Application of Chi-square test statistic, an intuitive introduction for biologists and physicians

Hans Stocker, Schaffhausen

Version October 2023

Contents

1 Introduction

Parts of these explanations follow the calculations as presented by Boris Burkov, June 17, 2021.

1. Motivation

This article adresses to biologists and physicians which are using the Chi-square test in many situations comparing results between groups of patients and biological entities.

Many questions where observed frequencies are compared with expected frequencies, can be assessed and verified by means of Chi-square tests. This article takes a closer look to the mathematical background applying the Chi-square test statistic. In particular three points are considered which are important for the insight how to apply the Chi-square test statistics:

1. The test statistic $X^2 = \sum(O - E)^2/E$, introduced by Pearson has a Chi-square distribution.

2. With the observation of k categories a significance test with (k-1) degrees of freedom will be performed.

3. Because the Hardy-Weinberg equilibrium is introduced quite early in the field of natural sciences, the Chi-square test is treated in particular.

Example

Example for goodness-of-fit test statistic: Do the observed relations of mobbing victims between men and women in a city match with the relation of 3 :1 from a bigger population?

Example 1:

sex	women	men
observed	203	45
expected	188	62
$(O-E)^2/E$	1.554	4.661

$$X^2 = \sum (O-E)^2/E = 1.554 + 4.661 = 6.215$$

general formulation

Deviations are assessed by the Chi-square statistic:

$$X^2 = \sum_i \frac{(O-E_i)^2}{E_i} \tag{1}$$

O ... observed values
E ... expected values

2. Principal mathematical assumptions

A binomial random variable M follows M $\sim$ B(n,p).

n is the number of trials (occurrences).
x is the number of successful trials.
p (x/n) is the probability of success in a single trial.
Parameters are: $\mu = np$, $\sigma = \sqrt{npq}$.

Sentence 1:

M is a continuous random variable with normal distribution $N(\mu,\, \sigma)$. By transformation $\frac{M-\mu}{\sigma}$ we achieve standardisation to $N(0,\, 1)$.

Sentence 2:

The sentence of Moivre-Laplace states that a binomially distributed random variable K $\sim$ Bin(p,n) with n$\to \infty$ converges to a standard normally distribution.

In particular the following holds:

$$\lim_{k\to\infty} \mathcal{P}\left(\frac{K-np}{\sqrt{np(1-p)}} \leq k\right) = \Phi(k) \tag{2}$$

whereby $\Phi(k)$ is the distribution function of the standard normally distribution.

Therefore, there is:

$$\frac{M-\mu}{\sigma} = \frac{K-np}{\sqrt{np(1-p)}} = N(0,\, 1). \tag{3}$$

Sentence 3:

if $Z_1, \ldots, Z_n$ are stochastic independent and random variables with standard normally distribution, then its sum of squares Q are Chi-square distributed with n degrees of freedom;

$$Q = Z_1^2 + \cdots + Z_n^2 \sim X^2(n). \tag{4}$$

2 Goodness-of-fit application

1. Chi-square for random variables with two categories

Calculation of the coin tossing

A coin is tossed 100 times and the number of heads and tails is listed. There should be clarified whether the coin is fair whereby heads and tails are equally distributed to be 50% each.

Example 2:

result toss	head	tail	p (head)
observed	60	40	0.6
expected	50	50	0.5
$(O-E)^2/E$	2	2	

$X^2 = \sum_i (O-E)^2/E^2 = 2 + 2 = 4$

Random variable"result coin toss" is binomially distributed, therefore:

p(head) = 0.5, p(tail) = 0.5

Expected(head) E= np = 100 x 0.5 = 50.

Standard deviation= $\sqrt{npq}$.

