OSIR
Outbreak, Surveillance and Investigation Reports
Field Epidemiology Training Program, Bureau of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625901734-5, Fax: +6625918581, Email: osireditor@osirjournal.net, http://www.osirjournal.net

What do you "Expect"?

Jaranit Kaewkungwal*

Mahidol University, Thailand

* Corresponding author, email address: jaranit.kae@mahidol.ac.th

Since "Statistics 101" course, we all learned that chi-square (χ^2) test is used when analyzing the association between exposure and outcome that are categorical variables, for examples: association between smoking ("smoke" vs. "not smoke") and lung cancer ("yes" or "no"), or association between treatment ("Drug A" vs. "Drug B") and treatment outcomes ("worsen", "stable", "improved"). Many might still remember that there are ***Pearson's chi-square test*** and ***Fisher's exact test***; and we would prefer to use Fisher's exact test rather than Pearson's chi-square when small ***"Expected Counts"*** are presented. So what do you "expect" in Pearson chi-square or Fisher's exact?

Back to Basic of "Probability"

When a teacher starts his/her statistic course, he/she will talk about tossing coins, rolling a dice, drawing cards out of a deck, and then "probability" theory. Many students start getting lost from there. But, in fact, it is not that difficult and it is the basic of most statistical methods. Let's look at some terms.[1-3]

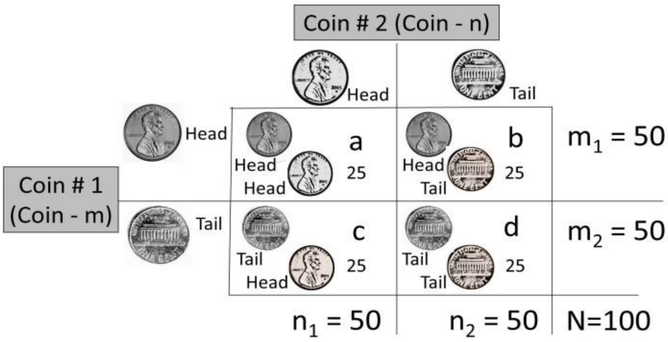
"Probability" or "Probable" derives from Latin "Probabilis" which means plausible or generally approved. "Probability", or another common term "Chance", deals with the stochastic (random) processes which lie behind data or outcomes. It could be considered as a measure of how some events will likely occur; it is usually expressing as the proportion of the number of cases of interest happening among the whole number of cases possible, for example, "the probability that you will get number 3 face landing after rolling a dice is 1 in 6 (or 0.1666..) as each dice has six faces".

Probabilities may be calculated either as marginal, joint or conditional functions. Most statistical methods rely on this concept. ***Marginal probability***, p(A), can be considered as an unconditional probability; that is, an event A that occurs is not conditioned on any other events. As an example in tossing an unbiased coin, the probability that a "Head" side will fall is unconditioned to chance that the "Tail" side will fall; thus p(head) = 1 in 2 (or 0.5). (The two sides of a coin are expressed as "Head" or "Tail" because head and tail has been historically considered as opposite body parts.) ***Joint probability***, p(A and B) or p(A ∩ B), refers to the probability of event A and event B are occurring together; it is the likelihood of two independent events happening at the time frame of interest (of note, it could be the probability of the intersection of two or more events). But wait – there are conditions that we have to take into consideration here: (a) the events A and B must be able to happen within the certain time frame and (b) the events A and B must be independent of each other. As an example, tossing two coins at the same time is independent events as the outcome of tossing one coin has no influence on the outcome of tossing the other coin. With the independent events, we can use the joint probability formula to calculate a chance of getting the jointed outcome of interest by the simple formula: p(A ∩ B = p(A) x p(B). As shown in figure 1, in tossing two unbiased coins, the joint probability to get "Tail" and "Tail" of the two coins will be 0.25.

