6.2 The Concept of a Hypothesis Test

Even though you will not have gone through a formal exercise of stating a pair of hypotheses about the value of a parameter and structuring a formal decision rule for how to reject the null, the group, through its reaction to the unexpected experimental results, has, in fact, performed a basic hypothesis test. By reacting with surprise to the high percentage of zeros observed (identifying a result as highly unlikely, in the tail of the sampling distribution) and suggesting an alternative answer to the question of what the chance of a 0 is, they have gone through the equivalent of a test of the null hypothesis that the probability is one-tenth.

You can build on this, explaining that this exercise is typical of one of the major types of activities carried out in statistics, using the results in a sample as the basis for drawing some conclusions about the characteristics of a population of interest, and giving brief explanations of what hypothesis testing and point estimation are.

6.3 Philosophical Issues in Model Development

Finally, the experiment can be used as the basis for a philosophical discussion of the process of developing a model for a particular real-world situation. In particular, it can be used to illustrate the problems associated with one's perceptions, rather than reality, as the basis for a model and the need to revise models as information is collected and compared with the predictions of the original model. Then, by revealing the actual contents of the box, you can respond to the initial questions that you may have had to duck and illustrate the m/n approach to assessing probabilities.

7. SUMMARY

A variation on the standard lottery-type box-of-balls experiment has been presented. This experiment has shown itself to be useful, in my experience, in introducing various concepts associated with the assessment of probability, the ideas of point estimation and hypothesis testing, and some of the issues one needs to be aware of in developing a model for an unknown system.

[*Received September 1987.*]

The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations

RALPH B. D'AGOSTINO, WARREN CHASE, and ALBERT BELANGER*

For testing the equality of two independent binomial populations the Fisher exact test and the chi-squared test with Yates's continuity correction are often suggested for small and intermediate size samples. The use of these tests is inappropriate in that they are extremely conservative. In this article we demonstrate that, even for small samples, the uncorrected chi-squared test (i.e., the Pearson chi-squared test) and the two-independent-sample t test are robust in that their actual significance levels are usually close to or smaller than the nominal levels. We encourage the use of these latter two tests.

KEY WORDS: Fisher exact test; Ill-use of the continuity correction; Robustness of the t test; Small sample testing.

1. INTRODUCTION

A traditional solution for testing the equality of two independent binomial distributions is to use (a) the Pearson chi-squared test for large samples, (b) the chi-squared test with Yates's continuity correction for intermediate size samples, and (c) the Fisher exact test for small samples. The major statistical software packages (e.g., SAS, SPSS, and BMDP) have all incorporated this solution. Some even internally determine when the sample sizes are small and produce only the Fisher test result in those cases and suppress the output for the chi-squared test. All give dire warnings about small expected values, and the implication of the potential lack of validity of the chi-squared test is obvious.

Occasionally authors have questioned this solution (e.g., Berkson 1978; Conover 1974; Grizzle 1967; Kempthorne 1979; Upton 1982). In particular, Upton (1982), in a careful study of 22 alternative tests, concluded that "the exact test of Fisher, and the corresponding Yates correction to Pearson's test, give tests which are both extremely conservative and inappropriate" (p. 86). He noted that the uncorrected chi-squared test performs well and suggested that a preferred test is the scaled test $(N - 1)X^2/N$, where N is the combined sample size and X^2 is the Pearson chi-squared test for the 2×2 contingency table.

The results published by Upton and others apparently have not had a significant or widespread impact on statistical practitioners. The reasons are not clear. Possibly the word *exact* in Fisher's exact test creates an impression that exact

*Ralph B. D'Agostino is Chairman, Mathematics Department, Boston University, Boston, Massachusetts 02215. Warren Chase is Professor, Mathematics Department, Framingham State College, Framingham, Massachusetts 01701. Albert Belanger is Statistician, Statistics and Consulting Unit, Boston University, Boston, Massachusetts 02215.