Term $X^2 = \sum_i (O-E)^2/E$ can now be reformulated, because

$O_2 = n - O_1$ und q = 1 - p:

$$\begin{aligned} X^2 &= \sum_{j=1}^{2} (O_j - E)^2/E^2 \\ &= \frac{(O_1 - np)^2}{np} + \frac{((n-O_1) - n(1-p))^2}{n(1-p)} \\ &= \frac{(O_1 - np)^2}{np} + \frac{(O_1 - np)^2}{n(1-p)} \\ &= \frac{(O_1 - np)^2(1-p) + (O_1 - np)^2 p}{np(1-p)} \end{aligned}$$

$$= \frac{(O_1 - np)^2(1 - \cancel{p} + \cancel{p})}{np(1-p)}$$
$$= \frac{(O_1 - np)^2}{np(1-p)} \sim X_1^2. \tag{5}$$

$$\text{As } \frac{K - np}{\sqrt{npq}} \sim N(0,\, 1)$$
$$\Rightarrow \text{(according 2) } \frac{(O_1 - np)^2}{np(1-p)} \sim X_1^2. \tag{6}$$

By this way it can nicely be demonstrated that testing with 2 categories leads to a Chi-square test with one degree of freedom, i.e. if one class is known the second class in this context is fixed and there is no further degree of freedom.

Testing a random variable with k = 2 categories $\Rightarrow \sim X_1^2$ (Chi-square with 1 DF)

2. Chi-square for random variables with three categories

It shall be demonstrated that by testing a random variable with 3 categories leads to a Chi-square test with 2 degrees of freedom.

First Pearson's formula with 3 squared terms will first be expanded by another term and afterwards reduced by this term.

Procedure

$X^2 = (O_1 - E_1)^2/E_1$ is expanded by $\frac{(O_2+O_3-n(p_2+p_3))^2}{n(p_2+p_3)}$, which are the collapsed classes 2 and 3.

$$X^2 = \sum_{i=1}^{3}(O_i - E_i)^2/E_i^2$$
$$= \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 + O_3 - n(p_2+p_3))^2}{n(p_2+p_3)} \tag{7}$$
$$- \frac{(O_2 + O_3 - n(p_2+p_3))^2}{n(p_2+p_3)} + \frac{(O_2 - np_2)^2}{np_2} + \frac{(O_3 - np_3)^2}{np_3} \tag{8}$$

As next it will be demonstrated that each (7) and (8) result in a Chi-square with 1 DF each.

Elaboration of (7):

With $O_2 + O_3 = n - O_1$ and $p_2 + p_3 = 1 - p_1$:

$$\frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 + O_3 - n(p_2+p_3))^2}{n(p_2+p_3)}$$

$$
\begin{aligned}
&= \frac{(O_1 - np_1)^2}{np_1} + \frac{(n - O_1 - n(1-p_1))^2}{n(1-p_1)} \\
&= \frac{(O_1 - np_1)^2}{np_1} + \frac{(\cancel{n} - O_1 - \cancel{n} + p_1)^2}{n(1-p_1)} \\
&= \frac{(O_1 - np_1)^2(1-p_1) + (O_1 - np_1)^2 p_1}{np_1(1-p_1)} \\
&= \frac{(O_1 - np_1)^2(1 - p_1 + p_1}{np_1(1-p_1)} \\
&= \frac{(O_1 - np_1)^2}{np_1(1-p_1)^*} \qquad (9)
\end{aligned}
$$

According to (6), for term (9) $\rightarrow \sim X_1^2$.

$(1-p_1)^*$ is probability of collapsed categories 2 + 3.

Elaboration of (8)

Formula part (8) is $\sim X_1^2$ as well. This will be shown in the following:

$$
\begin{aligned}
&-\frac{(O_2 + O_3 - n(p_2+p_3))^2}{n(p_2+p_3)} + \frac{(O_2 - np_2)^2}{np_2} + \frac{(O_3 - np_3)^2}{np_3} \\
&= -\frac{(O_2 + O_3 - n(p_2+p_3))^2 \cdot p_2 p_3}{n(p_2+p_3)\cdot p_2 p_3} + \frac{(O_2 - np_2)^2 \cdot (p_2+p_3)p_3}{np_2 \cdot (p_2+p_3)p3} + \frac{(O_3 - np_3)^2 \cdot (p_2+p_3)p_2}{np_3 \cdot (p_2+p_3)p2} \\
&= \big[(O_2^2 - 2O_2 np_2 + n^2 p_2^2)(p_2 p_3 + p_3^2) + (O_3^2 - 2O_3 np_3 + n^2 p_3^2)(p_2^2 + p_2 p_3) \\
&\quad -((O_2+O_3)^2 - 2(O_2+O_3)n(p_2+p_3) + n^2(p_2^2 + 2p_2p_3 + p_3^2))p_2p_3\big] \\
&\quad /[np_2p_3(p_2+p_3)]. \qquad (10)
\end{aligned}
$$

