# Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3,4]

# Statistical Inference for Proportions

Received November 5, 2004; accepted after revision November 10, 2004.

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 16th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[2]Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W, Montreal, QC H3A 1A2, Canada. Address correspondence to L. Joseph (Lawrence.Joseph@mcgill.ca).

[3]Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, QC H3G 1A4, Canada.

[4]Synarc Inc., 575 Market St., San Francisco, CA 94105.

*AJR* 2005;184:1057–1064

0361–803X/05/1844–1057



This module will discuss the most commonly used statistical procedures when the parameters of interest arrive in the form of proportions. Understanding these methods is especially important to radiologists because so much radiologic research and clinical work involves dichotomous (e.g., yes or no, present or absent) outcomes summarized as proportions. For example, a given disease or condition may be present or absent in any given subject, and any time a diagnostic tool is used, test characteristics such as sensitivity, specificity, and positive and negative predictive values are all summarized as proportions.

We will continue to use the three basic methods for statistical inferences, including *p* values and confidence intervals (CIs) from a frequentist viewpoint, and posterior distributions leading to credible intervals from a Bayesian viewpoint. We will only briefly review the basic principles behind these generic inferential principles, so readers may wish to ensure they have a good understanding of the previous module [1] in this series before tackling this one. It may also be useful to recall the basic properties of the binomial distribution [2] because it is the central distribution used for inferences involving proportions.

We begin with inferences for single proportions, which are covered in the next section. Then we discuss inferences for two or more proportions from independent groups, inferences for dependent proportions, sample size determination for studies involving one or two proportions, and Bayesian methods for proportions. Finally, we will summarize what we have learned in this module.

## Inferences for Single Proportions

### *Standard Frequentist Hypothesis Testing*

Suppose a new computer-aided automated system for the detection of lung nodules on chest radiographs has been developed [3]. Suppose further that one wishes to investigate whether this new system provides improved sensitivity compared with standard detection via non-computer-aided methods of analyzing chest radiographs. In other words, suppose that chest radiographs are taken from a series of subjects who all truly have lung nodules, and we know that using standard (non-computer-aided) methods 90% of them will be found to have lung nodules and 10% of these cases will be missed. Is there evidence that the new computer-aided automated system provides increased sensitivity compared with the standard method of detection?

To look for evidence of improved sensitivity in the new automated system, we might wish to test the null hypothesis ($H_0$) that the automated system is in fact not better than standard detection, versus an alternative hypothesis ($H_A$) that it is better. Formally, we can state these hypotheses as:

$$H_0: p \leq 0.9$$
$$H_A: p > 0.9$$

where $p$ represents the unknown true probability of success of the new automated system in detecting lung nodules.

Suppose that we observe the results from 10 subjects with lung nodules, and all 10 test positively with the new automated system. Recalling the correct definition of a *p* value [1] (it is the probability of obtaining a result as extreme as or more extreme than the result observed, given that the null hypothesis is exactly correct), how would we calculate the *p* value in this case? For our example of the new automated technique, the definition implies that we need to calculate the probability of obtaining 10 (or more, but in this case more than 10 is impossible) successful

lung nodule detections in the 10 patients to whom the technique was applied, given that the true rate of success is exactly 90%. Recall [2] that if $x$ follows a binomial distribution with probability of success $p$, then $Pr$ ($x$ successes in $n$ trials) = $[n!/(x!(n–x)!)]p^x(1-p)^{n-x}$, where $x!$ is read as "$x$ factorial" and is equal to $x(x-1)$ $(x-2)\ldots$ (2) (1). For example, 5! = (5) (4) (3) (2) (1) = 120, and by convention 0! = 1. Using this binomial probability function, we can calculate the probability of 10 successes in a row with $p$ = probability of success = 0.9 as shown in equation 1:

$$\frac{10!}{10!0!}0.9^{10}(1-0.9)^0 = 0.9^{10} = 0.3487 \quad (1)$$

So there is about a 34.9% chance of obtaining results as extreme as or more extreme than the 10 of 10 results observed, if the true rate for the new technique is exactly 90%. Therefore, the observed result is not unusual, and hence compatible with the null hypothesis, so we cannot reject $H_0$.

This calculation could be done exactly, because the sample size was quite small. For larger sample sizes, the normal approximation to the binomial distribution [2] could be used. Also, this test was one-sided, but two-sided hypotheses are also of interest. For example, suppose we wish to test a similar null hypothesis as above ($H_0$: $p$ = 0.9) but against a two-sided alternative ($H_A$: $p \neq 0.9$). Suppose we observed 98 successes in 100 trials. Because our test is two-sided, according to the definition of a $p$ value we need to calculate the probability of obtaining data as extreme as or more extreme than the observed 98 of 100. Now, 98 is 8 higher than the 90 expected under the null hypothesis, so that to be as extreme as or more extreme than the 98 observed, we need to be 8 or more above or below the expected 90. That is, we need to calculate the probability of 98, 99, or 100 successes on one side, and 82, 81, 80, …, 2, 1, 0 on the other side. This lengthy calculation, involving the sum of 85 binomial calculations, can be well approximated by using the normal approximation to the binomial distribution [2]. Let our estimate of the unknown proportion be $\hat{p}$ = 98 / 100 = 0.98. We can calculate equations 2–4:

$$z = \left|\frac{\hat{p}-0.9}{\sqrt{\frac{p(1-p)}{n}}}\right| \quad (2)$$

$$= \frac{.98-0.9}{\sqrt{.9(1-.9)/100}} \quad (3)$$

$$= 2.67 \quad (4)$$

Looking up 2.67 on normal tables, we find 0.004, and doubling this value gives us our two-sided $p$ value, which is 0.008. It is unlikely that rates of 98% or more extreme will be observed in 100 trials if the true rate is in fact only 90%. Therefore, in this case, sufficient evidence exists to reject the null hypothesis in favor of the alternative.