implies best. For the problem at hand "exact" is not best. Or perhaps it is because the standard textbooks do not discuss the conservative nature or the poor power of Fisher's exact test and Yates's correction to the chi-squared test. Moreover, although the statistical software firms attempt to keep pace with developments in sophisticated graphical and multivariate procedures, they seem no longer to review the essential univariate procedures. On the other hand, maybe the works of Upton and others are not complete enough to satisfy and convince. For example, Upton only examined seven sample configurations for "small" samples. These were (7, 7), (8, 6), (9, 5), (10, 9), (11, 9), (12, 9), and (5, 10), where the values in parentheses are the sample sizes (m, n). These may not be considered enough to justify removal of the Fisher and Yates tests.

In this article we focus on the problem of comparing two independent binomial populations, where the sample sizes are small (m, $n \leq 15$). We confirm Upton's results that both the Fisher exact and Yates continuity correction chi-squared tests are extremely conservative and inappropriate. Further, we demonstrate that the uncorrected chi-squared test and the two-independent-sample t test with pooled variance are robust in that their actual levels of significance are, in most situations, close to or smaller than the nominal levels. In addition, their maximum actual levels are close to the nominal levels. The nominal level is the stated or a priori selected significance level.

2. THE UNDERLYING MODEL AND NOTATION

Say we have two independent dichotomous or binomial populations with parameters p_1 and p_2 for the first and second populations, respectively. A random sample is extracted from each. The usual manner of presenting these data in contingency-table form is as follows:

	Sample 1	Sample 2	
success	a	b	s
failure	c	d	f
totals	m	n	N

For Samples 1 and 2, respectively, m and n are the sample sizes, a and b are the numbers of successes, and c and d are the numbers of failures. N, s, and f are the total sample size, successes, and failures, respectively. The null hypothesis under consideration is $H_0: p_1 = p_2$. The alternative can be one-sided or two-sided. In the following we present results for the two-sided alternative. Results for the one-sided alternative follow usually from the symmetry of the sampling distribution under the null hypothesis.

The Fisher exact test, the Pearson chi-squared test (or, equivalently, the z test for the difference in proportions), and the chi-squared test with Yates's continuity correction are well known and will not be reproduced here. In the following we refer to these as Fisher, X^2, and X_c^2, respectively. For completeness refer to Mantel (1974) for a discussion of the correct procedure needed to use X_c^2 as a one-sided or two-sided test.

2.1 The Two-Independent-Sample t Test With Pooled Variance

This test can be written in several equivalent ways. One such way is

$$t = \left[\frac{N-2}{N}\right]^{1/2} \frac{ad - bc}{[nac + mbd]^{1/2}}. \tag{1}$$

An important point to make is that with the use of computer software, simplifications of the usual formulas are not needed. The computer can work extremely fast with the usual formula

$$t = (\bar{x} - \bar{y})/S\sqrt{1/m + 1/n}, \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the means of the first and second samples, respectively, and S is the pooled standard deviation. Values of 0 and 1 are typical variable values to use. In this way the sample means are the sample proportions of successes (i.e., $\bar{x} = a/m$ and $\bar{y} = b/n$).

2.2 Relation of the t Test to the X^2 Test

Readers may be surprised by the inclusion of the t test in an article concerning testing binomial data. In fact it is very close in computation to the X^2 test. This is best seen by considering the z test for testing the difference in proportions. Its test statistic is

$$z = \left(\frac{a}{m} - \frac{b}{n}\right) \Big/ \sqrt{\frac{s}{N}\left(1 - \frac{s}{N}\right)\left(\frac{1}{m} + \frac{1}{n}\right)} \tag{3}$$

and $z = \pm(X^2)^{1/2}$. The z and X^2 tests are equivalent. Any result for one applies to the other. The z and t tests differ only in their use of estimates of the variance [compare (2) and (3)]. One of us has already noted this relation (D'Agostino 1972).