Calculate:

$$
\begin{aligned}
&= \big[(O_2^2 p_2 p_3 + O_2^2 p_3^2 - 2O_2 np_2^2 p_3 - 2O_2 np_2 p_3^2 + n^2 p_2^3 p_3 + n^2 p_2^2 p_3^2) \\
&\quad +(O_3^2 p_2^2 + O_3^2 p_2 p_3 - 2O_3 np_3 p_2^2 - 2O_3 np_2 p_3^2 + n^2 p_2^2 p_3^2 + n^2 p_2 p_3^3) \\
&\quad -(O_2^2 p_2 p_3 + 2O_2 O_3 p_2 p_3 + O_3^2 p_2 p_3 - 2O_2 np_2^2 p_3 - 2O_2 np_2 p_3^2 - 2O_3 np_2^2 p_3 \qquad (11) \\
&\quad -2O_3 np_2 p_3^2 + n^2 p_2^3 p_3 + 2n^2 p_2^2 p_3^2 + n^2 p_2 p_3^3)\big] \\
&\quad /[np_2p_3(p_2+p_3)].
\end{aligned}
$$

With exception of the 3 terms $O_2^2 p_3^2$, $O_3^2 p_2^2$, $-2O_2O_3p_2p_3$ all other terms cancel away.

$$
\Rightarrow\ (9) = \frac{O_2^2 p_3^2 + O_3^2 p_2^2 - 2O_2 O_3 p_2 p_3}{np_2p_3(p_2+p_3)} = \frac{(O_2 p_3 - O_3 p_2)^2}{np_2p_3(p_2+p_3)} \qquad (12)
$$

Follwing (2) und (3) term (12) is Chi-square distributed, if the square root of it is standard normally distributed.

$$
\text{also } \xi = \frac{O_2 p_3 - O_3 p_2}{\sqrt{np_2p_3(p_2+p_3)}} \sim N(0,1). \qquad (13)
$$

****

Begin of proof for (13)

Proof (13)

It must be demonstrated, that variables O_2 and O_3 are independent, therefore
1. Var[ξ] = 1 und Cov(O_2, O_3) = 0, sowie
2. $E[\xi] = 0$.

O_2 and O_3 following (2) are standard normally distributed random variables.

Expectations are $E[O_2] = np_2$ and $E[O_3] = np_3$

Variances are$Var[O_2] = np_2(1 - p_2]$ and $Var[O_2] = np_2(1 - p_2]$

****

Sentence 4

The sum of two normally distributed random variables are normally distributed with paramters:

$N[(\mu_X + \mu_Y);\ (\sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y))]$
$\mu_{X+Y} = \mu_X + \mu_Y;\ \ \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y)$

****

Calculation $E[\xi] = 0$

$$\begin{aligned} E[\xi] &= \frac{O_2 p_3}{\sqrt{np_2p_3(p_2+p_3)}} - \frac{O_3 p_2}{\sqrt{np_2p_3(p_2+p_3)}} \\ &= \frac{np_2p_3}{\sqrt{np_2p_3(p_2+p_3)}} - \frac{np_3p_2}{\sqrt{np_2p_3(p_2+p_3)}} \\ &= \frac{np_2p_3 - np_3p_2}{\sqrt{np_2p_3(p_2+p_3)}} = 0. \end{aligned} \tag{14}$$

Calculation $Var(\xi)$

Under assumption, that O_2 and O_3 are not independent, term
2Cov(O_2, O_3) must be included.

