

Vassilly Voinov
Mikhail Nikulin
Narayanaswamy Balakrishnan

Chi-Squared Goodness of Fit Tests with Applications

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$



Chi-Squared Goodness of Fit Tests with Applications

V. Voinov

KIMEP University;
Institute for Mathematics and Mathematical Modeling
of the Ministry of Education and Science,
Almaty, Kazakhstan

M. Nikulin

University Bordeaux-2,
Bordeaux, France

N. Balakrishnan

McMaster University, Hamilton,
Ontario, Canada



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an Imprint of Elsevier



Academic press is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

© 2013 Elsevier, Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

Application submitted

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN: 978-0-12-397194-4

Printed in the United States of America

13 14 15 16 17 10 9 8 7 6 5 4 3 2 1

**Working together to grow
libraries in developing countries**

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

For information on all BH publications visit our website at store.elsevier.com

Chi-Squared Goodness of Fit Tests with Applications

Dedication

To the memory of my father,

VV

To the memory of Antonina, Alexandra and Nikolai,

MN

**To the loving memory of my sister, Mrs. Chitra
Ramachandran, who departed too soon leaving a huge void
in my life!**

NB

Many parametric models, possessing different characteristics, shapes, and properties, have been proposed in the literature. These models are commonly used to develop parametric inferential methods. The inference developed and conclusions drawn based on these methods, however, will critically depend on the specific parametric model assumed for the analysis of the observed data. For this reason, several model validation techniques and goodness of fit tests have been developed over the years.

The oldest and perhaps the most commonly used one among these is the chi-squared goodness of fit test proposed by Karl Pearson over a century ago. Since then, many modifications, extensions, and generalizations of this methodology have been discussed in the statistical literature. Yet, there are some misconceptions and misunderstandings in the use of this method even at the present time.

The main aim of this book is, therefore, to provide an in-depth account of the theory, methods, and applications of chi-squared goodness of fit tests. In the process, pertinent formulas for their use in testing for some specific prominent distributions, such as normal, exponential, and Weibull, are provided. The asymptotic properties of the tests are described in detail, and Monte Carlo simulations are also used to carry out some comparisons of the power of these tests for different alternatives.

To provide a clear understanding of the methodology and an appreciation for its wide-ranging application, several well-known data sets are used as illustrative examples and the results obtained are then carefully interpreted. In doing so, some of the commonly made mistakes and misconceptions with regard to the use of this test procedure are pointed out as well.

We hope this book will serve as an useful guide for this popular methodology to theoreticians and practitioners alike. As pointed out at a number of places in the book, there are still many open problems in this area, and it is our sincere hope that the publication of this book will rejuvenate research activity, both theoretical and applied, in this important topic of research.

Preparation of a book of this nature naturally requires the help and co-operation of many individuals. We acknowledge the overwhelming support we received from numerous researchers who willingly shared their research publications and ideas with us. The editors of Academic Press/Elsevier were greatly supportive of this project from the start, and their production department were patient and efficient while working on the final production stages of the book. Our sincere thanks also go to our respective families for

their emotional support and patience during the course of this project, and to Ms. Debbie Iscoe for her diligent work on the typesetting of the entire manuscript.