Although $p$ values are still often found in the literature, several major problems are associated with their use, as we have previously discussed [1]. Briefly, the null hypothesis is virtually never exactly true (is it possible that the true underlying sensitivity is exactly 90%, as opposed to, say, 89.9999% or 90.0001%?), so we know it should be rejected regardless of the data we observe. Furthermore, the $p$ value says nothing about the effect size, which is crucial to clinical decision making, with large sizes usually implying a more clinically important effect than small sizes. A much more interesting question is to estimate the rate or proportion of interest, together with a measure of the accuracy of the estimate. CIs are one answer to this question, and we discuss them next. The Bayesian solution—credible intervals—is discussed later.

### *Confidence Intervals for Single Proportions*

Continuing the previous example, we have observed rates of 100% (10/10 in our smaller sample) or 98% (98/100 in our larger sample), but we know that these are estimates only, not guaranteed (in fact, unlikely) to exactly equal the true rates. On the basis of these data, however, what can we say about what we would expect the true rate to be?

One way to answer this question is with a CI. CIs usually have the form

$$\text{estimate} \pm k \times \text{standard error}$$

where the estimate and SE are calculated from the data, and where $k$ is a constant dependent on the width of the CI desired. The value of $k$ is usually near 2 (e.g., $k$ is 1.96 for a 95% CI).

If one observes $x$ = 98 positive tests in $n$ = 100 subjects known to have lung nodules, a point estimate of the success rate is $\hat{p} = x / n = 0.98$ or 98%. We use the notation $\hat{p}$ rather than $p$ to indicate that this is an estimated rate, not necessarily equal to the true rate, which we denote by $p$. Following this generic formula, a CI for a binomial probability of success parameter is given by the formula in equation 5,

$$\left(\hat{p} - z \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}},\ \hat{p} + z \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}\right) \quad (5)$$

where $z$ is derived from normal tables, and is given by $z$ = 1.96 for the usual 95% CI ($z$ = 1.64 for a 90% CI and $z$ = 2.56 for a 99% CI). Therefore, the 95% CI in our example is calculated as shown in equation 6,

$$\left(0.98 - 1.96 \times \sqrt{\frac{0.98\times0.02}{100}}, 0.98 + 1.96 \times \sqrt{\frac{0.98\times0.02}{100}}\right) \quad (6)$$

which here gives (0.930–0.994).

Technical note: This formula uses the normal approximation to the binomial distribution [2]. Exact formulae are also available [4], which are especially useful for small sample sizes or for estimates $\hat{p}$ near 0 or 1. For example, using an exact approach to this CI yields (0.930–0.998), which is very close to but not identical to that given by the indicated normal approximated interval. In addition, when $\hat{p}$ equals 0 or 1 exactly, the normal approximation breaks down, because the variance is estimated to be 0. Here one has no choice but to use a different procedure. The exact method yields a wider 95% CI of (0.741–1.000) in the case of our smaller data set, where 10 positive values were found in 10 subjects. There is also an easy-to-use and reasonably accurate rule of thumb when calculating a binomial CI and one observes 0 events. The rule is this: If you observe $n$ patients, and none of these patients have an event, then a 95% CI for the probability of the event goes from 0 to 3 / $n$. For example, if you observe 0 events in 10 binomial trials, then an approximate 95% CI would go from 0 to 3 / 10 = 0.3. By symmetry, the rule would say that if you observe only events in $n$ trials, then the 95% CI would go from (1 – 3 / $n$) to 1. For example, if you observe 10 events in 10 trials, then the 95% CI would go from 0.7 to 1, which is reasonably close to the exact solution of (0.741–1.000) given here.

How does one interpret this CI? Recall from the previous module [1] that the 95% confidence value (often called the confidence coefficient) is a long-run probability over repeated uses of the CI procedure. In practice, there are five different interpretations associated with CIs, depending on where the upper and lower CI limits fall with respect to clinical cut points of interest (see Fig. 2 of Joseph and Reinhold [1]). The formula displayed in equation 5 of this article provides a procedure that, when used repeatedly across different problems, will capture the true value of $p$ 95% of the time and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any single application, of course, the interval either does or does not contain the true proportion $p$.

For our smaller data set, with 10 subjects found to be positive in 10 trials, the 95% CI ranges from 74.1% to 100%, providing a large and inconclusive interval, because it may well be better or worse than the standard diagnosis, which is assumed to be successful 90% of the time. In our larger data set, the 95% CI ranged from 93.0% to 99.4%, so we can be quite certain that it is better than standard diagnoses. However, it can be as little as 3% better (90% compared with the lower CI

**TABLE 1 Data from a Two-Group Study**

| Diagostic Method | Test Positive | Test Negative | Total |
|---|---|---|---|
| Automated system | 285 | 15 | 300 |
| Standard diagnosis | 265 | 45 | 310 |
| Total | 550 | 60 | 610 |

**TABLE 2 Expected Data for the Example in Table 1 Under the Null Hypothesis**

| Diagnostic Method | Test Positive | Test Negative | Total |
|---|---|---|---|
| Automated system | 270.49 | 29.51 | 300 |
| Standard diagnosis | 279.51 | 30.49 | 310 |
| Total | 550 | 60 | 610 |

limit of 93%). Whether this is enough evidence to switch to the new automated system or not depends on clinical judgment. This in turn depends on many factors, including the cost and availability of the new automated system and the average clinical benefits that will accrue to those diagnosed earlier by the more sensitive diagnostic method.

