CHAPTER 6

CONTINGENCY TABLES

Contingency tables are data arrays containing counts of observations that are recorded in cross classifications by a number of discrete factor predictors. The basic probability model for count data is the Poisson model; if N_C is the count in cross-category C, then

$$N_C \sim Poisson\,(\lambda_{0C})$$

say, and a natural GLM has

$$\mu_C = \log \lambda_{0C} = x^T \beta$$

that is a linear predictor to the canonical parameter μ_C, which is log of the naive parameter. This model is the **log-linear** model. The assumed Poisson model can be shown to be linked to the binomial/multinomial model.

6.1 CONTINGENCY TABLES AND SAMPLING MODELS

Contingency tables are summaries of response data where the predictor variables (exposures, confounders) are **discrete factors** or **categorical variables**, and where the responses are **counts**. We have studied simple 2 × 2 tables in epidemiology extensively: the data are counts of cases and controls in exposed and unexposed groups.

	E	E'	TOTAL
F	n_{11}	n_{12}	$n_{1\cdot}$
F'	n_{21}	n_{22}	$n_{2\cdot}$
TOTAL	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

(6.1.1)

In a cohort study, the sampling model is **product binomial**, that is, the **column totals** are fixed

$$n_{11} \sim Binomial\,(n_{\cdot 1}, \pi_1) \qquad\qquad n_{12} \sim Binomial\,(n_{\cdot 2}, \pi_0) \tag{6.1.2}$$

with the two sub-samples independent binomials with probabilities $\pi_1 = P\,(F|E)$ and $\pi_0 = P\,(F|E')$ respectively. In fact, in cohort studies we could also regard the cases and controls as independent binomial cohorts with the **row totals** fixed, and instead model

$$n_{11} \sim Binomial\,(n_{1\cdot}, \gamma_1) \qquad\qquad n_{21} \sim Binomial\,(n_{2\cdot}, \gamma_0) \tag{6.1.3}$$

where $\gamma_1 = P\,(E|F)$ and $\gamma_0 = P\,(E|F')$.

In cohort studies, a sample or population of size $n_{\cdot\cdot}$ is identified, and the exposure status for the individuals in the sample determined; these individuals are then followed up until their disease or morbidity status is established. Thus, we have an interpretation that implies either fixed cohort size, $n_{\cdot\cdot}$, or fixed column totals $(n_{\cdot 1}, n_{\cdot 2})$, or, indeed fixed row totals $(n_{1\cdot}, n_{2\cdot})$ and we have random Bernoulli/binomial/multinomial sampling.

In **case-control** studies, however, the situation is somewhat different: we only have a product binomial sampling model in the **rows** of the table, as implied by (6.1.3). As we have seen previously, it is the **row totals** that are fixed in the design, and permits inferences about (γ_1, γ_0) to be made; inference about (π_1, π_0) is not possible as the column totals are **not fixed** by the experimenter, and we do not have independent Bernoulli/binomial random sampling.

6.2 SAMPLING MODELS FOR $I \times J$ TABLES

In this section, we study the equivalence of Poisson sampling models and binomial/multinomial sampling models. Consider first the 2×2 table with cell entries N_{ij}, $i, j = 1, 2$, and an independent Poisson sampling model

$$N_{ij} \sim Poisson\left(\lambda_{ij}\right).$$

Suppose that the pattern amongst the Poisson rates is of interest. Note that the MLE of λ_{ij} in this unrestricted model is

$$\widehat{\lambda}_{ij} = n_{ij}$$

6.2.1 CONDITIONING AND NUISANCE PARAMETERS

In the standard epidemiological analysis of 2×2 tables, the main focus of interest are the conditional probabilities in (6.1.2) and (6.1.3), but not the marginal probabilities $P(E)$, $P(F)$ and so on. These parameters are essentially **nuisance** parameters, and little useful inference is lost by replacing them in an analysis by their **sufficient statistics**, that is, the row and column totals. Similarly, if only inference about the **degree of association** between the row and column factors is required, then the conditional probabilities are not directly of interest, but the odds-ratio

$$\psi = \frac{\pi_1/\left(1-\pi_1\right)}{\pi_0/\left(1-\pi_0\right)}$$

is, and then the extra conditioning on the **grand total** may be used to reduce the inference focus to a one-parameter problem.

