hat{\pi}_{10} - \hat{\pi}_{00})/\sqrt{\hat{\pi}_0\hat{\pi}_1/X_{+0} + \hat{\pi}_0\hat{\pi}_1/X_{+1}}]^2 \sim \chi^2_1 \\ &= \frac{X_{++}X_{0+}X_{1+}}{X_{+0}X_{+1}}(X_{00}/X_{0+} - X_{10}/X_{1+})^2 \\ &= \frac{X_{++}X_{0+}X_{1+}}{X_{+0}X_{+1}}(X_{00}(1/X_{0+} + 1/X_{1+}) - X_{+0}/X_{1+})^2 \\ &= \frac{X_{++}X_{0+}X_{1+}}{X_{+0}X_{+1}}(X_{00}X_{++}/[X_{0+}X_{1+}] - X_{+0}/X_{1+})^2 \\ &= (X_{00} - E_{00})^2 v \end{aligned}$$

iii. For $E_{kj} = X_{j+}X_{+k}/X_{++}$

iv. For $v = (X_{+1}X_{0+}X_{+0}X_{1+})^{-1}X^3_{++}$

$$\begin{aligned} &= \frac{X_{++}}{X_{+0}X_{0+}} + \frac{X_{++}}{X_{+0}X_{1+}} + \frac{X_{++}}{X_{+1}X_{0+}} + \frac{X_{++}}{X_{+1}X_{1+}} \\ &= \sum E^{-1}_{kj} \end{aligned}$$

v. Working backwards through the above calculations, v is inverse of variance of $X_{00} - E_{00}$

vi. Keep in mind that E_{00} is random.

vii. Note $(X_{kj} - E_{kj})^2$ is the same for all pairs i, j

viii. Use χ^2 test statistic as before: $T = \sum_{j,k=0}^{1}(X_{kj} - E_{kj})^2/E_{kj}$

- Expectation satisfies $E_{j+} = X_{j+}$ $E_{+k} = X_{+k}$, (3 equations, 4 unknowns) R Code SAS Code

5. Conditional Moments of Cell Counts

a. WOLOG calculate moment sfor first row and column.

b. $\mathrm{E}_{\theta=1}\left[X_{jk}|\text{margins}\right] = X_{+k}X_{j+}/X_{++}$

i. $\mathrm{E}\left[X_{00}|\text{margins}\right] = \sum_{X_{00}=\max(0,x_{+0}+x_{0+}-x_{++})}^{\min(X_{+0},X_{0+})} x_{00}\mathsf{P}\left[X_{00} = x_{0}\right.$

ii. Remove term with $x_{00} = 0$

iii. Cancel factors x_{00} in numerator and denominator.

iv. Reparameterize sum to $y = x_{00} - 1$.

v. Note that terms are $X_{+0}X_{0+}/X_{++}$ times hypergeometric probabilities with one fewer observations in first row and column.

c. $\mathrm{Var}_{\theta=1}\left[X_{jk}|\text{margins}\right] = \frac{(X_{++}-X_{j+})X_{j+}X_{+k}(X_{++}-X_{+k})}{X_{++}^2(X_{++}-1)}$

i. Consider $\mathrm{E}\left[X_{00}(X_{00} - 1)|\text{margins}\right]$

ii. Treat as in $\mathrm{E}\left[X_{00}|\text{margins}\right]$, except now cancelling two factors in numerator and denominator.

iii. Use $\mathrm{Var}\left[X_{00}|\text{margins}\right] = \mathrm{E}\left[X_{00}(X_{00}-1)|\text{margins}\right] + \mathrm{E}\left[X_{00}|\text{margins}\right] - \mathrm{E}\left[X_{00}|\text{margins}\right]^2$

d. For $j \neq \ell$, Cov $\left[X_{jk}X_{\ell k}\right] = -X_{j+}X_{\ell+}X_{+k}(X_{++} - X_{+k})/(X_{++}^2(X_{++}-1))$

i. Already know $\mathrm{Var}\left[X_{jk}\right]$

ii. Summation trick gives covariances for two entries in the same column.

