Categorical Data Analysis

Models, likelihoods, conditioning and paired observations

Statistical Inference 2026–27
Lecture Notes

Organization and hospital setting

These two sessions follow a hospital network's quality-of-care questions. Each test is developed through its sampling model, population null, statistic, null calibration, explicit procedure, and interpretation.

Session 1	Multinomial likelihood; specified goodness of fit; Pearson and likelihood-ratio tests; independence and homogeneity; composite goodness of fit and fitted dimensions. This session ends with the milestone model and the derivation of the $k - 1 - d$ degrees of freedom.
Session 2	Fisher's exact test for independent binary groups; McNemar's test for paired binary outcomes; Simpson's paradox and standardization for baseline severity.

A hospital classifies each patient's day-30 outcome as fully recovered (F), partially recovered (P), or not recovered (N). Session 1 compares a cohort's counts $(110, 50, 40)$ with the benchmark $(.50, .30, .20)$, compares recovery across sites and independent pathways, and tests a fitted model in which recovery depends on two milestones. These are distinct questions about category probabilities, not interchangeable labels for one analysis.

Session 2 uses three further, illustrative hospital datasets:

- **Small independent pilot.** Ten patients receive each pathway. Full recovery occurs in one patient under A and eight under B. Does the recovery probability differ when a large-sample reference law is doubtful?
- **Paired follow-up.** The same 100 patients are assessed at discharge and day 30. Eighteen lose symptom control and six gain it. Has the marginal probability of control changed, allowing dependence within a patient?
- **Baseline severity.** In a separate observational cohort, pathway A has better recovery within both severity groups but worse recovery overall. Which population comparison does each calculation describe?

The pilot concerns independent groups, the follow-up concerns paired outcomes, and the severity example concerns different mixtures of patients. These design distinctions determine the analysis. Full recovery is a prespecified binary endpoint for the pilot and severity comparison; symptom control is a separate binary endpoint measured using the same criterion at both visits. The notes finish with sequential **Base R procedures** and a **Method-selection checklist** covering both sessions.

Session 1: multinomial models and goodness of fit

1 From individual outcomes to the multinomial likelihood

Let $Y_1, \dots, Y_n$ be iid categorical observations with $\mathbb{P}(Y_i = j) = p_j$, $j = 1, \dots, k$. Define $O_j = \sum_i 1\{Y_i = j\}$ and $\widehat{p}_j = O_j/n$. Then

$$\mathbb{P}_p(O = o) = \frac{n!}{\prod_j o_j!} \prod_j p_j^{o_j}, \qquad \sum_j o_j = n, \quad \sum_j p_j = 1.$$

The count vector is sufficient by factorization. The log-likelihood, apart from a constant, is $\ell(p) = \sum_j O_j \log p_j$. Lagrange equations give $O_j/p_j = \lambda$, so $\lambda = n$ and $\widehat{p}_j = O_j/n$. If counts vanish, the same expression maximizes the likelihood on the closed simplex. We use $0 \log 0 = 0$ where justified by continuity.

For one patient define the vector $U_i = (1\{Y_i = 1\}, \dots, 1\{Y_i = k\})^T$. Distinct indicators cannot both equal one, so

$$\mathbb{E}U_i = p, \qquad \mathrm{Cov}(U_i) = D - pp^T, \qquad D = \mathrm{diag}(p).$$

Independence across patients gives

$$\mathbb{E}O_j = np_j, \quad \mathrm{Var}(O_j) = np_j(1 - p_j), \quad \mathrm{Cov}(O_j, O_l) = -np_jp_l \quad (j \neq l).$$

The counts are dependent even though patients are independent. With positive probabilities the covariance has rank $k - 1$: there is precisely one linear constraint, the count sum.

2 Pearson goodness of fit: geometry and testing

Test $H_0 : p = p^0$ against $p \neq p^0$, with fixed k and known positive p_j^0. Expected counts are $E_j = np_j^0$, and

$$X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}.$$

The denominator is not the binomial variance $np_j^0(1 - p_j^0)$. The joint covariance structure, not an assumption of independent contributions, explains the choice.

Proposition 2.1. *Under the stated null and sampling assumptions,* $X^2 \Rightarrow \chi^2_{k-1}$.

Proof. Write $p = p^0$ and define $D = \mathrm{diag}(p_1^0, \dots, p_k^0)$, the $k \times k$ diagonal matrix of null probabilities. Since every null probability is positive, $D^{-1/2} = \mathrm{diag}(1/\sqrt{p_1^0}, \dots, 1/\sqrt{p_k^0})$ exists. Set $q = (\sqrt{p_1}, \dots, \sqrt{p_k})^T$. The multivariate CLT gives $\sqrt{n}(\widehat{p} - p) \Rightarrow N_k(0, D - pp^T)$. Therefore

$$Z_n = D^{-1/2}\sqrt{n}(\widehat{p} - p) \Rightarrow N_k(0, I - qq^T).$$

Since $q^Tq = 1$, $P = I - qq^T$ is symmetric and idempotent with rank $k - 1$. Choose a $k \times (k - 1)$ matrix H with orthonormal columns spanning $q^\perp$. Then $HH^T = P$, and the limiting vector has the distribution of HW with $W \sim N_{k-1}(0, I)$. Hence $X^2 = Z_n^T Z_n \Rightarrow W^T H^T HW = W^T W$. The constraint already holds exactly: $q^T Z_n = \sqrt{n} \sum_j (\widehat{p}_j - p_j) = 0$. □

2.1 Recovery benchmark and procedure

Suppose a known reference distribution is $(.5, .3, .2)$. In a new cohort of 200 patients, observed F/P/N counts are $(110, 50, 40)$, compared with expected counts $(100, 60, 40)$. Then $X^2 = 100/100 + 100/60 = 8/3$. With two degrees of freedom the p-value is $e^{-4/3} \approx .2636$.

The testing procedure is: prespecify categories, null probabilities and α; calculate expected counts and the statistic; obtain the upper-tail chi-square probability; reject if it is at most α. Here we do not reject at 5%. This does not establish clinical equivalence. Equivalence requires an explicit tolerance and a suitable inferential formulation. If the benchmark is itself estimated from a finite historical cohort, treating it as known ignores uncertainty; a two-sample model may be required.

2.2 Fixed and local alternatives

Under a fixed alternative $p^* \neq p^0$,

$$X^2/n \xrightarrow{P} \sum_j (p_j^* - p_j^0)^2/p_j^0 > 0,$$

so power tends to one. Under local alternatives $p_n = p^0 + h/\sqrt{n}$ with $\sum_j h_j = 0$, the standardized limit has mean $D_0^{-1/2}h$, where $D_0 = \mathrm{diag}(p^0)$. That mean lies in $q^\perp$. Thus

$$X^2 \Rightarrow \chi^2_{k-1}(\lambda), \qquad \lambda = \sum_j h_j^2/p_j^0.$$

To justify this, apply the triangular-array CLT to bounded categorical indicators with $p_n \to p^0$, then add the deterministic mean shift. The projection argument reduces the limit to the squared norm of $k-1$ independent normals with nonzero means. This describes detectability at the $n^{-1/2}$ scale and motivates approximate power calculations.

2.3 Support and approximation

Expected-count rules such as five per cell are heuristics, not guarantees. The theorem assumes fixed positive probabilities and fixed dimension. Growing numbers of categories or vanishing probabilities require other arguments. If $p_j^0 = 0$ and $O_j > 0$, the data are impossible under the null. If both are zero, remove the impossible category from the support rather than dividing by zero. Pooling categories should have scientific justification and preferably be prespecified.

3 Likelihood ratios and information divergence

3.1 Practical problem: how much likelihood does the benchmark lose?

The hospital observes $(110, 50, 40)$ recoveries in the F/P/N categories and wants to assess the benchmark $(.5, .3, .2)$. Pearson measures weighted squared discrepancies. An alternative question is: how much less well does the benchmark explain these data than the best-fitting unrestricted distribution? This leads to a likelihood-ratio test of exactly the same population null. We first derive its statistic, then establish its reference distribution rather than simply invoking a chi-square approximation.

3.2 Deriving the likelihood ratio

Let $\mathcal{S}_k = \{p : p_j \geq 0, \sum_j p_j = 1\}$. Define

$$\Lambda(o) = \frac{\sup_{p \in H_0} L(p; o)}{\sup_{p \in \mathcal{S}_k} L(p; o)} = \frac{L(p^0; o)}{L(\widehat{p}; o)}, \qquad \widehat{p}_j = o_j/n.$$

The multinomial coefficient is the same in numerator and denominator, so it cancels. Restricting products and logarithms to positive observed counts avoids writing 0^0:

$$\Lambda(o) = \prod_{j: o_j > 0} \left(\frac{p_j^0}{o_j/n} \right)^{o_j},$$

$$-2 \log \Lambda(o) = 2 \sum_{j: o_j > 0} o_j \log \frac{o_j/n}{p_j^0} = 2 \sum_{j=1}^{k} o_j \log \frac{o_j}{E_j} =: G^2.$$

Here the last sum defines a zero-count contribution as zero. This agrees with $\lim_{x \downarrow 0} x \log(x/E_j) = 0$. Since the denominator is a maximum over a set containing p^0, $0 < \Lambda \leq 1$, hence $G^2 \geq 0$. Small Λ, equivalently large G^2, supports rejection.

Define $D_{\mathrm{KL}}(u \| v) = \sum_{j: u_j > 0} u_j \log(u_j/v_j)$ for $v_j > 0$. Substitution gives $G^2 = 2n D_{\mathrm{KL}}(\widehat{p} \| p^0)$. The divergence measures the log-likelihood loss per observation from imposing the benchmark. Individual summands can be negative, although their sum cannot. The equality $G^2 = 0$ holds exactly when the empirical probability vector equals the benchmark.