### Inferences for Two or More Independent Proportions

Let us continue with our example comparing the diagnostic properties of a new automated system for the detection of lung nodules on chest radiographs compared with standard detection via non-computer-aided methods. Earlier we assumed that the rate in the standard diagnosis group was exactly known before the study, but this is somewhat unrealistic. We will now relax this assumption, and consider the data from the two-group study shown in Table 1 (presented in the form of a 2 × 2 table of data because we have two possible outcomes in each of the two groups being compared).

Again, we assume that all 610 subjects studied are truly positive, so that one would like to draw inferences about whether the automated system has increased sensitivity compared with the usual diagnosis group. Although one observes $\hat{p}_1$ = [285 / 300] = 0.95 sensitivity for the automated system compared with $\hat{p}_2$ = [265 / 310] = 0.855 sensitivity using standard diagnosis, for a 9.5% observed difference, a CI will provide us with a range of values compatible with the data that will help draw a better conclusion than simply looking at the observed point estimates. To calculate a CI for this difference in proportions, we can use the formula in equation 7,

$$\left(\hat{p}_1 - \hat{p}_2 - z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \hat{p}_1 - \hat{p}_2 + z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) \quad (7)$$

which extends equation 5 to the case of two proportions. In this formula, $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions in the two groups out of sample sizes $n_1$ and $n_2$, respectively, and $z$ is the relevant percentile from normal tables, chosen according to the desired level of the CI. For example, for a 95% CI $z$ = 1.96, for a 90% interval $z$ = 1.64, and so on. Using this formula for the diagnosis data given, one finds that a 95% CI for the difference in sensitivity is (0.049–0.141). This interval suggests that the automated system is indeed better, likely by at least as much as 0.049. Unless cost is a prohibitive factor, from these data it looks like the automated system is worthwhile (at least in these hypothetical data).

Although CIs are preferred for reasons we have briefly discussed here and which were more extensively discussed in a previous module in this series [1], we will also discuss hypothesis testing for proportions, because one often sees such tests in the literature. Suppose we wish to test the null hypothesis that $p_1 = p_2$—that is, the null hypothesis states that the success rates are identical in the two units. Because we hypothesize $p_1 = p_2$, we expect to observe, on average, the data in Table 2.

Why do we expect to observe this table of data if the null hypothesis is true? We have observed a total of 550 "successes" divided among the two groups. If $p_1 = p_2$ and if the sample sizes were equal in the two groups, we would have expected (550 / 2) = 275 successes in each group. However, because the sample sizes are not equal, we expect 550 × (300 / 610) = 270.49 to go to the automated system group, and 550 × (310 / 610) = 279.51 to go the standard diagnosis group. Similarly, expected values for the 60 negatively testing patients can be calculated. Observed discrepancies from these expected values are evidence against the null hypothesis. To perform a chi-square test, we now calculate as shown in equations 8–10:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (8)$$

$$= \frac{(285 - 270.49)^2}{270.49} + \frac{(15 - 29.51)^2}{29.51} + \frac{(265 - 279.51)^2}{279.51} + \frac{(45 - 30.49)^2}{30.49} \quad (9)$$

$$= 15.57. \quad (10)$$

Comparing the $\chi^2$ = 15.57 value on chi-square tables with 1 degree of freedom (*df*) (see Armitage and Berry [4] or almost any basic textbook on statistics to find such tables), we find that $p \approx 0.0001$ so that we have strong evidence to reject the null hypothesis. This coincides with our conclusion from the CI, but note that the CI is more informative than simply looking at the $p$ value from the chi-square test, because a range for the difference in sensitivities is provided by the CI. Thus, the clinical importance of any differences can be more easily evaluated.

The chi-square test can be extended to include tables larger than the so-called 2 × 2 table of this example. For instance, a 3 × 2 table could arise if, rather than classifying patients as positive or negative, we included a third outcome category, such as "chest radiograph is inconclusive." A 3 × 2 table could also arise if we considered comparing a third method of diagnosis rather than the two considered here. In these cases we would sum over 3 × 2 = 6 terms rather than the four terms of a 2 × 2 table. Although for 2 × 2 tables the *df* is always equal to 1, in general the *df* for chi-square tests is given by $(r - 1) \times (c - 1)$, where the number of rows in the table is $r$ and the number of columns is $c$. Thus, in the case of a 3 × 2 table, we would have (3 – 1) × (2 –1) = 2 *df*. In general, cases with arbitrary numbers of rows and columns can be constructed and analyzed using the chi-square test.

In order for the chi-square test to be valid, one needs to ensure that the expected value for each cell in the table is at least 5. This was satisfied in the previous example, in which our smallest expected table value was 29.51, much larger than 5. Fisher's exact test [4] is often used if this criterion is not satisfied for a particular table. The Fisher's exact test is valid for tables of any size, in particular for small sample sizes.

**TABLE 3 Generic Setup of a 2 × 2 Table**

| Second Test | First Test | | Total |
|---|---|---|---|
| | Positive | Negative | |
| Positive | *a* | *b* | *a* + *b* |
| Negative | *c* | *d* | *c* + *d* |
| Total | *a* + *c* | *b* + *d* | *N* = *a* + *b* + *c* + *d* |

### Inferences for Dependent Proportions

A two-group clinical trial, where $n_1$ subjects receive treatment A and $n_2$ different subjects receive treatment B, usually results in independent samples. That is, the results under treatment regimen A (number of successful outcomes among the $n_1$ subjects given treatment A) do not depend on the outcomes in group B (number of successful outcomes among the $n_2$ subjects given treatment B).

