Hypothesis Testing: Proportions

Kimberlee Gauvreau

Circulation 2006;114;1545-1548
DOI: 10.1161/CIRCULATIONAHA.105.586487
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7322. Online ISSN: 1524-4539

Hypothesis Testing
Proportions

Kimberlee Gauvreau, ScD

The process of drawing conclusions about an entire population on the basis of the information contained in a random sample drawn from that population is known as statistical inference. Methods of statistical inference fall into 2 general categories: estimation and hypothesis testing. With estimation, our goal is to describe or estimate some characteristic of a population of interest, such as the mean pulmonary regurgitation fraction of all patients alive 10 years after repair of tetralogy of Fallot or the proportion of children with acute Kawasaki disease who develop coronary artery abnormalities. With hypothesis testing, we begin by claiming that the population parameter of interest is equal to some postulated value (or, in the situation in which we are comparing 2 populations, that the 2 parameters are equal to each other). This statement about the value of the population parameter is called the null hypothesis (H_0). The alternative hypothesis (H_A) is a second statement that contradicts the null. Together, the null and alternative hypotheses account for all possible values of the population parameter; consequently, 1 of the 2 statements must be true. After formulating the hypotheses needed to answer our study question, we draw a random sample from the population of interest and use the information in this sample to calculate a test statistic. The test statistic is compared with the critical values of an appropriate probability distribution. If there is evidence that the sample could not have come from a population with the postulated value of the parameter, as determined by a comparison of the magnitude of the test statistic with the critical values of the probability distribution, we reject the null hypothesis. This occurs when the probability value of the test is sufficiently small, usually <0.05. The probability value is the probability of observing a test statistic as large as we got, or even larger, given that the null hypothesis is true. In this case we conclude that the data are not compatible with the null hypothesis; they are more supportive of the alternative. Such a test result is said to be statistically significant. If the probability value of the test is large, we fail to reject the null hypothesis.

The Sample Proportion

With dichotomous or binary data, values fall into 2 unordered categories or classes that are mutually exclusive; examples of dichotomous variables include gender and survival to hospital discharge after a surgical procedure yes/no. With this type of data, the proportion of times that a particular outcome occurs is the parameter of interest.

If we wish to calculate the proportion of times that some outcome occurs in a population, we count the number of subjects in the population who experience the outcome and divide by the total number of individuals in the population. The population proportion is represented by p. For a random sample, we count the number of subjects in the sample who experience the outcome and divide by the total number in the sample. The proportion of outcomes in the sample, called the sample proportion, is denoted by $\hat{p}$.

When analyzing proportions, we often rely on the binomial distribution. Suppose that we randomly select a sample of 20 patients from the population of children with acute Kawasaki disease. How many of the children in this sample will develop coronary artery abnormalities? The outcome development of coronary artery abnormalities is a dichotomous variable; a child either develops abnormalities or does not. We assume that each child in the population has the same probability of developing abnormalities, denoted by p. In this case, the number of children out of 20 who develop coronary artery abnormalities follows a binomial distribution.[1]

In practice, the binomial distribution can be cumbersome to work with if the sample size n is large. As an alternative, we often use approximate procedures based on the normal distribution. If n is large, then the sample proportion $\hat{p}$ has a normal distribution.[1]

One-Sample Test for a Proportion

In some clinical studies, a single sample of patients is collected. The goal of the study is to determine whether the proportion of times that an outcome occurs in the population from which the sample was drawn is equal to the proportion of times it occurs in an appropriate standard or reference population.

For example, suppose that we are interested in examining cognitive function as measured by the intelligence quotient (IQ) score for individuals who have survived a Fontan procedure. The Fontan procedure is an operation performed on patients with complex congenital heart defects that result in 1 functional ventricle rather than 2.[2] In the general population, IQ scores are scaled to have a normal distribution with mean 100 and SD 15.[3–5] Approximately 2.5% of the values in a normal distribution lie >2 SDs below the mean; therefore, $\approx$2.5% of IQ scores in

From the Department of Cardiology, Children's Hospital, Boston, Mass.
Correspondence to Kimberlee Gauvreau, ScD, Department of Cardiology, Children's Hospital, 300 Longwood Ave, Boston, MA 02115. E-mail gauvreau@tch.harvard.edu
(*Circulation.* 2006;114:1545-1548.)