$$\begin{aligned} \mathrm{Var}(\xi) &= \mathrm{Var}\left[\frac{O_2p_3 - O_3p_2}{\sqrt{np_2p_3(p_2+p_3)}}\right] \\ &= \frac{1}{n^2p_2^2p_3^2(p_2+p_3)^2}\mathrm{Var}[p_3O_2 - p_2O_3] \end{aligned}$$

$$\text{as } \mathrm{Var}(c \cdot X) = c^2 \mathrm{Var}(X) \rightarrow$$
$$= \frac{1}{n^2 p_2^2 p_3^2 (p_2+p_3)^2} \Big(\mathrm{Var}[O_2] \cdot p_3^2 + \mathrm{Var}[O_3] \cdot p_3^2 - 2\mathrm{Cov}[O_2, O_3] \Big)$$
$$= \frac{np_2(1-p_2)p_3^2 + np_3(1-p_3)p_2^2 - 2p_2p_3 \cdot \mathrm{Cov}[O_2, O_3]}{n^2 p_2^2 p_3^2 (p_2+p_3)^2}$$
$$= \frac{np_2p_3(p_2+p_3) - 2np_2^2p_3^2 - 2p_2p_3 \cdot \mathrm{Cov}[O_2, O_3]}{n^2 p_2^2 p_3^2 (p_2+p_3)^2} \tag{15}$$

How can now $\mathrm{Cov}[O_2, O_3]$ be calculated?

Let's consider result (R) by conducting an experiment. There are the following possibilities:

$$\text{ZV } O_2 = \begin{cases} 0, \text{ R} \neq 2 \\ 1, \text{ R} = 2 \end{cases} \qquad \text{ZV } O_3 = \begin{cases} 0, \text{ R} \neq 3 \\ 1, \text{ R} = 3 \end{cases} \tag{16}$$

$$\begin{aligned} \mathrm{Cov}[O_2, O_3] &= \mathrm{E}\big[(O_2 - \mathrm{E}[O_2]) \cdot (O_3 - \mathrm{E}[O_3])\big] \\ &= \mathrm{E}\big[O_2O_3 - O_2\mathrm{E}[O_3] - O_3\mathrm{E}[O_2] + \mathrm{E}[O_2]\mathrm{E}[O_3]\big] \\ &= \mathrm{E}[O_2O_3] - \cancel{\mathrm{E}[O_2]\mathrm{E}[O_3]} - \mathrm{E}[O_3]\mathrm{E}[O_2] + \cancel{\mathrm{E}[O_2]\mathrm{E}[O_3]} \\ &= \mathrm{E}[O_2O_3] - \mathrm{E}[O_3]\mathrm{E}[O_2] \end{aligned} \tag{17}$$

As shown in (16) $\mathrm{E}[O_2O_3] = 0$ (O_2 und O_3 cannot occur simultaneously), and $\mathrm{E}[O_2]\mathrm{E}[O_3] = n \cdot p_2 + n \cdot p_3$, it follows:

$$\begin{aligned} \mathrm{Cov}[O_2, O_3] &= 0 - \mathrm{E}[O_2]\mathrm{E}[O_3] \\ &= -n \cdot p_2 + n \cdot p_3 \end{aligned} \tag{18}$$

Therefore resulting for $\mathrm{Var}(\xi)$:

$$\begin{aligned} \mathrm{Var}(\xi) &= \frac{np_2p_3(p_2+p_3) - 2np_2^2p_3^2 - 2p_2p_3 \cdot (-np_2p_3)}{np_2p_3(p_2+p_3)} \\ &= \frac{np_2p_3(p_2+p_3) - \cancel{(2np_2^2p_3^2)} + \cancel{(2np_2^2p_3^2)}}{np_2p_3(p_2+p_3)} \\ &= \frac{np_2p_3(p_2+p_3)}{np_2p_3(p_2+p_3)} \\ &= 1. \end{aligned} \tag{19}$$

with (14) and (19) it could be shown, that assumption (13)

$\xi = \frac{O_2p_3 - O_3p_2}{\sqrt{np_2p_3(p_2+p_3)}} \sim N(0,1)$ is correct.

