

Fisher's exact test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables.^{[1][2][3]} Although in practice it is employed when sample sizes are small, it is valid for all sample sizes. It is named after its inventor, Ronald Fisher, and is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

Fisher is said to have devised the test following a comment from Muriel Bristol, who claimed to be able to detect whether the tea or the milk was added first to her cup. He tested her claim in the "lady tasting tea" experiment.^[4]

Contents

Purpose and scope

Example

Controversies

Alternatives

See also

References

External links

Purpose and scope

The test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification. So in Fisher's original example, one criterion of classification could be whether milk or tea was put in the cup first; the other could be whether Bristol thinks that the milk or tea was put in first. We want to know whether these two classifications are associated—that is, whether Bristol really can tell whether milk or tea was poured in first. Most uses of the Fisher test involve, like this example, a 2×2 contingency table. The p-value from the test is computed as if the margins of the table are fixed, i.e. as if, in the tea-tasting example, Bristol knows the number of cups with each treatment (milk or tea first) and will therefore provide guesses with the correct number in each category. As pointed out by Fisher, this leads under a null hypothesis of independence to a hypergeometric distribution of the numbers in the cells of the table.



A teapot, a creamer and teacup full of tea with milk—can a taster tell if the milk went in first?

With large samples, a chi-squared test (or better yet, a G-test) can be used in this situation. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution. The approximation is inadequate when sample sizes are small, or the data are

very unequally distributed among the cells of the table, resulting in the cell counts predicted on the null hypothesis (the “expected values”) being low. The usual rule of thumb for deciding whether the chi-squared approximation is good enough is that the chi-squared test is not suitable when the expected values in any of the cells of a contingency table are below 5, or below 10 when there is only one degree of freedom (this rule is now known to be overly conservative^[5]). In fact, for small, sparse, or unbalanced data, the exact and asymptotic *p*-values can be quite different and may lead to opposite conclusions concerning the hypothesis of interest.^{[6][7]} In contrast the Fisher exact test is, as its name states, exact as long as the experimental procedure keeps the row and column totals fixed, and it can therefore be used regardless of the sample characteristics. It becomes difficult to calculate with large samples or well-balanced tables, but fortunately these are exactly the conditions where the chi-squared test is appropriate.

For hand calculations, the test is only feasible in the case of a 2×2 contingency table. However the principle of the test can be extended to the general case of an $m \times n$ table,^{[8][9]} and some statistical packages provide a calculation (sometimes using a Monte Carlo method to obtain an approximation) for the more general case.^[10]

Example

For example, a sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently studying for a statistics exam on the other. We hypothesize, for example, that the proportion of studying individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant. The data might look like this:

	Men	Women	Row total
Studying	1	9	10
Not-studying	11	3	14
Column total	12	12	24

The question we ask about these data is: knowing that 10 of these 24 teenagers are studying, and that 12 of the 24 are female, and assuming the null hypothesis that men and women are equally likely to study, what is the probability that these 10 teenagers who are studying would be so unevenly distributed between the women and the men? If we were to choose 10 of the teenagers at random, what is the probability that 9 or more of them would be among the 12 women, and only 1 or fewer from among the 12 men?

Before we proceed with the Fisher test, we first introduce some notations. We represent the cells by the letters *a*, *b*, *c* and *d*, call the totals across rows and columns *marginal totals*, and represent the grand total by *n*. So the table now looks like this:

	Men	Women	Row Total
Studying	<i>a</i>	<i>b</i>	<i>a + b</i>
Non-studying	<i>c</i>	<i>d</i>	<i>c + d</i>
Column Total	<i>a + c</i>	<i>b + d</i>	<i>a + b + c + d (=n)</i>

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b} \binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! n!}$$

where $\binom{n}{k}$ is the binomial coefficient and the symbol ! indicates the factorial operator. This can be seen as follows. If the marginal totals (i.e. $a+b$, $c+d$, $a+c$, and $b+d$) are known, only a single degree of freedom is left: the value e.g. of a suffices to deduce the other values. Now, $p = p(a)$ is the probability that a elements are positive in a random selection (without replacement) of $a+c$ elements from a larger set containing n elements in total out of which $a+b$ are positive, which is precisely the definition of the hypergeometric distribution.