- $\mathrm{Var}\left[X_{jk} + X_{\ell k}\right] = (X_{j+} + X_{\ell+})(X_{++} - X_{+j} - X_{+\ell})X_{+k}(X_{++} - X_{k+})/(X_{++}^2(X_{++}-1))$
- Cov $\left[X_{jk}, X_{\ell k}\right] = (\mathrm{Var}\left[X_{jk} + X_{\ell k}\right] - \mathrm{Var}\left[X_{jk}\right] - \mathrm{Var}\left[X_{\ell j}\right])/2$

e. For $m \neq k$, Cov $\left[X_{jk}, X_{jm}\right] = -X_{+k}X_{+m}X_{j+}(X_{++} - X_{j+})/(X_{++}^2(X_{++}-1))$

i. by symmetry.

f. For $j \neq \ell$, $k \neq m$, Cov $\left[X_{jk} + X_{\ell m}\right] = X_{+k}X_{+m}X_{j+}X_{\ell+}/(X_{++}^2(X_{++}-1))$

i. Expanding $\mathrm{Var}\left[X_{ij} + X_{il} + X_{kj} + X_{kl}\right]$ gives equation for Cov $\left[X_{ij}, X_{kl}\right]$ + Cov $\left[X_{kj}, X_{il}\right]$.

- These two covariances are the same, but I don’t see how to show this symmetry without brute-force calculation.

ii. Without loss of generality, take $k = j = 1$ and $i = l = 2$.

iii. For $\boldsymbol{y}$ and $\boldsymbol{z}$ three-component vectors of non-negative integers, let

- $\mathcal{A}(\boldsymbol{y}, \boldsymbol{z}) = \{(x_{00}, \ldots, x_{22}) | x_{ij} \geq 0, \sum_{i=0}^{2} x_{ij} = y_j \,\forall j, \sum_{j=0}^{2} x_{ij} = z_i \forall i\}$.
- $\mathcal{B}(\boldsymbol{y}, \boldsymbol{z}) = \{(x_{00}, \ldots, x_{22}) | x_{ij} \geq 0, x_{11} \geq 1, x_{22} \geq 1, \sum_{i=0}^{2} x_{ij} = y_j \forall j, \sum_{j=0}^{2} x_{ij} = z_i \,\forall i\}$
- $c(\boldsymbol{y}, \boldsymbol{z}) = \sum_{\boldsymbol{x} \in \mathcal{A}(\boldsymbol{y}, \boldsymbol{z})} \frac{1}{\prod_{i=0}^{2} \prod_{j=0}^{2} x_{ij}!}$.

iv. Note that

$$\mathsf{P}\left[X_{ij} = x_{ij} \,\forall i, j\right] = \frac{y_0! y_1! y_2! z_0! z_1! z_2!}{(y_0 + y_1 + y_2)! \prod_{i=0}^{2} \prod_{j=0}^{2} x_{ij}!}.$$

- Then $c(\boldsymbol{y}, \boldsymbol{z}) = \frac{(y_0+y_1+y_2)!}{y_0! y_1! y_2! z_0! z_1! z_2!}$.
- Let $\boldsymbol{e}_i$ be the three-component vector of all zeros except for 1 in component i.
- Then $\sum_{\boldsymbol{x} \in \mathcal{A}(\boldsymbol{y}, \boldsymbol{z})} \frac{x_{11} x_{22}}{\prod_{i=0}^{2} \prod_{j=0}^{2} x_{ij}!}$ is

$$= \sum_{\boldsymbol{x}\in\mathcal{B}(\boldsymbol{y},\boldsymbol{z})} \frac{1}{x_{00}!x_{10}!x_{20}!x_{01}!(x_{11}-1)!x_{21}!x_{02}!x_{12}!(x_{22}-1)!}$$

$$= \sum_{\boldsymbol{x}\in\mathcal{A}(\boldsymbol{y}-\boldsymbol{e}_1-\boldsymbol{e}_2,\boldsymbol{z}-\boldsymbol{e}_1-\boldsymbol{e}_2)} \frac{1}{x_{00}!\cdots x_{22}!}$$

$$= c(\boldsymbol{y}-\boldsymbol{e}_1-\boldsymbol{e}_2, \boldsymbol{z}-\boldsymbol{e}_1-\boldsymbol{e}_2)$$