3.3 Why the relative discrepancies are small under the null

Let

$$\delta_{nj} = \frac{O_j - E_j}{E_j} = \frac{\widehat{p}_j - p_j^0}{p_j^0}.$$

Under H_0, $\sqrt{n}(\widehat{p}_j - p_j^0) = O_P(1)$ by the CLT; division by a fixed positive p_j^0 gives $\delta_{nj} = O_P(n^{-1/2})$. Because k is fixed, $M_n = \max_j |\delta_{nj}| = O_P(n^{-1/2})$ as well. In particular, the event $A_n = \{M_n \leq 1/2\}$ has probability tending to one. Taylor expansion can therefore be applied simultaneously to every cell on this event. This step matters: an observed zero count gives $\delta_{nj} = -1$, outside a neighborhood of zero, but the probability of encountering such a cell tends to zero under the fixed-positive-probability assumptions.

3.4 Taylor expansion with an explicit remainder bound

Put $f(t) = (1 + t) \log(1 + t)$ for $t > -1$.

Specifically, we expand around $t = 0$ to second order. Taylor's theorem with Lagrange remainder states that, for some ξ between 0 and t,

$$f(t) = f(0) + f'(0)t + \frac{f''(0)}{2!}t^2 + \underbrace{\frac{f^{(3)}(\xi)}{3!}t^3}_{r(t)}.$$

For this function,

$$f(0) = 0, \qquad f'(t) = 1 + \log(1 + t), \qquad f''(t) = \frac{1}{1+t}, \qquad f^{(3)}(t) = -\frac{1}{(1+t)^2}.$$

Thus $f'(0) = f''(0) = 1$, giving

$$f(t) = t + \frac{t^2}{2} + r(t).$$

To bound the remainder, note that if $|t| \le \frac{1}{2}$, then the intermediate point ξ also satisfies $|\xi| \le \frac{1}{2}$. Hence

$$1 + \xi \ge \frac{1}{2} \quad \Longrightarrow \quad |f^{(3)}(\xi)| = \frac{1}{(1+\xi)^2} \le \frac{1}{(1/2)^2} = 4.$$

Therefore,

$$\boxed{|r(t)| = \left| \frac{f^{(3)}(\xi)}{3!} t^3 \right| \le \frac{4}{6}|t|^3.}$$

The 4 comes from the bound on the third derivative; the 6 is 3! in Taylor's remainder formula. Since $O_j = E_j(1 + \delta_{nj})$, on A_n,

$$\begin{aligned} G^2 &= 2 \sum_j E_j f(\delta_{nj}) \\ &= 2 \sum_j E_j \delta_{nj} + \sum_j E_j \delta_{nj}^2 + R_n, \\ |R_n| &\le \frac{4}{3} \sum_j E_j |\delta_{nj}|^3 \le \frac{4}{3} n M_n^3. \end{aligned}$$

The last bound is $O_P(n^{-1/2})$, hence tends to zero in probability. More explicitly, for any $\varepsilon > 0$,

$$\mathbb{P}(|R_n| > \varepsilon) \le \mathbb{P}(A_n^c) + \mathbb{P}\left(\tfrac{4}{3} n M_n^3 > \varepsilon\right) \longrightarrow 0,$$

where outside A_n we interpret R_n as the difference between the exact statistic and the displayed linear and quadratic terms. The bound is needed only on the high-probability event.

3.5 Cancellation, limit, and test

The linear term is exactly zero, not merely asymptotically negligible:

$$\sum_j E_j \delta_{nj} = \sum_j (O_j - E_j) = n - n = 0.$$

The quadratic term is exactly Pearson's statistic:

$$\sum_j E_j \delta_{nj}^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = X^2.$$

Consequently $G^2 - X^2 \xrightarrow{P} 0$. Section 2 gives $X^2 \Rightarrow \chi^2_{k-1}$; Slutsky's theorem now gives $G^2 \Rightarrow \chi^2_{k-1}$. Reject when G^2 exceeds the $(1 - \alpha)$ chi-square quantile, or report its upper-tail p-value. This derivation proves the null calibration; it does not claim equality of the two statistics for a finite sample or under an arbitrary fixed alternative.

3.6 Return to the hospital data

For the benchmark cohort,

$$G^2 = 2\{110\log(1.1) + 50\log(5/6) + 40\log(1)\} \approx 2.7361.$$

The two-degree-of-freedom p-value is $e^{-G^2/2} \approx .2546$, compared with Pearson's .2636. Both fail to reject at 5%. The small difference is consistent with asymptotic equivalence; it is not a reason to select whichever statistic gives the preferred conclusion. Neither calculation demonstrates clinical equivalence. The proof also shows why sparse expected counts are problematic: relative fluctuations need not be small in the available sample, so the remainder can matter. Switching from Pearson to deviance does not by itself fix this problem.

4 Independence in one cross-classified sample

The quality team draws a single random sample of 150 patients from the network and records each patient's site and day-30 recovery. Site A contributes F/P/N counts $(30, 20, 10)$ and site B contributes $(20, 30, 40)$. The team asks whether knowing a patient's site changes the population distribution of recovery. Site sample sizes were not fixed beforehand; they arose from the network sample. We need a test of association, allowing both the site proportions and the overall recovery probabilities to be unknown.

4.1 Sampling model and independence hypothesis

Let

$$p_{ij} = \mathbb{P}(\text{site} = i,\ \text{recovery category} = j), \qquad i = 1, \dots, r, \quad j = 1, \dots, c.$$

The rc cell counts form one multinomial sample with total n. Write

$$O_{i+} = \sum_{j=1}^{c} O_{ij}, \qquad O_{+j} = \sum_{i=1}^{r} O_{ij}.$$

Define the marginal probabilities

$$a_i = \mathbb{P}(\text{site} = i), \qquad b_j = \mathbb{P}(\text{recovery category} = j).$$

By definition, site and recovery are independent if and only if their joint probabilities factor into the product of their marginals. Thus the null hypothesis is

$$H_0 : p_{ij} = a_i b_j \quad \text{for every } i, j, \qquad \sum_i a_i = 1, \quad \sum_j b_j = 1.$$

This is a composite null: it specifies independence but leaves the marginal distributions unknown.

4.2 Fitted expected counts and test statistics

Under the null, the log-likelihood is

$$\begin{aligned} \ell(a, b) &= C + \sum_{i,j} O_{ij} \log(a_i b_j) \\ &= C + \sum_i O_{i+} \log a_i + \sum_j O_{+j} \log b_j, \end{aligned}$$

where C does not depend on the parameters.

The two terms can be maximized separately, subject to their respective sum constraints. This gives

$$\widehat{a}_i = \frac{O_{i+}}{n}, \qquad \widehat{b}_j = \frac{O_{+j}}{n}.$$

Consequently, the fitted expected count in cell (i, j) is

$$E_{ij} = n\widehat{a}_i\widehat{b}_j = \frac{O_{i+}O_{+j}}{n}.$$

The Pearson and likelihood-ratio statistics are

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \qquad G^2 = 2\sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}},$$

with zero-count contributions to G^2 defined as zero.

4.3 Why are there $(r-1)(c-1)$ degrees of freedom?

The dimension of a model is the number of freely varying parameters needed to specify a distribution in that model.

Unrestricted model. Without the independence restriction, there are rc cell probabilities, subject to the single constraint

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1.$$

Once $rc - 1$ probabilities are specified, the remaining probability is determined. Hence the unrestricted model has dimension $rc - 1$. At an interior point, where all cell probabilities are positive, nonnegativity restricts the allowable values but does not further reduce the dimension.

Independence model. Under H_0, the entire table is determined by the two marginal probability vectors:

$$(a_1, \ldots, a_r) \quad \text{and} \quad (b_1, \ldots, b_c).$$

The row probabilities sum to one, so they contribute $r-1$ free parameters. The column probabilities also sum to one, contributing $c-1$ free parameters. These two marginal distributions can be chosen separately. No additional normalization constraint is needed, since

$$\sum_{i,j} a_i b_j = \left(\sum_i a_i\right)\left(\sum_j b_j\right) = 1.$$

Moreover, the margins are uniquely recoverable from the joint probabilities:

$$\sum_j p_{ij} = a_i, \qquad \sum_i p_{ij} = b_j.$$

Thus this parameterization has no additional redundancy, and the null model has dimension $(r - 1) + (c - 1)$.

The difference between the unrestricted and null dimensions is

$$(rc-1)-\big[(r-1)+(c-1)\big]=(r-1)(c-1).$$

Under H_0, with fixed table dimensions and positive population cell probabilities, the regular asymptotic theory therefore gives

$$X^2 \Rightarrow \chi^2_{(r-1)(c-1)}, \qquad G^2 \Rightarrow \chi^2_{(r-1)(c-1)}.$$

For the 2×3 hospital table, the unrestricted model has $6-1=5$ free parameters. Under independence, one parameter specifies the site distribution and two specify the recovery distribution. The null dimension is therefore $1+2=3$, leaving $5-3=2$ degrees of freedom.

4.4 Return to the hospital data

The observed table is

Site	F	P	N	Total
A	30	20	10	60
B	20	30	40	90
Total	50	50	50	150

The fitted expected rows are $(20, 20, 20)$ and $(30, 30, 30)$. Therefore,

$$X^2=\frac{(30-20)^2}{20}+\frac{(10-20)^2}{20}+\frac{(20-30)^2}{30}+\frac{(40-30)^2}{30}=\frac{50}{3}.$$

Using the χ^2_2 reference distribution, the approximate p-value is

$$\mathbb{P}\left(\chi^2_2 \geq \frac{50}{3}\right)=e^{-25/3} \approx 0.000240.$$

There is strong evidence that site and recovery are associated. This need not be a causal site effect: patient case mix can differ. Independence of patients is a sampling assumption; independence of site and recovery is the hypothesis being tested.

4.5 Locating the discrepancies

Pearson residuals

$$R_{ij}=\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}}$$

show the direction and contribution of each cell, since $X^2=\sum_{i,j} R_{ij}^2$. They are correlated and are not individually standard normal after the margins have been estimated.

Under the single-multinomial independence model, an asymptotically standardized residual is

$$R_{ij}^{\text{adj}}=\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}(1-\widehat{p}_{i+})(1-\widehat{p}_{+j})}}, \qquad \widehat{p}_{i+}=\frac{O_{i+}}{n}, \quad \widehat{p}_{+j}=\frac{O_{+j}}{n}.$$

These adjusted residuals remain correlated. Formal cellwise follow-up tests require attention to multiple comparisons; a global rejection does not establish that every cell differs from its null expectation.

