SMOOTH TESTS OF GOODNESS OF FIT
SMOOTH TESTS OF GOODNESS OF FIT
USING R
SECOND EDITION

J.C.W. Rayner
University of Newcastle, Australia

O. Thas
Ghent University, Belgium

D.J. Best
University of Newcastle, Australia

John Wiley & Sons (Asia) Pte Ltd
My thanks to my colleagues, the young man from the old world and the older man from the new, for the immense pleasure it has been working with both. And my thanks and love to Carol, as always, to our sons, Glen and Eric, and to our parents, all now gone, but certainly not forgotten.

J.C.W. Rayner

To Ingeborg, and to my parents.

O. Thas

To Helen, Rohan, Warwick, Jo, Matthew and Lilly

D.J. Best
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>xiii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The Problem Defined</td>
<td>1</td>
</tr>
<tr>
<td>1.2 A Brief History of Smooth Tests</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Monograph Outline</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Examples</td>
<td>10</td>
</tr>
<tr>
<td>2 Pearson’s $X^2$ Test</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Foundations</td>
<td>17</td>
</tr>
<tr>
<td>2.3 The Pearson $X^2$ Test – an Update</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1 Notation, Definition of the Test, and Class Construction</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 Power Related Properties</td>
<td>21</td>
</tr>
<tr>
<td>2.3.3 The Sample Space Partition Approach</td>
<td>24</td>
</tr>
<tr>
<td>2.4 $X^2$ Tests of Composite Hypotheses</td>
<td>26</td>
</tr>
<tr>
<td>2.5 Examples</td>
<td>27</td>
</tr>
<tr>
<td>3 Asymptotically Optimal Tests</td>
<td>33</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>33</td>
</tr>
<tr>
<td>3.2 The Likelihood Ratio, Wald, and Score Tests for a Simple Null Hypothesis</td>
<td>34</td>
</tr>
<tr>
<td>3.3 The Likelihood Ratio, Wald and Score Tests for Composite Null Hypotheses</td>
<td>38</td>
</tr>
<tr>
<td>3.4 Generalized Score Tests</td>
<td>47</td>
</tr>
<tr>
<td>4 Neyman Smooth Tests for Simple Null Hypotheses</td>
<td>53</td>
</tr>
<tr>
<td>4.1 Neyman’s $\Psi^2$ test</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Neyman Smooth Tests for Uncategorized Simple Null Hypotheses</td>
<td>55</td>
</tr>
<tr>
<td>4.3 The Choice of Order</td>
<td>59</td>
</tr>
<tr>
<td>4.4 Examples</td>
<td>61</td>
</tr>
<tr>
<td>4.5 EDF Tests</td>
<td>63</td>
</tr>
</tbody>
</table>
5 Categorized Simple Null Hypotheses 65
  5.1 Smooth Tests for Completely Specified Multinomials 65
  5.2 $X^2$ Effective Order 69
  5.3 Components of $X^2_p$
    5.3.1 Construction of the Components 71
    5.3.2 Power Study 72
    5.3.3 Diagnostic Tests 75
    5.3.4 Cressie and Read Tests 75
  5.4 Examples 76
  5.5 Class Construction 81
    5.5.1 The Alternatives 82
    5.5.2 Results of the Simulation Study 85
    5.5.3 Discussion 88
  5.6 A More Comprehensive Class of Tests 89
  5.7 Overlapping Cells Tests 91

6 Neyman Smooth Tests for Uncategorized Composite Null Hypotheses 95
  6.1 Neyman Smooth Tests for Uncategorized Composite Null Hypotheses 95
  6.2 Smooth Tests for the Univariate Normal Distribution 102
    6.2.1 The Construction of the Smooth Test 102
    6.2.2 Simulation Study 103
    6.2.3 Examples 105
    6.2.4 Relationship with a Test of Thomas and Pierce 109
  6.3 Smooth Tests for the Exponential Distribution 109
  6.4 Smooth Tests for Multivariate Normal Distribution 112
  6.5 Smooth Tests for the Bivariate Poisson Distribution 122
    6.5.1 Definitions 122
    6.5.2 Score Tests for the Bivariate Poisson Model 123
    6.5.3 A Smooth Covariance Test 126
    6.5.4 Variance Tests 127
    6.5.5 A Competitor for the Index of Dispersion Test 128
    6.5.6 Revised Index of Dispersion and Crockett Tests 130
  6.6 Components of the Rao–Robson $X^2$ Statistic 134