Sometimes, however, subjects or data points may come in pairs, so that dependencies among the groups are naturally induced. Consider, for example, the frequently occurring situation in which two diagnostic tests are given to each of a series of subjects. Each subject may test positively or negatively on each of the two tests, so that the data arising from such a study may be summarized in a 2 × 2 table, as seen in Table 3.

Thus, we observe *a* number of subjects who are positive on both tests, *b* subjects who are negative on the first test but positive on the second test, *c* subjects who are positive on the first test but negative on the second, and *d* subjects who test negatively on both tests. The cells with *a* and *d* contain concordant pairs, because the two test results agree with each other, whereas the cells with *b* and *c* contain discordant pairs.

Similar data can arise from a matched case–control study. In this type of study design, cases (e.g., those with a particular disease) are first found and then matched to a particular control case with similar characteristics but without the condition of interest.

As a concrete example, suppose we wish to investigate whether impaired renal function is related to diminished renal size. Because we would otherwise require large numbers of subjects to be followed up over a long period of time, a case–control design may be considered. Thus, one finds patients with impaired renal function and control subjects without impaired renal function, and discovers whether there is a tendency of those with impaired renal function to show diminished renal size on sonography compared with those without impaired renal function. Of course, patients with impaired renal function may tend to be different from subjects without (control subjects) in many ways, so to minimize possible confounding one may want to control for age, sex, height, hypertension, diabetes, and so on. For each patient, one may want to find a control subject with similar age, sex, height, and other characteristics, thus forming a series of matched pairs. Within each of these pairs, one then classifies each patient and control subject into whether they have diminished renal size at sonography or not.

Within each matched pair are four possibilities: Both the patients and control subjects show diminished renal size, or both may not show diminished renal size. These two possibilities form concordant pairs (introduced in previous text) because similar renal size is shown for each subject forming the pair. Of course, the other two possibilities are that the patient shows diminished renal size and the control does not, and vice versa, forming the nonconcordant pairs. As was the case with diagnostic test studies, the data may be formed into a 2 × 2 table, as shown in Table 4.

Note that there are a total of *N* pairs of subjects in this study, meaning that we in fact have 2*N* individuals (similarly, in the diagnostic test case, we have 2*N* tests, but only *N* subjects). We have *a* subjects in whom both the patient and the matched control subject showed diminished renal size, *b* subjects in whom the control but not the patient showed diminished renal size, and so on.

Suppose we would like to test the null hypothesis that diminished renal size is unrelated to impaired renal function versus the alternative hypothesis that a relation exists between diminished renal size and impaired renal function. The McNemar test focuses on the discordant pairs, represented in Table 4 by *b* and *c*. We can formulate the statistic shown in equation 11,

$$X^2 = \frac{(|b-c|-1)^2}{b+c} \qquad (11)$$

which approximately follows a chi-square distribution with 1 *df*. Thus, a *p* value can be calculated for this test.

For example, suppose we observe the following data: $a = 200$, $b = 100$, $c = 75$, and $d = 300$. According to the McNemar test, we calculate as shown in equation 12:

$$X^2 = \frac{(|100-75|-1)^2}{175} = 3.29 \qquad (12)$$

Looking up 3.29 on chi-square tables yields a *p* value of 0.069, so that it is close to but does not cross the (admittedly arbitrary) threshold of 0.05. Thus, at least at the type 1 error level of 0.05, we do not have evidence to reject the null hypothesis.

Of course, the McNemar test can also be used for testing hypotheses relating to diagnostic test data of the type described at the beginning of this section.

The general criticisms relating to hypothesis testing and *p* values carry over the particular case of testing dependent proportions through the McNemar test. Odds ratios and associated CIs can be calculated from matched pair studies, and these will be covered in a future module in this series.

### Sample Size Determination for One and Two Proportions

As previously discussed [1], there has been a strong trend away from hypothesis testing and *p* values toward the use of CIs in the reporting of results from biomedical research. Because the design phase of a study should synchronize with the analysis that will be eventually performed, sample size calculations should be performed on the basis of ensuring adequate numbers for

**TABLE 4 Data in a Case Control Study**

| Diminished Renal Size | Diminished Renal Size | | |
|---|---|---|---|
| | Patient Has | Patient Does Not Have | Total |
| Control has | *a* | *b* | *a* + *b* |
| Control does not have | *c* | *d* | *c* + *d* |
| Total | *a* + *c* | *b* + *d* | *N* = *a* + *b* + *c* + *d* |

accurate estimation of important quantities that will be estimated in the study, rather than by power calculations. For one- and two-sample problems, the formulae are as given in the following paragraphs.

### *Single Sample*

Let $p$ be the proportion that is to be estimated, and assume that we wish to estimate $p$ to an accuracy of a total CI width of $w = 2 \times h$, where $h$ is half the total CI width.

Then we can perform the calculation shown in equation 13,

$$n = \frac{z^2}{h^2}p(1-p) = \frac{4z^2}{w^2}p(1-p), \quad (13)$$

where, again, $z$ is the appropriate normal quantile (e.g., $z = 1.96$ for a 95% CI).

### *Two Sample*

Let $p_1$ and $p_2$ be the two proportions whose difference we would like to estimate to a total CI width of $w = 2 \times h$.

Then we can perform the calculation shown in equation 14,

$$n = \frac{4 \times (p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{w^2} = \frac{(p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{h^2} \quad (14)$$

where $n$ represents the required sample size for each group.