Chi-square and "Expected Counts"

Historically, Pearson's paper of 1900 introduced what subsequently became known as the chi-square test of goodness of fit. In series of tossing of ten shillings at a time "frequently in the open air", Pearson's analysis of these artificial experiments led to the concept of "deviations from the most probable" or "a criterion of the probability"[4].

Let's look at an example of a simple case of flipping a coin. If the coin is unbiased, meaning that it is fair and balanced, then the "most probable" or "expected" frequency of to get head is 0.5 or 50%. If we toss a coin 100 times and we get 45 or 55 heads, we may be not suspicious as the "deviations from the most probable" seems to be acceptable. But if only 31 heads occur in 100 flips, we would be now skeptical and



Example: If tossing two unbiased coins for 100 times,

- Probability of getting “Tail” of Coin # 1 = (m_2/N) = 50/100 = 0.5
- Probability of getting “Tail” of Coin # 2 = (n_2/N) = 50/100 = 0.5
- Probability of getting “Tail” of Coin # 1 AND “Tail” of Coin # 2 = (m_2/N) x (n_2/N) = 0.5 x 0.5 = 0.25

Thus, in tossing 100 times, one should get “Tail-Tail” for

[$(m_2/$ N) x (n/N) x N] = 0.25x100 = 25 times

Figure 1. Outcomes of tossing two unbiased coins

suspect that the coin is somehow unfair or weighted to come up with tails. This is the concept of Pearson’s chi-square test, the test that compares the observed distribution of counts against the expected distribution from some theoretical baseline which allow us to quantify the probability of such an event[5]. The size of the difference between observed and expected distributions is reflected in the test statistic.

The statistical null hypothesis is that the number of observed counts in each category is equal to that expected or predicted by a probability theory, and the alternative hypothesis is that the observed numbers are different from the expected. Then we will use a mathematical relationship, in this case the chi-square distribution, to estimate the probability of obtaining that value of the test statistic[6-8]. The chi-square test statistic is calculated by using the formula:

$$x^2 = \sum \frac{\left(O - E\right)^2}{E}$$

where O represents the observed frequency (counts). E is the expected frequency (counts} under the null hypothesis.

As an example of a study to determine association between exposure (E– vs. E+) and outcome (D– or D+), such as smoking (yes vs. no) and lung cancer (yes or no), we can generate a 2x2 table as shown in figure 2. The observed counts (from data collection in the study) would be: a, b, c, d as shown in each category (cell). Then how we do get the expected counts? Back to our joint probability concept - if “exposure” and “outcome” are independent (not associated), then we can calculate the probability of the joint event in each cell. As shown in figure 2, the probability of “not exposed, E-” and “not having outcome, D-” can be calculated and then compared against its observed count, d. The chi-square test statistic is then based on the combination of Os and Es of all categories in the table.

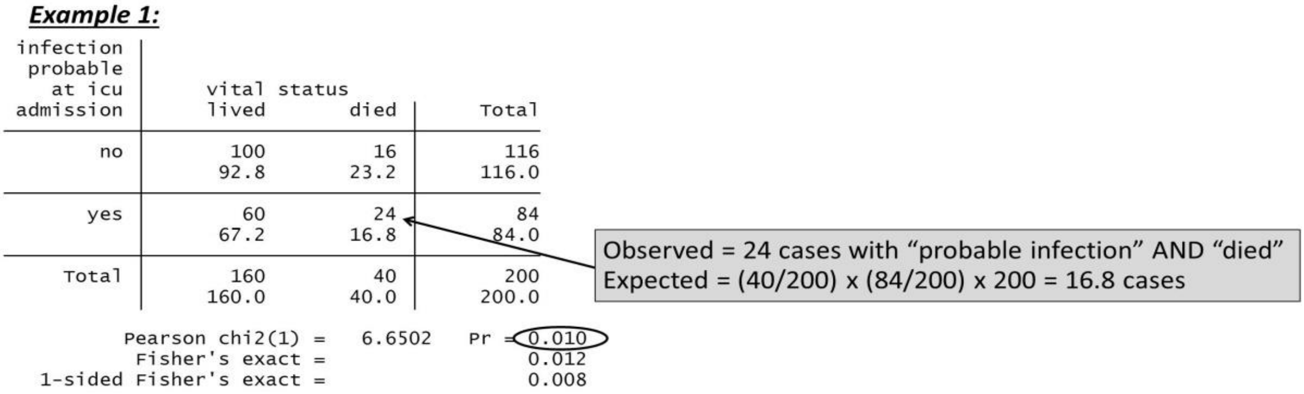
		Outcome		
		D +	D -	Total
Exposure	E +	a	C	m_1
	E -	c	D	m_2
	Total	n_1	n_2	N