With the data above (using the first of the equivalent forms), this gives:

$$p = \binom{10}{1} \binom{14}{11} / \binom{24}{12} = \frac{10! 14! 12! 12!}{1! 9! 11! 3! 24!} \approx 0.001346076$$

The formula above gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that men and women are equally likely to be studiers. To put it another way, if we assume that the probability that a man is a studier is p , the probability that a woman is a studier is also p , and we assume that both men and women enter our sample independently of whether or not they are studiers, then this hypergeometric formula gives the conditional probability of observing the values a, b, c, d in the four cells, conditionally on the observed marginals (i.e., assuming the row and column totals shown in the margins of the table are given). This remains true even if men enter our sample with different probabilities than women. The requirement is merely that the two classification characteristics—gender, and studier (or not)—are not associated.

For example, suppose we knew probabilities P, Q, p, q with $P + Q = p + q = 1$ such that (male studier, male non-studier, female studier, female non-studier) had respective probabilities (Pp, Pq, Qp, Qq) for each individual encountered under our sampling procedure. Then still, were we to calculate the distribution of cell entries conditional given marginals, we would obtain the above formula in which neither p nor P occurs. Thus, we can calculate the exact probability of any arrangement of the 24 teenagers into the four cells of the table, but Fisher showed that to generate a significance level, we need consider only the cases where the marginal totals are the same as in the observed table, and among those, only the cases where the arrangement is as extreme as the observed arrangement, or more so. (Barnard's test relaxes this constraint on one set of the marginal totals.) In the example, there are 11 such cases. Of these only one is more extreme in the same direction as our data; it looks like this:

	Men	Women	Row Total
Studying	0	10	10
Non-studying	12	2	14
<i>Column Total</i>	<i>12</i>	<i>12</i>	<i>24</i>

For this table (with extremely unequal studying proportions) the probability is $p = \binom{10}{0} \binom{14}{12} / \binom{24}{12} \approx 0.000033652$.

In order to calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the null hypothesis is true, we have to calculate the values of p for both these tables, and add them together. This gives a one-tailed test, with p approximately $0.001346076 + 0.000033652 = 0.001379728$. For example, in the R statistical computing environment, this value can be obtained as `fisher.test(rbind(c(1,9),c(11,3)), alternative="less")$p.value`. This value can be interpreted as the sum of evidence provided by the observed data—or any more extreme table—for the null hypothesis (that there is no difference in the proportions of studiers between men and women). The smaller the value of p , the greater the evidence for rejecting the null hypothesis; so here the evidence is strong that men and women are not equally likely to be studiers.

For a two-tailed test we must also consider tables that are equally extreme, but in the opposite direction. Unfortunately, classification of the tables according to whether or not they are 'as extreme' is problematic. An approach used by the `fisher.test` function in R is to compute the p -value by summing the probabilities for all tables with probabilities less than or equal to that of the observed table. In the example here, the 2-sided p -value is twice the 1-sided value—but in general these can differ substantially for tables with small counts, unlike the case with test statistics that have a symmetric sampling distribution.

As noted above, most modern statistical packages will calculate the significance of Fisher tests, in some cases even where the chi-squared approximation would also be acceptable. The actual computations as performed by statistical software packages will as a rule differ from those described above, because numerical difficulties may result from the large values taken by the factorials. A simple, somewhat better computational approach relies on a gamma function or log-gamma function, but methods for accurate computation of hypergeometric and binomial probabilities remains an active research area.

Controversies

Despite the fact that Fisher's test gives exact p -values, some authors have argued that it is conservative, i.e. that its actual rejection rate is below the nominal significance level.^{[11][12][13]} The apparent contradiction stems from the combination of a discrete statistic with fixed significance levels.^{[14][15]} To be more precise, consider the following proposal for a significance test at the 5%-level: reject the null hypothesis for each table to which Fisher's test assigns a p -value equal to or smaller than 5%. Because the set of all tables is discrete, there may not be a table for which equality is achieved. If α_e is the largest p -value smaller than 5% which can actually occur for some table, then the proposed test effectively tests at the α_e -level. For small sample sizes, α_e might be significantly lower than 5%.^{[11][12][13]} While this effect occurs for any discrete statistic (not just in contingency tables, or for Fisher's test), it has been argued that the problem is compounded by the fact that Fisher's test conditions on the marginals.^[16] To avoid the problem, many authors discourage the use of fixed significance levels when dealing with discrete problems.^{[14][15]}

The decision to condition on the margins of the table is also controversial.^{[17][18]} The p -values derived from Fisher's test come from the distribution that conditions on the margin totals. In this sense, the test is exact only for the conditional distribution and not the original table where the margin totals may change from experiment to experiment. It is possible to obtain an exact p -value for the 2×2 table when the margins are not held fixed. Barnard's test, for example, allows for random margins. However, some authors^{[14][15][18]} (including, later, Barnard himself)^[14] have criticized Barnard's test based on this property. They argue that the marginal success total is an (almost^[15]) ancillary statistic, containing (almost) no information about the tested property.