5 Homogeneity across independent groups

A new comparison enrolls 60 independent patients under pathway A and 90 under pathway B, with these group sizes set in advance. Suppose their full/partial/no recovery counts are $(30, 20, 10)$ and $(20, 30, 40)$.

The numerical table matches the preceding site example, but the sampling design differs. We now sample separately from two specified groups and ask whether their recovery distributions are the same. The common recovery distribution under the null is unknown and must be estimated.

5.1 Sampling model and null hypothesis

More generally, suppose there are r groups and c outcome categories. For group i, define

$$p_{ij} = \mathbb{P}(\text{outcome category} = j \mid \text{group} = i), \qquad j = 1, \dots, c.$$

These are within-group probabilities, so each row satisfies

$$\sum_{j=1}^{c} p_{ij} = 1.$$

Unlike the preceding independence model, p_{ij} here does not denote a joint probability of group membership and outcome.

Let n_i be the fixed sample size in group i. Assuming iid observations within each group and independence across groups,

$$O_{i\cdot} = (O_{i1}, \dots, O_{ic}) \sim \text{Mult}(n_i; p_{i1}, \dots, p_{ic}), \qquad i = 1, \dots, r,$$

independently. Write

$$n = \sum_{i=1}^{r} n_i, \qquad O_{+j} = \sum_{i=1}^{r} O_{ij}.$$

The null hypothesis of homogeneity is

$$H_0 : p_{ij} = p_j \quad \text{for every } i, j, \qquad \sum_{j=1}^{c} p_j = 1.$$

Thus every group has the same category distribution, although that distribution is unspecified. The alternative is that at least two groups differ in at least one category probability.

For the hospital study, the null states that pathways A and B have the same probabilities of full, partial, and no recovery. It does not require these three probabilities to be equal to one another.

5.2 Unrestricted maximum likelihood estimates

Independence of the group samples gives the product-multinomial likelihood

$$L(\{p_{ij}\}; O) = \prod_{i=1}^{r} \left[\frac{n_i!}{\prod_{j=1}^{c} O_{ij}!} \prod_{j=1}^{c} p_{ij}^{O_{ij}} \right].$$

Ignoring terms that do not depend on the parameters,

$$\ell(\{p_{ij}\}) = C + \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log p_{ij}.$$

Without the null restriction, each row can be maximized separately, subject to its own constraint $\sum_j p_{ij} = 1$. For row i, the Lagrange equations are

$$\frac{O_{ij}}{p_{ij}} - \lambda_i = 0.$$

Thus $O_{ij} = \lambda_i p_{ij}$; summing over j gives $\lambda_i = n_i$. The unrestricted MLE is therefore

$$\widehat{p}_{ij} = \frac{O_{ij}}{n_i}.$$

If a count is zero, this formula remains the likelihood maximizer on the closed probability simplex.

5.3 Maximum likelihood under homogeneity

Under H_0, substitute $p_{ij} = p_j$:

$$\begin{aligned}\ell_0(p) &= C + \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} \log p_j \\ &= C + \sum_{j=1}^{c} O_{+j} \log p_j.\end{aligned}$$

Maximizing subject to $\sum_j p_j = 1$ gives the pooled estimate

$$\widetilde{p}_j = \frac{O_{+j}}{n}.$$

Here $\widetilde{p}_j$ denotes the null-constrained MLE, whereas $\widehat{p}_{ij}$ denotes the unrestricted group-specific MLE.

The pooled estimate is a sample-size-weighted average:

$$\widetilde{p}_j = \sum_{i=1}^{r} \frac{n_i}{n} \widehat{p}_{ij}.$$

It is generally not the unweighted average of the group proportions.

Under the fitted null model, the expected count in group i and category j is

$$\boxed{E_{ij} = n_i \widetilde{p}_j = \frac{n_i O_{+j}}{n}.}$$

Each group's expected counts sum to its fixed sample size:

$$\sum_j E_{ij} = n_i.$$

5.4 Deriving the likelihood-ratio statistic

The likelihood ratio compares the best-fitting common distribution with the best-fitting separate distributions:

$$\Lambda = \frac{L(\{\widetilde{p}_j\}; O)}{L(\{\widehat{p}_{ij}\}; O)}.$$

The multinomial coefficients cancel, leaving

$$\Lambda = \prod_{i,j:O_{ij}>0} \left(\frac{\widetilde{p}_j}{\widehat{p}_{ij}} \right)^{O_{ij}}.$$

Consequently,

$$\begin{aligned} G^2 &= -2 \log \Lambda \\ &= 2 \sum_{i,j} O_{ij} \log \frac{\widehat{p}_{ij}}{\widetilde{p}_j} \\ &= 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}/n_i}{O_{+j}/n} \\ &= \boxed{2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}}. \end{aligned}$$

Zero-count contributions are defined as zero.

Equivalently,

$$G^2 = 2 \sum_{i=1}^{r} n_i D_{\mathrm{KL}}(\widehat{p}_{i\cdot} \| \widetilde{p}),$$

where $\widehat{p}_{i\cdot}$ is the empirical probability vector in group i. Thus the deviance measures how far the separate empirical distributions lie from their fitted common distribution, with each group weighted by its sample size.

Pearson's statistic uses the same fitted expected counts:

$$\boxed{X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.}$$

5.5 Why are there $(r-1)(c-1)$ degrees of freedom?

Unrestricted model. Each group has its own vector of c probabilities satisfying one sum constraint. Therefore, each group contributes $c-1$ free parameters. Across r groups, the unrestricted dimension is

$$r(c-1).$$

There are no parameters for group membership: the sample sizes $n_1, \ldots, n_r$ were fixed by design.

Null model. Under homogeneity, every group shares one probability vector $(p_1, \ldots, p_c)$. This vector has $c-1$ free parameters, so the null dimension is

$$c-1.$$

The dimension difference is therefore

$$r(c-1) - (c-1) = \boxed{(r-1)(c-1)}.$$

For the two-pathway, three-category study, the unrestricted model has two free recovery probabilities per pathway, giving four parameters. The null model has only two free probabilities shared by both pathways. The difference is $4 - 2 = 2$.

For the usual asymptotic result, assume that r, c remain fixed, that all common null probabilities are positive, and that

$$\frac{n_i}{n} \longrightarrow w_i > 0 \qquad \text{for every } i.$$

Thus every group contributes an increasing number of observations. Under these regularity conditions, likelihood-ratio theory gives

$$G^2 \Rightarrow \chi^2_{(r-1)(c-1)}.$$

5.6 Why Pearson has the same limiting distribution

Under H_0, both $\widehat{p}_{ij}$ and $\widetilde{p}_j$ converge to the same positive p_j. Moreover, because $n_i/n \to w_i > 0$,

$$\widehat{p}_{ij} - \widetilde{p}_j = O_P(n^{-1/2}).$$

Define

$$\delta_{ij} = \frac{O_{ij} - E_{ij}}{E_{ij}} = \frac{\widehat{p}_{ij} - \widetilde{p}_j}{\widetilde{p}_j}.$$

Then $\delta_{ij} = O_P(n^{-1/2})$. With fixed table dimensions, $\max_{i,j} |\delta_{ij}| = O_P(n^{-1/2})$.

Applying the Taylor expansion established in Section 3,

$$\begin{aligned} G^2 &= 2 \sum_{i,j} E_{ij}(1 + \delta_{ij}) \log(1 + \delta_{ij}) \\ &= 2 \sum_{i,j} E_{ij}\delta_{ij} + \sum_{i,j} E_{ij}\delta_{ij}^2 + o_P(1). \end{aligned}$$

The linear term vanishes exactly:

$$\sum_{i,j} E_{ij}\delta_{ij} = \sum_{i,j} (O_{ij} - E_{ij}) = 0.$$

The quadratic term is X^2. The remainder is bounded, on an event whose probability tends to one, by a constant times

$$\sum_{i,j} E_{ij}|\delta_{ij}|^3 \le n \max_{i,j} |\delta_{ij}|^3 = O_P(n^{-1/2}).$$

Hence

$$G^2 - X^2 \xrightarrow{P} 0, \qquad X^2 \Rightarrow \chi^2_{(r-1)(c-1)}.$$

5.7 Return to the pathway comparison

The observed counts are

Pathway	F	P	N	Total
A	30	20	10	60
B	20	30	40	90
Total	50	50	50	150

The pooled null estimate is

$$\widetilde{p} = (1/3, 1/3, 1/3),$$

giving expected counts

$$E_{A\cdot} = (20, 20, 20), \qquad E_{B\cdot} = (30, 30, 30).$$

Thus

$$X^2 = \frac{100}{20} + \frac{100}{20} + \frac{100}{30} + \frac{100}{30} = \frac{50}{3}.$$

With two degrees of freedom, the approximate p-value is

$$\mathbb{P}\left(\chi^2_2 \geq \frac{50}{3}\right) = e^{-25/3} \approx 0.000240.$$

We reject the hypothesis that the pathways have the same recovery distribution.

The estimated distributions are

$$\widehat{p}_{A\cdot} = (0.500, 0.333, 0.167), \qquad \widehat{p}_{B\cdot} = (0.222, 0.333, 0.444).$$

The observed difference is concentrated in full recovery and no recovery; the observed partial-recovery proportions coincide. The global test nevertheless concerns the complete distributions. It does not by itself provide separate significance tests for each category.

5.8 Relation to the independence test

The fitted counts and test statistics coincide numerically because the observed row totals in the independence analysis equal the fixed group sizes in the homogeneity analysis:

$$E_{ij} = \frac{O_{i+}O_{+j}}{n} = \frac{n_i O_{+j}}{n}.$$

Their sampling models remain different:

- **Independence:** one sample is cross-classified by two variables. Only the grand total is fixed, and the model includes probabilities of group membership.
- **Homogeneity:** separate samples are drawn from specified groups. Every group size is fixed, and only the within-group outcome distributions are modeled.