Example:

- Observed value = cases that not being exposed (E-) AND not having outcome (D-) among N people = d
- Expected value = (Prob. of being not exposed, E-) AND (Prob..of not having outcome, D-) of N people = (m_2/N) x (n_2/N) x N

Figure 2. Observed and expected frequencies (counts) in a 2x2 table

The chi-square statistic is a non-parametric (distribution free); that means it is robust with respect to the distribution of the data. Specifically, it does not require equality of variances among the study groups or homoscedasticity in the data[9]. Chi-square test can be used for both dichotomous independent variables (a shown in 2x2 table above) and multiple groups/outcomes. However, the chi-square test does not provide an exact calculation of the p-value but rather an approximation of the p-value. But no need to worry - when the assumptions of the test are met, it is like all probability density functions, the chi-square distribution is a continuous function whose area sums to one[5]. Just a note for the reader who is interested in mathematical foundation, the chi-square distribution is based on the summing of the square values of k standard normal distributions, whereas k is corresponding to the degrees of freedom for the chi-square distribution. Degree of freedom for chi-square is equal to (r-1)x(c-1), where r is the number of levels of one categorical variable and c is the number of levels of another categorical variable. As shown in figure 3, the observed counts vs. expected counts in the 2x2 table (4 cells) were compared in Pearson's chi-square test statistic and the p-value was calculated basing on chi-square distribution. The degree of freedom as shown next to the chi-square is 1, chi2(1), because we have 2 levels of exposure and 2 levels of outcome. Based on the p-value, we can then conclude that there is statistically significant association between exposure (infection at ICU admission) and outcome (vital status).



Example 1:

```
  infection |
   probable |
     at icu |       vital status
  admission |     lived       died |     Total
------------+----------------------+----------
         no |       100         16 |       116
            |      92.8       23.2 |     116.0
------------+----------------------+----------
        yes |        60         24 |        84
            |      67.2       16.8 |      84.0
------------+----------------------+----------
      Total |       160         40 |       200
            |     160.0       40.0 |     200.0

          Pearson chi2(1) =   6.6502   Pr = 0.010
           Fisher's exact =                 0.012
   1-sided Fisher's exact =                 0.008
```

Observed = 24 cases with “probable infection” AND “died”
Expected = (40/200) x (84/200) x 200 = 16.8 cases