- $\mathrm{E}\left[X_{11}X_{22}\right] = X_{1+}X_{+1}X_{2+}X_{+2}/(X_{++}(X_{++}-1))$

6. Use covariances to build correct quadratic form.

a. Define standardized quantities.

i. $Y_{ij} = (X_{ij} - X_{i+}X_{+j}/X_{++})\sqrt{X_{++}-1}/\sqrt{X_{i+}X_{+j}}$

ii. $\beta_i = \sqrt{X_{i+}/X_{++}}$

iii. $\gamma_j = \sqrt{X_{+j}/X_{++}}$

iv. $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

b. Cov $\left[Y_{ij}, Y_{kl}\right] = (\delta_{ik} - \beta_i\beta_k)(\delta_{jl} - \gamma_j\gamma_l)$

c. In matrix terms, Cov $\left[Y_{ij}, Y_{kl}\right] = (\boldsymbol{I} - \boldsymbol{\beta}^\top\boldsymbol{\beta}) \otimes (\boldsymbol{I} - \boldsymbol{\gamma}^\top\boldsymbol{\gamma})$

i. Operator $\otimes$ represents Kronecker product.

d. Hence covariance matrix for standardized cell counts is Kronecker product of matrices with same form as variance matrices for one-dimensional multinomial counts.

i. Presumes that

- the matrix Y_{ij} is turned into a vector,
- four-dimensional variance array compacted to two dimensions.

e. Take (generalized) inverse by inverting separate factors.

7. Here we approximate discrete distribution by continuous distribution

a. Probability of observed outcome must be added to the p value

b. On the raw obs scale, the lump has width 1

c. Again move upper corner by $\frac{1}{2}$ before calculating T

d. Normal approx. works poorly unless $E_{kj} \geq 5 \forall j, k$. See Fig. 10.

e. Unbalanced example. See Fig. 11.

