Article

TN:**653505**

LIBRARY ANNEX
HA1 .A51 - -

Borrower: 56130
ILL/ID#:

201.147.111.104

12\1

Shipping Address:
CIMMYT, Interlibrary Loan
RECIPROCAL ONLY ARTICLES - 5/WK
Km. 45, Carretera Mexico-Veracruz
56130 El Batan
Texcoco, Estado de Mexico
Fax: 525558047558

Patron:
Maxcost: $30 I

Journal Title: American statistician

Vol/Issue/Date/Pages: 26 #3 30-32 Jun 1972 30-32

Article: D'Agostino, R. Relation between the chi-squared and ANOVA test for testing

Imprint:

4

NEED A RESEND? PLEASE SEND THIS FORM AND FILL IN THE FOLLOWING:

PAGES TO RESEND: ______________________________

REASON: ______________________________

From:
Interlibrary Services
Olin Library
Central Ave
Cornell University Library
Ithaca, NY 14853-5301

Phone: 607-255-5293 Fax: 607-255-9091
Ariel: 128.253.70.20
olin-ils-lending@cornell.edu

Unaffiliated 11/25/2008 09:36:06 AM

Interlibrary Services
Cornell University Library
Olin Library
Ithaca, NY 14853-5301
U.S.A.

Phone: (607) 255-5293
Fax: (607) 255-9091
E-mail: olin-ils-lending@cornell.edu
IP: 128.253.70.20
FEIN: 15-0532-082

Invoice No. O- 653505

Customer Information
56130 CIMMYT, Interlibrary Loan RECIPROCAL ONLY ARTICLES - 5/WK Km. 45, Carretera Mexico-Veracruz 56130 El Batan Texcoco, Estado de Mexico Phone: 525558042004 Fax: 525558047558

Date	December 1, 2008
Billing Category	
ILL Number	
System ID	LWeb

Description	Base Price	Unit Price	Quantity	Total
Standard Charges for Article: American statistician Relation between the chi-squared and ANOVA test for testing	$20.00		1	$20.00
Service Level - Regular				
Shipping Fees - Ariel				
No Additional Charges (1)				
No Additional Charges (2)				
No Additional Charges (3)				
TOTAL				$20.00

PLEASE RETURN A COPY OF THIS INVOICE WITH YOUR PAYMENT

Mail payment to:	Make check or money order payable to:
Cornell University Library **Accounting Office** **234 Olin Library** **Ithaca, NY 14853-5301** **USA**	**CORNELL UNIVERSITY LIBRARY**

THE AMERICAN STATISTICIAN

Published by
The American Statistical Association
Founded 1839

JUNE 1972
Vol. 26 • Number 3

Annual Dues
American Statistical Association

Regular Membership (Residents of U. S. and Territories, Canada, APO and FPO addresses and North Americans on overseas assignment)	$ 20.00*
Residents outside North America (not including Americans or Canadians on overseas assignment)	15.00*
Student Membership (Full-time students only)	9.00*
Student Affiliate (Full-time undergraduate students only)	2.00
Corporate Membership	200.00*
Technometrics (special member rate)	10.00

* Includes subscription to the **Journal of the American Statistical Association** and to **The American Statistician.**

Journal Subscription rate: $25.00 per year. Prices for back issues available on request.

For information about advertising, membership and subscriptions, write to ASA, 806 15th St., N. W., Washington, D. C. 20005.

The Editorial Committee welcomes the submission of manuscripts for possible publication. Three copies, double-spaced, should be sent to the Editor, **Donald F. Morrison, Department of Statistics and Operations Research, Wharton School, University of Pennsylvania, Philadelphia, Pa. 19104.**

Correspondence concerning the separate departments should be sent to the Department Editor at the American Statistical Association, 806 15th Street, N.W., Washington, D. C. 20005.

Second class postage paid at Washington, D.C., and at additional mailing offices. **The American Statistician** is published five times a year—February, April, June, October and December—by the American Statistical Association, Business Office: 806 15th St., N.W., Washington, D.C. 20005. Subscription rate: five dollars a year or one dollar per copy. Anyone wishing to change his mailing address should allow eight weeks notice. A copy of the address taken from an issue of the periodical should accompany the change-of-address request.