As an example, suppose we want to design a study to measure the difference in diagnostic accuracy for two types of imaging techniques, say MRI versus CT for staging cervical carcinoma. Suppose that CT is thought to be successful in staging patients with cervical carcinoma, with probability $p_1 = 0.70$, and MRI may improve this to $p_2 = 0.80$. We would like to estimate the true difference to within $h = 0.05$, so that not only will we be able to detect any differences of 10%, but the 95% CI will be far enough away from 0 (if our predicted rates are correct) so that we can make a more definitive conclusion as to the clinical usefulness of MRI. We calculate as shown in equations 15 and 16

$$n = \frac{(p_1 \times (1-p_1) + p_2 \times (1-p_2)) \times z^2}{h^2} \quad (15)$$

$$= \frac{(0.7 \times (1-0.7) + 0.8 \times (1-0.8)) \times 1.96^2}{0.05^2} = 569 \quad (16)$$

so that 569 patients are required in each group.

The main practical difficulty with equations 13 and 14 is assigning appropriate values for $p$, $p_1$, and $p_2$. It is therefore useful to note that equation 13 is maximized when $p = 0.5$, so using this value is conservative in the sense that the desired CI width will be respected regardless of the estimated value of $p$ that will be observed in the study. This conservative value, however, may provide too large a sample size and therefore be wasteful of resources if the true proportion is far from 0.5. A conservative rule of thumb is to use the value of $p$ that is closest to 0.5, selected from the set of all plausible values. Similarly, equation 14 is maximized for $p_1 = p_2 = 0.5$, so a similar rule of thumb applies for each of $p_1$ and $p_2$.

## Bayesian Inference for Proportions

Consider again the problem introduced in the section called Inferences for Single Proportions. Recall that in that example the sensitivity of standard interpretation of radiographs is assumed to be 90%, whereas the small data set collected so far for the new automated radiograph interpretation system indicates a 100% success rate but is based on only 10 subjects. The frequentist CI was very wide, ranging from 74.1% to 100%. Therefore, the data themselves have not been particularly helpful in making a decision as to which technique to use for the next patient, because values indicating a new test that is both more and less sensitive than the standard diagnostic method have not been ruled out by the CI. At this point, with the data being relatively uninformative, the radiologist may decide to be conservative and remain with the standard method until more information becomes available about the new automated technique, or may go with his or her "gut feeling" as to the likelihood that the new therapy is truly better or not better. If there have been data from animal experiments or strong theoretic reasons why the new technique may be better, the radiologist may be tempted to try the new one. Can anything be done to aid in this decision-making process?

Bayesian analysis has several advantages over standard or frequentist statistical analyses. These advantages include the following:

First is the ability to address questions of direct clinical interest, such as direct probability statements about hypotheses of interest and credible intervals with similarly easy interpretations [1]. Hence, results of Bayesian analyses are straightforward to interpret, in contrast to the obscure and difficult-to-understand (and frequently misinterpreted) inferences provided by $p$ values and CIs [1].

Second is the ability to incorporate relevant information not directly contained in the data into any statistical analysis. This enters in the form of prior information about parameters of interest.

The third advantage is that Bayesian analysis is a natural way to update statistical analyses as new information becomes available.

A main theoretic difference between frequentist and Bayesian statistical analyses is that Bayesian analysis permits parameters of interest (binomial probabilities, population means, and so on) to be considered as random quantities, so that probabilities can be attached to the possible values that they may attain. On the other hand, frequentists consider these parameters to be fixed (albeit possibly unknown) constants, so they have no choice but to attach their probabilities to the data that could arise from the experiment, rather than to the parameters. This distinction is the main reason Bayesian analysis can answer direct questions of interest, whereas frequentist analyses must settle for answering more obscure questions in the form of $p$ values and CIs.

The ability to address questions of direct interest, however, comes at the cost of having to do a bit more work. Not only do Bayesians have to collect data from their experiments, but they also have to quantify the state of knowledge of all parameters before their collecting this data. This nontrivial step is summarized in a prior distribution. The information in the prior distribution is updated by the information in the data to arrive at a posterior distribution, which summarizes all available information, past and current. We will apply a Bayesian analysis to our radiologist's decision later in this section, but first we need to recall the basic elements of all Bayesian analyses and see how they are applied to drawing inferences about our parameter of interest here, the binomial success rate of the new automated radiographic technique.

Let us generically denote our parameter of interest as $\theta$. Hence, $\theta$ can be a binomial parameter, a set of two independent or dependent binomial parameters, or the mean and variance from a normal distribution, or an odds ratio, or a set of regression coefficients, and so on. Note in particular that $\theta$ can be two- or more dimensional. The parameter of interest is sometimes usefully thought of as the "true state of nature." As discussed in more detail in the previous module in this series [1], the basic elements of a Bayesian analysis then are as follows:

First is the prior probability distribution, $f(\theta)$. This subjective prior distribution summarizes what is known about $\theta$ before the experiment is performed.

Second is the likelihood function, $f(x \mid \theta)$. The likelihood function provides the distribution of the data, $x$, given the parameter value $\theta$.

For instance, for proportions it may be a binomial likelihood, as in equation 17:

$$l(x|p) = Pr\{x \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-x)!\,x!}p^x(1-p)^{(N-x)} \quad (17)$$

Third is the posterior distribution, $f(\theta \mid x)$. The posterior distribution summarizes the information in the data, $x$, together with the information in the prior distribution, $f(\theta)$. Thus, it summarizes what is known about the parameter of interest $\theta$ after the data are collected.