8. Example of Eliminating Tables through Conditioning

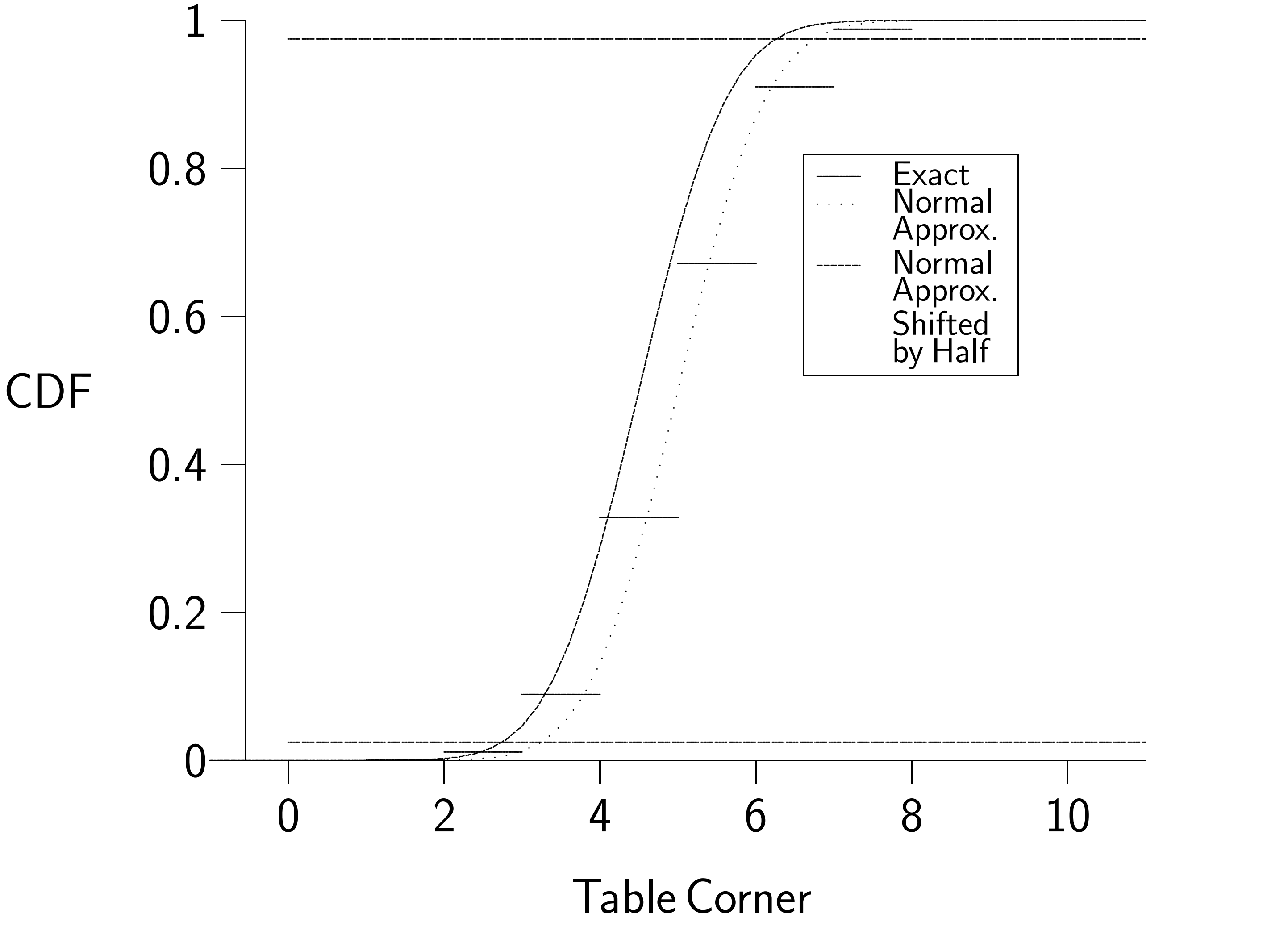
a. Observe table with 1,1 on diagonal, 0 elsewhere:

i. Sample space:

$$\begin{array}{cc|c} 0&0&0\\ 0&0&0\\ \hline 0&0&0 \end{array}\quad \begin{array}{cc|c} 1&0&1\\ 0&0&0\\ \hline 1&0&1 \end{array}\quad \begin{array}{cc|c} 0&1&1\\ 0&0&0\\ \hline 0&1&1 \end{array}\quad \begin{array}{cc|c} 0&0&0\\ 1&0&1\\ \hline 1&0&1 \end{array}\quad \begin{array}{cc|c} 0&0&0\\ 0&1&1\\ \hline 0&1&1 \end{array}\quad \begin{array}{cc|c} 2&0&2\\ 0&0&0\\ \hline 2&0&2 \end{array}$$

$$\begin{array}{cc|c} 0&2&2\\ 0&0&0\\ \hline 0&2&2 \end{array}\quad \begin{array}{cc|c} 0&0&0\\ 2&0&2\\ \hline 2&0&2 \end{array}\quad \begin{array}{cc|c} 0&0&0\\ 0&2&2\\ \hline 0&2&2 \end{array}\quad \begin{array}{cc|c} 1&1&2\\ 0&0&0\\ \hline 1&1&2 \end{array}\quad \begin{array}{cc|c} 1&0&1\\ 1&0&1\\ \hline 2&0&2 \end{array}\quad \begin{array}{cc|c} 1&0&1\\ 0&1&1\\ \hline 1&1&2 \end{array}$$