Vassilly Voinov, Kazakhstan

Mikhail Nikulin, France

Narayanaswamy Balakrishnan, Canada

Preface	xi
1. A Historical Account	1
2. Pearson's Sum and Pearson-Fisher Test	11
2.1 Pearson's chi-squared sum	11
2.2 Decompositions of Pearson's chi-squared sum	12
2.3 Neyman-Pearson classes and applications of decompositions of Person's Sum	17
2.4 Pearson-Fisher and Dzhaparidze-Nikulin tests	19
2.5 Chernoff-Lehmann Theorem	24
2.6 Pearson-Fisher test for random class end points	24
3. Wald's Method and Nikulin-Rao-Robson Test	27
3.1 Wald's Method	27
3.2 Modifications of Nikulin-Rao-Robson test	34
3.3 Optimality of Nikulin-Rao-Robson test	37
3.4 Decomposition of Nikulin-Rao-Robson Test	37
3.5 Chi-squared tests for multivariate normality	38
3.5.1 Introduction	38
3.5.2 Modified chi-squared tests	39
3.5.3 Testing for bivariate circular normality	41
3.5.4 Comparison of different tests	45
3.5.5 Conclusions	47
3.6 Modified chi-squared tests for the exponential distribution	48
3.6.1 Two-parameter exponential distribution	48
3.6.2 Scale-exponential distribution	51
3.7 Power generalized Weibull distribution	52
3.7.1 Estimation of parameters	52
3.7.2 Modified chi-squared test	53
3.7.3 Evaluation of power	54
3.8 Modified chi-squared goodness of fit test for randomly right censored data	58
3.8.1 Introduction	58
3.8.2 Maximum likelihood estimation for right censored data	59
3.8.3 Chi-squared goodness of fit test	63
3.8.4 Examples	68
3.9 Testing normality for some classical data on physical constants	79
3.9.1 Cavendish's measurements	80

3.9.2	Millikan's measurements	82
3.9.3	Michelson's measurements	85
3.9.4	Newcomb's measurements	87
3.10	Tests based on data on stock returns of two Kazakhastani companies	89
3.10.1	Analysis of daily returns	90
3.10.2	Analysis of weekly returns	94
4.	Wald's Method and Hsuan-Robson-Mirvaliev Test	97
4.1	Wald's method and moment-type estimators	97
4.2	Decomposition of Hsuan-Robson-Mirvaliev test	99
4.3	Equivalence of Nikulin-Rao-Robson and Hsuan-Robson-Mirvaliev tests for exponential family	100
4.4	Comparisons of some modified chi-squared tests	103
4.4.1	Maximum likelihood estimates	103
4.4.2	Moment-type estimators	105
4.5	Neyman-Pearson classes	107
4.5.1	Maximum likelihood estimators	108
4.5.2	Moment-type estimators	109
4.6	Modified chi-squared test for three-parameter Weibull distribution	110
4.6.1	Parameter estimation and modified chi-squared tests	111
4.6.2	Power evaluation	112
4.6.3	Neyman-Pearson classes	114
4.6.4	Discussion	114
4.6.5	Concluding remarks	115
5.	Modifications Based on UMVUEs	117
5.1	Test for Poisson, binomial, and negative binomial distributions	117
5.2	Chi-squared test for one-parameter exponential family	121
5.3	Revisiting Clarke's data on flying bombs	124
6.	Vector-Valued Tests	127
6.1	Introduction	127
6.2	Vector-valued tests: An artificial example	128
6.3	Example of Section 2.3 revisited	132
6.4	Combining nonparametric and parametric tests	134
6.5	Combining nonparametric tests	135
6.6	Concluding comments	137
7.	Applications of Modified Chi-Squared Tests	139
7.1	Poisson versus binomial: Appointment of judges to the US supreme court	139
7.1.1	Introduction	139
7.1.2	Data to be analyzed	140

7.1.3	Statistical analysis of the data	142
7.1.4	Revisiting the analyses of Wallis and Ulmer	148
7.1.5	Comments about King's exponential Poisson regression model	150
7.1.6	Concluding remarks	151
7.2	Revisiting Rutherford's data	152
7.2.1	Analysis of the data	152
7.2.2	Concluding remarks	158
7.3	Modified tests for the logistic distribution	159
7.4	Modified chi-squared	163
7.4.1	Introduction	163
7.4.2	The NRR, DN and McCulloch tests	164
8.	Probability Distributions of Interest	167
8.1	Discrete probability distributions	167
8.1.1	Binomial, geometric, and negative binomial distributions	167
8.1.2	Multinomial distribution	171
8.1.3	Poisson distribution	171
8.2	Continuous probability distributions	173
8.2.1	Exponential distribution	174
8.2.2	Uniform distribution	180
8.2.3	Triangular distribution	180
8.2.4	Pareto model	180
8.2.5	Normal distribution	181
8.2.6	Multivariate normal distribution	182
8.2.7	Chi-square distribution	187
8.2.8	Non-central chi-square distribution	188
8.2.9	Weibull distribution	191
8.2.10	Generalized power Weibull distribution	194
8.2.11	Birnbaum-Saunders distribution	194
8.2.12	Logistic distribution	195
9.	Chi-Squared Tests for Specific Distributions	197
9.1	Test for Poisson, binomial, and "binomial" approximation of Feller's distribution	197
9.2	Elements of matrices K, B, C, and V for the three-parameter Weibull distribution	203
9.3	Elements of matrices J and B for the generalized power Weibull distribution	206
9.4	Elements of matrices J and B for the two-parameter exponential distribution	208
9.5	Elements of matrices B, C, K, and V to test the logistic distribution	209
9.6	Testing for normality	211
9.7	Testing for exponentiality	211