Bayes' theorem relates the above three quantities:

posterior distribution =
[likelihood of the data × prior distribution] /
a normalizing constant,

or using our notation and omitting the normalizing constant, as shown in equation 18,

$$f(\theta|x) \propto f(x|\theta) \times f(\theta) \quad (18)$$

where $\propto$ indicates "is proportional to."

Thus, we update the prior distribution to a posterior distribution after seeing the data via Bayes' theorem. The current posterior distribution can be used as a prior distribution for the next study; hence, Bayesian inference provides a natural way to represent the learning that occurs as science progresses.

The prior distribution is subjective and chosen by each investigator according to his or her appreciation of the past literature regarding the unknown parameters of interest. Hence, the prior distribution is not unique to each experiment but can vary from investigator to investigator. This can be seen as accurately reflecting clinical reality. Different clinicians can have different initial opinions about a parameter value, although these opinions tend to concentrate about a constantly narrowing range of values as more data accumulate. This is how Bayes' theorem operates, because the prior becomes a less important contributor to the posterior distribution as more data become available. See the previous module for more discussion about prior distributions [1].

We now will apply the general Bayesian technique we have described to the specific problem of inferences for binomial proportions.

Suppose that in a given experiment $x$ "successes" are observed in $N$ binomial trials. Let $\theta = p$ denote the parameter of interest—the true but unknown probability of success—and suppose that the problem is to find an interval that covers the most likely locations for $p$ given the data.

The Bayesian solution to this problem follows the usual pattern, as outlined previously. Hence, the main steps can be summarized as first, write down the likelihood function for the data. Second, write down the prior distribution

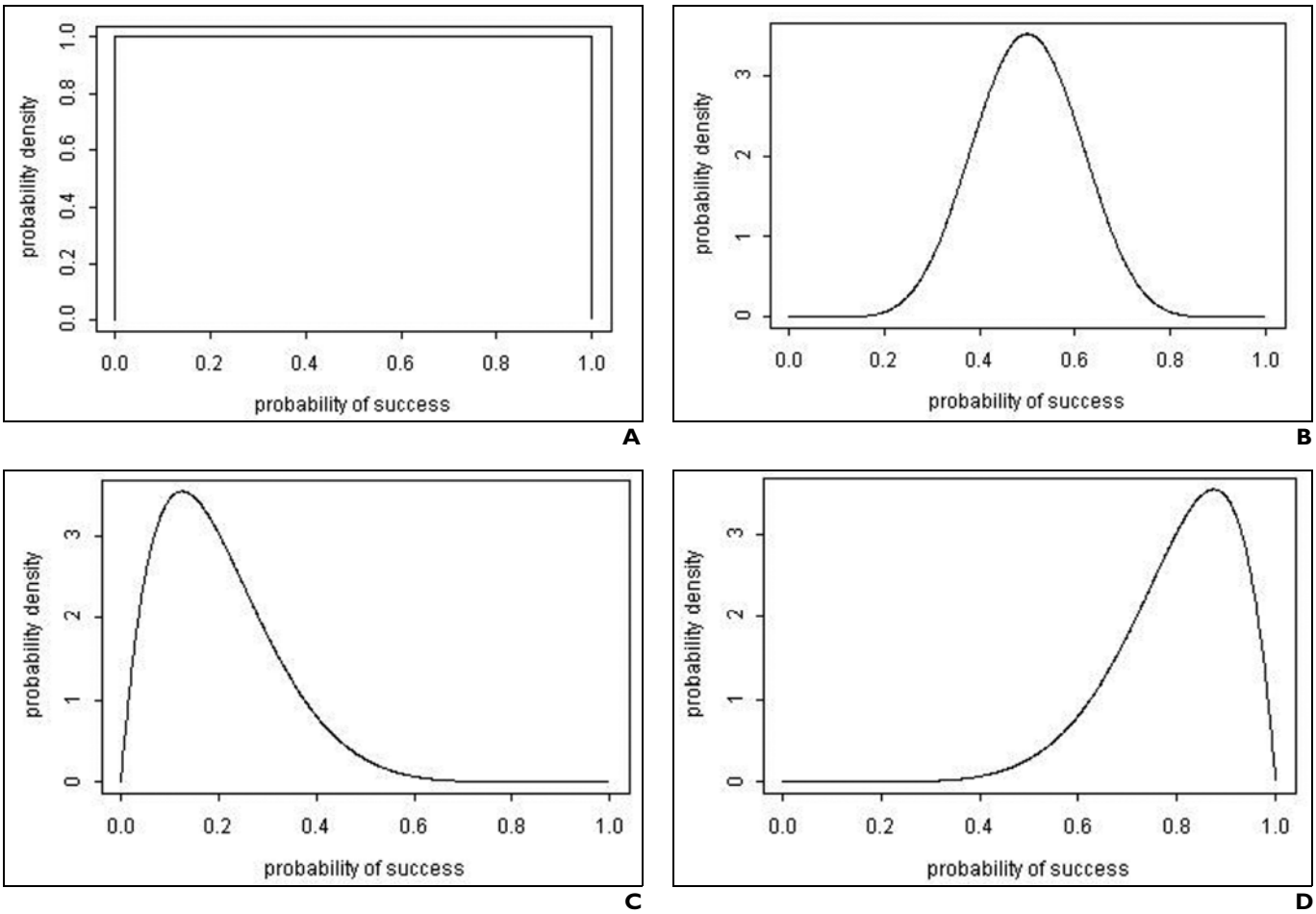


**Fig. 1.**—Series of four beta densities.
**A–D,** Graphs show beta(1,1) (**A**), beta(10,10) (**B**), beta(2,8) (**C**), and beta(8,2) (**D**) densities. Beta(1,1) distribution (**A**) is also known as the uniform density.

for the unknown parameter $p$. Third, use Bayes' theorem (i.e., multiply the equation for the likelihood function of the data by the prior distribution) to derive the posterior distribution. Use this posterior distribution, or summaries of it like 95% credible intervals, for statistical inferences. Credible intervals are the Bayesian analogues to frequentist CIs.