7 Neyman Smooth Tests for Categorized Composite Null Hypotheses 137
  7.1 Neyman Smooth Tests for Composite Multinomials 137
  7.2 Components of the Pearson–Fisher Statistic 142
  7.3 Composite Overlapping Cells and Cell Focusing $X^2$ Tests 144
  7.4 A Comparison between the Pearson–Fisher and Rao–Robson $X^2$ Tests 147

8 Neyman Smooth Tests for Uncategorized Composite Null Hypotheses: Discrete Distributions 151
  8.1 Neyman Smooth Tests for Discrete Uncategorized Composite Null Hypotheses 151
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2 Smooth and EDF Tests for the Univariate Poisson Distribution</td>
<td>155</td>
</tr>
<tr>
<td>8.2.1 Definitions</td>
<td>155</td>
</tr>
<tr>
<td>8.2.2 Size and Power Study</td>
<td>157</td>
</tr>
<tr>
<td>8.2.3 Examples</td>
<td>160</td>
</tr>
<tr>
<td>8.3 Smooth and EDF Tests for the Binomial Distribution</td>
<td>163</td>
</tr>
<tr>
<td>8.3.1 Definitions</td>
<td>163</td>
</tr>
<tr>
<td>8.3.2 Size and Power Study</td>
<td>165</td>
</tr>
<tr>
<td>8.3.3 Examples</td>
<td>169</td>
</tr>
<tr>
<td>8.4 Smooth Tests for the Geometric Distribution</td>
<td>170</td>
</tr>
<tr>
<td>8.4.1 Definitions</td>
<td>170</td>
</tr>
<tr>
<td>8.4.2 Size and Power Study</td>
<td>171</td>
</tr>
<tr>
<td>8.4.3 Examples</td>
<td>175</td>
</tr>
<tr>
<td>9 Construction of Generalized Smooth Tests: Theoretical Contributions</td>
<td>179</td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>179</td>
</tr>
<tr>
<td>9.2 Smooth Test Statistics with Informative Decompositions</td>
<td>180</td>
</tr>
<tr>
<td>9.2.1 Sufficient Condition for ‘Convenient’ Test Statistics</td>
<td>180</td>
</tr>
<tr>
<td>9.2.2 Testing for an Exponential Family of Distributions</td>
<td>181</td>
</tr>
<tr>
<td>9.2.3 Testing for Distributions not from an Exponential Family</td>
<td>183</td>
</tr>
<tr>
<td>9.3 Generalized Smooth Tests with Informative Decompositions</td>
<td>183</td>
</tr>
<tr>
<td>9.3.1 Uncategorized Distributions</td>
<td>183</td>
</tr>
<tr>
<td>9.3.2 Categorized Distributions</td>
<td>186</td>
</tr>
<tr>
<td>9.3.3 A Note on the Efficient Score Test</td>
<td>187</td>
</tr>
<tr>
<td>9.4 Efficiency</td>
<td>187</td>
</tr>
<tr>
<td>9.5 Diagnostic Component Tests</td>
<td>189</td>
</tr>
<tr>
<td>9.5.1 Are Smooth Tests and Their Components Diagnostic?</td>
<td>189</td>
</tr>
<tr>
<td>9.5.2 Properly Rescaled Tests</td>
<td>190</td>
</tr>
<tr>
<td>9.5.3 Rescaling Outside Exponential Families</td>
<td>191</td>
</tr>
<tr>
<td>9.5.4 A Simulation Study</td>
<td>193</td>
</tr>
<tr>
<td>10 Smooth Modelling</td>
<td>199</td>
</tr>
<tr>
<td>10.1 Introduction</td>
<td>199</td>
</tr>
<tr>
<td>10.2 Model Selection through Hypothesis Testing</td>
<td>201</td>
</tr>
<tr>
<td>10.2.1 Forward Selection and Backward Elimination</td>
<td>201</td>
</tr>
<tr>
<td>10.2.2 Smooth Tests for Improved Models</td>
<td>202</td>
</tr>
<tr>
<td>10.2.3 Examples</td>
<td>203</td>
</tr>
<tr>
<td>10.3 Model Selection Based on Loss Functions</td>
<td>206</td>
</tr>
<tr>
<td>10.3.1 Loss Functions and Expected Loss</td>
<td>206</td>
</tr>
<tr>
<td>10.3.2 AIC and BIC</td>
<td>208</td>
</tr>
<tr>
<td>10.4 Goodness of Fit Testing after Model Selection</td>
<td>211</td>
</tr>
<tr>
<td>10.4.1 Motivation</td>
<td>211</td>
</tr>
<tr>
<td>10.4.2 Theory</td>
<td>212</td>
</tr>
<tr>
<td>10.4.3 Examples</td>
<td>214</td>
</tr>
<tr>
<td>10.4.4 A Final Note</td>
<td>218</td>
</tr>
<tr>
<td>10.5 Correcting the Barton Density</td>
<td>218</td>
</tr>
</tbody>
</table>
Preface

