

Lesson I Methods of Analysis for a Single 2x2 Table

The typical two by two table will be identified in epidemiology as have an exposure variable (E) with 2 levels: exposed, unexposed (represented by + and – or 1 and 2 respectively) and a disease variable (D) with 2 levels: disease present and disease not present (represented in a similar fashion to exposure).

Disease	Exposure		
	1 +	2 -	
1 +	A	B	n ₁
2 -	C	d	n ₂
Total	m ₁	m ₂	N

Sampling Framework

1. No margins fixed and only N is known in advance of observing the pair (D, E) jointly. The null hypothesis of independence is appropriate:

$$H_0 : P(D = i, E = j) = P(D = i) \cdot P(E = j) \quad i = 1, 2 \quad j = 1, 2$$

2. One marginal total is fixed (either D or E). If E is fixed and we observe D then this is called a *cohort study*. If D is fixed and we observe E then this is called a *case-control study*. The null hypothesis is now homogeneity of the marginals:

$$H_0 : P(E = j | D = 1) = P(E = j | D = 2) \quad \text{or} \quad P(D = j | E = 1) = P(D = j | E = 2)$$

case- control

cohort

3. All margins fixed. This situation will form the basis of our exact analysis of a single 2x2 table. Here the data follow a central-hypergeometric distribution and the null hypothesis is concerned with cell probabilities:

$$H_0 : P(A = a) = \frac{\binom{n_1}{a} \binom{n_2}{m_1 - a}}{\binom{N}{m_1}}$$

For frameworks 1 and 2 above, the usual large sample analysis uses the following test statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

where n_{ij} = observed count in cell $D = i, E = j$ (i, j)

\hat{n}_{ij} = estimated count under H_0 in cell (i, j)

Estimation

It is easy to show using the principle of maximum likelihood that the estimated expected values for the above are given by:

1 Independence

$$\hat{n}_{ij} = N \hat{P}(D = i, E = j) = N \hat{P}(D = i) \cdot \hat{P}(E = j) = N \cdot \frac{n_i}{N} \cdot \frac{m_j}{N} = \frac{n_i m_j}{N}$$

2 Homogeneity (in the case-control study)

$$\hat{n}_{ij} = n_i \cdot \hat{P}(E = j) = n_i \frac{m_j}{N}$$

where $\hat{n}_{ij} = \frac{(\text{Row total}) \cdot (\text{Column total})}{\text{Grand total}}$

3 Hypergeometric

$$\hat{P}(A = a) = \frac{\binom{n_1}{n_{11}} \binom{n_2}{n_{21}}}{\binom{N}{m_1}}$$

Assessing significance in a 2×2 table

Asymptotic Distribution (based on adequate (?) sample size)

One may show that in the case of a 2×2 table

$$X^2 = \frac{(ad - bc)^2 \cdot N}{n_1 n_2 m_1 m_2} \quad p = \Pr\{\chi^2(1) \geq X^2\}$$

where $\chi^2_{1-\alpha}(\nu)$ = upper α percent point of the chi-square distribution with ν degrees of freedom

When will the p-value be accurate?

- i. If you have an independent identically distributed sample (iid)
- ii. $P(D = i, E = j) > 0$ for all i, j
- iii. N is large as measured via \hat{n}_{ij} where a good rule of thumb is
 - (a) $\hat{n}_{ij} > 5$ or
 - (b) not more than 10% of the \hat{n}_{ij} 's < 1

Then $X^2 \xrightarrow{H_0} \chi^2(1)$

If \hat{n}_{ij} are too small

- 1) X^2 is not distributed as a chi-square under H_0
- 2) the p-value will be incorrect
- 3) χ^2 tends to be too large

The solutions are either to use an exact test (hypergeometric) or use a continuity correction to obtain p-values. The hypergeometric is difficult to calculate by hand and therefore the continuity correction was used more frequently in the pre-computer days. The continuity correction is a smoothing technique used to smooth out the cell counts. The rationale is to deal with small expected values to get a better approximation to the correct p-value.

$$X_c^2 = \frac{\left(|ad - bc| - \frac{N}{2}\right)^2 \cdot N}{n_1 n_2 m_1 m_2} \quad \Pr_c(\chi^2(1) > X_c^2) = P_c$$

Since we have the availability of computers, we won't be using continuity corrections. In the event that we have small cell sizes we'll use exact tests.

Example:

OBSERVED

Disease	Exposure		
	1	2	
1	20	10	30
2	15	25	40
total	35	35	70

EXPECTED VALUES

Disease	Exposure	
	1	2
1	15	15
2	20	20

$$X^2 = \frac{(20 \times 25 - 15 \times 10)^2 \times 70}{35 \times 35 \times 30 \times 40} = 5.83 \quad p = P(\chi^2(1) \geq 5.83) = 0.016$$

since $0.016 < 0.05$, we would reject H_0 at the $\alpha = .05$ level.

Exact Distribution – using Fisher’s exact test

Fisher proposed using the central hypergeometric distribution for evaluating the significance of a 2×2 table. Consider the table:

A		n_1
		n_2
m_1	m_2	N

under H_0

$$P(A = a | n_1, n_2, m_1, m_2) = \frac{\binom{n_1}{a} \cdot \binom{n_2}{m_1 - a}}{\binom{N}{m_1}}$$

To assess the significance, we must calculate this probability for all possible values of A leaving all the marginals fixed and see how extreme our observed table is.

Example:

a		5
		6
4	7	11

$a = 0, 1, 2, 3, 4$

$$P(a) = \frac{\binom{5}{a} \cdot \binom{6}{4-a}}{\binom{11}{4}}$$

a =	0	1	2	3	4
P(a)	.0435	.3030	.4546	.1818	.0151
X^2	5.238	1.061	.052	2.213	7.543
$\Pr\{\chi^2(1) \geq X^2\}$.0221	.3030	.8196	.1369	.0060
Exact p-value*					
$\sum P(a)$.0606	.5464	1.0	.2424	.0151

Example of how to obtain the exact p-value:

$$P(A = a_i) = \sum_{all P(a) \leq P(a_i)} P(a) \quad P(A = 3) = P(3) + P(0) + P(4) = .1818 + .0455 + .0151$$

Notice how the asymptotic approximation is not very good for this example when comparing it to the exact value – that’s due to the small sample size (this was done so that the illustration wouldn’t get unwieldy)

* Remember, these are two tailed p-values which are the sum of the values of P(a) for all tables with a deviation from equal proportions which is equal to or greater than the observed in absolute value.

Here are what some of the tables would look like:

0	5	5
4	2	6

$$\frac{0}{5} - \frac{4}{6} = -.67$$

1	4	5
3	3	6

$$\frac{1}{5} - \frac{3}{6} = -.3$$

2	3	5
2	4	6

$$\frac{2}{5} - \frac{2}{6} = .067$$

If a one-tailed test were desired, we would pay attention to the sign of the deviations.