POINTS OF SIGNIFICANCE

Tabular data

Tabulating the number of objects in categories of interest dates back to the earliest records of commerce and population censuses.

So far in this column, we have discussed mainly the statistical analysis of continuous variables. We now turn to categorical data, which appear in experimental designs that enumerate how samples are distributed among different categories. There are two types of categorization: ordinal, for which there is an implied order (healthy, asymptomatic, symptomatic) and nominal, for which there is no order (gender or blood type). This month, we discuss nominal data.

Categorical data are typically counts (or percentages) of samples in a category and can be tabulated to help assess and compare proportions. One-way tables assess whether counts in a category match a predetermined distribution, provided either by theory or by historical trends. Multiway tables show more than one categorical variable and assess whether there is an association between counts when cross-classified by both variables. In both kinds of tables, statistical testing is performed comparing the counts (not the percentages) to the number expected under a null hypothesis.

Analyzing one-way tables addresses what is called the 'goodness of fit' problem, because it is a way to assess how well the hypothesized distribution fits the observed counts. Mendelian genetics provides a classic example. Assume two alleles A and a with frequencies p and $1 - p$, respectively. Under random mating and survival, we expect offspring genotypes AA, Aa and aa with probabilities p^2, $2p(1 - p)$ and $(1 - p)^2$, respectively. In **Table 1** we simulate this for 100 randomly selected individuals, together with expected counts based on $p = 0.25$ and the differences between observed and expected counts. The expected counts do not need to be whole numbers, and the sum of differences always adds up to zero, which imposes a constraint. How do these frequencies match our expectation, and what kind of inferences can we make?

In the goodness-of-fit approach, we ask whether differences as large as those observed could be due to random variation between samples, or whether they provide evidence against the null hypothesis of random mating and survival. For randomly sampled count data, statistical theory shows that the population variance in counts is equal to the expected value. When the counts are sufficiently large (the usual rule of thumb is that the expected number of counts in each category is greater than 5), they are approximately normally distributed. In this case, we can use a z test to determine whether the difference between the observed (O) and expected (E) counts is larger than predicted compared to the s.d., using $z = (O - E)/\sqrt{E}$. For example, for AA the z score of the difference is $3.75/\sqrt{6.25} = 1.5$, yielding $P = 0.13$. However, for aa the same computation yields $-16.25/\sqrt{56.25} = 1.5$ and $P = 0.03$.

To avoid multiple testing and account for the constraint that the differences must sum to zero, we test all the differences simultaneously by summing the squared z scores using the chi-squared goodness-of-fit test[1], $\chi^2 = \Sigma(O - E)^2/E$. Under the null hypothesis, the χ^2 test statistic will be distributed according to the χ^2 distribution on $k - 1$ degrees of freedom (d.f.), where k is the number of categories.

Table 1 | Distribution of two alleles, A and a, among 100 individuals

	Genotype		
	AA	Aa	aa
Observed	10	50	40
Expected	6.25	37.5	56.25
Difference	3.75	12.5	−16.25

The expected values are based on the known fraction of A in the population, $p = 0.25$.

Using differences as shown in **Table 1**, we obtain $\chi^2 = 11.11$ (d.f. = 2) and $P = 0.0039$, suggesting that the observed counts are highly unlikely to arise from random mating and survival with $p = 0.25$.

Typically, the distribution of counts is characterized by one or more parameters. For our case, this is the value of p. In **Table 1**, we used $p = 0.25$, which we might have based on previous evidence; but usually we don't know what these parameter values are and must estimate them from the samples. In doing so, the degrees of freedom of the test are reduced by the number of parameters estimated. For our observed counts in **Table 1**, we can estimate the proportion of A in the sample as $p = (2 \times 10 + 1 \times 50)/200 = 0.35$, where 200 is twice the sample size because the population is diploid. Using this estimate of p changes the expected and difference cells in our table (**Table 2**).