Circulation is available at http://www.circulationaha.org DOI: 10.1161/CIRCULATIONAHA.105.586487

the general population lie <70. We wish to know whether the proportion of Fontan survivors who have an IQ score <70 is also equal to 0.025, the proportion for the general population.

To conduct a hypothesis test, we begin by claiming that p, the proportion of Fontan survivors with an IQ score <70, is in fact equal to the proportion in the general population. This postulated proportion is represented by p_0. Therefore, we test the null hypothesis

$$H_0: p = p_0 = 0.025$$

against the alternative

$$H_A: p \neq 0.025.$$

Together, these 2 hypotheses account for all possible values of the population proportion p; 1 and only 1 of the hypotheses must be true.

We draw a random sample from the population of Fontan survivors, measure IQ score for each patient in the sample, and compare the sample proportion of individuals with an IQ score <70 to the postulated proportion $p_0=0.025$. In a sample of size $n=128$, 10 patients had an IQ score <70; therefore, $\hat{p}=10/128=0.078$. Note that $\hat{p}$ is a random variable; if we were to select a different sample of size 128, we would almost surely get a different value for the sample proportion because of sampling variability. How much variability is allowed? In other words, is the difference between the observed sample proportion $\hat{p}$ and the postulated proportion p_0 too large to be attributed to sampling variability alone?

To answer this question, we must quantify the amount of variability expected in the sample proportion; we do this using the SD of $\hat{p}$, defined as follows

$$\sqrt{p \times (1-p)/n}.$$

This is also called the standard error of $\hat{p}$. The difference between $\hat{p}$ and p_0 divided by the standard error gives us the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 \times (1-p_0)/n}}.$$

The denominator of the test statistic,

$$\sqrt{p_0 \times (1-p_0)/n},$$

is the value of the standard error given that the null hypothesis is true and $p=p_0$. If the null hypothesis is true, this test statistic has a standard normal distribution with mean 0 and SD 1. The larger the absolute value of the test statistic, meaning the farther it is from 0, the stronger is the evidence that the null hypothesis is not true. For the standard normal distribution, 2.5% of the values lie below the critical value −1.96 (≈2 SDs below the mean), and 2.5% lie above 1.96. Therefore, if we are conducting a 2-sided hypothesis test at the 0.05 level of significance, we reject H_0 when $z < -1.96$ or $z > 1.96$.

The probability value of the test for Fontan survivors, defined as the probability of observing a sample proportion as far from the postulated value of 0.025 as 0.078, or even farther, given that the null hypothesis is true and p really is 0.025, is $P<0.001$ (Figure 1). Because this probability is <0.05, we reject the null hypothesis; the data in the sample are more compatible with the alternative that $p \neq 0.025$. In fact, it appears that the proportion of Fontan survivors who have an IQ score <70 is >2.5%, the proportion in the general population.

$$H_0 : p = p_0 = 0.025$$
$$H_A : p \neq 0.025$$

$$n = 128, \hat{p} = 10/128 = 0.078$$

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 \times (1-p_0)/n}}$$

$$= \frac{0.078 - 0.025}{\sqrt{0.025 \times (1-0.025)/128}}$$

$$= 3.84$$

Figure 1. One-sample test for a proportion compared to a known value. The test statistic tells us that the sample proportion $\hat{p}$ is 3.84 standard errors above the postulated proportion p_0; when the standard normal distribution is used, the probability that this occurs given that the null hypothesis is true is $P<0.001$. Therefore, we reject the null hypothesis.