Fig. 10: Approximations to the Hypergeometric Distribution



Row margins are 10 , 10 and column margins are 10 , 10

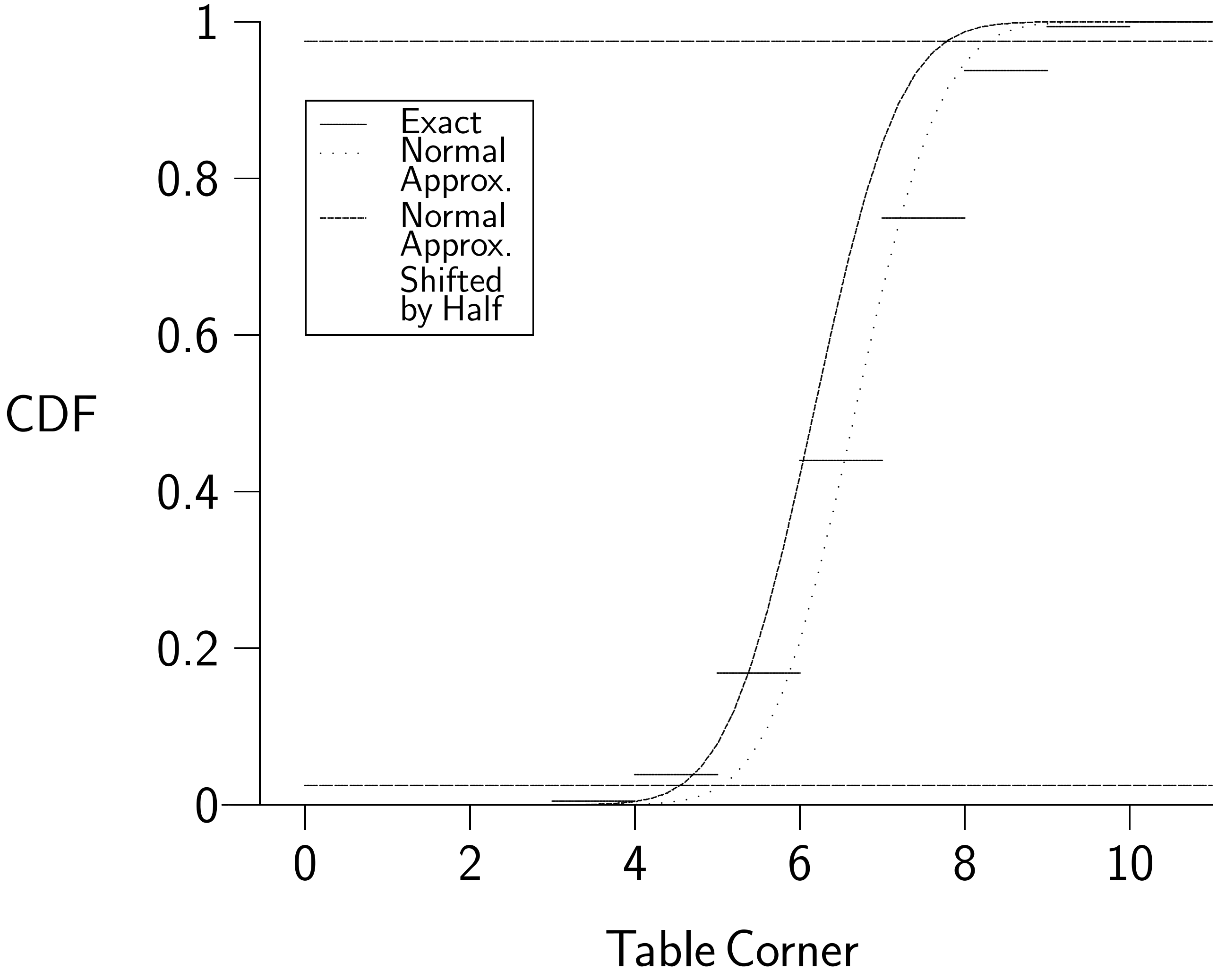
$$\begin{array}{cc|c} 0&1&1\\ 1&0&1\\ \hline 1&1&2 \end{array} \quad \begin{array}{cc|c} 0&1&1\\ 0&1&1\\ \hline 0&2&2 \end{array} \quad \begin{array}{cc|c} 0&0&0\\ 1&1&2\\ \hline 1&1&2 \end{array} \quad \begin{array}{cc|c} 3&0&3\\ 0&0&0\\ \hline 3&0&3 \end{array} \quad \begin{array}{cc|c} 0&3&3\\ 0&0&0\\ \hline 0&3&3 \end{array} \quad \begin{array}{cc|c} 0&0&0\\ 3&0&3\\ \hline 3&0&3 \end{array}$$