Conditioning the single-multinomial model on its row totals produces a product-multinomial model. Under independence, the conditional outcome distribution is the same in every row. This provides a probabilistic explanation for the connection between the two tests.

Finally, evidence of different recovery distributions does not automatically establish a causal pathway effect. That interpretation depends on treatment allocation, follow-up, and other design assumptions. In an observational comparison, differences in baseline severity or other patient characteristics may explain some or all of the association.

6 Composite goodness of fit and fitted dimensions

The preceding tests share a common structure. We specify a restricted model for the category probabilities, estimate its unknown parameters, and compare the observed counts with the fitted expected counts. The question is whether the discrepancies remaining after fitting are too large to attribute to sampling variation.

In the independence test, we estimated the row and column marginal probabilities before calculating expected counts. In the homogeneity test, we estimated a common recovery distribution by pooling the groups. In both cases, the null hypothesis specified a family of distributions rather than a single distribution.

We now develop this idea more generally. The guiding question is:

> After estimating the parameters allowed by a proposed model, is there evidence of lack of fit that those parameters cannot explain?

6.1 A new healthcare question: does a milestone model explain recovery?

The clinical team proposes that recovery requires two milestones: restored mobility and adequate pain control. For this example, define full recovery as achieving both milestones, partial recovery as achieving exactly one, and no recovery as achieving neither.

The team proposes a simple model: the milestones are independent, and each is achieved with the same unknown probability θ. The resulting recovery probabilities are

$$p_F(\theta) = \theta^2, \qquad p_P(\theta) = 2\theta(1-\theta), \qquad p_N(\theta) = (1-\theta)^2, \qquad 0 < \theta < 1.$$

The observed counts are again 110 fully recovered, 50 partially recovered, and 40 unrecovered patients. The question is not whether θ equals a specified value. It is whether *any* value of θ makes this model compatible with the recovery distribution.

We must therefore distinguish two tasks:

1. Find the member of the proposed family that best fits the observations.
2. Assess whether the remaining discrepancy is consistent with sampling variation.

Estimating θ improves the fit to these same data. The reference distribution must account for that fitting benefit.

6.2 The general multinomial formulation

Let

$$O = (O_1, \dots, O_k) \sim \text{Mult}(n; p_1, \dots, p_k), \qquad \widehat{p}_j = \frac{O_j}{n}.$$

Consider the composite null

$$H_0 : p = p(\theta) \quad \text{for some } \theta \in \Theta \subseteq \mathbb{R}^d,$$

against the alternative that the probability vector does not belong to this family.

The unrestricted MLE is the empirical vector $\widehat{p}$. The null-constrained MLE is obtained by maximizing

$$\ell_0(\theta) = C + \sum_{j=1}^{k} O_j \log p_j(\theta).$$

Denote its maximizer by $\widehat{\theta}$, and define

$$\widetilde{p} = p(\widehat{\theta}), \qquad E_j = n\widetilde{p}_j.$$

Here $\widehat{p}$ is the unrestricted empirical distribution, while $\widetilde{p}$ is the best-fitting distribution within the null family.

The fitted Pearson statistic and likelihood-ratio deviance are

$$X^2 = \sum_{j=1}^{k} \frac{(O_j - n\widetilde{p}_j)^2}{n\widetilde{p}_j}, \qquad G^2 = 2\sum_{j=1}^{k} O_j \log \frac{O_j}{n\widetilde{p}_j}.$$

The deviance follows by comparing $L(\widetilde{p})$ with $L(\widehat{p})$, with zero-count terms defined by continuity.

6.3 Dimension counting: the common principle

The unrestricted multinomial model has $k-1$ free parameters. If the null family has d locally identifiable parameters, the difference in dimensions is

$$(k-1)-d.$$

Under suitable regularity conditions, this is the asymptotic chi-square degrees of freedom.

This principle includes the earlier tests:

Problem	Unrestricted dimension	Null dimension	Difference
Specified probabilities	$k-1$	0	$k-1$
Independence	$rc-1$	$(r-1)+(c-1)$	$(r-1)(c-1)$
Homogeneity	$r(c-1)$	$c-1$	$(r-1)(c-1)$
Milestone model	2	1	1

Independence is directly a special case of the single-multinomial formulation, with $k=rc$ and the marginal probabilities as the null parameters. Homogeneity uses a product of multinomials, so its unrestricted dimension is different. It follows the same dimension-difference principle, but should not be treated as a single unrestricted multinomial over all rc cells.

Dimension counting identifies the candidate degrees of freedom. The derivation below explains why this count produces a chi-square limit in the regular multinomial case.

6.4 Assumptions and notation

Suppose the null is true at θ_0. Assume:

1. The number of categories k is fixed, and observations are iid.
2. Every component of $p(\theta_0)$ is positive.
3. The true parameter θ_0 is an interior point of Θ, and $p(\theta)$ is sufficiently smooth near θ_0.
4. The $k \times d$ Jacobian
$$J = \left.\frac{\partial p(\theta)}{\partial \theta^T}\right|_{\theta=\theta_0}$$
has full column rank d.
5. The null MLE is consistent and satisfies the usual regular likelihood conditions for a score expansion.

For the derivation, write

$$p = p(\theta_0), \qquad D = \operatorname{diag}(p_1, \ldots, p_k), \qquad q = (\sqrt{p_1}, \ldots, \sqrt{p_k})^T,$$

and

$$Z_n = D^{-1/2}\sqrt{n}(\widehat{p} - p).$$

From the multinomial CLT,

$$Z_n \Rightarrow N_k(0, P), \qquad P = I - qq^T.$$

Define the standardized Jacobian

$$A = D^{-1/2}J.$$

Its columns describe, to first order, the directions in which the standardized probability vector can move when the null parameters change.

6.5 Deriving the score and information

For general θ, let

$$J(\theta) = \frac{\partial p(\theta)}{\partial \theta^T}, \qquad D(\theta) = \operatorname{diag}\{p(\theta)\}.$$

Differentiating the null log-likelihood gives

$$U_n(\theta) = \frac{\partial \ell_0(\theta)}{\partial \theta} = J(\theta)^T D(\theta)^{-1} O.$$

Because $\sum_j p_j(\theta) = 1$ for every θ,

$$J(\theta)^T \mathbf{1} = 0.$$

Also, $D(\theta)^{-1} p(\theta) = \mathbf{1}$. We may therefore rewrite the score as

$$U_n(\theta) = J(\theta)^T D(\theta)^{-1} \{O - np(\theta)\}.$$

At the truth,

$$\frac{U_n(\theta_0)}{\sqrt{n}} = J^T D^{-1} \sqrt{n}(\widehat{p} - p) = A^T Z_n.$$

The information per observation is

$$\mathcal{I} = J^T D^{-1} J = A^T A.$$

To verify this from the score covariance, note that

$$\frac{1}{n} \operatorname{Var}\{U_n(\theta_0)\} = J^T D^{-1} (D - pp^T) D^{-1} J = J^T D^{-1} J,$$

because $J^T D^{-1} p = J^T \mathbf{1} = 0$. Full column rank of J makes $\mathcal{I}$ positive definite.

6.6 The first-order effect of estimating the parameters

The MLE satisfies $U_n(\widehat{\theta}) = 0$ with probability tending to one. An integral form of the score expansion gives

$$0 = U_n(\theta_0) - n\overline{\mathcal{I}}_n(\widehat{\theta} - \theta_0),$$

where

$$\overline{\mathcal{I}}_n = -\frac{1}{n} \int_0^1 \frac{\partial U_n}{\partial \theta^T}\big(\theta_0 + t(\widehat{\theta} - \theta_0)\big)\, dt.$$

Under the regularity assumptions, $\overline{\mathcal{I}}_n \xrightarrow{P} \mathcal{I}$. Consequently,

$$\begin{aligned} \sqrt{n}(\widehat{\theta} - \theta_0) &= \overline{\mathcal{I}}_n^{-1} \frac{U_n(\theta_0)}{\sqrt{n}} \\ &= \mathcal{I}^{-1} A^T Z_n + o_P(1). \end{aligned}$$

Next, expand the fitted probability vector:

$$p(\widehat{\theta}) = p + J(\widehat{\theta} - \theta_0) + o_P(n^{-1/2}).$$

The standardized residual after fitting is therefore

$$\begin{aligned} R_n &= D^{-1/2}\sqrt{n}\{\widehat{p} - p(\widehat{\theta})\} \\ &= Z_n - A\sqrt{n}(\widehat{\theta} - \theta_0) + o_P(1) \\ &= \left\{I - A(A^TA)^{-1}A^T\right\} Z_n + o_P(1). \end{aligned}$$

Define

$$P_A = A(A^TA)^{-1}A^T.$$

This is the orthogonal projection onto the column space of A. Thus

$$\boxed{R_n = (I - P_A)Z_n + o_P(1).}$$

The interpretation is concrete: fitting removes the component of the empirical fluctuation that can be explained by changing the null parameters. The remaining component measures lack of fit.

6.7 The residual covariance and its rank

Since

$$q^TA = q^TD^{-1/2}J = \mathbf{1}^TJ = 0,$$

every column of A lies in $q^\perp$. Hence

$$PP_A = P_AP = P_A, \qquad P = I - qq^T.$$

The limiting covariance of R_n is

$$\begin{aligned} (I - P_A)P(I - P_A) &= P - P_A \\ &= I - qq^T - P_A =: Q. \end{aligned}$$

The matrix Q is symmetric. Moreover,

$$Q^2 = (P - P_A)^2 = P - 2P_A + P_A = P - P_A = Q.$$

It is therefore an orthogonal projection. Its image consists of vectors perpendicular to both q and the d columns of A. These are $d + 1$ independent directions, so

$$\operatorname{rank}(Q) = k - 1 - d.$$

Thus

$$R_n \Rightarrow N_k(0, Q).$$

An orthonormal basis of the image of Q transforms this limiting vector into $k - 1 - d$ independent standard-normal coordinates. Its squared norm consequently has distribution χ^2_{k-1-d}.