The mathematical derivation of the 1-sample test statistic assumes that the sample size n is large enough that the binomial distribution can be approximated by a normal distribution. In general, this assumption is satisfied if n is large and p is not too close to either 0 or 1. One rule of thumb states that we should have both $n \times \hat{p} \geq 5$ and $n \times (1-\hat{p}) \geq 5$.[1]

If the sample size is not large enough, an exact method of hypothesis testing uses the binomial distribution itself rather than relying on the normal approximation.[6] This test is more computationally intensive than the normal theory method but can be performed by many statistical software packages. For large sample sizes, the 2 methods produce nearly identical probability values. For small samples, the exact binomial test is preferred.

Two-Sample Tests for Proportions

Rather than compare the proportion of times that an outcome occurs in a single population with the known proportion for some standard or reference population, it is more common to compare the proportions in 2 different populations, neither of which is known. Most often we want to know whether the 2 proportions are equal. The hypothesis test we use depends on whether our data come from independent or paired samples.

Independent Samples

When samples are drawn from 2 independent populations, the normal theory method described for the 1-sample test can be generalized to compare the proportions of times an outcome occurs in each of 2 populations. The null hypothesis claims that the 2 population proportions are identical, or

$$H_0: p_1 = p_2$$

whereas the alternative hypothesis says that they are not

$$H_A: p_1 \neq p_2.$$

We draw a random sample from each population and calculate 2 sample proportions $\hat{p}_1$ and $\hat{p}_2$. If the null hypothesis is true, we expect the 2 sample proportions to be fairly close to

TABLE 1. Data for 500 Patients Undergoing the Fontan Operation

		Early Failure		
		Yes	No	Total
Heterotaxy Syndrome	Yes	9	32	41
	No	75	384	459
	Total	84	416	500

each other. We reject H_0 if they are too far apart. The test statistic takes the form

$$z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}\times(1-\hat{p})\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $\hat{p}$ is the proportion of times that the outcome occurs in the 2 samples combined. Again this test statistic follows a standard normal distribution; therefore, H_0 will be rejected if $z<-1.96$ or $z>1.96$. The probability value of the test is the probability of observing 2 sample proportions as far apart or even farther apart than the observed values $\hat{p}_1$ and $\hat{p}_2$, given that the null hypothesis is true and $p_1=p_2$. Although this technique is straightforward, when presented with a comparison of proportions from 2 independent populations, it is more common to apply contingency table methods.

To illustrate, a study evaluated factors associated with early failure of the Fontan procedure. Early failure was defined as death, takedown of the Fontan circulation, or cardiac transplantation within 30 days of the operation or before hospital discharge.[7] One research question was as follows: Is there any difference in the proportion of early failures for patients with and without a diagnosis of heterotaxy syndrome, which involves abnormal left/right placement of 1 or more organs in the body? The null hypothesis that the 2 population proportions of early failure are the same, H_0: $p_1=p_2$, implies that there is no association between heterotaxy syndrome and early failure.

Data from a sample of 500 patients are arranged in a tabular format known as a contingency table (Table 1). In its simplest form, the 2×2 contingency table, 2 dichotomous variables are involved. The rows of the table represent the values of one variable (eg, presence of heterotaxy syndrome), and the columns the other (eg, early failure). The entries in the cells of the table are the counts that correspond to a particular combination of categories.

The χ^2 test compares the observed frequencies in each category or cell of the table (O) with the expected frequencies given that the null hypothesis is true (E). It is used to determine whether the deviations or differences between observed and expected counts in the 4 cells are too large to be attributed to sampling variability alone. The test statistic takes the form

$$\chi^2=\sum_{i=1}^{4}(O_i-E_i)^2/E_i.$$

This statistic has a χ^2 distribution with 1 degree of freedom; the larger the test statistic, the stronger is the evidence that the null hypothesis is not true. If we are conducting the test at the 0.05 level of significance, we reject the null hypothesis when $\chi^2>3.84$. Note that this is mathematically equivalent to the hypothesis test based on the standard normal distribution; the critical value of 3.84 for the χ^2 test is actually (1.96).[2] The probability value for the test (Figure 2) is the probability of observing differences $O-E$ as large as or even larger than those obtained given that the null hypothesis is true. Because this probability is >0.05, we fail to reject the null hypothesis. The data are more compatible with the null hypothesis of no association between heterotaxy syndrome and early failure than they are with the alternative hypothesis.

$$H_0: p_1 = p_2$$
$$H_A: p_1 \neq p_2$$

or

H_0: There is no association between heterotaxy syndrome and early failure.