The act of conditioning on the marginal success rate from a 2×2 table can be shown to ignore some information in the data about the unknown odds ratio.^[19] The argument that the marginal totals are (almost) ancillary implies that the appropriate likelihood function for making inferences about this odds ratio should be conditioned on the marginal success rate.^[19] Whether this lost information is important for inferential purposes is the essence of the controversy.^[19]

Alternatives

An alternative exact test, Barnard's exact test, has been developed and proponents of it suggest that this method is more powerful, particularly in 2×2 tables.^[20] Another alternative is to use maximum likelihood estimates to calculate a p-value from the exact binomial or multinomial distributions and reject or fail to reject based on the p-value.

For stratified categorical data the Cochran–Mantel–Haenszel test must be used instead of Fisher's test.

Choi et al.^[19] propose a p-value derived from the likelihood ratio test based on the conditional distribution of the odds ratio given the marginal success rate. This p-value is inferentially consistent with classical tests of normally distributed data as well as with likelihood ratios and support intervals based on this conditional likelihood function. It is also readily computable.^[21]

See also

- Bernoulli trial

References

- Fisher, R. A. (1922). "On the interpretation of χ^2 from contingency tables, and the calculation of P" (<https://zenodo.org/record/1449484>). *Journal of the Royal Statistical Society*. **85** (1): 87–94. doi:10.2307/2340521 (<https://doi.org/10.2307%2F2340521>). JSTOR 2340521 (<https://www.jstor.org/stable/2340521>).
- Fisher, R.A. (1954). *Statistical Methods for Research Workers*. Oliver and Boyd. ISBN 0-05-002170-2.
- Agresti, Alan (1992). "A Survey of Exact Inference for Contingency Tables". *Statistical Science*. **7** (1): 131–153. CiteSeerX 10.1.1.296.874 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.296.874>). doi:10.1214/ss/1177011454 (<https://doi.org/10.1214%2Fss%2F1177011454>). JSTOR 2246001 (<https://www.jstor.org/stable/2246001>).
- Fisher, Sir Ronald A. (1956) [*The Design of Experiments* (1935)]. "Mathematics of a Lady Tasting Tea" (<https://books.google.com/?id=oKZwtLQTmNAC&pg=PA1512&dq=%22mathematics+of+a+lady+tasting+tea%22>). In James Roy Newman (ed.). *The World of Mathematics, volume 3*. Courier Dover Publications. ISBN 978-0-486-41151-4.
- Larntz, Kinley (1978). "Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics". *Journal of the American Statistical Association*. **73** (362): 253–263. doi:10.2307/2286650 (<https://doi.org/10.2307%2F2286650>). JSTOR 2286650 (<https://www.jstor.org/stable/2286650>).
- Mehta, Cyrus R; Patel, Nitin R; Tsiatis, Anastasios A (1984). "Exact significance testing to establish treatment equivalence with ordered categorical data". *Biometrics*. **40** (3): 819–825. doi:10.2307/2530927 (<https://doi.org/10.2307%2F2530927>). JSTOR 2530927 (<https://www.jstor.org/stable/2530927>). PMID 6518249 (<https://pubmed.ncbi.nlm.nih.gov/6518249>).
- Mehta, C. R. 1995. SPSS 6.1 Exact test for Windows. Englewood Cliffs, NJ: Prentice Hall.