Using the expected and difference counts based on $p = 0.35$ gives $\chi^2 = 0.98$. Since we estimated one parameter, one degree of freedom is lost; we use the χ^2 distribution on d.f. = 1 and find $P = 0.32$. Now we fail to reject the hypothesis that the allele distribution is due to random mating and survival, in contrast to our rejection of the null hypothesis when using $p = 0.25$. Hence we can conclude that our rejection of the null hypothesis in the previous test was due to the proposed value for the parameter, rather than to an unusual distribution of genotypes given the observed proportion of A alleles.

If we have two or more categorical variables, we use multiway tables such as **Table 3**, which now also includes the incidence of hypertension (H). The hypothesis most usually tested using cross-tabulated data is whether there is an association between the two categorical variables. For example, we can see that although 14% of the individuals have hypertension overall, individuals with AA have a 70% incidence of hypertension, but those with aa have only a 5% incidence. How likely is this to arise by chance under the null hypothesis that there is no association between genotype and hypertension?

The two most common tests of association are Fisher's exact test[2] and the χ^2 test of independence[3]; both compute the probability of observing a table that is as extreme as or more extreme than the one observed when there is no association between variables and the row and column totals are fixed. For example, a table with more extreme observations would be one with all the same entries in the Aa and aa columns as in **Table 3** but with values of 2 and 8 in the AA

Table 2 | Distribution of two alleles, A and a, among 100 individuals shown in **Table 1**

	Genotype		
	AA	Aa	aa
Observed	10	50	40
Expected	12.25	45.5	42.25
Difference	−2.25	4.5	2.25

Here the expected values are based on the fraction of A in the population estimated from the sample, $p = 0.35$.

Table 3 | A multiway table reports two or more categorical variables

	Genotype			
Hypertension	**AA**	**Aa**	**aa**	**Total hypertension**
No	3 (8.6)	45 (43)	38 (34.4)	86
Yes	7 (1.4)	5 (7)	2 (5.6)	14
Total	10	50	40	100

Here, the counts of each genotype is broken down by the incidence of hypertension within each category. Values in parentheses represent expected values based on column and row totals and are used in the χ^2 test of independence.

Table 4 | A scenario with larger expected counts than in **Table 3**.

	Genotype			
Hypertension	**AA**	**Aa**	**aa**	**Total hypertension**
No	3 (5)	27 (25)	8 (20)	50
Yes	7 (5)	23 (25)	30 (20)	50
Total	10	50	40	100

When the smallest expected count is larger than 5, the *P* values estimated from the χ^2 and Fisher's tests are similar. Expected values are shown in parentheses.

column for H = No and H = Yes, respectively. Given the table totals and under the null hypothesis, these observations are more extreme, since we expect only 1.4 individuals (14% of 10) with this genotype to have hypertension.

Fisher's exact test is based on the hypergeometric distribution. It computes the probability of allocating each of the 86 normotensive individuals and each of the 14 hypertensive individuals at random among the 3 genotypes or, equivalently, allocating the 10 AA individuals, 50 Aa individuals and 40 aa individuals at random to the hypertensive and normotensive groups. For **Table 3**, Fisher's exact test gives $P = 2.6 \times 10^{-5}$, and we can conclude that it is highly unlikely that a table this extreme or more extreme would be observed if there is no association between hypertension and genotype.

The χ^2 test uses the fact that when two events are independent, the probability of both occurring is the product of the probabilities of each, for example, $P(\text{AA}\cap\text{H}) = P(\text{AA})P(\text{H})$. For example, since $P(\text{AA}) = 10\%$ and $P(\text{H}) = 14\%$, then $P(\text{AA}\cap\text{H}) = 1.4\%$. Thus, among 100 individuals we expect a count of 1.4 for AA and H. The test applies this substitution to all the cells in the table: replacing the counts with their expected values calculated by assuming independence, calculating probabilities for each cell and multiplying them by the total sample size (**Table 3**, values in parentheses). The χ^2 statistic is then computed as for the goodness-of-fit test, and for **Table 3** we find $\chi^2 = 29.4$ and $P = 4.1 \times 10^{-7}$. The degrees of freedom for this test is d.f. $= (r-1)(c-1) = 2$, for a table with $r = 2$ rows and $c = 3$ columns. This value reflects that, for given totals, we need $(r-1)(c-1)$ expected counts in different rows and columns to determine all of the rc expected counts.