The effect of different forms of conditioning may be studied first by regarding the overall rate parameter for inclusion in the study as a nuisance parameter (often it will not be the focus of interest). Note first that

$$N_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij} \sim Poisson\left(\lambda_{..}\right) \qquad \text{where} \ \ \lambda_{..} = \lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22}$$

We condition on the grand total $N_{..} = n_{..}$, and inspect the conditional distribution of the remaining three table entries;

$$f_{N_{11},N_{12},N_{21}|N_{..}}\left(n_{11}, n_{12}, n_{21}|n_{..}\right) = \frac{P\left[N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{..} = n_{..}\right]}{P\left[N_{..} = n_{..}\right]}$$

note that fixing $N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{..} = n_{..}$ automatically fixes $N_{22} = n_{..} - (n_{11} + n_{12} + n_{21}) = n_{22}$ say. Then

$$\begin{aligned} f_{N_{11},N_{12},N_{21}|N_{..}}\left(n_{11}, n_{12}, n_{21}|n_{..}\right) &= \frac{P\left[N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}\right]}{P\left[N_{..} = n_{..}\right]} \\ &= \frac{\left[\dfrac{e^{-\lambda_{11}}\lambda_{11}^{n_{11}}}{n_{11}!} \times \dfrac{e^{-\lambda_{12}}\lambda_{12}^{n_{12}}}{n_{12}!} \times \dfrac{e^{-\lambda_{21}}\lambda_{21}^{n_{21}}}{n_{21}!} \times \dfrac{e^{-\lambda_{22}}\lambda_{22}^{n_{22}}}{n_{22}!}\right]}{\dfrac{e^{-\lambda_{..}}\lambda_{..}^{n_{..}}}{n_{..}!}} \\ &= \binom{n_{..}}{n_{11}, n_{12}, n_{21}, n_{22}} \theta_1^{n_{11}}\theta_2^{n_{12}}\theta_3^{n_{21}}\theta_4^{n_{22}} \end{aligned}$$

where

$$\theta_1 = \frac{\lambda_{11}}{\lambda_{..}} \qquad \theta_2 = \frac{\lambda_{12}}{\lambda_{..}} \qquad \theta_3 = \frac{\lambda_{21}}{\lambda_{..}} \qquad \theta_4 = \frac{\lambda_{22}}{\lambda_{..}} = 1 - \theta_1 - \theta_2 - \theta_3$$

as all otther terms cancel, and where the first term is the multinomial coefficient

$$\binom{n_{..}}{n_{11}, n_{12}, n_{21}, n_{22}} = \frac{n_{..}!}{n_{11}!n_{12}!n_{21}!n_{22}!}$$

Hence, conditional on $N_{..} = n_{..}$, we have that

$$N_{11}, N_{12}, N_{21} | N_{..} = n_{..} \sim Multinomial\,(n_{..}, \theta_1, \theta_2, \theta_3)$$

Other forms of conditioning lead to different probability models For example, conditional on having the column totals fixed

$$N_{.1} = n_{.1} \qquad N_{.2} = n_{.2}$$

we have, using exactly similar techniques, we have that

$$\begin{aligned} N_{11} | N_{.1} &= n_{.1} \sim Binomial\left(n_{.1}, \frac{\lambda_{11}}{\lambda_{11} + \lambda_{21}}\right) \\ N_{12} | N_{.2} &= n_{.2} \sim Binomial\left(n_{.2}, \frac{\lambda_{12}}{\lambda_{12} + \lambda_{22}}\right) \end{aligned}$$

and also, conditional on having the row, column (and hence grand) totals fixed

$$N_{.1} = n_{.1} \qquad N_{1.} = n_{1.} \qquad N_{..} = n_{..}$$

we have that

$$N_{11} | N_{.1} = n_{.1}, N_{1.} = n_{1.}, N_{..} = n_{..} \sim HyperGeometric\,(n_{.1}, n_{1.}, n_{..}, \psi_\lambda)$$

so that

$$P\left[N_{11} = n_{11} | N_{.1} = n_{.1}, N_{1.} = n_{1.}, N_{..} = n_{..}\right] = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{.1} - n_{11}}}{\binom{n_{..}}{n_{.1}}} \psi_\lambda^{n_{11}}$$

where

$$\psi_\lambda = \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}$$

These results are proved in detail in section 6.5.2.

6.2.2 SUMMARY OF POSSIBLE MODELS OBTAINED BY CONDITIONING

The results from the previous section have profound implications for how we fit models to observational data. We have four possible sampling models that might arise in the context of contingency tables

1. **Model 1:** Independent **Poisson** samples within the cells of the table

$$N_{ij} \sim Poisson(\lambda_{ij})$$

2. **Model 2:** Independent **Poisson** samples within the cells of the table

$$N_{ij} \sim Poisson(\lambda_{ij})$$

but with the grand total

$$N_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$$

presumed **fixed**, and equal to $n_{..}$ say. Then, given $N_{..} = n_{..}$, $\{N_{ij}\}$ have a **multinomial** distribution with parameters $n_{..}$ and

$$\lambda_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}$$

3. **Model 3:** Independent **Poisson** samples within the cells of the table

$$N_{ij} \sim Poisson(\lambda_{ij})$$

but with the **row** totals

$$N_{i.} = \sum_{j=1}^{J} N_{ij}$$

or the **column** totals

$$N_{.j} = \sum_{i=1}^{I} N_{ij}$$

presumed **fixed**, and equal to $(n_{1.}, ..., n_{I.})$ and $(n_{.1}, ..., n_{.J})$ say. Then, given these row totals, $\{N_{ij} : j = 1, ..., J\}$ have a **multinomial** distribution within the rows with parameters $n_{i.}$ and