6.8 From the residual limit to the test statistics

Pearson's statistic uses fitted probabilities in its denominators, whereas R_n was defined using the true probabilities. Write

$$\widetilde{D} = \mathrm{diag}\{p(\widehat{\theta})\}.$$

Then

$$X^2 = R_n^T D^{1/2} \widetilde{D}^{-1} D^{1/2} R_n.$$

Consistency and positivity imply

$$D^{1/2} \widetilde{D}^{-1} D^{1/2} \xrightarrow{P} I.$$

Since $R_n = O_P(1)$,

$$X^2 = R_n^T R_n + o_P(1) \Rightarrow \chi^2_{k-1-d}.$$

For the deviance, define

$$\delta_j = \frac{O_j - E_j}{E_j} = \frac{\widehat{p}_j - \widetilde{p}_j}{\widetilde{p}_j}.$$

The residual expansion and positivity give $\max_j |\delta_j| = O_P(n^{-1/2})$. Applying the Taylor argument used for the specified-null test,

$$G^2 = 2\sum_j E_j \delta_j + \sum_j E_j \delta_j^2 + o_P(1).$$

The linear term is zero because both observed and fitted counts sum to n, and the quadratic term is X^2. Hence

$$\boxed{X^2 \Rightarrow \chi^2_{k-1-d}, \qquad G^2 \Rightarrow \chi^2_{k-1-d}.}$$

6.9 Return to the milestone model

For the clinical model, the null likelihood is proportional to

$$\{\theta^2\}^{O_F} \{2\theta(1-\theta)\}^{O_P} \{(1-\theta)^2\}^{O_N}.$$

Dropping factors that do not depend on θ,

$$\ell_0(\theta) = C + (2O_F + O_P)\log\theta + (O_P + 2O_N)\log(1-\theta).$$

The score equation is

$$\frac{2O_F + O_P}{\theta} - \frac{O_P + 2O_N}{1-\theta} = 0,$$

giving

$$\widehat{\theta} = \frac{2O_F + O_P}{2n}.$$

For $(O_F, O_P, O_N) = (110, 50, 40)$,

$$\widehat{\theta} = \frac{270}{400} = 0.675,$$

and the fitted expected counts are

$$(E_F, E_P, E_N) = 200\big(0.675^2,\, 2(0.675)(0.325),\, 0.325^2\big) = (91.125, 87.75, 21.125).$$

Therefore,

$$X^2 = \frac{(110 - 91.125)^2}{91.125} + \frac{(50 - 87.75)^2}{87.75} + \frac{(40 - 21.125)^2}{21.125} \approx 37.01.$$

There are $k - 1 - d = 3 - 1 - 1 = 1$ degrees of freedom. The discrepancy is far beyond the 5% critical value $\chi^2_{1,0.95} \approx 3.84$, so we reject the proposed model.

The observed distribution has more patients at both extremes and fewer partially recovered patients than the fitted model predicts. This is evidence against the combined assumptions defining the milestone model. It does not identify a unique explanation: dependence between milestones, unequal achievement probabilities, or patient heterogeneity can each violate the model.

6.10 When the dimension rule needs additional care

The rule "subtract the number of fitted parameters" is a consequence of the regular likelihood and projection arguments, not an unconditional counting recipe.

In particular:

- Consistency of an arbitrary estimator does not guarantee the efficient score expansion used above.
- Parameters estimated from ungrouped observations need not behave like MLEs fitted to the grouped multinomial counts.
- Boundary parameters, unidentifiable parameters, structural zeros, and singular information matrices can change the limiting distribution.
- Data-dependent categories require analysis of the actual selection and fitting procedure.

When using a parametric bootstrap, simulate under the fitted null and repeat the same estimation and processing steps in every replicate. Finally, if $d = k - 1$, no residual directions remain in this first-order analysis. A saturated multinomial model fits the empirical proportions exactly and therefore provides no goodness-of-fit test against the unrestricted multinomial model.

Session 2: exact tests, paired outcomes, and aggregation

We now move from fitted chi-square reference laws to two conditional exact laws, then examine how aggregation can reverse a comparison. The central questions are what to condition on, which observations are paired, and whose recovery rate is being compared.

7 Fisher's exact test: eliminating a nuisance parameter

Only ten patients receive each pathway in an initial independent-groups pilot. Pathway A has one full recovery and nine other outcomes; B has eight full recoveries and two other outcomes. Under equality, fitted success counts are only 4.5 in each group. The team wants a finite-sample test that does not rely on the chi-square approximation, but the common null recovery probability is unknown. Conditioning will remove that nuisance parameter and provide an exact reference distribution.

7.1 Sampling model and conditioning

Let $A \sim \mathrm{Bin}(n_1, p_1)$ and $C \sim \mathrm{Bin}(n_2, p_2)$ independently, and write $n = n_1 + n_2$. Under $p_1 = p_2 = p$, condition on $M = A + C$. The joint mass is

$$\binom{n_1}{x}\binom{n_2}{m-x}p^m(1-p)^{n-m}.$$

Since $M \sim \mathrm{Bin}(n, p)$, division cancels p:

$$h_m(x) = \mathbb{P}(A = x \mid M = m, H_0) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}}{\binom{n}{m}}, \quad \max(0, m - n_2) \le x \le \min(n_1, m).$$

Conditioning fixes both margins; one cell then determines the table. Both margins need not have been fixed by the experiment. Here row sizes are fixed and total successes are conditioned upon.

7.2 Hypotheses, table orientation, and testing procedure

Write the observed table as

	Full recovery	Other	Total
A	a	b	n_1
B	c	d	n_2
Total	m	$n-m$	n

The null is $H_0 : p_1 = p_2$. The alternative is $p_1 \neq p_2$, or a direction specified before inspecting the table. The statistic is the upper-left success count A conditional on the margins.

1. Check that groups contain independent patients and outcomes use the same binary definition. Fix the row/column order and significance level.

2. Calculate $n_1, n_2, m = a+c$ and the feasible integer range $L = \max(0, m-n_2)$ to $U = \min(n_1, m)$.

3. For every feasible x, compute $h_m(x)$. The corresponding table has rows $(x, n_1 - x)$ and $(m - x, n_2 - m + x)$.

4. For $p_1 < p_2$, sum $h_m(x)$ over $x \le a$; for $p_1 > p_2$, sum over $x \ge a$. For the two-sided probability-ordering convention, sum over all x with $h_m(x) \le h_m(a)$, including probability ties.

5. Reject if the selected p-value is at most α. Report the conditioning, alternative, observed risks and effect estimate. If the support contains a single table, the conditional p-value is one.

7.3 Directional alternatives and likelihood ordering

Define the population odds ratio

$$\omega = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

For $0 < p_1, p_2 < 1$, the null $p_1 = p_2$ is equivalent to $\omega = 1$. For observed rows (a, b) and (c, d), the ordinary sample odds ratio is $ad/(bc)$ when defined. This is a ratio of odds, not a ratio of risks. Under general probabilities the conditional mass is proportional to $w(x)\omega^x$, where $w(x) = \binom{n_1}{x}\binom{n_2}{m-x}$. For $\omega_2 > \omega_1$, the likelihood ratio is a positive constant times $(\omega_2/\omega_1)^x$, increasing in x. Thus larger A supports $p_1 > p_2$, and smaller A supports $p_1 < p_2$. This supplies a monotone-likelihood-ratio conditional testing problem. It does not establish unrestricted optimality among all unconditional tests with nuisance parameters.

For observed a, use $\sum_{x \ge a} h_m(x)$ or $\sum_{x \le a} h_m(x)$ for a prespecified direction. A common two-sided convention is

$$p_{\text{two}} = \sum_{x : h_m(x) \le h_m(a)} h_m(x).$$

It need not equal twice the smaller tail. State the convention used rather than treating all two-sided exact tests as identical.

7.4 Why exactness holds

Group outcomes by their ordering values and include ties. At p-value threshold α, the rejection set can include only groups with cumulative null mass at most α. Thus $\mathbb{P}(\text{reject} \mid M = m) \le \alpha$. Averaging gives

$$\mathbb{P}_p(\text{reject}) = \sum_m \mathbb{P}(\text{reject} \mid M = m)\mathbb{P}_p(M = m) \le \alpha$$

for every common p. Exact means finite-sample control, not equality of rejection probability to α. Discreteness typically causes conservativeness. Boundary randomization can attain exact conditional size; ordinary reported tests generally avoid this. Mid-p modifications discount part of the observed mass but do not retain the same universal size guarantee.

7.5 A small recovery pilot

Suppose success/failure counts under A and B are $(1, 9)$ and $(8, 2)$. For a prespecified alternative $p_A < p_B$,

$$p_L = \frac{\binom{10}{0}\binom{10}{9} + \binom{10}{1}\binom{10}{8}}{\binom{20}{9}} = 460/167960 \approx .002739.$$

The two-sided probability-ordered set is $\{0, 1, 8, 9\}$ and its p-value is .005477. Symmetry explains doubling here only. Uncorrected Pearson gives $X^2 = 9.8990$ and approximate p-value .00165. These

finite-sample calibrations can differ without a computational error. Choosing a favorable one-sided direction after inspecting the data invalidates the prespecified one-sided analysis.

7.6 Permutation versus randomization

Uniformly assigning n_1 pooled binary observations to group 1 gives a hypergeometric count. This is a permutation interpretation under exchangeability. In a randomized experiment the assignment calculation can instead test a sharp null that every patient's outcome is unaffected by treatment. That sharp causal null and equality of population proportions are distinct hypotheses even when calculations coincide. Exactness does not mean universally highest power among every design-specific alternative procedure.

8 McNemar: marginal change within patients

Suppose now that the hospital follows the same 100 patients from discharge to day 30. At both visits, symptom control is recorded as yes or no using the same clinical criterion. The results are:

	Day 30: yes	Day 30: no	Total
Discharge: yes	40	18	58
Discharge: no	6	36	42
Total	46	54	100

The observed proportion with controlled symptoms falls from $58/100 = 0.58$ to $46/100 = 0.46$. The question is whether this 12-percentage-point decrease provides evidence of a change in the population probability of symptom control.