$$\begin{array}{cc|c} 0&0&0\\ 0&3&3\\ \hline 0&3&3 \end{array} \quad \begin{array}{cc|c} 2&1&3\\ 0&0&0\\ \hline 2&1&3 \end{array} \quad \begin{array}{cc|c} 2&0&2\\ 1&0&1\\ \hline 3&0&3 \end{array} \quad \begin{array}{cc|c} 2&0&2\\ 0&1&1\\ \hline 2&1&3 \end{array} \quad \begin{array}{cc|c} 1&2&3\\ 0&0&0\\ \hline 1&2&3 \end{array} \quad \cdots$$

b. Removed tables with total not 2, moving from 0 to 1,

c. Removed tables with row totals not 1,1, moving from 1 to 2.

Fig. 11: Approximations to the Hypergeometric Distribution



Row margins are 20 , 10 and column margins are 10 , 20

d. Removed tables with column totals not 1,1, moving from 2 to 3.

R Code SAS Code

A: 2.3.3

9. Likelihood ratio

a. $L(\theta)$ is probability for table as function of θ

b. Compare value at 1 to highest value it takes

c. $2 \times \log(L) \sim \chi^2$

d. Stratified cohort study (ie., condition on row totals).

i. Estimate π_j under H_0 as $\hat{\pi_j} = X_{+j}/X_{++}$.

A: 2.5

D. Ordered rows, unordered columns

1. Test using scores.

a. Test null hypothesis of equality of distribution using sum of squared columwise score statistics.

b. Rows scored to reflect ordering.

c. Assign row j a score u_j

d. Calculate columnwise sum $T_k = \sum_{j=0}^{K-1} u_j(X_{jk} - E_{jk})$

i. $\mathrm{E}\left[T_k\right] = 0$.

ii. Variance of scored statistic uses conditional covariances of table entries: $\mathrm{Var}\left[T_k\right] =$

$$= \frac{(X_{++} - X_{+k})X_{+k}}{X_{++}(X_{++} - 1)} \left\{ \sum_{j=0}^{K-1} u_j^2 \frac{X_{j+}}{X_{++}} - \left(\sum_{j=0}^{K-1} u_j \frac{X_{j+}}{X_{++}} \right)^2 \right\}$$

2. Squaring and rescaling makes columnwise sum $\approx \chi_1^2$

a. Rescaling is done using exact variance

b. Covariances of T_k use Cov $\left[X_{jk}, X_{li}\right]$.

c. Properly rescaled, $S = \sum_k T_k^2 / c_k \sim \chi^2_{K-1}$

i. Since $\sum_k T_k = 0$, the T_k are not independent.

d. S gives test of H_0 independence vs. H_A: some rows have column probabilities putting more weight on higher columns than low rows

3. Some alternative existing procedures.

a. Treating this as standard least–squares regression gives you reasonable SE for test statistic

i. Regresssing scores on 0 and 1 gives standard two–sample pooled t test

ii. Squaring $\hat{\beta}$/SE gives χ^2_1 statistic

4. Choice of score:

a. Additive constant washes out of test statistic when one subtracts expectation.

b. Spacing washes out of test statistic when one divides by the standard error.

c. By default these are equally spaced

d. Alternatively, one can use $Ridit\, scores$ $u_k = [\sum_{i<k} X_{i+} + (X_{k+} + 1)/2]/X_{++}$

i. Gives Mann–Whitney–Wilcoxon test

ii. Interpret test statistic as estimated probability that a random individual from one group scores higher than random individual from the other.

