SPECIAL ARTICLE

Basic Statistics for Comparing Categorical Data From 2 or More Groups

Matt Hall, PhD, Troy Richardson, PhD

In both clinical research and quality improvement, it is commonplace to compare groups of patients (eg, treatment versus control, pre versus post, hospital A versus hospital B) on a variety of characteristics. These characteristics usually take the form of (1) continuous data with comparisons made with *t* tests (for normal distributions) or Wilcoxon rank-sum tests (for nonnormal distributions) or (2) categorical data. Continuous data are data that can take almost any numeric value within a given range and can be subdivided into smaller and smaller increments without losing the meaning associated with the data. Examples of continuous data commonly found in health care include age, height, weight, temperature, or cost. Categorical data, as the name suggests, can be put into nonoverlapping categories, groups, or classes. Some examples of categorical data that frequently occur in health care are gender, disposition, and skill level (eg, RN, LPN, AHT). Antibiotic receipt, chest radiograph receipt, or admission from the emergency department also qualify because they can be categorized into "yes" or "no" responses. As long as people cannot be classified in >1 group, you are likely dealing with categorical data. There is, however, a special type of categorical data that is treated a little differently from the data we discuss in this article, and it is somewhere between categorical and continuous. It too can be put into nonoverlapping categories, but the categories have a logical ordering or sequence. This is called ordinal data, and a Likert scale commonly used on surveys (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree) is an example. Although experts do not always agree on the best approach to analyze ordinal data, it generally requires a different approach from the categorical data that we discuss in this here.

Knowing what type of data you have is always the first step in any analytical silique because it dictates the statistical approach that is taken. However, regardless of the type of data you are dealing with, our aim is to understand if our groups are different with respect to some characteristics. In the context of hospital quality, we may want to compare our hospital's rate of an event (eg, adverse drug events for ICU patients) with another hospital's rate to see whether we have an opportunity for improvement. Or we may need to compare our hospital's 30-day readmission rate before and after an intervention to determine if the intervention was successful.

In the context of a comparative effectiveness or randomized controlled study, comparisons like this are useful to assess the balance of the cohorts on key characteristics before analyzing the outcome. For example, if we determine that female patients were significantly more likely to receive one of the treatments, then we need to take this into account when comparing the effectiveness of the treatments. Otherwise, the comparison of the effectiveness may be confounded (ie, ignoring a variable that is related to both the dependent and independent variables). Different approaches to mitigate the effect of confounding are available and include stratifying the analysis, multivariable modeling, and matching through propensity scores or other means. In this article, we discuss how to compare

www.hospitalpediatrics.org
DOI:10.1542/hpeds.2015-0273

Address correspondence to Matt Hall, PhD, 6803 W. 64th St, Overland Park, KS 66202. E-mail: matt.hall@childrenshospitals.org

HOSPITAL PEDIATRICS (ISSN Numbers: Print, 2154-1663; Online, 2154-1671).

FINANCIAL DISCLOSURE: The authors have indicated they have no financial relationships relevant to this article to disclose.

FUNDING: No external funding.

Children's Hospital Association, Overland Park, Kansas

POTENTIAL CONFLICT OF INTEREST: The authors have indicated they have no potential conflicts of interest to disclose.

Drs Hall and Richardson conceptualized and designed the study, drafted the initial manuscript, and approved the final manuscript as submitted.

categorical data from ≥2 groups to detect these important differences.

SUMMARIZING AND ORGANIZING THE DATA

One common way of summarizing categorical data are to use proportions (or percent if we multiply the proportion by 100). Suppose we want to compare the proportion of patients in our hospital's ICU that experience an adverse drug event (ADE) with that of another hospital's ICU. In our hospital, 20 of 180 patients experienced an ADE (20/180 = 11.1%), whereas in another hospital, 25 of 155 patients (16.1%) experienced an ADE. It looks like these percentages are different (11.1% vs 16.1%), but do we have enough evidence to conclude that they are really different and not due to random chance? To get a better sense for whether the observed differences are due to random chance, we will perform a hypothesis test.

To facilitate our hypothesis test, we need to organize the data into a contingency table (ie, an $r \times c$ table where r is the number of rows corresponding to the number of levels for the categorical variable and c is the number of columns corresponding to the number of comparison groups). In our example, the data can be organized into a 2×2 table with 2 rows (experienced an ADE: yes vs no) and 2 columns (hospital A and hospital B) as in Table 1.