These are *paired observations*: each patient contributes a before-and-after outcome. The two sample proportions are therefore generally dependent. We should analyze the changes within patients, rather than treat the two visits as independent samples. The first and fourth groups are **concordant pairs**; their binary status does not change. The second and third groups are **discordant pairs**. McNemar's test assesses whether losses and gains are balanced in the population.

8.1 Sampling model and the meaning of marginal equality

For patient i, define

$$U_i = \begin{cases} 1, & \text{symptom control at discharge,} \\ 0, & \text{otherwise,} \end{cases} \qquad V_i = \begin{cases} 1, & \text{symptom control at day 30,} \\ 0, & \text{otherwise.} \end{cases}$$

Assume that the pairs

$$(U_1, V_1), \ldots, (U_n, V_n)$$

are iid. Dependence between U_i and V_i within a patient is allowed.

Write

$$\pi_{uv} = \mathbb{P}(U_i = u, V_i = v), \qquad u, v \in \{0, 1\},$$

and denote the observed counts by

	$V = 1$	$V = 0$	Total
$U = 1$	a	b	$a + b$
$U = 0$	c	d	$c + d$
Total	$a + c$	$b + d$	n

Thus

$$(A, B, C, D) \sim \text{Mult}\big(n; \pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\big).$$

The population probabilities of symptom control at the two visits are

$$p_{\text{before}} = \pi_{11} + \pi_{10}, \qquad p_{\text{after}} = \pi_{11} + \pi_{01}.$$

Consequently,

$$p_{\text{after}} - p_{\text{before}} = \pi_{01} - \pi_{10}.$$

The two-sided hypothesis is

$$H_0 : p_{\text{after}} = p_{\text{before}} \qquad \text{against} \qquad H_1 : p_{\text{after}} \neq p_{\text{before}},$$

or, equivalently,

$$\boxed{H_0 : \pi_{10} = \pi_{01}.}$$

This is called **marginal homogeneity**: the marginal binary distribution is the same at the two visits. It does not require individual patients to retain the same status. Many patients may change status under the null, provided gains and losses are equally likely.

For a prespecified improvement alternative,

$$H_1 : p_{\text{after}} > p_{\text{before}} \iff \pi_{01} > \pi_{10}.$$

For deterioration, the inequality is reversed.

8.2 Why the discordant pairs contain the relevant direction information

Define the within-patient change

$$W_i = V_i - U_i.$$

Then

$$W_i = \begin{cases} -1, & (U_i, V_i) = (1, 0) : \text{ loss}, \\ +1, & (U_i, V_i) = (0, 1) : \text{ gain}, \\ 0, & U_i = V_i : \text{ no change}. \end{cases}$$

Hence

$$\sum_{i=1}^{n} W_i = C - B, \qquad \widehat{p}_{\text{after}} - \widehat{p}_{\text{before}} = \frac{C - B}{n}.$$

The concordant patients contribute zero to the difference. The direction of the change depends on the balance between B and C, not on whether the concordant patients are mostly controlled or mostly uncontrolled.

There are nevertheless two distinct quantities:

$$s = \pi_{10} + \pi_{01} \quad \text{and} \quad q = \frac{\pi_{10}}{\pi_{10} + \pi_{01}},$$

where q is defined when $s > 0$.

Here,

- s is the probability that a patient changes status;
- q is the probability that a change is a loss, conditional on a change occurring.

Under marginal equality,

$$q = \frac{1}{2}.$$

Thus the testing question becomes:

Among patients whose status changes, are losses and gains equally likely?

8.3 Deriving the exact conditional distribution

Let

$$M = B + C$$

be the number of discordant pairs. Combining the two concordant categories gives

$$(B, C, n - M) \sim \text{Mult}(n; sq, s(1-q), 1-s).$$

Therefore, for $0 \le x \le m$,

$$\mathbb{P}(B = x, C = m - x) = \frac{n!}{x!(m-x)!(n-m)!}(sq)^x\{s(1-q)\}^{m-x}(1-s)^{n-m}.$$

Also,

$$M \sim \text{Bin}(n, s), \qquad \mathbb{P}(M = m) = \binom{n}{m} s^m (1-s)^{n-m}.$$

Dividing the joint probability by this marginal probability gives

$$\begin{aligned}\mathbb{P}(B = x \mid M = m) &= \frac{\frac{n!}{x!(m-x)!(n-m)!} s^m q^x (1-q)^{m-x}(1-s)^{n-m}}{\binom{n}{m} s^m (1-s)^{n-m}} \\ &= \binom{m}{x} q^x (1-q)^{m-x}.\end{aligned}$$

Thus

$$\boxed{B \mid M = m \sim \text{Bin}(m, q).}$$

Under the null,

$$\boxed{B \mid M = m, H_0 \sim \text{Bin}(m, 1/2).}$$

Conditioning eliminates the nuisance parameter s, which describes how frequently patients change status. The split between the two concordant categories is also irrelevant to this conditional law.

Equivalently, for $0 < s < 1$, write

$$r = \frac{\pi_{11}}{1-s}.$$

The full multinomial likelihood factors, apart from its coefficient, as

$$L(s, q, r) \propto s^{b+c}(1-s)^{a+d} q^b (1-q)^c r^a (1-r)^d.$$

The factor involving q is exactly the likelihood of a binomial experiment with $b + c$ trials and b losses.

8.4 Exact test: statistic, p-value, and procedure

For the two-sided alternative, use the imbalance

$$T = |B - C| = |2B - M|.$$

Conditional on the observed $m = b + c > 0$, let

$$B^* \sim \mathrm{Bin}(m, 1/2).$$

The exact two-sided p-value is

$$p_{\text{exact}} = \mathbb{P}\big(|2B^* - m| \geq |b - c|\big).$$

Because the null binomial distribution is symmetric,

$$\boxed{p_{\text{exact}} = \min\left\{1,\ 2\sum_{x=0}^{\min(b,c)} \binom{b+c}{x} 2^{-(b+c)}\right\}.}$$

The cap at one handles the central outcome when $b = c$.

The directional p-values are

$$p_{\text{improvement}} = \mathbb{P}(B^* \leq b), \qquad p_{\text{deterioration}} = \mathbb{P}(B^* \geq b).$$

Small B supports improvement because B counts losses, while $C = m - B$ counts gains.

Testing procedure.

1. Specify the binary outcome, the two occasions, the alternative, and the significance level α.
2. Construct the table using complete paired observations. Identify b, the number of losses, and c, the number of gains.
3. Calculate $m = b + c$. If $m = 0$, report the exact p-value as one. There is no observed directional discordance.
4. If $m > 0$, calculate the appropriate binomial tail probability under $\mathrm{Bin}(m, 1/2)$.
5. Reject H_0 if the p-value is at most α. Report the before and after proportions, the estimated change, and the discordant count alongside the test.

For each fixed m, the conditional test has rejection probability at most α. Therefore, under H_0,

$$\mathbb{P}(\text{reject}) = \sum_{m=0}^{n} \mathbb{P}(\text{reject} \mid M = m)\mathbb{P}(M = m) \leq \alpha.$$

This is finite-sample validity under the paired sampling model. Discreteness generally prevents the nonrandomized rejection probability from being exactly α.

8.5 Deriving the large-sample McNemar statistic

Under the conditional null,

$$\mathbb{E}(B \mid M = m) = \frac{m}{2}, \qquad \mathrm{Var}(B \mid M = m) = \frac{m}{4}.$$

As $m \to \infty$, the binomial CLT gives

$$\frac{B - m/2}{\sqrt{m/4}} \Rightarrow N(0, 1).$$

Since $C = m - B$,

$$\frac{B - m/2}{\sqrt{m/4}} = \frac{2B - m}{\sqrt{m}} = \frac{B - C}{\sqrt{B + C}}.$$

Squaring gives

$$\boxed{X_M^2 = \frac{(B - C)^2}{B + C} \Rightarrow \chi_1^2.}$$

The approximate two-sided test rejects when

$$X_M^2 > \chi_{1,1-\alpha}^2,$$

and its p-value is

$$p_{\text{asym}} = \mathbb{P}\left(\chi_1^2 \geq \frac{(b - c)^2}{b + c}\right).$$

The relevant count is $b + c$, not merely n. If the population discordance probability $s > 0$ is fixed, then $M/n \to s$, so $M \to \infty$ as $n \to \infty$. This supplies an unconditional large-sample justification. A large n with only a few discordant pairs does not ensure an accurate chi-square approximation.

An alternative derivation starts from $W_i = V_i - U_i$. Under H_0,

$$\mathbb{E}W_i = 0, \qquad \mathrm{Var}(W_i) = \mathbb{E}W_i^2 = s.$$

The CLT and the consistent estimate $\widehat{s} = (B + C)/n$ give

$$\frac{\sqrt{n}\,\overline{W}}{\sqrt{\widehat{s}}} = \frac{C - B}{\sqrt{B + C}} \Rightarrow N(0, 1),$$

leading to the same squared statistic.

Continuity correction. A common corrected statistic is

$$X_{M,\text{cc}}^2 = \frac{\{\max(0, |b - c| - 1)\}^2}{b + c}.$$

The subtraction of one reflects the spacing of the discordant-count imbalance when approximating a discrete distribution by a continuous one. The corrected procedure remains an approximation.

8.6 Likelihood-ratio formulation

Conditional on $M = m > 0$, the likelihood for q is

$$L(q) \propto q^b (1-q)^c.$$

The unrestricted MLE is $\widehat{q} = b/m$, whereas the null fixes $q = 1/2$. Thus

$$\Lambda = \frac{(1/2)^m}{(b/m)^b (c/m)^c},$$

with boundary terms interpreted by continuity. The deviance is

$$\boxed{G_M^2 = -2\log\Lambda = 2\left[b\log\frac{2b}{b+c} + c\log\frac{2c}{b+c}\right].}$$

Under the null and as $m \to \infty$, the same Taylor expansion used for binomial goodness of fit gives

$$G_M^2 - X_M^2 \xrightarrow{P} 0, \qquad G_M^2 \Rightarrow \chi_1^2.$$

The exact binomial test, Pearson-type McNemar statistic, and conditional deviance therefore address the same null but use different finite-sample calibrations.