$$\left(\frac{\lambda_{i1}}{\lambda_{i.}}, \frac{\lambda_{i2}}{\lambda_{i.}}, ..., \frac{\lambda_{iJ}}{\lambda_{i.}}\right) \qquad \text{where} \qquad \lambda_{i.} = \sum_{j=1}^{J} \lambda_{ij}$$

with the data within rows **independent across rows**. Also given these column totals, $\{N_{ij} : i = 1, ..., I\}$ have a **multinomial** distribution within the columns with parameters $n_{.j}$ and

$$\left(\frac{\lambda_{1j}}{\lambda_{.j}}, \frac{\lambda_{2j}}{\lambda_{.j}}, ..., \frac{\lambda_{Ij}}{\lambda_{.j}}\right) \qquad \text{where} \qquad \lambda_{.j} = \sum_{i=1}^{I} \lambda_{ij}$$

with the data within columns **independent across columns**.

4. **Model 4:** Independent **Poisson** samples within the cells of the table

$$N_{ij} \sim Poisson(\lambda_{ij})$$

but with the **row** totals and the **column** totals

$$N_{i.} = \sum_{j=1}^{J} N_{ij} \qquad N_{.j} = \sum_{i=1}^{I} N_{ij}$$

presumed fixed, equal to $(n_{1.}, ..., n_{I.})$ and $(n_{.1}, ..., n_{.J})$ say. Then the $\{N_{ij} : i = 1, ..., I, j = 1, ..., J\}$ have a **multivariate hypergeometric** distribution. Under the independence assumption, the distribution is a **common** or **central** hypergeometric, whereas under a more general assumption the distribution is the **general** or **non-central** hypergeometric distribution.

These different sampling distributions might

- reflect the **data collection** mechanisms
- be utilized in order to construct **a more powerful analysis**; if the sample size is small, exact results in an **unconditional analysis** (model 4) might be not available (not enough data to estimate the parameters in the model), so that we are forced to carry out a **conditional** analysis that effectively reduces the number of parameters to be estimated.
- be utilized to **facilitate the statistical inference;** effectively, models 2 and 3 demonstrate the **equivalence** of statistical inference in the unconditional analysis (to estimate λ_{ij}) and a conditional analysis given grand, or row or column totals (to estimate π_{ij}).

6.3 THE EQUIVALENCE OF DIFFERENT SAMPLING MODELS FOR INFERENCE

The third point above is a **crucial** one; effectively, by conditioning on estimates of the "nuisance" parameters ($\lambda_{..}$ estimated by $n_{..}$ in the case of model 2, $(\lambda_{1.}, ..., \lambda_{I.})$ or $(\lambda_{.1}, ..., \lambda_{.J})$ estimated by $(n_{1.}, ..., n_{I.})$ or $(n_{.1}, ..., n_{.J})$ in the case of model 3) we can perform inference for the parameters of interest, that is, the cell specific probabilities and how they depend on row/column classifications, using the usual GLM framework.

EXAMPLE: For model 1, we might use a Poisson GLM with the canonical log link for the cell counts

$$\mu_{ij} = \log \lambda_{ij} = \beta_{ij} = x_{ij}^T \beta$$

and obtain ML estimates of β_{ij}, $\widehat{\beta}_{ij}$ say. Then we would have

$$\widehat{\lambda}_{ij} = \exp\left\{\widehat{\beta}_{ij}\right\} \qquad \widehat{\lambda}_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} \exp\left\{\widehat{\beta}_{ij}\right\}$$

and hence

$$\widehat{\pi}_{ij} = \frac{\widehat{\lambda}_{ij}}{\widehat{\lambda}_{..}} = \frac{\exp\left\{\widehat{\beta}_{ij}\right\}}{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} \exp\left\{\widehat{\beta}_{ij}\right\}}$$

is an estimate of the cell-specific probability π_{ij}

6.4 INFERENCE AND TESTING IN $I \times J$ TABLES

For two discrete (factor, predictor, categorical variables) X and Y with I and J levels respectively. The simplest type of sampling model is the **multinomial** model, where the $I \times J$ cross-categories each have an attached probability, and where structured models (involving main-effects or interactions say) for these probabilities can be used.

In the multinomial model, the total number of observations, $n_{..}$, is presumed **fixed**. This sampling model is not the only one that can be used, as discussed below, but is appropriate for the analysis of some epidemiological and biostatistical studies.