THE HYPOTHESIS TEST

We would like to know whether the proportion of ICU patients experiencing ADEs is the same at our hospital compared with the other hospital. In statistical terminology, we are testing the following hypotheses:

$$\text{Null hypothesis}(H_0): \pi_{\text{Hospital A}} = \pi_{\text{Hospital B}}$$
$$\text{Alternative hypothesis}(H_1): \pi_{\text{Hospital A}} \neq \pi_{\text{Hospital B}}$$

where π represents the proportion of ICU patients experiencing an ADE. In statistical testing, we always assume that the null hypotheses is true and then determine if we have enough evidence from the data to reject the null in favor of the alternative hypothesis. In other words, we assume the proportion of ICU patients experiencing an ADE at the 2 hospitals is the same until we have sufficient evidence from the data to conclude otherwise.

TABLE 1 Observed Counts of ADEs in 2 Hospitals

	Hospital A	Hospital B	Total
Experienced an ADE	20	25	45
Did not experience an ADE	160	130	290
Total	180	155	335

TYPES OF TESTS

Generally, 2 main tests are used for comparing categorical data across ≥2 groups: χ^2 test[1] (sometimes referred to as Pearson's χ^2 test of independence) and Fisher's exact test.[2] The development of the χ^2 test is fairly intuitive. At a high level, we decide how the data would look in our table if the null hypothesis was true (ie, the 2 proportions were equal) and then measure how far off the actual data are from these expected counts. By default, most statisticians use the χ^2 test because it performs well under most circumstances. However, the validity of the test can come into question when you have small observed numbers in the cells of your table. If this is the case, then you should consider using Fisher's exact test.

CALCULATING EXPECTED CELL COUNTS

If the null hypothesis is true and the proportions are equal, then we can calculate the counts that we would expect to see in the 4 cells of our table. We can do this by using the total count of cases that we are trying to redistribute (335 in our example), and the marginal proportions displayed in Table 2. To determine the expected number of patients who experienced an ADE in Hospital A (ie, the upper left cell of the table), we simply take the overall population size (335) and multiply it by the proportion of patients who we expect in Hospital A (0.5373) and the proportion of all patients who we would expect to experience an ADE (0.1343). That is, of the 335 patients, we expect 53.73% to be in Hospital A and 13.43% of patients to experience an ADE ($335 \times 0.5373 \times 0.1343 = 27.18$ patients). Similar calculations are performed for each of the cells until the expected table is completed, as in Table 3. Notice that the marginal totals are exactly as they were in the original Table 1 but that the percentage of patients in each hospital experiencing an ADE is now equal (24.18/180 = 13.4% and 20.82/155 = 13.4%).

TABLE 2 Observed Counts of ADEs in 2 Hospitals With Marginal Proportions

	Hospital A	Hospital B	Total (Column Proportion)
Experienced an ADE	20	25	45 (0.1343)
Did not experience an ADE	160	130	290 (0.8657)
Total (row proportion)	180 (0.5373)	155 (0.4627)	

MEASURING OBSERVED FROM EXPECTED

We now know what the table would have looked like if the null hypothesis was true and the proportions were equal (ie, the expected in Table 3), and we know what the data actually looked like (ie, the observed in Table 1). Next, we measure how far the observed was from what was expected and decide if we have enough evidence to reject the null hypothesis. To do this, we calculate the following measure for each cell:

$$\frac{(Observed - Expected)^2}{Expected}$$

and then sum them to form the χ^2 statistic: $X^2 = 0.722 + 0.839 + 0.112 + 0.130 = 1.803$

STATISTICAL SIGNIFICANCE

Finally, we need to determine if the X^2 we calculated is "large enough" to reject the null hypothesis. This is why we need to compute a P value. It tells us the probability of getting a test statistic (X^2) at least this extreme, assuming that the null hypothesis is true. The test statistic has a χ^2 distribution, so we compare the value we calculated for X^2 to a χ^2 distribution. One of the nuances of the χ^2 distribution is that you need to know the degrees of freedom

TABLE 3 Expected Counts of ADEs in 2 Hospitals

	Hospital A	Hospital B	Total
Experienced an ADE	24.18	20.82	45
Did not experience an ADE	155.82	134.18	290
Total	180	155	

(df; ie the number of data points available to estimate a parameter of the population) for the test statistic. For the χ^2 test, the df is (number of rows − 1) × (number of columns − 1). For a 2 × 2 table, the df = 1. Comparing our $X^2 = 1.803$ to the χ^2 distribution with df = 1 [using an online calculator like Social Science Statistics: http://www.socscistatistics.com/tests/chisquare/default2.aspx or the EXCEL function = 1-CHISQ.DIST(1.803,1,TRUE)], we get $P = .179$. Here, the probability of getting a test statistic at least as extreme as the one we calculated is rather high ($P > .05$), so we fail to reject the null hypothesis and conclude that the proportion of patients experiencing ADEs between hospital A (11.1%) and hospital B (15.6%) are not significantly different.