End of proof for (13)

****

It could be demonstrated that by testing a random variable with 3 categories leads to a Chi-square test with 2 degrees of freedom, whereas (7) and (8) contribute each with 1 DF.
Testing a random variable with k = 3 categories $\Rightarrow \sim X_2^2$ (Chi-square with 2 DF)

General rule:

By much more sophisticated mathematics it could be demonstrated that by testing a random variable with k categories leads to a Chi-square test with k-1 degrees of freedom.
Testing a random variable with k categories $\Rightarrow \sim X_{k-1}^2$ (Chi-square with k-1 DF)

3 Hardy - Weinberg equilibrium - one degree of freedom

The Hardy–Weinberg principle relates allele frequencies to genotype frequencies in a randomly mating population. Imagine that you have a population with two alleles (A and B) that segregate at a single locus. The frequency of allele A is denoted by p and the frequency of allele a is denoted by q. The Hardy–Weinberg principle states that after one generation of random mating genotype frequencies will be p^2, 2pq, and q^2 and the following equation is fulfilled:

$$p^2 + 2pq + q^2 = 1. \tag{20}$$

When in an experiment the allele are counted for the three phenotypes p^2, 2pq, and q^2, the deviation from an equilibrium can be tested by a Chi-square test. In the following it is demonstrated that the deviations are tested by a Chi-square test with one degree of freedom and not with two degrees as probably it could be suspected because we have three phenotypes: n_{AA} = observed number of phenotype AA. n_{Aa} = observed number of phenotype Aa. n_{aa} = observed number of phenotype aa. e_{AA}, e_{Aa}, e_{aa} = corresponding expected values.

the Chi-square statistics is computed as:

$$X^2 = \frac{(n_{AA} - e_{AA})^2}{e_{AA}} + \frac{(n_{Aa} - e_{Aa})^2}{e_{Aa}} + \frac{(n_{aa} - e_{aa})^2}{e_{aa}} \tag{21}$$

this X^2 is asymptically distributed $\sim X_1^2$.

Procedure

Following R.V. Rohlfs and B.S. Weir, 2018. For genotypes AA, Aa, aa, the sample counts for the genotypes n_{AA}, n_{Aa}, n_{aa} summing to n. The counting of the genotypes can be represented as a 2 x 2 contingency table.

As $n_{AA} + n_{Aa}/2 + n_{Aa}/2 + n_{aa} = n$ we rather you use the allele countings $2n_{AA} + n_{Aa} + 2n_{aa} = 2n$.

Allele	A	a	Sum
A	$2n_{AA}$	n_{Aa}	n_A
a	n_{Aa}	$2n_{aa}$	n_a
	n_A	n_a	2n

The general formula applying for the deviation between observed and expected in a 2 x 2 contingency table shows:

$$X^2 = \frac{N(ad - bc)^2}{(a + b)(a + c)(c + d)(b + d)} \tag{22}$$

Proof that (22) is asymptotically Chi-square distributed with 1 DF is separately presented by Hans Stocker: The Chi-square test statistic of 2 x 2 tables, October 2023.

By inserting the corresponding allele frequencies from the above table we get:

$$X^2 = n \left(\frac{4n_{AA}n_{aa} - n_{Aa}^2}{(2n_{AA} + n_{Aa})(2n_{aa} + n_{Aa})} \right)^2 \tag{23}$$

Observe that we need to multiply by n because we doubled the countings above for convenience.

4 Literature

Boris Burkov, June 17, 2021 Pearson's Chi-square tests - intuition and derivation. https://www.borisburkov.net/2021-06-17-1.

R.V. Rohlfs and B.S. Weir, Genetics, **180**: 1609 - 1616 (November 2018).

Hans Stocker: The Chi-square test statistic of 2 x 2 tables, October 2023. https://tianshan-tours.ch/wp-content/uploads/2023/11/Chi-square-of-2-x-2-tables.pdf

Contact: hstock@bluewin.ch