6.4.1 MULTINOMIAL MODELS AND INFERENCE

The multinomial model for an $I \times J$ table specifies a joint distribution over the discrete factors X and Y; here we might have a predictor variable (exposure, confounder) and a response, or merely two discrete variables. The joint mass function therefore is given by

$$P\left[X = i, Y = j\right] = \pi_{ij} \qquad i = 1, ..., I, \; j = 1, ..., J$$

where the marginal mass functions are given by

$$P\left[X = i\right] = \sum_{j=1}^{J} \pi_{ij} = \pi_{i.} \qquad (i = 1, ..., I) \qquad\qquad P\left[Y = j\right] = \sum_{i=1}^{I} \pi_{ij} = \pi_{.j} \qquad (j = 1, ..., J)$$

and

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1$$

Inference for the multinomial distribution is to be based in the cell counts $\{n_{ij} : i = 1, ..., I, j = 1, ..., J\}$. The likelihood for the multinomial model is

$$f_{n|\pi}\left(n; \pi\right) = \binom{n_{..}}{n_{11}, n_{12}, ..., n_{IJ}} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}} \tag{6.4.4}$$

so the log-likelihood is

$$\log l\left(\pi\right) = \log \binom{n_{..}}{n_{11}, n_{12}, ..., n_{IJ}} + \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \pi_{ij} \tag{6.4.5}$$

The functions in (6.4.4) and (6.4.5) will be the basis of likelihood and Bayesian inference.

Models for subsets of the data can also be deduced conditional on the total number of observations $n_{..}$. For an individual cell count,

$$n_{ij} | n_{..} \sim Binomial\left(n_{..}, \pi_{ij}\right)$$

For row i or column j

$$n_{i.} | n_{..} \sim Binomial\left(n_{..}, \pi_{i.}\right) \qquad\qquad n_{.j} | n_{..} \sim Binomial\left(n_{..}, \pi_{.j}\right)$$

and

$$n_{i1}, ..., n_{iJ} | n_{i.} \sim Multinomial\left(n_{i.}, \frac{\pi_{i1}}{\pi_{i.}}, ..., \frac{\pi_{iJ}}{\pi_{i.}}\right) \qquad n_{1j}, ..., n_{Ij} | n_{.j} \sim Multinomial\left(n_{.j}, \frac{\pi_{1j}}{\pi_{.j}}, ..., \frac{\pi_{Ij}}{\pi_{.j}}\right)$$

6.4.2 CHI-SQUARED AND LIKELIHOOD-RATIO TESTS

The multinomial model can be used in standard statistical hypothesis testing; two methods are commonly used for carrying out hypothesis tests of specific hypotheses. Suppose that a model implies (returns) fitted cell probabilities $\widehat{\pi}_{ij}$ and fitted cell entries $\widehat{n}_{ij} = n_{..}\pi_{ij}$. The following two test statistics can be used to test for model adequacy

- **PEARSON CHI-SQUARED GOODNESS-OF-FIT STATISTIC**:
  This is defined by

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \widehat{n}_{ij}\right)^2}{\widehat{n}_{ij}} \tag{6.4.6}$$

- **LIKELIHOOD RATIO/DEVIANCE STATISTIC**
  This Likelihood ratio statistic for a model against the saturated alternative, or deviance, is defined in the usual way by computing the maximized log-likelihood under the fitted model with that under the saturated model

$$LR = -2\log\frac{l_M\left(\widehat{\pi}_M\right)}{l_S\left(\widehat{\pi}_S\right)}.$$

  Here, the fit under the saturated model gives

$$\widehat{\pi}_{ij} = \frac{n_{ij}}{n_{..}} \qquad \widehat{n}_{ij} = n_{ij}$$

  and, using the log-likelihood for 6.4.5) we have

$$LR = -2\log\frac{l_M\left(\widehat{\pi}_M\right)}{l_S\left(\widehat{\pi}_S\right)} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\frac{n_{ij}}{\widehat{n}_{ij}} \tag{6.4.7}$$

  where n_{ij} is the fitted cell entry under the model M.

Under the model (if the model is correct), we complete the tests by noting that both (6.4.6) and (6.4.7) have asymptotic Chisquared distribution with $IJ - d$ degrees of freedom, where d is the total number of parameters fitted.

EXAMPLE: TESTS FOR INDEPENDENCE

Under the an **independence** model, we have a simple form for the cell probabilities in that, if X and Y are independent,

$$\pi_{ij} = \pi_{i.}\pi_{.j}$$

and the two marginal distributions are estimated separately from the data. The independence model presumes, for example, that the cell entries **within row** i, say, are independent *Binomial* $(n_{..}, \pi_{i.})$ random variables, as the likelihood becomes

$$f_{n|\pi}\left(n;\pi\right) \propto \prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{n_{ij}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\pi_{i.}\pi_{.j}\right)^{n_{ij}} = \left\{\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{i.}^{n_{ij}}\right\}\left\{\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{.j}^{n_{ij}}\right\} = \left\{\prod_{i=1}^{I}\pi_{i.}^{n_{i.}}\right\}\left\{\prod_{j=1}^{J}\pi_{.j}^{n_{.j}}\right\}$$

The row and column totals are therefore **sufficient statistics** for estimating the row and column probabilities, and

$$\widehat{\pi}_{i.} = \frac{n_{i.}}{n_{..}} \qquad \widehat{\pi}_{.j} = \frac{n_{.j}}{n_{..}}$$

and thus

$$\widehat{\pi}_{ij} = \widehat{\pi}_{i.}\widehat{\pi}_{.j} = \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} = \frac{n_{i.}n_{.j}}{n_{..}^2} \qquad \widehat{n}_{ij} = n_{..}\widehat{\pi}_{i.}\widehat{\pi}_{.j} = n_{..} \times \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} = \frac{n_{i.}n_{.j}}{n_{..}}.$$