The χ^2 test is commonly used because Fisher's exact test requires substantially more computation. While Fisher's exact test provides an exact evaluation of the *P* value, when the smallest expected count is 5 or greater, the χ^2 statistic is an adequate approximation to the *P* value. The accuracy of the approximation is better for larger expected values and less extreme *P* values.

In **Table 3**, because the smallest expected count is only 1.4 and the *P* value is very small, the two *P* values differ substantially. If instead we observed the counts shown in **Table 4**, with a smallest expected count of 5, the *P* value from Fisher's exact test is $P = 0.0031$ while the *P* value from the χ^2 test is $P = 0.0036$. For less extreme *P* values, the approximate value from the χ^2 test is even closer to the exact value from Fisher's exact test.

The χ^2 test uses the fact that if events A and B are independent, $P(\text{AB}) = P(\text{A})P(\text{B})$ or, alternatively, $\log(P(\text{AB})) = \log(P(\text{A})) + \log(P(\text{B}))$. Letting $x = P(\text{AB})/(P(\text{A})P(\text{B}))$, we have $\log(P(\text{AB}) = \log(P(\text{A})) + \log(P(\text{B})) + \log(x)$. Thus the null hypothesis that A and B are independent is equivalent to a test that $x = 1$ or $\log(x) = 0$. This is similar to the testing for the presence of an effect in ANOVA or additive models. Another way of testing the independence hypothesis that is particularly useful for three- and higher-way tables is to test whether the log(counts) can be accounted for by an additive model (independence) or whether non-additive (interaction) terms are needed[4].

What is often not appreciated in assessing hypotheses using the χ^2 test is that, although the χ^2 distribution is continuous, for any set of row and column count totals, only a finite set of *P* values is possible. In addition, because there is a table that is the least extreme and that can occur with finite probability, $P = 1$ is an exact and valid value. Because of this, in multiple testing problems such as genome-wide association studies in which the association between a trait and multiple genotypes is assessed, the histogram of *P* values tends to have peaks at many values, including at $P = 1$. This phenomenon does not occur for continuous measurement data. However, simulation studies have shown[5] that methods for adjusting *P* values to control the false discovery rate (FDR) are conservative for tabular data (and therefore have at most the desired FDR).

Historically, tabular data were generated when samples were classified into categories defined by a small number of nominal variables. In high-throughput analysis, tabular data arise from counting reads (classified by gene), assigning samples to genotypes (often at many loci) or other types of markers. The availability of software for computing Fisher's exact test and log-linear models with high-throughput data have made it feasible to test for dependence between these types of data and a categorical phenotype such as disease status. The χ^2 goodness-of-fit test may also be used to determine whether the counts follow distributions suggested by prior knowledge about the biological system.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

Naomi Altman is a professor of statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.

1. Snedecor, G.W. & Cochran, W.G. *Statistical Methods* 8th edn. (Iowa State Univ. Press, 1989).
2. Fisher, R.A. *J. R. Stat. Soc. Ser. A* **85**, 87–94 (1922).
3. Pearson, K. *Philosophical Magazine* **50**, 157–175 (1900).
4. Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. *Discrete Multivariate Analysis* (MIT Press, 1975).
5. Dialsingh, I., Austin, S. & Altman, N.S. *Bioinformatics* **31**, 2303–2309 (2015).

nature methods

AMENDMENTS

https://doi.org/10.1038/s41592-019-0474-z

Author Correction: Tabular data

Naomi Altman and Martin Krzywinski

Correction to: *Nature Methods* https://doi.org/10.1038/nmeth.4239, published online 30 March 2017

The published version of this article contains a mathematical error. In the discussion of z tests in the fourth paragraph, the same z score is incorrectly given for both AA and aa; the last two sentences of the paragraph should read as follows: "For example, for AA the z score of the difference is $3.75/\sqrt{6.25} = 1.5$, yielding $P = 0.13$. However, for aa the same computation yields $-16.25/\sqrt{56.25} = -2.167$ and $P = 0.03$."

Published online: 12 June 2019
https://doi.org/10.1038/s41592-019-0474-z