Statistics for Nuclear Medicine
Part 3: A. Comparing Two Proportions (The Relative Deviate Test and Chi-Square Equivalent)
B. Counting Data

Peter C. O'Brien, Marc A. Shampo, and James S. Robertson

Mayo Clinic and Mayo Foundation, Rochester, Minnesota

J Nucl Med 24: 269–272, 1983

A. COMPARING TWO PROPORTIONS (THE RELATIVE DEVIATE TEST AND CHI-SQUARE EQUIVALENT)

The previous paper presented a method for comparing observations of two continuous variables. Such variables are called "continuous" because they can have a continuum of values; and the measurement of interest was the level of free thyroxine.

Formulation of the problem. We now consider how to compare dichotomous variables, which are observed as yes-no, alive-dead, normal-abnormal, and so on. For an example, let us compare the incidence (yes-no) of a side effect (headache) in association with each of two drugs: 15 of 50 cases with drug F and 8 of 50 cases with drug G.

Note that the dichotomous observations of each group can be summarized by a proportion, which will express the incidence within the group as a degree on a continuous scale of possibilities. Let π_F and π_G represent the proportions (true but unknown) of the incidence of headache associated with drugs F and G, respectively, in the population. For an estimate of π_F, we can use the sample proportion $p_F = 15/50 = 0.30$; and for π_G we can use the sample proportion $p_G = 8/50 = 0.16$.

Using these terms, we state the familiar questions: (1) Is there a real difference between these groups—that is, does $\pi_F = \pi_G$?—and (2) if so, how large may the difference be?

Received Sept. 30, 1982; revision accepted Dec. 3, 1982.

For reprints contact: Dr. O'Brien, Mayo Clinic and Mayo Foundation, 200 First Street SW, Rochester, MN 55905.

Question 1: Is there a difference? If $\pi_F = \pi_G$ (if the proportions π_F and π_G are the same), we can write this unknown common proportion as π_0. To obtain a corresponding sample statistic (p_0) in accord with the null hypothesis that there is no underlying difference between the samples (that the apparent difference is only random variation), we pool the samples:

$$p_0 = \frac{15 + 8}{50 + 50} = 0.23$$

This resulting value of 0.23 is an estimate of the common proportion assumed (for test purposes) to satisfy the hypothesis in question.

Again we compute the ratio (here we use the test statistic z) of the difference between the two data sets to the standard error of the difference (the variability within each data set as calculated with the sample statistic p_0). Still assuming that the null hypothesis is true (no underlying difference between the samples), we use the common proportion p_0 in the denominator for this calculation.

$$z = \frac{p_F - p_G}{\sqrt{p_0(1 - p_0)\left(\frac{1}{N_F} + \frac{1}{N_G}\right)}}$$

$$= \frac{0.30 - 0.16}{\sqrt{0.23(0.77)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$

$$= 1.663$$

In this example we will reject the null hypothesis ($\pi_F = \pi_G$) if either drug is found to cause fewer headaches than the other. (This differs from the interpretations in the two preceding sections. There we asked, a priori, "Is A superior to B?" Here we are asking, "Is either F or G superior to the other?") Hence we look for the probability of getting a value of z that is either 1.663 or higher (signifying more headaches with drug F) or -1.663 or lower (signifying more headaches with G). From appropriate tables, this probability $P = 0.096$; so we remain unsure that either drug excels the other in regard to incidence of headache.

Question 2: How large may the difference be? An approximate 95% confidence interval for $\pi_F - \pi_G$ can be calculated with this formula:

$$95\%\ \text{CI} = p_F - p_G \pm 1.96 \cdot \sqrt{\frac{p_F(1-p_F)}{N_F} + \frac{p_G(1-p_G)}{N_G}}$$

Note that because the confidence interval will contain values of π_F and π_G that are unequal, we can no longer use p_0 in our estimate of the standard error:

$$95\%\ \text{CI} = 0.14 \pm 1.96 \cdot \sqrt{\frac{0.30 \cdot 0.70}{50} + \frac{0.16 \cdot 0.84}{50}}$$
$$= 0.14 \pm 0.163$$

Thus the 95% confidence limits are -0.023 and $+0.303$.

Even though the P value (0.096) is greater than 0.05 and the 95% confidence interval for $\pi_F - \pi_G$ contains 0, we still might conclude that the data provide suggestive evidence of a superiority for drug G. Our large confidence interval (reflecting the somewhat small sample size) indicates that drug G may offer a substantial advantage despite the lack of statistical significance.