Thus the LR statistic (deviance) for the independence model is

$$LR = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\frac{n_{ij}}{\widehat{n}_{ij}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\frac{n_{..}n_{ij}}{n_{i.}n_{.j}} = 2n_{..}\log n_{..} + 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\frac{n_{ij}}{n_{i.}n_{.j}}$$

and the number of fitted parameters is

$$I + (J-1) = I + J - 1$$

as we have, essentially, used the **main effects only** model $X + Y$, and

$$\log\pi_{ij} = \log\pi_{i.} + \log\pi_{.j}$$

The independence model , therefore, assumes that there is **no interaction** between row and column classification.

6.5 EXACT HYPOTHESIS TESTING

An **Exact** hypothesis test is a test which proceeds by calculating the exact probability distribution of the chosen test statistic under the null hypothesis, rather than using approximations to construct an approximate null distribution. For example, if we take the chi-squared statistic of equation (6.4.6), and consider corresponding random variable X^2 say, (with **random cell entries** N_{ij} replacing observed values n_{ij})

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij} - M_{ij})^2}{M_{ij}} \tag{6.5.8}$$

where the M_{ij} are themselves random quantities defined in accordance with the model being fitted; for example, under **independence**

$$\widehat{n}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}} = \frac{\left\{\sum\limits_{i=1}^{I} n_{ij}\right\}\left\{\sum\limits_{j=1}^{J} n_{ij}\right\}}{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} n_{ij}} \qquad \Longrightarrow \qquad M_{ij} = \frac{\left\{\sum\limits_{i=1}^{I} N_{ij}\right\}\left\{\sum\limits_{j=1}^{J} N_{ij}\right\}}{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} N_{ij}}.$$

In any case the random variable X^2 in (6.5.8) is the result of a multivariate transformation of the IJ variables $\{N_{ij}\}$, and its probability distribution is difficult to compute. Normal distribution theory, however, tells us that X^2 is approximately $Chi-squared((I-1)(J-1))$ distributed.

In summary, exact testing is merely a special type of hypothesis testing in which no approximations are made in the construction of the null distribution of the test statistic. Many tests for Normal samples and the Normal Linear Model (Z,T,F and ANOVA-F) are also exact tests, because the null distribution is available analytically in each case. Contrast this, for example, with the chi-squared goodness of fit test, the Score and Wald Tests, the Likelihood Ratio/Deviance test; in these cases, the null distribution is only available via a normal approximation.

Exact testing relies on being able to compute, either analytically or numerically, the probability distribution of the chosen test statistic under the assumptions of the null hypothesis. In the following sections we will examine some techniques for exact testing.

6.5.1 FISHER'S EXACT TEST IN 2×2 TABLES

In the special case where $I = J = 2$, exact methods lead to a specific test of the independence model. Suppose that the table in (6.1.1) is to be used to attempt to find evidence against the independence model. An **exact (conditional) test of independence** is based on examination of the cell entry n_{11} **conditional on fixed row, column and grand totals** $(n_{1.}, n_{.1}, n_{..})$; under the null hypothesis of independence the probability of observing n_{11} in cell $(1, 1)$ conditional on $(n_{1.}, n_{.1}, n_{..})$ is given by the **hypergeometric** formula

$$\frac{\binom{n_{1.}}{n_{11}}\binom{n_{..}-n_{1.}}{n_{.1}-n_{11}}}{\binom{n_{..}}{n_{.1}}} = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n_{.1}}} \tag{6.5.9}$$

(as, of course, we can deduce the values of $n_{2.}$ and n_{21}). In the more common notation of the hypergeometric distribution, we identify the values of (N, R, n, r) as $(n_{..}, n_{1.}, n_{.1}, n_{11})$. Conditional on the values of $(n_{1.}, n_{.1}, n_{..})$ we know that n_{11} must take a value that satisfies

$$\max\{0, n_{.1} - (n_{..} - n_{1.})\} \leq n_{11} \leq \min\{n_{1.}, n_{.1}\} \tag{6.5.10}$$

The hypothesis test for independence based on this conditioning on row and column totals leading to the hypergeometric distribution is known as **Fisher's Exact Test.**