H_A: There is an association between heterotaxy syndrome and early failure.

$$O_1 = 9, O_2 = 32, O_3 = 75, O_4 = 384$$
$$E_1 = 6.9, E_2 = 34.1, E_3 = 77.1, E_4 = 381.9$$

$$\chi^2 = \sum_{i=1}^{4}(O_i - E_i)^2 / E_i$$
$$= \left[\frac{(9-6.9)^2}{6.9}\right] + \left[\frac{(32-34.1)^2}{34.1}\right] + \left[\frac{(75-77.1)^2}{77.1}\right] + \left[\frac{(384-381.9)^2}{381.9}\right] = 0.85$$

$$n_1 = 41, \hat{p}_1 = 9/41 = 0.220$$
$$n_2 = 459, \hat{p}_2 = 75/459 = 0.163$$
$$\hat{p} = 84/500 = 0.168$$

$$z = \frac{\hat{p} - p_2}{\sqrt{\hat{p}\times(1-\hat{p})\times(1/n_1 + 1/n_2)}}$$
$$= \frac{0.220 - 0.163}{\sqrt{0.168\times(1-0.168)\times(1/41 + 1/459)}}$$
$$= 0.921$$

Figure 2. Tests for equality of proportions in 2 independent samples. The first test statistic has a χ^2 distribution with 1 degree of freedom; the second has a standard normal distribution. Note that $(0.921)^2=0.85$. The probability of obtaining a test statistic this large or larger given that the null hypothesis is true is $P=0.36$ in both cases. Therefore, we fail to reject the null hypothesis.

In addition to assuming that the observations or subjects are independent, the χ^2 test is based on an approximation that works best when the samples are fairly large and the proportions being compared are neither too big nor too small. As a conservative guideline, no cell in a 2×2 table should have an expected count <5.[1] If this assumption is violated, then the Fisher exact test can be used instead.[6] It is never wrong to use the exact test, and it is preferable for small sample sizes.

Paired Samples

We now consider the situation in which the dichotomous data of interest come from paired rather than independent samples. The defining characteristic of paired dichotomous data is that for each observation in the first sample, there is a corresponding observation in the second sample.

A study was conducted to investigate the association between a diagnosis of cardiac enlargement based on chest x-ray and the same diagnosis based on echocardiogram.[8] The same group of study subjects had both tests performed; therefore, each individ-

TABLE 2. Data for 95 Patients Undergoing Evaluation for Cardiac Enlargement

		Echocardiogram		
		Normal	Enlarged	Total
Chest x-ray	Normal	72	7	79
	Enlarged	6	10	16
	Total	78	17	95

ual had 2 diagnoses. Is one test more likely than the other to result in a diagnosis of cardiac enlargement? The null hypothesis is that there is no association between a diagnosis of cardiac enlargement and the particular testing modality used; the alternative hypothesis is that there is an association, meaning that one test is more likely than the other to produce a diagnosis of cardiac enlargement.

A sample of 95 subjects underwent both testing procedures; diagnosis was assessed independently by 2 different physicians. By chest x-ray, 16 patients had a diagnosis of cardiac enlargement. By echocardiogram, 17 patients had this diagnosis. Ten patients received the diagnosis on both tests. The data are summarized in Table 2. Each entry in the table corresponds to the pair of results for a single individual. Therefore, the sample size is 95 pairs rather than 190 measurements.

