The Chi-square test statistic of 2 x 2 tables

Hans Stocker, Schaffhausen

Version October 2023

Introduction

This notice shows within the context of a 2 x 2 contingency table how the well-known Pearson's test statistic $X^2 = \sum(O-E)^2/E$ can be transfered to the the form $X^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$ with 1 degree of freedom.

Proof

Example:

The succes of two treatments A and B is tabulated.

healed	yes	no	sum
A	80	40	120
B	20	100	120
sum	100	120	240

General notation of a 2 x 2 contingency table:

a	b	a + b
c	d	c + d
a + c	b + d	N

Pearson's Chi-square is summation over all 4 cells:

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i \tag{1}$$

O_i ... observed values.

E_i ... expected values, whereas:

$E_i = \frac{r_T \ x \ c_T}{N}$, ($r_T$...row total, c_T...column total)

Expected values:

$$E(a) = \frac{(a+b)(a+c)}{N} \qquad E(c) = \frac{(c+d)(a+c)}{N}$$
$$E(b) = \frac{(a+b)(b+d)}{N} \qquad E(d) = \frac{(c+d)(b+d)}{N} \tag{2}$$

$$X^2 = \frac{(a-E(a))^2}{E(a)} + \frac{(b-E(b))^2}{E(b)} + \frac{(c-E(c))^2}{E(c)} + \frac{(d-E(d))^2}{E(d)} \tag{3}$$

a - E(a) can be expressed:

$$
\begin{aligned}
a - E(a) &= a - \frac{[(a+b)(a+c)]}{N} \\
&= \frac{[aN - (a+b)(a+c)]}{N} \\
&= \frac{[a(a+b+c+d)] - [(a+b)(a+c)]}{N} \\
&= \frac{a(a+b+c+d) - (a^2+ac+ab+bc)}{N} \\
&= \frac{\cancel{a^2} + \cancel{ab} + \cancel{ac} + ad - \cancel{a^2} - \cancel{ac} - \cancel{ab} - bc}{N} \\
&= \frac{ad-bc}{N} \\
&\Rightarrow \boxed{\text{a -E(a) =} \frac{ad-bc}{N}}
\end{aligned}
\tag{4}
$$

Similarly we will get:

$$
b - E(b) = \frac{ad-bc}{N}, c - E(c) = \frac{ad-bc}{N}, d - E(d) = \frac{ad-bc}{N} \tag{5}
$$

When aubstituting now (4) into (3) we have:

$$
\frac{(ad-bc)^2}{N^2 \cdot E(a)} = \frac{(ad-bc)^2}{N^2} \cdot \frac{1}{E(a)} \tag{6}
$$

and similarly for b, c and we will get:

$$
\begin{aligned}
X^2 &= \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)}\right] \\
&= \frac{(ad-bc)^2}{N}\left[\frac{N}{(a+b)(a+c)} + \frac{N}{(a+b)(b+d)} + \frac{N}{(c+d)(a+c)} + \frac{N}{(c+d)(b+d)}\right] \\
&= \frac{(ad-bc)^2}{N} N \left[\underbrace{\frac{1}{(a+b)(a+c)} + \frac{1}{(a+b)(b+d)}}_{\text{A}} + \underbrace{\frac{1}{(c+d)(a+c)} + \frac{1}{(c+d)(b+d)}}_{\text{B}}\right]
\end{aligned}
$$

Both terms A and B with common denominater result in:

$$= \frac{(ad-bc)^2}{N}\left[\frac{(b+d+a+c)}{(a+b)(a+c)(b+d)} + \frac{(b+d+a+c)}{(a+c)(c+d)(b+d)}\right]$$
$$= \frac{(ad-bc)^2}{N}N\left[\frac{(1)}{(a+b)(a+c)(b+d)} + \frac{(1)}{(a+c)(c+d)(b+d)}\right]$$
$$= \frac{(ad-bc)^2}{N}\left[\frac{(c+d+a+b)}{(a+b)(a+c)(c+d)(b+d)}\right]$$

$$\Rightarrow \boxed{X^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}} \tag{7}$$

Proof that (7) is asymptotically X^2 distributed

The notation of the 2 x 2 table is expanded by probabilities.
Let $p_1 := \frac{a}{n_1}$, the observed proportion of a and
let $p_2 := \frac{b}{n_2}$, the observed proportion of a, and
similarly q_1 and q_2 the observed propotions of c and d.
The 2 x 2 table results now in:

healed	yes	no	
A	a = $n_1 p_1$	b = $n_2 p_2$	
B	c = $n_1 q_1$	c = $n_2 q_2$	
	n_1	n_2	N

Substitute these values into (7) and simplify:

$$X^2 = \frac{Nn_1n_2(p_1q_2 - p_2q_1)^2}{(n_1p_1 + n_2p_2)(n_1q_1 + n_2q_2)(n_1)(n_2)}$$
$$= \frac{N(n_1p_1n_2q_2 - n_2p_2n_1q_1)^2}{(n_1p_1 + n_2p_2)(n_1 + n_2 - (n_1p_1 + n_2p_2))} \tag{8}$$

We note that
$N = n_1 + n2$ and

$p_1 q_1 - p_2 q_1 = p_1(1-p_1) - p_2(1-p_1) = p_1 - p_2$ and we define
$p(a+b) = p := (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$.

Inserting in (8) results in:

$$\frac{(n_1+n_2)n_1 n_2 (p_1-p_2)^2}{(n_1+n_2)p(n_1+n_2)(1-p)} = \left(\frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}} \right)^2 \tag{9}$$

which is square of the **z test statistic for comparing the sampling distribution of two proportions**.

This means that if the square root of (9) is asymptotically standard normally distributed the squared term is asymptotically Chi-square distributed.

Due to the known fact: If $X \sim$ N(0,1) $\Rightarrow$ $X^2 \sim X^2$, we can conlude that (9) is Chi-square distributed with 1 DF.

Contact: hstock@bluewin.ch