6.5.2 PROOF OF EXACT CONDITIONING RESULTS

1. **Hypergeometric:** Suppose that $\{N_{ij} : i = 1, 2 \text{ and } j = 1, 2\}$ are independent Poisson random variables with parameters $\{\lambda_{ij} : i = 1, 2 \text{ and } j = 1, 2\}$respectively to fill the cells in a 2×2 table. Consider the new random variables (Y_1, Y_2, Y_3, Y_4) defined to be

$$\left.\begin{array}{l} Y_1 = N_{11} \\ Y_2 = N_{11} + N_{12} \\ Y_3 = N_{11} + N_{21} \\ Y_4 = N_{11} + N_{12} + N_{21} + N_{22} \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} N_{11} = Y_1 \\ N_{12} = Y_2 - Y_1 \\ N_{21} = Y_3 - Y_1 \\ N_{22} = (Y_4 - Y_3) - (Y_2 - Y_1) \end{array}\right.$$

that is, Y_2 is the row total for row 1, Y_3 is the column total for column 1, and Y_4 is the grand total. The joint mass function of $(N_{11}, N_{12}, N_{21}, N_{22})$ is

$$f_{N_{11},N_{12},N_{21},N_{22}}(n_{11}, n_{12}, n_{21}, n_{22}) = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{\exp\{-\lambda_{ij}\}\lambda_{ij}^{n_{ij}}}{n_{ij}!}$$

which, after a variable transformation, gives the joint distribution for (Y_1, Y_2, Y_3, Y_4) is given by

$$\frac{\exp\{-\lambda_{11}\}\lambda_{11}^{y_1}}{y_1!} \times \frac{\exp\{-\lambda_{12}\}\lambda_{12}^{y_2-y_1}}{(y_2-y_1)!} \times \frac{\exp\{-\lambda_{21}\}\lambda_{21}^{y_3-y_1}}{(y_3-y_1)!} \times \frac{\exp\{-\lambda_{22}\}\lambda_{22}^{(y_4-y_3)-(y_2-y_1)}}{((y_4-y_3)-(y_2-y_1))!}$$

or, on rearrangement

$$\frac{\left(\frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}\right)^{y_1}\left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2}\left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_3}\lambda_{22}^{y_4}}{y_1!\,(y_2-y_1)!\,(y_3-y_1)!\,((y_4-y_3)-(y_2-y_1))!}\exp\{-\lambda_{..}\} \tag{6.5.11}$$

where $\lambda_{..} = (\lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22})$. The marginal distribution for (Y_2, Y_3, Y_4) is then obtained from (6.5.11) by summing out over y_1 on the range given by (6.5.10). Now under independence, for each (i, j)

$$\lambda_{ij} = \lambda_{i.}\lambda_{.j} \qquad \Longrightarrow \qquad \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}} = 1$$

$$\begin{aligned} \lambda_{..} &= \lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22} = \lambda_{1.}\lambda_{.1} + \lambda_{1.}\lambda_{.2} + \lambda_{2.}\lambda_{.1} + \lambda_{2.}\lambda_{.2} \\ &= (\lambda_{1.} + \lambda_{2.})(\lambda_{.1} + \lambda_{.2}) \qquad (6.5.12) \end{aligned}$$

so the sum required is

$$\sum_{y_1} \frac{1}{y_1!\,(y_2 - y_1)!\,(y_3 - y_1)!\,((y_4 - y_3) - (y_2 - y_1))!}.$$

Now, for convenience, we will introduce the hypergeometric notation, with $r = y_1, R = y_2, n = y_3, y_4 = N$, so that the sum becomes

$$\begin{aligned} \sum_{y_1} \frac{1}{r!\,(R - r)!\,(n - r)!\,((N - n) - (R - r))!} &= \frac{1}{R!(N - R)!} \sum_{y_1} \binom{R}{r} \binom{N - R}{n - r} \\ &= \frac{1}{R!(N - R)!} \binom{N}{n} \end{aligned}$$

from the hypergeometric distribution properties. Hence

$$\begin{aligned} f_{Y_2,Y_3,Y_4}(y_2, y_3, y_4) &= \frac{1}{y_2!(y_4 - y_2)!} \binom{y_4}{y_3} \left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2} \left(\frac{\lambda_{21}}{\lambda_{22}}\right)^{y_3} \lambda_{22}^{y_4} \exp\{-\lambda_{..}\} \qquad (6.5.13) \\ &= \binom{y_4}{y_2} \binom{y_4}{y_3} \left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2} \left(\frac{\lambda_{21}}{\lambda_{22}}\right)^{y_3} \frac{\lambda_{22}^{y_4} \exp\{-\lambda_{..}\}}{y_4!} \qquad (6.5.14) \end{aligned}$$

with constraints on the variables $0 \leq y_2, y_3 \leq y_4$. Hence the conditional distribution of Y_1 given Y_2, Y_3, Y_4 is

$$\begin{aligned} f_{Y_1|Y_2,Y_3,Y_4}(y_1|y_2, y_3, y_4) &= \frac{f_{Y_1,Y_2,Y_3,Y_4}(y_1, y_2, y_3, y_4)}{f_{Y_2,Y_3,Y_4}(y_2, y_3, y_4)} \\ &= \frac{1}{y_1!\,(y_2 - y_1)!\,(y_3 - y_1)!\,((y_4 - y_3) - (y_2 - y_1))!} \frac{y_2!(y_4 - y_2)!}{\binom{y_4}{y_3}} \\ &= \frac{\binom{y_2}{y_1}\binom{y_4 - y_2}{y_3 - y_1}}{\binom{y_4}{y_3}} \qquad \max\{0, y_3 - (y_4 - y_2)\} \leq y_1 \leq \min\{y_2, y_3\} \end{aligned}$$

that is, a hypergeometric distribution.