3. METHOD

Our investigation proceeded as follows. First, we selected a representative set of binomial populations under the null hypothesis. For H_0: $p_1 = p_2 = p$ we examined $p = .05(.05).50$. For $p > .50$ results follow from symmetry. Second, for each null hypothesis selection we considered all possible sample sizes with $m = 5(1)15$ and $n = 5(1)15$. There were 660 configurations or combinations of p, m, and n. Third, for each p, m, and n we generated the exact sampling distributions of the X^2, X_c^2, Fisher, and t statistics. Note that this is *not* a simulation study. For each statistic we generated a complete enumeration of all possible sample configurations with their corresponding probabilities and from these we then derived the exact sampling distribution of the statistics. Consequently, we calculated the *exact* probability that the statistic under consideration leads to rejection of the null hypothesis in favor of the two-sided alternative with the rejection rule based on the standard critical values for that statistic. For example, for the t statistic with $m = n = 10$ and a nominal level of .05, we obtained the probability that the absolute value of the t statistic exceeds 2.101 (the 97.5th percentile of the t distribution with 18 df). This probability is the actual significance level. The commonly

used nominal levels of .01, .02, .05, and .10 were examined.

4. COMPARISONS OF THE TESTS

4.1 Fisher and X_c^2 Tests

As Upton (1982) noted, the X_c^2 test approximates very well the Fisher exact test, and both tests are extremely conservative. For nominal significance levels of .10, .05, .02, and .01 the maximum actual levels observed in the 660 configurations of p, m, and n were .061, .029, .008, and .004, respectively. Usually the actual levels were much smaller. In all cases these tests had actual significance levels smaller than the X^2 and t tests. For the two-independent-binomial case neither the Fisher test nor the X_c^2 test should be used.

4.2 X^2 and t Tests

Tables 1 and 2 summarize the results for the X^2 and t tests. Robustness can be evaluated in several different ways. Tables 1 and 2 offer some of these. First, there is the mean actual significance level over the 660 configurations of p, m, and n. For both the X^2 and t tests the mean actual levels are smaller than the nominal levels for nominal levels .10, .05, .02, and .01. For the X^2 test they are .097, .043, .015, and .007, respectively. For the t test they are .082, .040, .016, and .008, respectively. On the average, both of these tests are robust.

Examination of the relationship of p and the mean actual significance level reveals an increase in the levels with the maximum attained at $p = .50$. Tables 1 and 2 contain the mean actual levels for $p = .05(.05).50$. Recall that for $p > .50$ results are obtained by symmetry. Evaluating robustness by comparing the mean actual level as a function of the null hypothesis p to nominal levels indicates robustness for both tests. Even for the maximum actual levels at $p = .50$ we have good agreement with the nominal. For example, for nominal level .05 the mean actual levels for the chi-squared and t tests are .057 and .053, respectively.

Another way to evaluate robustness is to compare the maximum actual levels to the nominal levels. Tables 1 and 2 contain the maximum actual levels both over all p values and for the specific p values [$p = .05(.05).50$]. These results are extremely satisfying. For the t test the maximum actual significance levels are .145, .077, .039, and .022 for nominal levels .10, .05, .02, and .01, respectively. Very few configurations of p, m, and n even approach these maximum values (e.g., less than .5%) and these maximum values are, for all practical purposes, very close to the actual levels. For example, with the use of the t test and nominal significance level .01, the absolute worst situation possible is to have an actual level of .022. For a nominal level of .05 the most extreme actual level is .077. The X^2 test is similar with respect to the maximum actual significance levels. For nominal levels .10 and .05 the X^2 test's maximum levels exceed those of the t test and probably make it less attractive to use.

The last criterion for evaluating robustness that we present deals with finding the percentage of cases where the actual significance levels exceed the nominal levels by some fixed set of percentages. Upton (1982) used a suggestion of W. G. Cochran—namely, that exceeding the nominal level is acceptable up to a 20% error (e.g., for a nominal level of .05 an actual level less than or equal to .06 can be considered as an indication of robustness). Tables 1 and 2 give the percentages of cases where the actual significance levels exceed the nominal levels by 10%, 20%, 30%, 40%, and 50%. Recall that a 50% error for nominal level .05 only means an actual significance level of .075.