5. Ordered row and column categories

   a. Give scores for second dimension as well v_k

   b. Called *Mantel–Haenszel test*.

   c. When $J = 2$ or $K = 2$, called *Cochran–Armitage test*. R Code SAS Code

      i. this is the same as the previous example, with any second dimension scores

   d. Test is a multiple of correlation betw. row and column scores (1 for column k, and 0 for all other coluns):

   e. Calculate $T = \sum_k v_k T_k = \sum_{j=0}^{K-1} \sum_{k=0}^{J-1} v_k u_j (X_{jk} - E_{jk})$

   f. Multiple of correlation betw. row and column scores

   g. Squaring and rescaling makes it $\approx \chi_1^2$

      i. T gives test of H_0 independence vs. H_A: higher rows have column probabilities putting more weight on higher columns than low rows

ii. Since $\sum_k T_k = 0$, the T_k are not independent. R Code SAS Code

A: 2.7–2.7.3

IV. Controling for additional variables

A. Introduction

1. Additional variable provides an alternative explanation for association between disease and exposure

a. Add superscript i to tell which table

b. Phenomenon is called $confounding$.

c. Definition: distortion of disease/exposure association by other factor

i. Other factor related to exposure

ii. Other factor causally related to disease

$$\begin{tikzcd} C \arrow[r] \arrow[d] & D \\ E & \end{tikzcd}$$

d. Can change direction of relationship: $Simpson's\,Paradox$ (See example)

A: 2.7.4–2.7.5

e. Define the effect of exposure to be that with everything else held constant.

2. Definitions

   a. Split contingency table into separate tables defined by confounder

   b. Separate odds ratios are called $conditional\ odds\ ratios$

   c. Over-all odds ratio is called $marginal\ odds\ ratio$

   d. If distribution of exposure and disease are independent in each separate table, they are $conditionally\ independent$ $\iff$ conditional odds ratios are all 1.

   e. If conditional odds ratios are all the same, association between disease and exposure is $homogeneous$, even if the common odds ratio is not 1.

3. Example

   a. Aspirin is associated with stomach upset

   b. Does aspirin cause stomach upset?

   c. Alternative explanation: stress causes

      i. stomach upset

      ii. diseases like headaches for which aspirin is likely treatment.

   d. Direction of causation not indicated in an observational study

A: 2.7.6

B. Common odds ratios

1. Testing common odds ratio

   a. Hypotheses:

      i. Null hypothesis: all tables have a common odds ratio 1

      ii. Alternative hypothesis: All tables have a common odds ratio that is not 1.

A: 4.3.4

   b. Use $T = \sum_{i=1}^{I} w_i(X_{11}^i - E_{11}^i)$

      i. Intuition might suggest $w_i = 1/\sqrt{\mathrm{Var}\left[X_{00}^i|\text{margins}\right]}$

      ii. We will use $w_i = 1$

      iii. Use as standard error sum of exact variances.

      - Implies assumption that tables are independent.

   c. Called *Mantel–Haenszel test*.

   d. Is a score test for the stratified binomial model. R Code SAS Code

A1: 3.2.3

2. Estimation of the common odds ratio

   a. *Mantel–Haenszel estimator* $\hat{\theta} = \frac{\sum_{i=1}^{I} X_{00}^i X_{11}^i / X_{++}^i}{\sum_{i=1}^{I} X_{10}^i X_{01}^i / X_{++}^i}$

      i. ∞ only if all bottom products are 0

b. *logit estimator*

$$\hat{\theta} = \exp\left(\frac{\sum_{i=1}^{I} w_i \log(X_{00}^i X_{11}^i / [X_{10}^i X_{01}^i])}{\sum_{i=1}^{I} w_i}\right)$$

i. $w_i = (\frac{1}{X_{00}^i} + \frac{1}{X_{01}^i} + \frac{1}{X_{10}^i} + \frac{1}{X_{11}^i})^{-1}$

ii. Omit term i if $X_{jk}^i = 0$ for some j, k

- $w_i = 0$
- Corresponding logit will be ∞
- Acceptable since $\lim_{x \to 0} x \log(x) = 0$
- Alternative method is to add a bit to zero counts.

iii. This w_i minimizes variance

iv. SE of $\log(\hat{\theta})$ is $1/\sqrt{\sum_j w_j}$