8.7 Return to the hospital example

For

$$(a, b, c, d) = (40, 18, 6, 36),$$

there are

$$m = b + c = 24$$

discordant patients. Under the null, the expected numbers of losses and gains, conditional on these 24 changes, are 12 each.

The uncorrected McNemar statistic is

$$X_M^2 = \frac{(18-6)^2}{24} = 6,$$

giving

$$p_{\text{asym}} = \mathbb{P}(\chi_1^2 \geq 6) \approx 0.01431.$$

The exact two-sided p-value is

$$p_{\text{exact}} = 2\mathbb{P}\{\text{Bin}(24, 1/2) \leq 6\} \approx 0.02266.$$

The continuity-corrected statistic is

$$X_{M,\text{cc}}^2 = \frac{11^2}{24} \approx 5.0417,$$

with approximate p-value 0.02474.

All three reject marginal equality at the 5% level. There is evidence of a change, and its observed direction is deterioration: losses exceed gains.

The different p-values do not indicate a computational error. They reflect an exact discrete calculation and two continuous approximations to it.

For comparison, if $n = 1000$, $b = 3$, and $c = 1$, then

$$p_{\text{exact}} = 2\mathbb{P}\{\text{Bin}(4, 1/2) \leq 1\} = 2\frac{1+4}{16} = 0.625.$$

The 996 concordant patients do not turn four directional changes into a large conditional sample.

8.8 Estimating the magnitude of marginal change

The population change is

$$\Delta = p_{\text{after}} - p_{\text{before}} = \pi_{01} - \pi_{10} = \mathbb{E}W_i,$$

estimated by

$$\widehat{\Delta} = \overline{W} = \frac{c-b}{n}.$$

Because $W_i^2 = 1$ precisely for a discordant pair,

$$\mathbb{E}W_i^2 = \pi_{10} + \pi_{01} = s.$$

Therefore,

$$\operatorname{Var}(W_i) = s - \Delta^2, \qquad \operatorname{Var}(\widehat{\Delta}) = \frac{s - \Delta^2}{n}.$$

A plug-in standard error is

$$\widehat{\operatorname{se}}(\widehat{\Delta}) = \sqrt{\frac{(b+c)/n - \widehat{\Delta}^2}{n}}.$$

For the hospital data,

$$\widehat{\Delta} = -0.12, \qquad \widehat{\operatorname{se}} = \sqrt{\frac{0.24 - 0.0144}{100}} \approx 0.04750.$$

An approximate 95% normal interval is

$$-0.12 \pm 1.96(0.04750) \approx (-0.2131, -0.0269).$$

Thus the estimated decline is 12 percentage points, with an approximate interval from 2.69 to 21.31 percentage points of decline.

This interval is not obtained by inverting the exact conditional test. The exact binomial analysis concerns

$$q = \mathbb{P}(\text{loss} \mid \text{discordance}),$$

whereas the marginal effect satisfies

$$\Delta = s(1 - 2q).$$

Estimating Δ also requires information about the frequency of discordance s. A binomial confidence interval for q must not be reported as an interval for Δ.

With sparse discordance or boundary estimates, the elementary normal interval can perform poorly. In particular, an estimated standard error of zero when no changes are observed does not establish that the population change is known exactly.

8.9 Why independent-sample tests and Fisher's test answer different questions

The variance of the paired difference in proportions is

$$\operatorname{Var}(\overline{V} - \overline{U}) = \frac{p_{\text{after}}(1 - p_{\text{after}}) + p_{\text{before}}(1 - p_{\text{before}}) - 2\operatorname{Cov}(U_i, V_i)}{n},$$

where

$$\operatorname{Cov}(U_i, V_i) = \pi_{11} - p_{\text{before}} p_{\text{after}}.$$

An independent two-proportion analysis omits the covariance term. Depending on its sign, that can overstate or understate uncertainty.

Applying Fisher's exact test to the before-by-after table raises another question: are before and after status associated? Its independence null, for positive cell probabilities, is

$$\pi_{11}\pi_{00} = \pi_{10}\pi_{01}.$$

McNemar's null is instead

$$\pi_{10} = \pi_{01}.$$

Neither null implies the other.

For example,

$$(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}) = (0.45, 0.05, 0.05, 0.45)$$

gives equal before and after probabilities, both 0.5, but strong association between visits. McNemar's null is true while the independence null is false.

Conversely, independent before and after outcomes can have different success probabilities. Independence would then hold while marginal homogeneity would fail.

8.10 Assumptions and interpretation

The exact derivation assumes iid patient pairs with a common joint distribution. It permits arbitrary dependence within each pair. Additional clustering, informative missing follow-up, or non-identically distributed pairs may require a different analysis.

The outcome definition must be comparable at both visits, and the table must preserve the pairing. With missing follow-up, a complete-pair analysis directly describes the observed complete pairs; extension to all enrolled patients needs assumptions about the missingness mechanism.

Finally, McNemar's test identifies a change in marginal probabilities. A before-and-after change alone does not identify a causal effect of care: natural recovery, time trends, changes in measurement, and other interventions may also contribute.

9 Simpson's paradox: recovery and baseline severity

9.1 The hospital comparison

The dashboard reports overall recovery of 78.00% for pathway A and 82.57% for B. Before interpreting B as better, the team separates patients by *baseline* severity. In this illustrative observational cohort,

	Pathway A		Pathway B	
Severity	Recovered/total	Rate	Recovered/total	Rate
Mild	81/87	93.10%	234/270	86.67%
Severe	192/263	73.00%	55/80	68.75%
Combined	273/350	78.00%	289/350	82.57%

A has the higher recovery rate in each severity group but the lower rate when groups are combined. The combined counts are obtained by adding numerators and denominators, not by averaging the displayed percentages. Simpson's paradox is this reversal of an association after aggregation; it is not itself a hypothesis test or a contradiction in probability theory.

9.2 Deriving the reversal from unequal weights

Let R indicate recovery, $T \in \{A, B\}$ the pathway, and S baseline severity. Define

$$r_{ts} = \mathbb{P}(R = 1 \mid T = t, S = s), \qquad w_{ts} = \mathbb{P}(S = s \mid T = t).$$

The law of total probability gives

$$r_t = \mathbb{P}(R = 1 \mid T = t) = \sum_s w_{ts} r_{ts}.$$

In the observed table, pathway A's mild-case proportion is $87/350 = .2486$, whereas B's is $270/350 = .7714$. Thus A has 75.14% severe cases and B only 22.86%. Since severe cases have lower recovery under either pathway, the overall comparison weights different clinical populations. Explicitly,

$$\widehat{r}_A = \frac{87}{350}\frac{81}{87} + \frac{263}{350}\frac{192}{263} = \frac{273}{350},$$
$$\widehat{r}_B = \frac{270}{350}\frac{234}{270} + \frac{80}{350}\frac{55}{80} = \frac{289}{350}.$$

The within-stratum risk differences, A minus B, are 6.44 and 4.25 percentage points, while the pooled difference is -4.57 percentage points. One useful decomposition is

$$r_A - r_B = \underbrace{\sum_s w_{As}(r_{As} - r_{Bs})}_{\text{within-stratum contrasts using A's weights}} + \underbrace{\sum_s (w_{As} - w_{Bs}) r_{Bs}}_{\text{difference in patient composition}}.$$

The first term is positive when A is better within each stratum. The second can be negative and larger in magnitude, producing the observed reversal. The decomposition uses B's stratum rates as a reference; other reference choices give different decompositions of the same marginal difference.

9.3 Common-weight standardization and its interpretation

Choose a target severity distribution $w_s \geq 0$, $\sum_s w_s = 1$, and define

$$r_A^{\text{std}} = \sum_s w_s r_{As}, \qquad r_B^{\text{std}} = \sum_s w_s r_{Bs}.$$

Using the *same* weights ensures

$$r_A^{\text{std}} - r_B^{\text{std}} = \sum_s w_s (r_{As} - r_{Bs}).$$

If all within-stratum differences are nonnegative, the standardized difference cannot be negative. It is strictly positive if at least one positively weighted stratum has a strictly positive difference. This proves why the sign reversal above requires different weights when every stratum favors A.

For a hypothetical target population with half mild and half severe cases,

$$\widehat{r}_A^{\text{std}} = \tfrac{1}{2}(81/87) + \tfrac{1}{2}(192/263) = .83054,$$
$$\widehat{r}_B^{\text{std}} = \tfrac{1}{2}(234/270) + \tfrac{1}{2}(55/80) = .77708.$$

The standardized difference is 5.35 percentage points in favor of A. Equal weights are an explicit target-population choice. They are not the unique correct weights, nor are these point estimates hypothesis tests. Uncertainty in stratum-specific rates remains and would need its own interval or testing procedure.

9.4 What to analyze and report

1. Identify whether severity was measured before treatment and whether it plausibly influences pathway allocation and recovery.

2. Show pooled counts and rates, then stratum counts, rates, and sample sizes. Calculate the severity proportions separately for A and B.

3. Use the weighted-average identity to explain the pooled comparison. Do not choose pooled or stratified results according to which favors a pathway.

4. Specify the estimand: recovery in each pathway's actual patient mix, a within-severity comparison, or a comparison standardized to a stated mix. For standardization use common weights and report them.

5. Distinguish descriptive adjustment from causal identification. Interpretation as a treatment effect would require additional assumptions, such as no unmeasured confounding given the baseline variables, consistency, and positive treatment probabilities in the target strata.

Adjusting for a genuine baseline confounder can be useful. Conditioning on post-treatment mediators or colliders can answer a different question or introduce bias. Marginal and conditional comparisons need not agree, and neither automatically estimates a causal effect. Large within-stratum differences in effects should be reported rather than hidden in a common summary. Stratified association tests and regression are possible extensions, not prerequisites for understanding the reversal in this session.

10 Base R procedures

These blocks follow the hospital examples and can be run sequentially. They use only base R and the standard `stats` package. Table orientation is explicit because swapping a row or column can reverse a one-sided alternative. The accompanying R file contains the same code as one flat script.