Using the 20% exceedance as a reasonable measure we see again that both the X^2 and t tests have good robustness properties over all nominal significance levels considered. Of major interest here are the results for the t test for nominal levels .10 and .05. (For one-sided tests these correspond to nominal levels .05 and .025.) For the t test only 3.8% (or

Table 1. Comparison of Actual and Nominal Levels of Significance for the Chi-Squared Test (two-sided test)

	Nominal level; Mean actual level (overall)				Nominal level; Maximum actual level (overall)			
	.10; .097	.05; .043	.02; .015	.01; .007	.10; .165	.05; .081	.02; .033	.01; .021
Mean actual level for p								
.05	.039	.008	.002	.001	.124	.028	.015	.011
.10	.064	.022	.005	.002	.137	.045	.024	.020
.15	.087	.034	.010	.004	.139	.062	.028	.021
.20	.101	.042	.013	.006	.142	.066	.027	.019
.25	.108	.048	.016	.007	.150	.067	.025	.015
.30	.112	.052	.018	.008	.161	.073	.027	.012
.35	.113	.054	.020	.009	.165	.077	.030	.016
.40	.114	.056	.021	.010	.165	.079	.032	.019
.45	.114	.056	.022	.010	.164	.080	.033	.021
.50	.114	.057	.022	.010	.163	.081	.033	.021
Percent by which the actual level exceeds the nominal level*								
10%					35.9	28.2	18.5	14.1
20%					20.2	16.5	9.8	8.3
30%					6.5	6.7	5.6	2.7
40%					1.5	1.7	1.7	1.4
50%					.8	.5	.8	1.1

NOTE: One-sided test levels are found by dividing the nominal and actual levels by 2.
*Values are percentages of 660 configurations.

25 cases) and 9.1% (60 cases) of the 660 configurations examined had the actual significance levels exceed .12 and .06 for the nominal levels .10 and .05, respectively. The percent times that the actual levels exceed 30%, 40%, and 50% of the nominal levels for the t tests are trivial for nominal levels .10 and .05. For these nominal levels the t test performs better than the X^2 test in that it is less likely to have the actual level exceed the nominal level. The performance of the X^2 is, however, acceptable for practical use.

For nominal significance levels .02 and .01 the X^2 test performs better than the t test. Again, however, even here the t test performs very well.

We should mention that for null hypothesis $p = 0$ or 1, both the t and X^2 tests are conservative with an actual significance level equal to 0 for all nominal levels. Thus the preceding extends to all p values ($0 \leq p \leq 1$).

5. DISCUSSION AND CONCLUSIONS

At present, unconditional tests of significance for testing the equality of two independent binomial populations, which have exact significance levels, do not exist. The Fisher exact test is an exact test, but it is not unconditional. It is conditional on all four margin totals. Considered unconditionally this test is extremely conservative over the entire range of configurations examined ($m, n \leq 15$; $0 \leq p \leq 1$). The chi-squared test with Yates's continuity correction is an excellent approximation to the Fisher test and so is just as conservative. For small sample sizes and for comparing two independent binomial populations, neither the Fisher exact test nor the chi-squared test with Yates's continuity correction should be used. See, for example, Fisher (1956) and Kempthorne (1979) for appropriate applications of these tests.

In this article we have examined the robustness properties of the usual chi-squared test and the two-independent-sample t test with pooled variance. Our conclusion is that both can be considered robust, with the t test favored over the chi-squared test. Although both tests did have configurations of m, n, and p where the actual significance level exceeded the nominal level, *mean* actual levels (averaged over $m, n \leq 15$ and $.05 \leq p \leq .95$) were well below the nominal levels for the usual nominal levels of .01, .02, .05, and .10. Further, for both tests the maximum actual levels did not exceed greatly, from a practical point of view, the nominal levels. In particular, for either a one-sided or a two-sided t test at nominal level .05, less than 10% of the 660 configurations of m, n, and p examined produced actual levels greater than .06. The maximum actual level was only .077.