FISHER'S EXACT TEST

When you have cells in your table where the expected count is <5, you should be cautious of using the χ^2 test and instead use Fisher's exact test to be more conservative. Most statistical packages will give you a warning that the χ^2 test may be invalid and also compute Fisher's exact test for you. Although it is valid for all sample sizes, computing Fisher's exact test is substantially more mathematically involved than doing a χ^2 test and computationally challenging with tables beyond 2 × 2. Although programming is a bit tricky in Excel, there are online calculators such as Social Science Statistics (http://www.socscistatistics.com/tests/fisher/default2.aspx) that can be used to calculate Fisher's exact test.

Fisher developed his test as part of the classic "lady tasting tea" experiment that he proposed with Muriel Bristol, who claimed that she could tell if milk was added to a cup before or after her tea. He proposed a randomized experiment with 8 cups of tea (4 with the milk added first, and 4 with the tea added first) given to her in random order. He described his exact test using probability calculations to determine the likelihood that she was simply guessing. She supposedly got them all correct, and the likelihood of getting them all correct if she was guessing was 1.4%. He describes the experiment in his classic book on the design of experiments.[3]

EXPANDING TO TABLES LARGER THAN 2×2

The χ^2 test can easily be generalized to tables with >2 rows and/or 2 columns. The steps involved and the computations are nearly identical with the appropriate changes made to the df for the number of rows and columns. Table 4 shows a comparison of 30-day asthma readmission rates across 3 hospitals. In this case, the χ^2 test is evaluating whether the distribution across the 3 hospitals is the same or not. The $P = .009$ indicates that there are significant differences across the hospitals, but it does not necessarily tell us where these differences occur. To determine where these differences exist, we can break the table into three 2 × 2 tables and do pairwise comparisons. The first 2 × 2 table would compare Hospital A versus Hospital B ($P = .002$), the second would compare Hospital B versus Hospital C ($P = .488$), and the last would compare Hospital A versus C ($P = .023$). A little caution is necessary when performing pairwise comparisons. By doing 3 statistical tests on the 2 × 2 tables, you are increasing your likelihood of coming to an erroneous conclusion by committing a type I error (ie, erroneously rejecting a true null hypothesis of equal proportions). One easy way to protect from this is by doing a Bonferroni correction,[4] which reduces the significance level for each individual test from the traditional $P = .05$ level, and shares this across the multiple tests. Because we are performing 3 tests, we would reduce the significance level to 0.05/3 = 0.017 for each test. With this new level of significance, we would conclude that Hospitals A and B are statistically different, but none of the other pairs are. Although the Bonferroni correction can be overly conservative, it is intuitive and easy to perform. Another common approach is to control for the false discovery rate (ie, the expected proportion of tests that reject the null hypothesis erroneously) using the Benjamini-Hochberg procedure.[5]

TABLE 4 Comparison of 30-Day Asthma Readmission Rates Across 3 Hospitals

	Hospital A, *n* (%)	Hospital B, *n* (%)	Hospital C, *n* (%)	*P*
No readmission	701 (96.6)	810 (93.1)	625 (94.0)	.009
Readmission	25 (3.4)	60 (6.9)	40 (6.0)	

CONCLUSIONS

It is important to understand how populations differ on baseline characteristics before assessing outcomes so that these differences can be accounted for appropriately in the analysis. The χ^2 test is an intuitive test that is easy to calculate, and it is useful for comparing proportions across groups for categorical variables. With this test, you can decide if there are important differences that may confound your results and take appropriate steps to avoid this.

REFERENCES

1. Pearson KX. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*. 1900;50:157–175
2. Fisher RA. *Statistical Methods for Research Workers*. 5th ed. Edinburgh, Scotland: Oliver & Boyd; 1934
3. Fisher RA. *The Design of Experiments*. New York, NY: Hafner Publishing Company; 1935
4. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310(6973):170
5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B*. 1995;57:289–300

Basic Statistics for Comparing Categorical Data From 2 or More Groups

Matt Hall and Troy Richardson

Hospital Pediatrics 2016;6;383
DOI: 10.1542/hpeds.2015-0273 originally published online May 26, 2016;

Hospital Pediatrics®

AN OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

Basic Statistics for Comparing Categorical Data From 2 or More Groups

Matt Hall and Troy Richardson

Hospital Pediatrics 2016;6;383
DOI: 10.1542/hpeds.2015-0273 originally published online May 26, 2016;