$E(k \mid m)$, first calculate

$$E(n \mid m) = \sum_{n=0}^{2^{m-1}-1} np(n \mid m) = \sum_{n=1}^{2^{m-1}-1} np(n \mid m). \qquad (7)$$

Examine the product $np(n \mid m)$. From (3), each term of the numerator of this expression can be written as

$$n\binom{2^{m-1} - 1 + n}{n} = 2^{m-1}\binom{2^{m-1} + (n-1)}{n-1}. \qquad (8)$$

Let $n' = n - 1$. The sum of these terms for $n' = 0, 1, \ldots, 2^{m-1} - 2$, can be written as a constant times an expression in the form of (4), with $r = 2^{m-1} - 2$, and $M = 2^{m-1} + 1$. Therefore, from (3) and (4), it follows that

$$E(n \mid m) = 2^{m-1}\binom{2^m - 1}{2^{m-1} - 2} \bigg/ \binom{2^m - 1}{2^{m-1} - 1} \qquad (9)$$

$$= 2^{m-1}(2^{m-1} - 1)/(2^{m-1} + 1) \quad .$$

Recall that $k = 2^{m-1} - n + 1$, so that the expected value of k is given by

$$E(k \mid m) = 2^{m-1} - E(n \mid m) + 1$$

$$= (3 + 1/2^{m-1})/(1 + 1/2^{m-1}). \qquad (10)$$

Thus the expected ability rank of the runner-up approaches the rank three geometrically as m increases.

4. *Interpretation*

This analysis may help to explain the "dark horses" of some tournaments whose playing ability had never previously been noted because they had been knocked out of play early in the competition by better players. It may also help to explain the lack of consistency in performance by "flash in the pan" players whose final rank in a tournament may have been due to the fortuitous elimination of the better players earlier in the competition. The analysis does bring out clearly the inherent inequities of the knockout tournament as compared to a round robin tournament. In the former, the players are exposed to the vagaries of chance through both the uncertainty of the initial alignment and the uncertainty of the individual encounters, while in the latter, since each player plays each other player an equal number of times, only the second source of unpredictability is present.

REFERENCE

Feller, W. *An Introduction to Probability Theory and Its Applications, Vol. I*, Third Edition, New York: John Wiley and Sons, Inc. 1968.

Relation Between the Chi–Squared and ANOVA Tests for Testing the Equality of k Independent Dichotomous Populations

RALPH B. D'AGOSTINO
Boston University

It is oftentimes pointed out in statistics texts that the two-sample test of equal proportions (using the normal approximation) is equivalent to the chi-squared test on 2×2 contingency tables. Pointing out this equivalence aids in demonstrating a unity of large-sample procedures and in easing the student's or reader's mind by showing him that for exactly the same problem (assumptions and hypotheses) he really does not have two different procedures from which to select. (Of course, this is not to be confused with situations where, for example, either a parametric test or a non-parametric test may be used. In those situations the non-parametric test usually is an alternative procedure requiring *less* assumptions than the parametric test.) Similar situations as the above (although not large sample results) are in the demonstrations that the two sample t-tests for differences in means (both independent and dependent samples) are equivalent to analysis of variance (ANOVA) tests (one- and two-way ANOVA, respectively).

Another situation which is not as straightforward as the above but where it would prove instructive to point out relations between statistical methods is for the case of testing the equality of population proportions for k independent dichotomous populations. The statistical procedures are the one-way ANOVA test and the chi-squared test for $2 \times k$ contingency tables. It has been the author's experience that students and users are often very confused about this situation. It seems to them to be an appropriate place for a one-way ANOVA test, especially in light of the robustness of the ANOVA test. Yet most, if not all, texts handle the problem only in the section on chi-squared tests. A question often asked is, "What would happen if the ANOVA test is used?" In the following we demonstrate the algebraic and probabilistic similarities between these tests. They are almost algebraically equivalent-neither is easier to compute. Also for testing the null hypothesis they are asymptotically equivalent tests.