2. **Marginal Totals:** As noted above it can be deduced from elementary probability theory (mgfs, convolution) that

$$Y_4 \sim Poisson\left(\lambda_{..}\right) \qquad f_{Y_4}\left(y_4\right) = \frac{\exp\left\{-\lambda_{..}\right\}\lambda_{..}^{y_4}}{y_4!} \qquad y_4 \geq 0 \tag{6.5.15}$$

and hence, conditionally on $Y_4 = y_4$, we have that, from (6.5.14) and (6.5.15)

$$\begin{aligned} f_{Y_2,Y_3|Y_4}\left(y_2, y_3|y_4\right) &= \frac{f_{Y_2,Y_3,Y_4}\left(y_2,y_3,y_4\right)}{f_{Y_4}\left(y_4\right)} = \frac{\binom{y_4}{y_2}\binom{y_4}{y_3}\left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2}\left(\frac{\lambda_{21}}{\lambda_{22}}\right)^{y_3}\frac{\lambda_{22}^{y_4}\exp\left\{-\lambda_{..}\right\}}{y_4!}}{\frac{\exp\left\{-\lambda_{..}\right\}\lambda_{..}^{y_4}}{y_4!}} \\ &= \binom{y_4}{y_2}\binom{y_4}{y_3}\left(\frac{\lambda_{12}}{\lambda_{22}}\right)^{y_2}\left(\frac{\lambda_{21}}{\lambda_{22}}\right)^{y_3}\left(\frac{\lambda_{22}}{\lambda_{..}}\right)^{y_4} \\ &= \binom{y_4}{y_2}\binom{y_4}{y_3}\left(\frac{\lambda_{1.}}{\lambda_{2.}}\right)^{y_2}\left(\frac{\lambda_{.1}}{\lambda_{.2}}\right)^{y_3}\left(\frac{\lambda_{2.}\lambda_{.2}}{\left(\lambda_{1.}+\lambda_{2.}\right)\left(\lambda_{.1}+\lambda_{.2}\right)}\right)^{y_4} \end{aligned}$$

because of (6.5.12). This may be re-written by

$$\begin{aligned} f_{Y_2,Y_3|Y_4}\left(y_2, y_3|y_4\right) &= \binom{y_4}{y_2}\left(\frac{\lambda_{1.}}{\lambda_{1.}+\lambda_{2.}}\right)^{y_2}\left(\frac{\lambda_{2.}}{\lambda_{1.}+\lambda_{2.}}\right)^{y_4-y_2} \\ &\times\binom{y_4}{y_3}\left(\frac{\lambda_{.1}}{\lambda_{.1}+\lambda_{.2}}\right)^{y_2}\left(\frac{\lambda_{.2}}{\lambda_{.1}+\lambda_{.2}}\right)^{y_4-y_3} \end{aligned}$$

and hence, **given** $Y_4 = y_4$, Y_2 and Y_3 are **independent Binomial random variables**

$$Y_2|Y_4 = y_4 \sim Binomial\left(y_4, \frac{\lambda_{1.}}{\lambda_{1.}+\lambda_{2.}}\right) \qquad Y_3|Y_4 = y_4 \sim Binomial\left(y_4, \frac{\lambda_{.1}}{\lambda_{.1}+\lambda_{.2}}\right).$$

EXTENSIONS: Each of these results can be extended to the general $I \times J$ table case under independence; conditional on fixed row and column totals, the distribution of the $(I-1)(J-1)$ undetermined components is **multivariate hypergeometric;** conditional on the grand total, the row and column totals have independent **multinomial** distributions; the unconditional distribution of the grand total is **Poisson.**

For the 2×2 table:

Quantity of Interest	Parameters of Interest	Conditioning
Rates	$\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}$	None
Cell Probabilities	$\pi_{11}, \pi_{12}, \pi_{21}$	Grand Total
Column Conditional Probs	π_1, π_0	Column Totals
Row Conditional Probs	γ_1, γ_0	Row Totals
Association	ψ	Row and Column Totals

6.5.3 EXACT TESTING VIA SIMULATION: MONTE CARLO (EXACT) METHODS

Computing the null distribution, in fact the null cdf, is essential for carrying out exact tests. In Fisher's exact test for a 2 × 2 table, the null mass function is given by (6.5.9), and computing the null cdf is straightforward, if potentially laborious; we must evaluate the expression

$$\frac{\binom{n_{1.}}{x}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n_{.1}}}$$

for all values of x that satisfy $\max\{0, n_{.1} - (n_{..} - n_{1.})\} \leq x \leq \min\{n_{1.}, n_{.1}\}$. In practice, in a one-tailed test or to compute a p-value, we only have to evaluate this expression for a restricted range of values of x, that is, for x not greater than the observed cell entry n_{11}

In a general $I \times J$ table, and for conditional testing, the **multivariate** hypergeometric cdf must be computed for all relevant tables that are consistent with the fixed row and column totals; unfortunately, the number of such tables is potentially huge. More generally, for unconditional testing, the null cdf may be impossible/impracticable to compute. This is a dilemma if we cannot rely on the normal approximations that lead to Chi-squared type results; this is most commonly an issue when sample sizes used are small.