For the case of a single binomial parameter, these steps are realized in this manner:

*Step 1*

The likelihood function is the usual binomial probability formula shown in equation 17, where $l(x \mid p)$ represents the likelihood function for the success rate $p$ given data $x$.

*Step 2*

Although any prior distribution can be used, two distributions are of particular interest. The first prior distribution we will discuss is the uniform prior distribution, which specifies that all possible values (for proportions, this implies all values in the range of 0–0) are equally probable, a priori. See Figure 1A. The uniform distribution is suitable for use as a "diffuse" or a "noninformative" distribution, when little or no prior information is available or when one wishes to see the information contained in the data by itself.

A second particularly convenient prior distribution, for reasons to be explained, is the beta distribution. A random variable, $\theta$, has a distribution that belongs to the beta family if it has a probability density given by equation 19

$$f(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (19)$$

for $0 \le \theta \le 1$, and $\alpha, \beta > 0$. $B(\alpha,\beta)$ represents the beta function evaluated at $(\alpha,\beta)$. It is simply the normalizing constant that is necessary to make the total area under the curve equal to 1, but otherwise plays no role.

Some beta distributions are illustrated in Figure 1. For example, using a beta($\alpha = 1$, $\beta = 1$) distribution reproduces the perfectly flat or uniform distribution discussed previously. Thus, the uniform distribution is really just a special case of the beta distribution. On the other hand, a beta($\alpha = 10$, $\beta = 10$) density produces a curve similar in shape to a normal density centered at $\theta = 0.5$. If $\alpha > \beta$ the curve is skewed toward values near 1, whereas if $\alpha < \beta$ the curve is skewed toward values near 0.

The mean of the beta distribution is given by equation 20,

$$\mu = \frac{\alpha}{\alpha+\beta}, \quad (20)$$

and the SD is given by equation 21.

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad (21)$$

To choose a prior distribution, one needs only to specify values for $\alpha$ and $\beta$. This can be done by finding the $\alpha$ and $\beta$ values that give the correct prior mean and SD values. Solving these two equations in two unknowns, the formulae are shown in equations 22 and 23.

$$\alpha = -\frac{\mu\,(\sigma^2+\mu^2-\mu)}{\sigma^2} \quad (22)$$

$$\beta = \frac{(\mu-1)\,(\sigma^2+\mu^2-\mu)}{\sigma^2} \quad (23)$$

For example, if we wish to find a member of the beta family centered near $\mu = 0.9$ and with $\sigma = 0.05$, then plugging these values for $\mu$ and $\sigma$ into these two equations gives $\alpha = 31.5$ and $\beta = 3.5$, so that a beta(31.5, 3.5) will have the desired properties. This curve, pictured in Figure 2, may be an appropriate prior distribution for the problem introduced at the beginning of this section if the radiologist believes, a priori, that the new technique is likely to be successful between 80% and 100% of the time, and whose best guess of the rate is 90%. Note that this clinician has centered the prior around the rate thought to be equal to the standard treatment. Thus, this prior distribution would give equal a priori weight to both the null and alternative hypotheses given at the start of the section on Inferences for Single Proportions. We will return to this example again shortly.

*Step 3*

As always, Bayes' theorem says

posterior distribution ∝ prior distribution × likelihood function.

In this case, it can be shown (by relatively simple algebra) that if the prior distribution is beta($\alpha,\beta$) and the data are $x$ successes in $N$ trials, then the posterior distribution is again a beta distribution, beta($\alpha + x$, $\beta + N - x$). This simplicity arises from noticing that both the beta prior distribution as represented in equation 19 and the binomial likelihood as given in equation 17 have the general form $p^a \times (1-p)^b$, so that when multiplying them as required by Bayes' theorem, the exponents simply add, and the form is once again recognized to be from the beta family of distributions.

Hence, if we observe the new automated computer-aided radiologic method to correctly identify 10 patients in a row with lung nodules, and if we use the prior distribution discussed previously, then the posterior distribution is a beta(31.5 + 10, 3.5 + 0) = beta(41.5, 3.5) distribution, which is illustrated in Figure 2. The mean of this distribution is [41.5 / (41.5 + 3.5)] = 0.922, and the 95% posterior credible interval is (0.844–0.988). The