9.7.1	Test of Greenwood and Nikulin (see Section 3.6.1)	211
9.7.2	Nikulin-Rao-Robson test (see Eq. (3.8) and Section 9.4)	212
9.8	Testing for the logistic	212
9.9	Testing for the three-parameter Weibull	212
9.10	Testing for the power generalized Weibull	213
9.11	Testing for two-dimensional circular normality	213
Bibliography		215
Index		227

A Historical Account

The famous chi-squared goodness of fit test was proposed by Pearson (1900). If simple observations are grouped over r disjoint intervals Δ_j and $N_j^{(n)}$ denote observed frequencies corresponding to a multinomial scheme with $np_j(\theta)$ as the expected frequencies, for $j = 1, 2, \dots, r$, the Pearson's sum is given by

$$\chi^2 = \sum_{j=1}^r \frac{(N_j^{(n)} - np_j(\theta))^2}{np_j(\theta)} = \mathbf{V}^{(n)T}(\theta) \mathbf{V}^{(n)}(\theta), \quad (1.1)$$

where $\mathbf{V}^{(n)}(\theta)$ is the vector of standardized frequencies with components

$$v_j^{(n)}(\theta) = (N_j^{(n)} - np_j(\theta)) / (np_j(\theta))^{1/2}, \quad j = 1, \dots, r, \quad \theta \in \Theta \subset R^s.$$

If the number of sample observations $n \rightarrow \infty$, the statistic in (1.1) will follow the chi-squared probability distribution with $r - 1$ degrees of freedom. We know that this remarkable result is true only for a simple null hypothesis when a hypothetical distribution is specified uniquely (i.e. the parameter θ is considered to be known). Until 1934, Pearson believed that the limiting distribution of the statistic in (1.1) will be the same if the unknown parameters of the null hypothesis are replaced by their estimates based on a sample; see, for example, Baird (1983), Plackett (1983, p. 63), Lindley (1996), Rao (2002), and Stigler (2008, p. 266). In this regard, it is important to reproduce the words of Plackett (1983, p. 69) concerning E.S. Pearson's opinion: "I knew long ago that KP (meaning Karl Pearson) used the 'correct' degrees of freedom for (a) difference between two samples and (b) multiple contingency tables. But he could not see that χ^2 in curve fitting should be got asymptotically into the

same category.” Plackett explained that this crucial mistake of Pearson arose from Karl Pearson’s assumption “that individual normality implies joint normality.” Stigler (2008) noted that this error of Pearson “has left a positive and lasting negative impression upon the statistical world.” Fisher (1924) clearly showed that the number of degrees of freedom of Pearson’s test must be reduced by the number of parameters estimated from the sample. To this point, it must be added that Fisher’s result is true if and only if the parameters are estimated from the vector of frequencies minimizing Pearson’s chi-squared sum, using multinomial maximum likelihood estimates (MLEs), or by any other asymptotically equivalent procedure (Greenwood and Nikulin, 1996, p. 74). Such estimates based on a vector of frequencies, which is not in general the vector of sufficient statistics, are not asymptotically efficient, however, due to which the Pearson-Fisher test is not powerful in many cases. For a review on using minimum chi-squared estimators, one may refer to Harris and Kanji (1983). Nowadays, Pearson’s test with unknown parameters replaced by estimates $\hat{\theta}_n$ based on the vector of frequencies is referred to as Pearson-Fisher (PF) test given by

$$X_n^2(\hat{\theta}_n) = \sum_{j=1}^r \frac{(N_j^{(n)} - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)} = \mathbf{V}^{(n)T}(\hat{\theta}_n) \mathbf{V}^{(n)}(\hat{\theta}_n). \quad (1.2)$$