A possible solution involves **simulation-based methods**, such as **Monte Carlo** methods and **randomization/permutation tests.** Monte Carlo tests can be applied if the exact null distribution is known, but is difficult to compute**;** randomization or permutation tests can be used when the exact null distribution is not known analytically, but the null hypothesis implies some symmetry in the data. Some details of these methods are given below.

In **Monte Carlo** simulation, the null cdf and/or the p-value is computed numerically using a random sample produced from the relevant null distribution. This might be advantageous if the exact null distribution is difficult to calculate because the number of possible values that the test statistic can take is vast; the test for Fisher test for independence using the multivariate hypergeometric example above in the **Monte Carlo** strategy for hypothesis testing is outlined below:

1. Choose a test statistic T, and compute its mass function, f_T, analytically

2. Produce a simulated random sample of size N_T from f_T using stochastic simulation methods; label the simulated values $t_1, t_2, ..., t_{N_T}$

3. Estimate the null cdf from this sample as follows; for any value t, use the estimate

$$\widehat{F}_T(t) = \frac{1}{N_T}\sum_{i=1}^{N_T} I\{t_i \leq t\} \qquad\qquad I\{t_i \leq t\} = \left\{\begin{array}{ll} 1 & t_i \leq t \\ 0 & t_i > t \end{array}\right. \quad \text{is an indicator variable}$$

that is,

$$\widehat{F}_T(t) = \frac{\text{Number of } t_1, t_2, ..., t_{N_T} \text{ that are not greater than } t}{N_T}$$

4. Compute the (estimated) p-value for observed test statistic t^* by

$$\widehat{p} = 1 - \widehat{F}_T(t^*) = \frac{\text{Number of } t_1, t_2, ..., t_{N_T} \text{ more extreme than } t^*}{N_T}$$

If the sample size N_T is large enough, then $\widehat{p}$ will be an accurate, well-behaved estimate of the true p-value, by the usual laws of large numbers/Central Limit Theorems.

6.5.4 PERMUTATION TESTS

The central idea of permutation tests refers to rearrangements of the data. The null hypothesis of the test specifies that **the permutations are all equally likely**. The sampling distribution of the test statistic under the null hypothesis is computed by forming all (or many) of the permutations, calculating the test statistic for each and considering these values all equally likely.

Consider the following two group example, where we want to test for any significant difference between the groups.

$$\text{Group 1} \;:\; 55, 58, 60$$
$$\text{Group 2} \;:\; 12, 22, 34$$

Here are the steps we will follow to use a permutation test to analyze the differences between the two groups. For the original order the sum for Group 1 is 173. In this example, if the groups were truly equal (**and the null hypothesis was true**) then randomly moving the observations among the groups would make no difference in the sum for Group 1. Some of the sums would be a little larger than the original sum and some would be a bit smaller. For the six observations there are 720 permutations of which there are 20 distinct combinations for which we can compute the sum of Group 1.

ORDER	GROUP 1	GROUP 2	SUM	ORDER	GROUP 1	GROUP 2	SUM
1	55, 58, 60	12, 22, 34	173	11	12, 22, 60	55, 58, 34	94
2	55, 58, 12	60, 22, 34	125	12	12, 58, 22	55, 60, 34	92
3	55, 58, 22	12, 60, 34	135	13	55, 12, 22	12, 55, 58	89
4	55, 58, 34	12, 22, 34	148	14	12, 34, 60	55, 58, 34	106
5	55, 12, 60	58, 22, 34	127	15	12, 58, 34	55, 22, 60	104
6	55, 22, 60	12, 58, 34	137	16	55, 12, 34	12, 58, 60	101
7	55, 34, 60	12, 22, 58	149	17	22, 34, 60	55, 58, 34	116
8	12, 58, 60	55, 22, 34	130	18	22, 58, 34	55, 22, 60	114
9	22, 58, 60	12, 55, 34	140	19	55, 22, 34	12, 58, 60	111
10	34, 58, 60	12, 22, 55	152	20	12, 22, 34	55, 58, 60	68

Of these 20 different orderings only **one** has a Group 1 sum that greater than or equal to the Group 1 sum from our original ordering. Therefore the probability that a sum this large or larger would occur by chance alone is $1/20 = 0.05$ and can be considered to be statistically significant.

In the analysis of contingency tables, permutation tests for specific hypotheses would consist of producing permutations of the original data consistent with the null hypothesis, subject to the required constraints. For a permutation Chisquared or LR test, the statistic would be computed for all (or a large number of permutations of data with respect to row or column classifications, perhaps subject to row and/or column total sum restrictions