With this type of data, the concordant pairs, in which a patient has the same diagnosis on both the chest x-ray and the echocardiogram, provide no information about differences between the 2 tests. Therefore, we discard the concordant pairs and instead focus on the discordant pairs, in which a patient gets different diagnoses with the 2 different procedures. If the null hypothesis is true and there is no relationship between a diagnosis of cardiac enlargement and testing modality, then we would expect the numbers of each of the 2 different types of discordant pairs to be equal. In other words, the number of pairs in which the chest x-ray is normal but the echocardiogram indicates enlargement (represented by r) should be equal to the number of pairs in which the echocardiogram is normal and the chest x-ray indicates enlargement (represented by s). The McNemar test is used to determine whether the observed difference between r and s is larger than would be expected by sampling variability alone. The test statistic takes the form

$$\chi^2=\frac{(r-s)^2}{r+s}$$

and again has a χ^2 distribution with 1 degree of freedom. The probability value of the test is the probability of observing a difference as big as or bigger than the absolute value of $r-s$, given that H_0 is true. Because the probability value of this test is large (Figure 3; $P=0.78$), we fail to reject the null hypothesis; the data are compatible with the null hypothesis. Note that this hypothesis test does not assume that either diagnostic modality is a gold standard.

If a statistically significant result had been found, the conclusion drawn would necessarily be conditional on an observed difference in testing modalities; keep in mind that the concordant pairs of data were discarded. For example, if we had determined that there were more pairs in cases in which the chest x-ray is normal and the echocardiogram indicates cardiac enlargement than the other way around, we would have concluded that in situations in which the 2 tests produce different results, it is more likely that the echocardiogram will identify cardiac enlargement.

H_0: There is no association between testing modality and a diagnosis of cardiac enlargement.
H_A: There is an association between testing modality and a diagnosis of cardiac enlargement.

$$\chi^2=\frac{(r-s)^2}{r+s}$$
$$=\frac{(7-6)^2}{7+6}$$
$$=0.08$$

Figure 3. McNemar test for paired dichotomous data. The test statistic has a χ^2 distribution with 1 degree of freedom. The probability of obtaining a test statistic this large or larger given that the null hypothesis is true is $P=0.78$. Therefore, we fail to reject the null hypothesis.

The McNemar test is based on an approximation that works best when the number of discordant pairs is fairly large. If this is not the case, an exact binomial test is available for small samples.[6]

Summary

We have described elementary methods for testing hypotheses about proportions for 1 or 2 populations. These methods are used for dichotomous data and include a z test based on the normal distribution to compare 1 sample proportion with a postulated or reference value, as well as χ^2 tests to compare proportions in 2 independent or paired samples. With small samples or proportions close to 0 or 1, exact tests should be used instead of these large-sample approximate procedures. Methods to account for concomitant or confounding variables will be addressed later in the series.

Disclosures

None.

References

1. Pagano M, Gauvreau K. *Principles of Biostatistics*. 2nd ed. Pacific Grove, Calif: Duxbury/Thomson Learning; 2000.
2. Wernovsky G, Stiles KM, Gauvreau K, Gentles TL, du Plessis AJ, Bellinger DC, Walsh AZ, Burnett J, Jonas RA, Mayer JE, Newburger JW. Cognitive development following the Fontan operation. *Circulation*. 2000;102:883–889.
3. Wechsler D. *Wechsler Preschool and Primary Scale of Intelligence—Revised Manual*. San Antonio, Tex: The Psychological Corporation, Harcourt Brace Jovanovich Inc; 1989.
4. Wechsler D. *Wechsler Intelligence Scale for Children*. 3rd ed. San Antonio, Tex: The Psychological Corporation, Harcourt Brace Jovanovich Inc; 1991.
5. Wechsler D. *Wechsler Adult Intelligence Scale—Revised Manual*. San Antonio, Tex: The Psychological Corporation, Harcourt Brace Jovanovich Inc; 1981.
6. Rosner B. *Fundamentals of Biostatistics*. 6th ed. Belmont, Calif: Thomson Brooks/Cole; 2006.
7. Geggel RL, Gauvreau K, Lock JE. Balloon dilation angioplasty of peripheral pulmonary stenosis associated with Williams syndrome. *Circulation*. 2001;103:2165–2170.
8. Satou GM, Lacro RV, Chung T, Gauvreau K, Jenkins KJ. Heart size on chest x-ray as a predictor of cardiac enlargement by echocardiography in children. *Pediatr Cardiol*. 2001;22:218–222.

KEY WORDS: statistics ■ hypothesis testing