Dzhaparidze and Nikulin (1974) proposed a modification of the standard Pearson statistic (DN test), valid for any \sqrt{n} -consistent estimator $\tilde{\theta}_n$ of an unknown parameter, given by

$$U_n^2(\tilde{\theta}_n) = \mathbf{V}^{(n)T}(\tilde{\theta}_n)(\mathbf{I} - \mathbf{B}_n(\mathbf{B}_n^T \mathbf{B}_n)^{-1} \mathbf{B}_n^T) \mathbf{V}^{(n)}(\tilde{\theta}_n), \quad (1.3)$$

where \mathbf{B}_n is an estimate of the matrix \mathbf{B} with elements

$$b_{jk} = \frac{1}{\sqrt{p_j(\theta)}} \int_{\Delta_j} \frac{\partial f(x, \theta)}{\partial \theta_k} dx, \quad j = 1, \dots, r, \quad k = 1, \dots, s.$$

This test, being asymptotically equivalent to the Pearson-Fisher statistic in many cases, is not powerful for equiprobable cells (Voinov et al., 2009) but is rather powerful if an alternative hypothesis is specified and one uses the Neyman-Pearson classes for constructing the vector of frequencies.

Several authors, such as Cochran (1952), Yarnold (1970), Larntz (1978), Hutchinson (1979), and Lawal (1980), considered the problem of approximating the discrete distribution of Pearson’s sum if some expected frequencies become too small. Baglivo et al. (1992) elaborated methods for calculating the exact distributions and significance levels of goodness of fit statistics that can be evaluated in polynomial time. Asymptotically normal approximation of the chi-squared test valid for very large number of observations such that $n \rightarrow \infty$, $n/r \rightarrow \alpha$ was considered by Tumanyan (1956) and Holst (1972). Haberman (1988) noted that if some expected frequencies become too small and one does

not use equiprobable cells, then Pearson's test can be biased. Mann and Wald (1942) and Cohen and Sackrowitz (1975) proved that Pearson's chi-squared test will be unbiased if one uses equiprobable cells. Other tests, including modified chi-squared tests, can be biased as well. Concerning selecting category boundaries and the number of classes in chi-squared goodness of fit tests, one may refer to Williams (1950), the review of Kallenberg et al. (1985) and the references cited therein, Bajgier and Aggarwal (1987) and Lemeshko and Chimitova (2003). Ritchey (1986) showed that an application of the chi-squared goodness of fit test with equiprobable cells to daily discrete common stock returns fails, and so suggested a test based on a set of intervals defined by centered approach.

Even after Fisher's clarification, many statisticians thought that while using Pearson's test one may use estimators (such as MLEs) based on non-grouped (raw) data. Chernoff and Lehmann (1954) showed that replacing the unknown parameters in (1.1) by their MLEs based on non-grouped data would dramatically change the limiting distribution of Pearson's sum. In this case, it will not follow a chi-squared distribution and that, in general, it may depend on the unknown parameters and consequently cannot be used for testing. In our opinion, what is difficult to understand for those who use chi-squared tests is that an estimate is a realization of a random variable with its own probability distribution and that a particular estimate can be quite far from the actual unknown value of a parameter or parameters. This misunderstanding is rather typical for those who apply both parametric and nonparametric tests for compound hypotheses (Orlov, 1997). Erroneous use of Pearson's test under such settings is reproduced even in some recent textbooks; see, for example, Clark (1997, p. 273) and Weiers (1991, p. 602). While Chernoff and Lehmann (1954) derived their result considering grouping cells to be fixed, Roy (1956) and Watson (1958, 1959) extended their result to the case of random grouping intervals. Molinari (1977) derived the limiting distribution of Pearson's sum if moment-type estimators (MMEs) based on raw data are used, and like in the case of MLEs, it depends on the unknown parameters. Thus, the problem of deriving a test statistic whose limiting distribution will not depend on the parameters becomes of interest. Roy (1956) and Watson (1958) (also see Drost, 1989) suggested using Pearson's sum for random cells. Dahiya and Gurland (1972a) showed that, for location and scale families with properly chosen random cells, the limiting distribution of Pearson's sum will not depend on the unknown parameters, but only on the null hypothesis. Being distribution-free, such tests can be used in practice, but the problem is that for each specific null distribution, one has to evaluate the corresponding critical values. Therefore, two different ways of constructing distribution-free Pearson-type tests are: (i) to use proper estimates of the unknown parameters (e.g. based on grouped data) and (ii) to use specially constructed grouping intervals. Yet another way is to modify Pearson's sum such that its limiting distribution would not depend on the unknown parameters. Roy (1956), Moore (1971), and Chibisov (1971)