For routine data analysis comparing two groups, the t test is often used. Heeren and D'Agostino (1987) showed that it can be used robustly for ordinal data with 3, 4, or 5 ordinal categories even with small samples. The present investigation extends its robustness to binomial data. We encourage its use.

At a minimum the chi-squared test or, equivalently, the z test of (3) should replace the Fisher exact test, and the Yates continuity correction should not be used.

[*Received May 1987. Revised October 1987.*]

REFERENCES

Berkson, J. (1978), "In Dispraise of the Exact Test," *Journal of Statistical Planning and Inference*, 2, 27–42.

Conover, W. J. (1974), "Some Reasons for Not Using the Yates Continuing Correction on 2 × 2 Contingency Tables," *Journal of the American Statistical Association*, 69, 374–376.

D'Agostino, R. B. (1972), "Relation Between the Chi-Squared and ANOVA Tests for Testing the Equality of k Independent Dichotomous Populations," *The American Statistician*, 26, 30–32.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inferences*, Edinburg: Oliver & Boyd.

Grizzle, J. E. (1967), "Continuity Correction in the X^2 Test for 2 × 2 Tables," *The American Statistician*, 21, 28–32.

Table 2. Comparison of Actual and Nominal Levels of Significance for the t Test With Pooled Variance (two-sided test)

	Nominal Level; Mean actual level (overall)				Nominal level; Maximum actual level (overall)			
	.10; .082	.05; .040	.02; .016	.01; .008	.10; .146	.05; .077	.02; .039	.01; .022
Mean actual level for p								
.05	.025	.007	.002	.001	.119	.027	.015	.010
.10	.056	.020	.006	.002	.126	.054	.024	.020
.15	.076	.032	.010	.005	.121	.055	.025	.021
.20	.087	.041	.014	.007	.133	.063	.024	.019
.25	.093	.046	.017	.009	.130	.066	.025	.015
.30	.096	.049	.020	.010	.120	.068	.024	.014
.35	.098	.051	.021	.011	.125	.066	.030	.017
.40	.098	.052	.022	.012	.137	.071	.036	.020
.45	.098	.053	.023	.013	.144	.075	.038	.022
.50	.098	.053	.023	.013	.146	.077	.039	.022
Percent by which the actual level exceeds the nominal level*								
10%					11.1	17.1	23.8	30.8
20%					3.8	9.1	14.4	21.8
30%					.9	3.5	8.2	12.7
40%					.3	.5	2.6	9.7
50%					.0	.2	1.7	5.8

NOTE: One-sided test levels are found by dividing the nominal and actual levels by 2.
*Values are percentages of 660 configurations.

Heeren, T., and D'Agostino, B. (1987), "Robustness of the Two Independent Samples *t*-Test When Applied to Ordinal Scaled Data," *Statistics in Medicine*, 6, 79–90.

Kempthorne, O. (1979), "In Dispraise of the Exact Test: Reactions," *Journal of Statistical Planning and Inference*, 3, 199–213.

Mantel, N. (1974), "Comment and a Suggestion" on "Some Reasons for Not Using the Yates Continuity Correction on 2 × 2 Contingency Tables," by W. J. Conover, *Journal of the American Statistical Association*, 69, 378–380.

Upton, G. J. G. (1982), "A Comparison of Alternative Tests for the 2 × 2 Comparative Trial," *Journal of the Royal Statistical Society*, Ser. A, 145, 86–105.

A Central Limit Theorem for the Bootstrap Mean

SHIE–SHIEN YANG*

The central limit theorem of the bootstrap mean is proved by the elementary method of characteristic functions.

KEY WORDS: Characteristic function; Law of large numbers.

1. INTRODUCTION

Efron (1979) proposed a "bootstrap" method for approximating the distribution of a function of the observations and population. This method can be used to set confidence intervals and to estimate the bias and variance of an estimate.