Figure 3. Person's Chi-square test based in observed and expected frequencies

Karl Pearson vs. Ronald A. Fisher

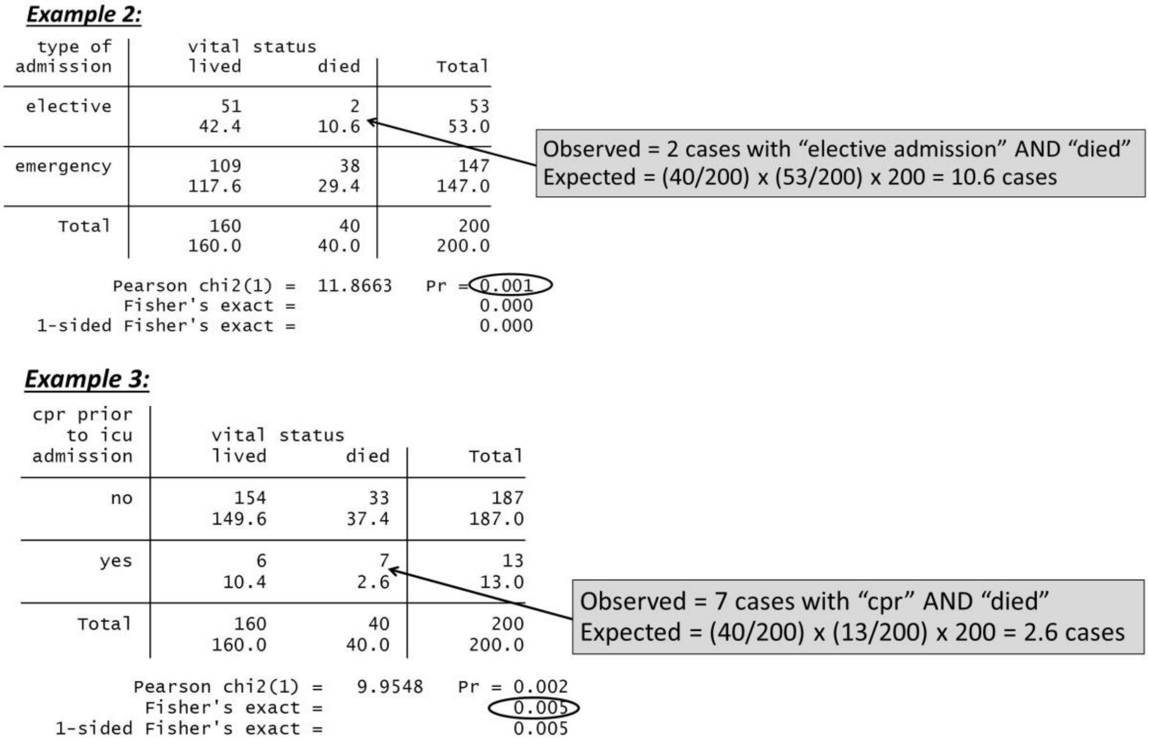
It is not a strange phenomenon to see a scientific controversy debating on certain issue publicly and privately. In 1935, Karl Pearson and R. A. Fisher exchanged hot letters in Nature, one of the most prestigious scientific journals, on testing statistical hypotheses. The disagreements and rivalry between Ronald A Fisher and Karl Pearson were also noted in history in many other statistical theories; after dying of Karl Pearson, Fisher even continued to argue with Ergon Pearson (Karl Pearson's son) and Jerzy Neyman on this hypothesis testing concept[10,11]. In fact, there has been another debate on philosophy of hypothesis testing from Bayesian approach which is based on stronger assumptions[10]. This is fun to read but it is beyond the purpose of this article.

Fisher argued that in all cases of applying the chi-square test it is mathematically necessary to take account of the number of degrees of freedom of the observations in relation to the expected distribution to which they are compared[12]. Fisher then developed the “Exact” test which means that we can calculate from the marginal totals and get exactly what is the probability of getting an observed result, in the same way that we can work out exactly the chance that we may get 55 heads out of 100 tosses of an unbiased coin. However, the method and formula for Fisher's exact test is not easy to write up; it is based on the "factorial" or successive multiplication by numbers in descending series[13].