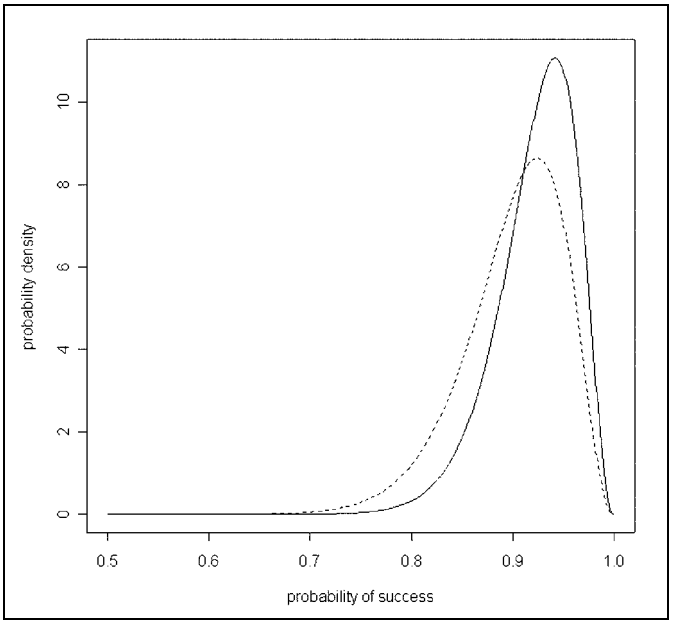


**Fig. 2.**—Prior (*dotted line*) and posterior (*solid line*) beta densities for automated radiology example.

probability of being greater than 90% is 0.748 (area under the curve to the right of 0.9 in Fig. 2). Therefore, the radiologist may or may not be tempted to try the automated technique on the next patient but should realize that this decision is mostly based on the prior information, to which the data contributed only a small amount of new information. Looking at Figure 2, we see that the prior density was shifted only a small amount by the data. If instead the radiologist "lets the data speak for themselves" by using a beta(1,1) or uniform prior distribution (Fig. 1), then the 95% interval is (0.773–0.971), very similar numerically to the frequentist CI of the section Inferences for Single Proportions, although their interpretations are quite different. Bayesian intervals (deliberately called credible intervals to distinguish them from frequentist confidence intervals) are interpreted directly as the posterior probability that $p$ is in the interval, given the data and the prior distribution. No references to long-run frequencies or other experiments are required, as is the case for CIs.

In general, one should usually perform a Bayesian analysis using a diffuse prior distribution like a beta(1,1) distribution, to examine what information the current data set provides. Then one or more Bayesian analyses with more informative prior distributions could be performed, depending on the available prior information. If opinions in the medical community are widely divergent concerning the parameters of interest, then several prior distributions should be used. If the data set is large, then similar conclusions will be reached no matter which prior distribution one starts with. On the other hand, with smaller data sets, diversity of opinions will still exist, even after the new data are analyzed. Bayesian analysis allows this situation to be accurately represented and assessed.

Although we discuss only the simple case of Bayesian inference for a single binomial proportion, these methods are easily extended to the case of two or more proportions. For a clinical example using Bayesian analysis to compare two proportions, see Brophy and Joseph [5]. This example also illustrates the use of a range of prior distributions and shows that Bayesian analysis can often come up with answers that are quite different from those obtained using a frequentist approach.

## Discussion

This module has introduced some of the major ideas behind statistical inference for proportions, with emphasis on the simple methods for one and two samples. Rather than a simple catalogue listing of which methods to use for which types of dichotomous data, we have tried to explain the logic behind the common statistical procedures seen for binary data in the medical literature, the correct way to interpret the results, and what their advantages and drawbacks may be. We have also introduced Bayesian inference as a strong alternative to standard frequentist statistical methods, for both its ability to incorporate the available prior information into the analysis and its ability to address questions of direct clinical interest.

For more information about inferences on proportions, see the books by Fleiss [6] for the frequentist perspective and by Gelman et al. [7] for the Bayesian view. General books on statistical inferences in medicine [8–10] all contain many techniques on inferences for proportions that are beyond the scope of this module.

Software is available that makes carrying out all the analyses discussed in this module relatively easy. From the frequentist viewpoint, there are literally dozens of statistical packages available for purchase, but much excellent free software is also available. For example, the *R* package [11] is freely available for most computer platforms, including Windows (Microsoft) and Linux PCs and MacOS (Apple). It is a comprehensive package that is constantly being updated. Free Bayesian software includes First Bayes [12] for simple problems and WinBUGS [13, 14] for more complicated problems.

The previous module covered similar techniques to those covered here for continuous data, and future modules in this series will cover techniques suitable for other types of study designs and questions that arise in radiology, including linear and logistic regression methods. The latter is especially relevant because logistic regression allows one to analyze dichotomous outcomes from one or more groups while adjusting the analysis for potential confounding factors.

## References


1. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists. Statistical inferences for continuous variables. *AJR* 2005;184:1047–1056
2. Joseph L, Reinhold C. Fundamentals of clinical research for radiologists. Introduction to probability theory and sampling distributions. *AJR* 2003;180:917–923
3. Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. *AJR* 2004;182:505–510
4. Armitage P, Berry G. *Statistical methods in medical research*, 3rd ed. Oxford, England: Blackwell Scientific Publications, 1994
5. Brophy J, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* 1995;273:871–875
6. Fleiss J. *Statistical methods for rates and proportions.* New York, NY: Wiley, 1981
7. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian data analysis,* 2nd ed. London, England: Chapman and Hall, 2003
8. Rosner B. *Fundamentals of biostatistics*. Belmont, MA: Duxbury, 1995
9. Bland M. *An introduction to medical statistics,* 3rd ed. Oxford, England: Oxford University Press, 2000
10. Le C. *Introductory biostatistics.* New York, NY: Wiley, 2003
11. R, version 1.8.0. Available at: cran.r-project.org/. Accessed February 2, 2004
12. O'Hagan A. First Bayes software. Available at: www.shef.ac.uk/~st1ao/1b.html. Accessed December 25, 2003
13. Spiegelhalter D, Thomas A, Best N. *WinBUGS version 1.4 user manual*. Cambridge, UK: MRC Biostatistics Unit, 2003
14. WinBUGS, version 1.4. Available at: www.mrc-bsu.cam.ac.uk/bugs/. Accessed February 2, 2004


The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004
13. Clinical Evaluation of Diagnostic Tests, January 2005
14. ROC Analysis, February 2005
15. Statistical Inference for Continuous Variables, April 2005