8. Mehta C.R.; Patel N.R. (1983). "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables". *Journal of the American Statistical Association*. **78** (382): 427–434. doi:10.2307/2288652 (https://doi.org/10.2307%2F2288652). JSTOR 2288652 (https://www.jstor.org/stable/2288652).
9. mathworld.wolfram.com (http://mathworld.wolfram.com/FishersExactTest.html) Page giving the formula for the general form of Fisher's exact test for $m \times n$ contingency tables
10. Cyrus R. Mehta; Nitin R. Patel (1986). "ALGORITHM 643: FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables". *ACM Trans. Math. Softw.* **12** (2): 154–161. doi:10.1145/6497.214326 (https://doi.org/10.1145%2F6497.214326).
11. Liddell, Douglas (1976). "Practical tests of 2×2 contingency tables". *The Statistician*. **25** (4): 295–304. doi:10.2307/2988087 (https://doi.org/10.2307%2F2988087). JSTOR 2988087 (http://www.jstor.org/stable/2988087).
12. Berkson, Joseph (1978). "In dispraise of the exact test". *Journal of Statistic Planning and Inference*. **2**: 27–42. doi:10.1016/0378-3758(78)90019-8 (https://doi.org/10.1016%2F0378-3758%2F90019-8).
13. D'Agostino, R. B.; Chase, W. & Belanger, A. (1988). "The appropriateness of some common procedures for testing equality of two independent binomial proportions". *The American Statistician*. **42** (3): 198–202. doi:10.2307/2685002 (https://doi.org/10.2307%2F2685002). JSTOR 2685002 (https://www.jstor.org/stable/2685002).
14. Yates, F. (1984). "Tests of significance for 2×2 contingency tables (with discussion)". *Journal of the Royal Statistical Society, Series A*. **147** (3): 426–463. doi:10.2307/2981577 (https://doi.org/10.2307%2F2981577). JSTOR 2981577 (https://www.jstor.org/stable/2981577).
15. Little, Roderick J. A. (1989). "Testing the equality of two independent binomial proportions". *The American Statistician*. **43** (4): 283–288. doi:10.2307/2685390 (https://doi.org/10.2307%2F2685390). JSTOR 2685390 (https://www.jstor.org/stable/2685390).
16. Mehta, Cyrus R.; Senchaudhuri, Pralay (4 September 2003). "Conditional versus unconditional exact tests for comparing two binomials" (https://www.statsols.com/hubfs/Resources/_Comparing-Two-Binomials.pdf) (PDF). Retrieved 20 November 2009.
17. Barnard, G.A (1945). "A new test for 2×2 tables". *Nature*. **156** (3954): 177. doi:10.1038/156177a0 (https://doi.org/10.1038%2F156177a0).
18. Fisher (1945). "A New Test for 2×2 Tables". *Nature*. **156** (3961): 388. doi:10.1038/156388a0 (https://doi.org/10.1038%2F156388a0).; Barnard, G.A (1945). "A new test for 2×2 tables". *Nature*. **156** (3974): 783–784. doi:10.1038/156783b0 (https://doi.org/10.1038%2F156783b0).
19. Choi L, Blume JD, Dupont WD (2015). "Elucidating the foundations of statistical inference with 2×2 tables" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4388855). *PLoS ONE*. **10** (4): e0121263. doi:10.1371/journal.pone.0121263 (https://doi.org/10.1371%2Fjournal.pone.0121263). PMC 4388855 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4388855). PMID 25849515 (https://pubmed.ncbi.nlm.nih.gov/25849515).
20. Berger R.L. (1994). "Power comparison of exact unconditional tests for comparing two binomial proportions". *Institute of Statistics Mimeo Series No. 2266*: 1–19.
21. Choi, Leena (2011). "ProfileLikelihood: profile likelihood for a parameter in commonly used statistical models; 2011. R package version 1.1" (https://cran.r-project.org/web/packages/ProfileLikelihood/index.html). See also: Likelihood Ratio Statistics for 2×2 Tables (http://statcomp2.vanderbilt.edu:37212/dalep/ProfileLikelihood/) Archived (https://web.archive.org/web/20160604023208/http://statcomp2.vanderbilt.edu:37212/dalep/ProfileLikelihood/) 4 June 2016 at the Wayback Machine (Online calculator).

External links

- Calculate Fisher's Exact Test Online (http://www.langsrud.com/stat/fisher.htm)
- Likelihood Ratio Statistics for 2×2 Tables (https://web.archive.org/web/20170616070646/http://statcomp2.vanderbilt.edu:37212/ProfileLikelihood/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Fisher%27s_exact_test&oldid=933086673"

This page was last edited on 29 December 2019, at 22:30 (UTC).

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.