obtained a very important result which showed that the limiting distribution of a vector of standardized frequencies with any efficient estimator (such as the MLE or the best asymptotically normal (BAN) estimator) instead of the unknown parameter would be multivariate normal and will not depend on whether the boundaries of cells are fixed or random. Nikulin (1973c), by using this result and a very general theoretical approach (nowadays known as Wald's method; see Moore (1977)) solved the problem completely for any continuous or discrete probability distribution if one uses grouping intervals based on predetermined probabilities for the cells (a detailed derivation of this result is given in Greenwood and Nikulin (1996, Sections 12 and 13)). A year later, Rao and Robson (1974), by using much less general heuristic approach, obtained the same result for a particular case of the exponential family of distributions. Formally, their result is that

$$Y1_n^2(\hat{\theta}_n) = X_n^2(\hat{\theta}_n) + \mathbf{V}^{(n)T}(\hat{\theta}_n)\mathbf{B}_n(\mathbf{J}_n - \mathbf{J}_{gn})^{-1}\mathbf{B}_n^T\mathbf{V}^{(n)}(\hat{\theta}_n), \quad (1.4)$$

where \mathbf{J}_n and $\mathbf{J}_{gn} = \mathbf{B}_n^T\mathbf{B}_n$ are estimators of Fisher information matrices for non-grouped and grouped data, respectively. Incidentally, this result is Rao and Robson (1974) and Nikulin (1973c). The statistic in (1.4) can also be presented as (see Nikulin, 1973b,c; Moore and Spruill, 1975; Greenwood and Nikulin, 1996)

$$Y1_n^2(\hat{\theta}_n) = \mathbf{V}^{(n)T}(\hat{\theta}_n)(\mathbf{I} - \mathbf{B}_n\mathbf{J}_n^{-1}\mathbf{B}_n^T)^{-1}\mathbf{V}^{(n)}(\hat{\theta}_n). \quad (1.5)$$

The statistic in (1.4) or (1.5), suggested first by Nikulin (1973a) for testing the normality, will be referred to in the sequel as Nikulin-Rao-Robson (NRR) test (Voinov and Nikulin, 2011). Nikulin (1973a,b,c) assumed that only efficient estimates of the unknown parameters (such as the MLEs based on non-grouped data or BAN estimates) are used for testing. Spruill (1976) showed that in the sense of approximate Bahadur slopes, the NRR test is uniformly at least as efficient as Roy (1956) and Watson (1958) tests. Singh (1987) showed that the NRR test is asymptotically optimal for linear hypotheses (see Lehmann, 1959, p. 304) when explicit expressions for orthogonal projectors on linear subspaces are used. Lemeshko (1998) and Lemeshko et al. (2001) suggested an original way of taking into account the information lost due to data grouping. Their idea is to partition the sample space into intervals that maximize the determinant of Fisher information matrix for grouped data. Implementation of the idea to NRR test showed that the power of the NRR test became superior. This optimality is not surprising because the second term in (1.4) depends on the difference between the Fisher information matrices for grouped and non-grouped data that possibly takes the information lost into account (Voinov, 2006). A unified large-sample theory of general chi-squared statistics for tests of fit was developed by Moore and Spruill (1975).

Hsuan and Robson (1976) showed that a modified statistic would be quite different in case of moment-type estimators (MMEs) of unknown parameters. They succeeded in deriving the limiting covariance matrix for standardized