It was suggested in literature that the Pearson's chi-square test involves using the chi-square distribution to approximate the underlying exact distribution. The main assumptions for Pearson's chi-square test include: (a) individual observations are independent of each other, and (b) individual cells contain sufficient counts. The approximation becomes better as the expected cell counts grow larger, and may be inappropriate for tables with very small expected cell counts[14]. There are many recommendations about the sufficient counts[5,14,15]. A standard (and conservative) rule of thumb is to avoid using the Pearson's chi-square test statistics for tables with expected cell counts <1, or when more than 20% of the table cells have expected cell counts <5. Another rule of thumb is that if the total number of observations is at least 10, the number categories is at least 3, and the square of the total number of observations is at least 10 times the number of categories, then the Pearson's chi-square approximation should be reasonable. Caution should be made when cell categories are combined (collapsed together) to fix problems of small expected cell frequencies as it may destroy evidence of non-independence[14].

So – when to use Fisher's exact test? According to the common rule of thumb, we should use Fisher's exact test when the Pearson's chi-square test is inappropriate due to small sample sizes and expected counts in the 20% of the table cells are <5 (for the 2x2 table, when the expected value in a cell is <5)[15]. Note that for some statistical software, Fisher's exact test is applied to only 2x2 table; but there are extensions that allow the test to be applied to cases with more than two categories per variable.[5]

As examples shown in figure 4, the decision to report p-value of Pearson's chi-square or Fisher exact test would generally be based on the expected counts in the table cells. In the scenario shown in example 2 representing the association between the exposure (type of ICU admission) and the outcome (vital status), the cell of "elective admission" and "died" contains 2 observed cases but 10.6 expected counts; the p-value of Pearson's chi-square test is thus applicable. In contrast, in the scenario shown in example 3 representing the association between the exposure (CPR prior to ICU admission) and the outcome (vital status), the cell of "having CPR" and "died" contains 7 observed cases but 2.6 expected counts; the p-value of Fisher's exact test is more appropriate.



Example 2:

```
   type of |     vital status
 admission |    lived       died |     Total
-----------+----------------------+----------
  elective |       51          2 |        53
           |     42.4       10.6 |      53.0
-----------+----------------------+----------
 emergency |      109         38 |       147
           |    117.6       29.4 |     147.0
-----------+----------------------+----------
     Total |      160         40 |       200
           |    160.0       40.0 |     200.0

          Pearson chi2(1) =  11.8663   Pr = 0.001
           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000
```

Observed = 2 cases with "elective admission" AND "died"
Expected = (40/200) x (53/200) x 200 = 10.6 cases

Example 3:

```
 cpr prior |
    to icu |     vital status
 admission |    lived       died |     Total
-----------+----------------------+----------
        no |      154         33 |       187
           |    149.6       37.4 |     187.0
-----------+----------------------+----------
       yes |        6          7 |        13
           |     10.4        2.6 |      13.0
-----------+----------------------+----------
     Total |      160         40 |       200
           |    160.0       40.0 |     200.0

          Pearson chi2(1) =   9.9548   Pr = 0.002
           Fisher's exact =                 0.005
   1-sided Fisher's exact =                 0.005
```

Observed = 7 cases with "cpr" AND "died"
Expected = (40/200) x (13/200) x 200 = 2.6 cases

Figure 4. Person's Chi-square test vs. Fisher's Exact Test

It should be noted that the Pearson's chi-square test would be more close to Fisher's exact test as the number of observations increases. As its name implies, Fisher's exact test gives an exact probability for all sample sizes. So, why don't we just use Fisher's exact test for all, and not using Pearson's chi-square at all? This is back to the debatable issue - some statisticians would argue that Fisher's exact test may give the exact answer to the wrong question and the test itself is based on experimental study with the assumption that the row and column totals are fixed, which is not quite fit to many other kinds of study[14].

In fact, there is another controversial idea against Pearson's chi-square test. That is the Yates's correction for continuity (or Yates's chi-square test) which was designed to make the Pearson's chi-square approximation better. However, many argued that it may adjust too far making the p-value too large (too 'conservative') and thus its use is limited. Moreover, with large sample sizes, Yates' correction makes little difference. Again, there were statisticians who agree and disagree on whether to use Yates's correction[16].