One-Way ANOVA

Let us say that we have k independent dichotomous populations. From each a random sample of size n_i $(i = 1, \ldots, k)$ is drawn. Each observation X_{ij}

TABLE 1

One-Way ANOVA Table

Sources of Variation	Sums of Squares	Degrees of Freedom	Mean Squares
Between samples	$\sum_i n_i(p_i - \bar{p})^2 = A$	$k - 1$	$A/(k - 1)$
Within samples	$\sum_i \sum_j (X_{ij} - p_i)^2 = B$	$N - k$	$B/(N - k)$
Total	$\sum_i \sum_j (X_{ij} - \bar{p})^2$	$N - 1$	

TABLE 2

Revised One-Way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
Between samples	$\sum_i n_i(p_i - \bar{p})^2 = A$	$k - 1$	$A/(k - 1)$
Within samples	$\sum_i n_i p_i(1 - p_i) = B$	$N - k$	$B/(N - k)$
Total	$N\bar{p}(1 - \bar{p}) = NC$	$N - 1$	

($i = 1, \ldots, k$ is the index of samples and $j = 1, \ldots, n_i$ is the index of observations within the ith sample.) is either a one (success) or zero (failure). Let a_i and p_i represent the number and the proportion of ones, respectively, for the ith sample. Let $\bar{p}$ be the overall sample proportion of ones. That is,

$$\bar{p} = \frac{\sum_i n_i p_i}{\sum_i n_i} \quad (1)$$

Further, let $N = \sum_i n_i$ (i.e., the total number of observations).

If the null hypothesis of equal proportions of successes for each population is to be tested and one views this as a one-way ANOVA problem the appropriate ANOVA table is given in Table 1.

The entries of Table 1 can be rewritten. First, the within-samples sum of squares, B, of Table 1

$$\begin{aligned} B &= \sum_i \sum_j (X_{ij} - p_i)^2 \\ &= \sum_i [a_i(1 - p_i)^2 + (n_i - a_i)p_i^2] \\ &= \sum_i n_i[p_i(1 - p_i)^2 + (1 - p_i)p_i^2] \\ &= \sum_i n_i p_i(1 - p_i). \end{aligned} \quad (2)$$

Similarly the total sum of squares can be rewritten

$$\begin{aligned} \sum_i \sum_j (X_{ij} - \bar{p})^2 &= N\bar{p}(1 - \bar{p}) \\ &= NC \end{aligned} \quad (3)$$

where $C = \bar{p}(1 - \bar{p})$. Table 2 is the ANOVA table incorporating these changes.

The ANOVA F statistic is

$$F = \frac{A/(k - 1)}{B/(N - k)} \quad (4)$$

From large sample theory and the known robustness of the F test for ANOVA problems (Scheffé, 1959, Chapter 10) the F of (4) has asymptotically an F distribution with $k - 1$ and $N - k$ degrees of freedom (numerator and denominator, respectively) under the null hypothesis.

$2 \times k$ *Contingency Table*

The problem given above fits exactly into the theory of contingency table analysis (here a $2 \times k$ contingency table). Letting O represent the observed and E the expected value the appropriate chi-squared statistic for the null hypothesis of equal population proportions of successes is

$$\begin{aligned} X^2 &= \sum (O - E)^2/E \\ &= \sum_i \frac{(a_i - n_i\bar{p})^2}{n_i\bar{p}} + \sum_i \frac{[(n_i - a_i) - n_i(1 - \bar{p})]^2}{n_i(1 - \bar{p})} \\ &= \sum_i \frac{(a_i - n_i\bar{p})^2}{n_i}\left[\frac{1}{\bar{p}} + \frac{1}{1 - \bar{p}}\right] \\ &= \sum_i \frac{n_i(p_i - \bar{p})^2}{\bar{p}(1 - \bar{p})} \\ &= \frac{A}{C} \end{aligned} \quad (5)$$

where A and C are defined in Table 2. For large samples X^2 of (5) has a chi-squared distribution with $k - 1$ degrees of freedom under the null hypothesis. If we divide X^2 of (5) by $k - 1$ then under the same conditions as above

$$\frac{X^2}{k-1} = \frac{A/(k-1)}{C} \tag{6}$$

has an F distribution with $k - 1$ and infinite degrees of freedom. A comparison of (4) and (6) will help clarify the relationship between the ANOVA and chi-squared tests. The algebraic similarities are obvious. Notice both require the same amount of computation.