It is a convention that P values are to be considered significant only if they are less than 0.05, and some investigators require P values less than 0.01 for convincing evidence against the null hypothesis. However, the distinction between significant and nonsignificant test results depends on circumstances in the individual study, and often an intermediate interpretation is appropriate, as it is here. More generally, a P value should be interpreted as a measure of the strength of the evidence against the null hypothesis. Such strength can have many degrees, and it offers more meaning than "enough" and "not enough."

Comment. 1. An additional lesson is concealed in this example. Suppose the investigators had not thought carefully about the problem of associated headaches until they saw that more occurred with F than with G. They might have formulated a hypothesis that G was superior in this regard and tested it looking only for a difference in one direction. The outcome would have been a statistically significant superiority for drug G ($P = 0.048$).

What is the probability that this approach to hypothesis testing will lead to an erroneous conclusion? Let us suppose that there is no real difference between F and G. The probability of erroneously concluding that G is superior is 0.048. However, it is equally likely that the sample results would favor F by the same amount; and this also would give $P = 0.048$. Thus the probability for error is the probability of concluding G superior to F plus the probability of concluding F superior to G, which is $0.048 + 0.048 = 0.096$.

In general, how do we determine whether to look for differences in just one direction (a one-sided test) or in both directions (a two-sided test)? The answer is to formulate the hypothesis clearly, and *before* the data are collected. The way the hypothesis is stated will determine how the test should be done. For example, when we ask the question "Is experimental drug A superior to placebo?" We clearly are looking for a difference in only one direction. If the experimental drug is found to perform either the same as or worse than placebo, the same negative conclusion will be reached. Since it is not our goal to establish that A is worse than placebo, a one-sided test is appropriate. (Notice that this was the situation in our previous examples, where all our tests were one-sided.)

Conversely, when comparing two drugs (as in the present example), we may ask: "Is either drug superior to the other?" In this instance we clearly desire to determine whether superiority exists in either direction, so a two-sided test is appropriate.

The decision as to whether a test should be one-sided or two-sided illustrates a very important principle in statistics: the study objectives and specific hypotheses to be tested should be formulated before the data are collected.

2. Various computational formulas are available for performing the test described in this section. Since they all give the same P value, they are equivalent. The formula that is simplest computationally and is used most commonly is called the chi-square (χ^2) test. (The number actually computed is z^2.) Although we have presented the computations in terms of the relative deviate statistic in order to provide a better understanding of the test, in practice the tests for comparing two proportions are most commonly referred to as chi-square tests.

3. In our previous examples, the computed test statistic was usually denoted by the letter t. Although any letter could have been used, t ordinarily is chosen for those situations because it corresponds to the name of the statistical tables used in obtaining the related P values. For the tables used in the relative deviate test for comparing two proportions, the letter z is commonly used. When the test is based on simple computational formulas

(which yield the square of the relative deviate z), the test statistic is denoted by the symbol χ^2.

B. COUNTING DATA

Evaluation of a single counting measurement. Frequently in nuclear medicine, we are concerned with counting particles (or photons) emitted due to radioactive decay. If it can be assumed that the probability that an emitted particle will be detected by the counting system is constant during the time interval of interest, some special statistical techniques may be used in evaluating the data.

Specifically, under these circumstances it is appropriate to focus attention on the total number of counts observed (usually denoted by N) during some specified period of time (t). Like measurements in our previous examples, N will vary from one sample to the next. How much of this change is mere random variation? In our previous examples, we made use of repeated measurements in estimating the amount of random variation. However, relying on our assumption of equiprobable events occurring over time, we can now estimate variability based on the observed value of N. We use the formula $s = \sqrt{N}$. Although we are estimating the same quantity as before, the computational algorithm for obtaining this estimate is different from that used in the previous papers.

Usually, the number of counts that should be expected is an unknown constant, which we will denote by N*. If N is sufficiently large (greater than 20), we can set confidence limits for N* using the formula:

$$95\%\ \text{CI} = N \pm 2\sqrt{N}.$$

(For smaller values of N, one can use special tables.)

In some applications, interest focuses on the count rate, $R = N/t$. For this, the standard deviation for R is obtained from the formula $s_R = \sqrt{R/t}$, and the 95% confidence interval is given by:

$$95\%\ \text{CI} = R \pm 2\sqrt{R/t}.$$

Rather than counting for a fixed period of time, it is sometimes more convenient to count until a predetermined total count is reached. Although the rate still equals N/t, time then is the random variable and N is the predetermined constant. That difference requires alteration of the rest of the statistical approach; but we shall not pursue it further.