Conclusion

The chi-square test is the most well-known statistics used to test the agreement between observed and expected counts while the probability to reject the null hypothesis is calculated based on the theoretical chi-square distribution. The hot arguments regarding the use and misuse of chi-square tests came from different schools of thought in the assumptions and applications of hypothesis testing[10,11,17]. Despite different approaches, there have also been studies suggesting that Fisher's exact and Pearson's chi-square tests are "asymptotically equivalent" (the statistics term meaning that the two tests are eventually becoming "essentially equal") and a formal similarity also exists in small samples[18]. In fact, Pearson's chi-square test even gave an excellent approximation to the actual Bayesian probability approach except for those with extremely disproportionate marginal frequencies[18]. So – the common practice among researchers to use Pearson's chi-square test or Fisher's exact test is still based main assumption – the sufficient "expected" counts!

Suggested Citation

Kaewkungwal J. What do you “expect”? OSIR. 2018 Mar;11(1):22-26.

References

1. Walsh J. Joint probability: definition, formula & Examples [cited 2017 Nov 4]. <http://study.com/academy/lesson/joint-probability-definition-formula-examples.html>.
2. Hildebrand AJ. Math 370/408, Actuarial Problemsolving [cited 2017 Nov 4]. <https://faculty.math.illinois.edu/~hildebr/370/370jointdistributions.pdf>.
3. Albright EA. Joint, marginal and conditional probabilities [cited 2017 Nov 4]. <http://sites.nicholas.duke.edu/statsreview/probability/jmc/>.
4. Plackett, RL. Karl Pearson and the Chi-square test. International Statistical Review. 1983;51:59-72.
5. Quigley D. Module 7.1: The binomial, chi-square and Fisher’s exact tests. 2016 [cited 2017 Nov 4]. <http://davidquigley.com/talks/2015/biostatistics/module_07.1.html>.
6. PennState Eberly College of Science. Chi-square test of independence [cited 2017 Nov 4]. <https://onlinecourses.science.psu.edu/statprogram/node/158>.
7. McDonald JH. Handbook of biological statistics. 3rd ed. Baltimore: Sparky House Publishing; 2014. p. 45-52 [cited 2017 Nov 4]. <http://www.biostathandbook.com/chigof.html>.
8. Buonocore A, Pirozzi E. On the Pearson-Fisher chi-square theorem. Applied Mathematical Sciences. 2014;8(134):6733-44.
9. McHugh ML. The chi-square test of independence. Biochemia Medica. 2013;23(2):143-9.
10. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory of two? Journal of the American Statistical Association. 1993;88(424):1242-9.
11. Inman HF. Karl Pearson and R. A. Fisher on statistical tests: a 1935 exchange from nature. The American Statistician. 1994;48(1):2-11.
12. Fisher RA. On the intepretation of $\chi 2$ from contigency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922;58:87-94.
13. BMJ. Exact probability test [cited 2017 Nov 4]. <http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/9-exact-probability-test>.
14. Feinberg School of Medicine. PROPHET StatGuide: do your data violate goodness of fit (chi-square) test assumptions? [cited 2017 Nov 4]. <http://www.basic.northwestern.edu/statguidefiles/gf-dist_ass_viol.html>.
15. Cochran WG. The $\chi 2$ test of goodness of fit. Ann Math Stat. 1952;25:315-45.
16. Graphpad Software. GraphPad statistics guide [cited 2017 Nov 4]. <https://www.graphpad.com/guides/prism/7/statistics/stat_chi-square_or_fishers_test.htm?toc=0&printWindow>.
17. Bolboacă SD, Jäntschi L, Sestraş AF, Sestraş RE, Pamfil DC. Pearson-Fisher chi-square statistic revisited. Information. 2011;2:28-545.
18. Camill G. The relationship between Fisher's exact test and Pearson's chi-square test: A bayesian perspective. Psychometrika. 1995;60(2):305-12.