The Relationship

The difference between the ANOVA F of (4) and the chi-squared $X^2/(k - 1)$ of (6) lies in the denominators—$B/(N - k)$ for F and C for $X^2/(k - 1)$. However, under the null hypothesis these are asymptotically equivalent. For from Table 2 we have

$$C = \frac{A+B}{N} = \frac{A}{k-1}\left[\frac{k-1}{N}\right] + \frac{B}{N-k}\left[\frac{N-k}{N}\right] \tag{7}$$

and because k is fixed we have as $N \to \infty$ and $n_i/N \to \lambda_i \neq 0$

$$\frac{A}{k-1}\left[\frac{k-1}{N}\right] \underset{P}{\to} 0 \tag{8}$$

and

$$\frac{B}{N-k}\left[\frac{N-k}{N}\right] \underset{L}{\to} \frac{B}{N-k} \tag{9}$$

where $\underset{P}{\to}$ and $\underset{L}{\to}$ mean convergence in probability and law, respectively. Thus

$$C \underset{L}{\to} \frac{B}{N-k} \tag{10}$$

and in turn the distributions of F and $X^2/(k - 1)$ are asymptotically equivalent. Actually under the null hypothesis both C and $B/(N - k)$ are estimators of $P(1 - P)$ where P is the true population proportion of successes and they converge in probability to $P(1 - P)$. This suffices to establish asymptotic equivalence. (7) is helpful for establishing results under alternative hypotheses.

With regard to the different number of degrees of freedom employed by these tests, i.e., $N - k$ for the denominator of F and infinity for $X^2/(k - 1)$, it is interesting to note that in an extensive simulation study (Lunney, 1970) with equal sample sizes (i.e., $n_i = n$) it was found that the use of the ANOVA test in the present situation produces a conservative test if the sample sizes are small. One way of removing this conservativeness would be to claim (as does the chi-squared test) more degrees of freedom for the denominator.

Conclusion

We have seen for the situation studied that the one-way ANOVA procedure and the standard chi-squared procedure are algebraically similar and under the null hypothesis asymptotically equivalent. Pointing this out to students and users of statistical methods may aid substantially in their understanding of statistical methodology. There really are not two distinct ways of handling this problem.

REFERENCES

Lunney, G. H. Using analysis of variance with a dichotomous dependent variable: an empirical study. *Journal of Educational Measurement*, 7, 263–69, 1970.

Scheffé, H. *The Analysis of Variance.* New York: John Wiley and Sons, Inc. 1959.

Snedecor, G. W. and Cochran, W. G. *Statistical Methods* (6th Edition). Ames, Iowa: Iowa State University Press, 1967.

Two Approaches to Polynomial Distributed Lags Estimation: An Expository Note and Comment

J. PHILLIP COOPER
University of Chicago

Introduction

This paper (1) discusses the approaches that have been used to explain and to implement the method suggested by Shirley Almon (2) for estimating the distributed lag model. Almon's original approach utilized Lagrangian interpolation polynomials; but, because of a simpler exposition and lack of general understanding among economists and economics students alike of the nature of Lagrangian interpolation techniques, a direct, polynomial approximation approach is gradually being adopted in the classroom and in computer programs (3). Unfortunately, the direct approach can be hampered by multicollinearity in the artificial variables created for computational purposes (4). Both techniques are explained here, along with a demonstration of their algebraic equivalence and computational differences. No attempt is made here to comment on the specification of parameters of the estimation technique—for example, the appropriate degree of polynomial, zero restrictions, etc. (5).