Comparing two counts. Suppose that a scan has two regions of equal area, but it is suspected that the emission rate should be higher in the second region. To test the null hypothesis that the observed difference reflects only randomness of the observed emissions, we compute the relative deviate.

To illustrate, suppose the counts observed were $N_1 = 4{,}225$ and $N_2 = 4{,}900$. The estimated standard deviations are $s_1 = 65$ and $s_2 = 70$, and the 95% confidence interval ranges from 4,095 to 4,355 in the first region and from 4,760 to 5,040 in the second. The lack of overlap of these two intervals suggests that the observed difference is statistically significant. To confirm this impression, we compute the test statistic:

$$z = \frac{N_2 - N_1}{\sqrt{N_1 + N_2}}$$

The corresponding P value is then obtained from special tables or computing equipment.

For the example data:

$$z = \frac{4{,}900 - 4{,}225}{\sqrt{4{,}900 + 4{,}225}}$$

$$= \frac{675}{95.52}$$

$$= 7.07$$

The corresponding P value, reflecting the probability that the observed difference would occur as the result of only random fluctuations in emissions, is less than 0.001. Thus we conclude that a real difference exists between regions.

Multiple counts. As we mentioned at the beginning of our discussion on "Counting Data," the validity of the statistical techniques that we have described depends critically on our equiprobable assumption—that detection (counting) of an emitted particle (or photon) remains equally probable throughout the period of observation. We can check this assumption by making a series of repeat measurements.

For example, suppose a series of 10 counts, all under the same conditions, produced the data in Table 1. Since we now have 10 measurements, we can estimate variability in two ways. The first is the method intro-

TABLE 1. HYPOTHETICAL DATA CORRESPONDING TO 10 REPEATED COUNTS

Count no.	Observed count
1	38
2	61
3	61
4	50
5	49
6	65
7	41
8	51
9	70
10	58
$\bar{x} = 54.4$	
$s = 10.4$	
$s^2/\bar{x} = 1.988$	

duced earlier—based on the observed deviations of the 10 counts from their average—providing a value of s = 10.4. The second method, based on the assumption of equiprobable counts in which $\sqrt{\bar{x}} = 7.4$, is also an estimation of variability. Thus, the ratio of these estimates, $s/\sqrt{\bar{x}}$, measures how much the observed variability exceeds the variability to be expected. In this case, the observed variability was 1.41 times the expected.

To test the null hypothesis that the equiprobable assumption is true, we compare $D = s^2/\bar{x}$ (called index of dispersion) with values found in special tables. In this case, we find that D = 1.988 and $P = 0.036$, indicating that the assumption may not be valid. The counting appears to be somewhat erratic.

BOOKS RECEIVED

Computed Body Tomography. J.K.T. Lee, S.S. Sagel, R.J. Stanley, Eds. New York, NY, Raven Press, 1982, 602 pp, $80.00

Noninvasive Techniques for Assessment of Atherosclerosis in Peripheral, Carotid, and Coronary Arteries. T.F. Budinger, A.S. Berson, I. Ringqvist, M.B. Mock, J.T. Watson, R.S. Powell, Eds. New York, NY, Raven Press, 1982, 271 pp, $49.00

Bone Metastasis. L. Weiss, H.A. Gilbert, Eds. Boston, MA, G.K. Hall & Co., 1981, 512 pp, $57.50

Cancer Therapy. D.S. Fischer, J.C. Marsh, Eds. Boston, MA, G.K. Hall & Co., 1982, 749 pp, $59.95

Digital Nuclear Medicine. J.J. Erickson, F.D. Rollo, Eds. Philadelphia, PA, J.B. Lippincott, 1982, 240 pp, $19.50

A Handbook of Nuclear Pharmacy. W.M. Hibbard. Springfield IL, Charles C. Thomas, Publisher, 1982, 66 pp, $18.75

Actualités en Radiodiagnostic: Nouvelles Technologies. M. Amiel, D. Doyon, H. Fischgold, R. Schmidt. Paris, Masson (available in U.S. through S.M.P.F. Corp., 485 Fifth Ave., NY 10017, 1982, 128 pp, $42.00

1001 Questions About Radiologic Technology Volume 2. R. Bell. Baltimore, MD, University Park Press, 1983, 175 pp, $9.95

Quality Control in Diagnostic Imaging. J.E. Gray, N.T. Winkler, J. Stears, E.D. Frank. Baltimore, MD, University Park Press, 1983, 249 pp, $34.95