Many recent developments of bootstrap can be found, for example, in Efron (1981), Bickel and Freedman (1981), Singh (1981), Athreya (1983), Beran (1984), Hall (1986), and Wu (1986). The asymptotic theory for the bootstrap in the literature, even for the simplest result such as the central limit theorem for the bootstrap mean, is too difficult for the average graduate students in statistics to understand. In this note, using characteristic function, an elementary proof of the central limit theorem for the bootstrap mean is given. The proof given here is similar to that given in theorem 1 of Singh (1981). Singh's proof, however, may be less appealing to average graduate students in statistics.

2. THE CENTRAL LIMIT THEORY

The following setting and the associated notations will be adopted throughout this note. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with common distribution function F. Let F_n be the usual sample distribution function of $X_1, \ldots, X_n$, putting probability $1/n$ at each X_i. Given $(X_1, \ldots, X_n)$, let $Y_1, \ldots, Y_n$ be a random sample drawn from F_n.

Theorem 1. Suppose that F has finite positive variance σ^2. Let $T_n = n^{1/2}(\bar{Y}_n - \bar{X}_n)$, where $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then along almost all sample sequences $X_1, X_2, \ldots$, the conditional distribution of T_n given $(X_1, \ldots, X_n)$ converges weakly to a normal distribution $N(0, \sigma^2)$ with mean 0 and variance σ^2 as $n \to \infty$.

Proof. It is known (see, e.g., Billingsley 1979, p. 298) that if X is a random variable with mean 0 and finite positive variance σ^2, then the characteristic function $\varphi(t)$ of X can be expressed as

$$\varphi(t) = 1 - t^2\sigma^2/2 + \theta(t),$$

where $|\theta(t)| \le |t|^2 \mathrm{E}[(|t| \cdot |X|^3) \wedge (X^2)]$. $a \wedge b$ means the smaller of a and b. Using this fact, the characteristic function of the conditional distribution of T_n given $(X_1, \ldots, X_n)$ can be written as

$$[\mathrm{E}\{\exp[it(Y_1 - \bar{X}_n)/n^{1/2}] \mid X_1, \ldots, X_n\}]^n = [1 - t^2\hat{\sigma}^2/2n + \theta_n(t)]^n, \quad (1)$$

where $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$ and

$$|\theta_n(t)| \le |t|^2 n^{-1} \sum_{i=1}^n [(|t|n^{-1/2}|X_i - \bar{X}|^3) \wedge (X_i - \bar{X})^2]/n. \quad (2)$$

Let $g_n(X_i, \bar{X})$ denote the ith term of the sum in (2). Using the fact that for $p > 1$, $|a + b|^p \le 2^{p-1}(|a|^p + |b|^p)$, we have

$$0 \le g_n(X_i, \bar{X}) \le \{|t|n^{-1/2}4(|X_i - \mu|^3 + |\bar{X} - \mu|^3)\} \wedge \{2(|X_i - \mu|^2 + |\bar{X} - \mu|^2)\}.$$

By the law of large numbers, $|\bar{X} - \mu|$ converges to 0 almost surely as $n \to \infty$. Hence, almost surely,

$$0 \le \limsup_{n\to\infty} n^{-1} \sum_{i=1}^n g_n(X_i, \bar{X}) \le \limsup_{n\to\infty} n^{-1} \sum_{i=1}^n h_n(X_i), \quad (3)$$

where

$$h_n(X_i) = \{|t|n^{-1/2}4(|X_i - \mu|^3 + 1)\} \wedge \{2(|X_i - \mu|^2 + 1)\}.$$

Clearly, $h_n(X_i)$ converges to 0 for any value of X_i as $n \to \infty$, and $|h_n(X_i)| \le 2(|X_i - \mu|^2 + 1)$. Hence, by the dominated convergence theorem, $\mathrm{E}[h_n(X_i)]$ converges to 0 as $n \to \infty$. Given $\varepsilon > 0$, let N be a positive integer such that $\mathrm{E}[h_N(X_i)] < \varepsilon$. For $n > N$, we have

*Shie-Shien Yang is Associate Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506.