10.1 Session 1: benchmark, tables, and fitted goodness of fit

```
# SESSION 1: a known recovery benchmark. Null probabilities are fixed.
observed <- c(Full = 110, Partial = 50, None = 40)
benchmark <- c(0.5, 0.3, 0.2)
gof <- chisq.test(observed, p = benchmark)
gof$expected
gof                     # X-squared = 2.6667, df = 2, p = 0.2636

# The likelihood-ratio deviance tests the same specified null.
expected <- sum(observed) * benchmark
positive <- observed > 0
deviance <- 2 * sum(observed[positive] *
                  log(observed[positive] / expected[positive]))
c(G2 = deviance, p_value = pchisq(deviance, 2, lower.tail = FALSE))

# Independence (one cross-classified sample) and homogeneity (separate
# independent groups) share this computation but have different designs.
site_table <- matrix(c(30, 20, 10, 20, 30, 40),
                     nrow = 2, byrow = TRUE,
                     dimnames = list(Site = c("A", "B"),
```

```
                                    Recovery = c("Full", "Partial", "None")))
site_test <- chisq.test(site_table, correct = FALSE)
site_test$expected
site_test               # X-squared = 16.6667, df = 2, p about 0.000240

# Fitted milestone model: estimate theta from the SAME category counts.
# chisq.test() does not know that one parameter was estimated. Therefore
# calculate the statistic and use df = 3 - 1 - 1 = 1 explicitly.
n <- sum(observed)
theta_hat <- (2 * observed[1] + observed[2]) / (2 * n)
fitted_prob <- c(theta_hat^2, 2 * theta_hat * (1 - theta_hat),
                 (1 - theta_hat)^2)
fitted_expected <- n * fitted_prob
fitted_X2 <- sum((observed - fitted_expected)^2 / fitted_expected)
fitted_expected        # 91.125, 87.750, 21.125
c(theta = theta_hat, X2 = fitted_X2,
  p_value = pchisq(fitted_X2, df = 1, lower.tail = FALSE))
```

10.2 Fisher: independent groups and hypergeometric enumeration

```
# SESSION 2: Fisher's exact test for two INDEPENDENT groups.
# Success is the FIRST column. With this orientation, "less" means
# recovery odds under pathway A are lower than under pathway B.
pilot <- matrix(c(1, 9, 8, 2), nrow = 2, byrow = TRUE,
                dimnames = list(Pathway = c("A", "B"),
                                Recovery = c("Yes", "No")))
fisher.test(pilot, alternative = "two.sided")  # p = 0.00547749
# Use this direction only if it was specified before seeing the data.
fisher.test(pilot, alternative = "less")       # p = 0.00273875

# Enumerate the null tables. Given both margins, one cell determines all.
# dhyper(x, m, n, k): x successes in k draws from m successes + n failures.
n_A <- sum(pilot[1, ])
n_B <- sum(pilot[2, ])
total_success <- sum(pilot[, 1])
N <- sum(pilot)
x <- seq.int(max(0, total_success - n_B), min(n_A, total_success))
null_prob <- dhyper(x, m = total_success, n = N - total_success, k = n_A)
observed_a <- pilot[1, 1]
observed_prob <- dhyper(observed_a, total_success, N - total_success, n_A)
data.frame(A_success = x, A_failure = n_A - x,
           B_success = total_success - x,
           B_failure = n_B - total_success + x, null_prob = null_prob)
# Include probability ties; the tolerance guards against roundoff.
p_two <- sum(null_prob[null_prob <= observed_prob * (1 + 1e-7)])
c(lower_tail = sum(null_prob[x <= observed_a]), two_sided = p_two)
# The sample odds ratio is NOT generally fisher.test()$estimate:
# that function reports a conditional MLE of the odds ratio.
sample_OR <- pilot[1, 1] * pilot[2, 2] / (pilot[1, 2] * pilot[2, 1])
sample_OR             # 1/36; also report observed risks 0.1 versus 0.8
```

10.3 McNemar: paired change and exact binomial calibration

```
# McNemar: rows are BEFORE and columns are AFTER for the SAME patients.
# b = loss of control; c = gain of control. Only b and c enter the test.
```

```
paired <- matrix(c(40, 18, 6, 36), nrow = 2, byrow = TRUE,
                 dimnames = list(Before = c("Yes", "No"),
                                 After = c("Yes", "No")))
b <- paired[1, 2]
c_gain <- paired[2, 1]
discordant <- b + c_gain
# mcnemar.test() is asymptotic, even when correct = TRUE.
# If there are no discordances, the exact p-value is 1; skip 0/0 ratios.
if (discordant > 0) {
  mcnemar.test(paired, correct = FALSE)   # X-squared = 6, p = 0.01431
  mcnemar.test(paired, correct = TRUE)    # p = 0.02474
  exact_change <- binom.test(b, discordant, p = 0.5)
  exact_change$p.value                    # 0.02265584
  # Improvement means gains exceed losses, hence a SMALL loss count b.
  # Select this alternative in advance if improvement is the question.
  binom.test(b, discordant, p = 0.5, alternative = "less")$p.value
} else {
  exact_change_p <- 1
  exact_change_p
}
# binom.test()'s interval concerns the probability of LOSS conditional
# on discordance; it is not a CI for the marginal before-after change.
n_paired <- sum(paired)
before_rate <- sum(paired[1, ]) / n_paired
after_rate <- sum(paired[, 1]) / n_paired
change <- (c_gain - b) / n_paired
se_change <- sqrt((discordant / n_paired - change^2) / n_paired)
c(before = before_rate, after = after_rate, change = change)
change + c(-1, 1) * qnorm(0.975) * se_change
# This normal interval is approximate and is not useful with sparse
# discordance or at boundaries. Fisher on 'paired' tests a DIFFERENT null.
```

10.4 Simpson: separate group weights and common target weights

```
# Simpson's paradox: a separate illustrative observational cohort.
# Use the SAME baseline-severity categories for both pathways.
severity <- c("Mild", "Severe")
success_A <- c(81, 192)
total_A <- c(87, 263)
success_B <- c(234, 55)
total_B <- c(270, 80)
risk_A <- success_A / total_A
risk_B <- success_B / total_B
weight_A <- total_A / sum(total_A)
weight_B <- total_B / sum(total_B)
data.frame(severity, risk_A, risk_B, difference = risk_A - risk_B,
           weight_A, weight_B)
# Within BOTH strata A has higher recovery. Overall, A has lower recovery.
pooled_A <- sum(success_A) / sum(total_A)
pooled_B <- sum(success_B) / sum(total_B)
c(A = pooled_A, B = pooled_B, difference = pooled_A - pooled_B)

# Equal weights describe a hypothetical target population with 50% in
# each severity stratum. They are an explicit choice, not uniquely correct.
target_weights <- c(0.5, 0.5)
standard_A <- sum(target_weights * risk_A)
standard_B <- sum(target_weights * risk_B)
```

```
c(A = standard_A, B = standard_B, difference = standard_A - standard_B)
# Common weights give A = 0.83054, B = 0.77708. Changing the weights changes
# the target population. Standardization alone does not establish causality.
barplot(rbind(A = risk_A, B = risk_B), beside = TRUE,
        names.arg = severity, ylim = c(0, 1),
        col = c("steelblue", "tan3"), ylab = "Observed recovery proportion",
        legend.text = c("Pathway A", "Pathway B"),
        args.legend = list(x = "bottomleft", bty = "n"))
```

Software conventions to recognize

`fisher.test` uses probability ordering for its two-sided p-value and reports a conditional MLE of the odds ratio, which can differ from $ad/(bc)$. Its exact interval concerns the odds ratio. `mcnemar.test` uses a chi-square approximation; its default continuity correction does not make it exact. Use `binom.test` on discordant counts for the exact paired test. The latter's confidence interval concerns the conditional direction probability q, not the marginal change Δ. For a fitted multinomial family, manually use the appropriate degrees of freedom; `chisq.test` does not infer how many parameters you estimated.

Documentation: Fisher's exact test, McNemar's test, and exact binomial test.

11 Method-selection checklist

1. **Identify the sampling unit and dependence.** Is each row a new patient, or do two measurements belong to one patient? Independence across patients does not imply independence of two measurements within a patient.

2. **State the population question before choosing a statistic.** Distinguish a specified distribution, a fitted family, association, equal group distributions, and equal paired marginals.

3. **Match the reference distribution to what is fixed or conditioned on.** Fisher conditions on margins; exact McNemar conditions on discordance; chi-square approximations rely on their stated regularity conditions.

4. **Check the relevant information count.** Sparse expected counts can impair Pearson or deviance calibration. For McNemar, inspect $b + c$, not only total n. Exact conditional methods can be conservative.

5. **Inspect group composition.** For an observational comparison, check prespecified baseline strata. If standardizing, state the common target weights and retain stratum-specific results.

6. **Report magnitude and uncertainty along with the p-value.** Specify the outcome definition, table orientation, alternative, correction or exact convention, and any missing-pair exclusions. Rejection does not establish causality or clinical importance; nonrejection does not prove equality.

Question and design	Method and calibration	Main check
One cohort versus specified recovery probabilities	Pearson or deviance; χ^2_{k-1} approximation	Benchmark fixed; positive probabilities; adequate expected counts.
One cohort versus fitted milestone family	Fitted Pearson/deviance; χ^2_{k-1-d} under regular fitting	Fit the null model; account for estimated dimensions.
Site and recovery in one cross-classified sample	Independence test; $\chi^2_{(r-1)(c-1)}$	Independent patients; distinguish association from independence of sampling units.
Recovery categories in independent pathway groups	Homogeneity test; $\chi^2_{(r-1)(c-1)}$	Groups independent; same categories; adequate expected counts.
Small independent binary groups	Fisher's exact conditional test; hypergeometric law	Fix orientation and alternative; condition on margins; state two-sided convention.
Before/after binary outcomes for the same patients	McNemar; exact $\mathrm{Bin}(b+c, 1/2)$ or approximate χ^2_1	Independent pairs; count discordances; do not substitute Fisher's independence null.
Pooled and severity-specific comparisons disagree	Display strata and weights; standardize to a common target if appropriate	Simpson's paradox is not a test; choose the estimand and justify adjustment.