Preface to the First Edition

The importance of probabilistic or statistical modeling in the modern world cannot be overrated. With the advent of high-speed computers, complex models for important processes can now be constructed and implemented. These models and the associated statistical analyses are of great assistance in making decisions in diverse fields, from marketing, medicine, and management, to politics, weapons systems, and food science. *Goodness of fit* is concerned with assessing the validity of models involving statistical distributions, an essential and sometimes forgotten aspect of the modeling exercise. One can only speculate on how many wrong decisions are made due to the use of an incorrect model.

Karl Pearson pioneered goodness of fit in 1900, when his paper introducing the $X^2$ test appeared. Since then, perhaps reflecting the needs and importance of the subject, a great many new tests have been constructed. The *smooth tests* are a class of goodness of fit tests that are informative, easy to apply, and generally applicable. Typically they can be derived as score tests, and hence are, in a sense, optimal for large sample sizes. For moderate and small samples they are very competitive in the cases we have examined. Pearson’s $X^2$ test is in fact a smooth test. In the formulation we prefer, components with simple graphic interpretations are readily available. We suggest that the properties of the smooth tests are such that a new goodness of fit test must be in some way superior to a corresponding smooth test if it is to be worthy of consideration.

This book is complementary to that by D’Agostino and Stephens (*Goodness of Fit Techniques*, 1986) in that they do not cover the smooth tests in any detail, while we do not cover in detail topics such as tests based on the empirical distribution function, and tests based on regression and correlation. There is some overlap in the coverage of $X^2$ tests. The tests that they discuss and our smooth tests are in competition with each other. We give some comparisons, and, not surprisingly, recommend use of the smooth tests. Usually, the smooth tests are more informative than their competitors. The D’Agostino and Stephens book covers a broad range of topics, generally omitting mathematical details and including many tables and examples so that it reads as a handbook of methods. Since our book concentrates on smooth methods, we have been able to present derivations and mathematical details that might have been omitted in a more comprehensive treatment of goodness of fit in its entirety. We consider this to be highly desirable because the development of the smooth tests of fit is far from complete. Indeed, we hope that researchers will read this book and be motivated to help with its further development.
In spite of this last statement, many economists, scientists and engineers who have taken an introductory mathematical statistics course, to the level of Larsen and Marx (1981), will be able to read this book. The more technical details are clearly signposted and are in sections that may be omitted or skimmed. Undergraduates with a sufficient background in statistics and calculus should be able to absorb almost everything. Practical examples are given to illustrate use of the techniques.

The smooth tests for the uniform distribution were introduced by Neyman (1937), but they were slow to gain acceptance because the computations are heavy by hand. This is no longer a barrier. Many of the techniques we discuss are readily implemented on modern computers, and we give some algorithms to assist in doing this. When used in conjunction with density estimate plots or Q–Q plots, the smooth tests can play an important part in many analysis.

Chapter 1 outlines the goodness of fit problem, gives a brief history of the smooth tests, outlines the monograph, and gives some examples of the sort of problems that arise in practice. A review of Pearson (1900), and an outline of the early development of the tests for simple and composite hypotheses is given in Chapter 2. In Chapter 3, tests that are asymptotically optimal are introduced; these include the score tests that are particularly important later in the book. Using score tests and smooth models, tests of completely specified null hypotheses are derived in Chapters 4 and 5. These chapters cover both uncategorized (discrete or continuous) and categorized null distributions. The tests are essentially tests for uniformity. Then, in Chapters 6 and 7, we consider tests for composite null hypotheses, again treating both the categorized and uncategorized cases. Chapters 4 to 7 emphasize the components our tests yield. In Chapter 6 we look at tests for the univariate and later for the multivariate normal, the Poisson, geometric and exponential distributions. These are extracted from a class of smooth goodness of fit tests. In Chapter 7, we discuss $X^2$ statistics for composite hypotheses. We conclude with a review and an examination of some of the other uses to which our techniques may be put.

Our interest in the subject of goodness of fit came about partly from questions from colleagues relating to the ‘Gaussian’ assumption in routine statistical analyses, and partly from work J. C. W. R. had begun as a student of H. O. Lancaster. Our approach is based on the use of orthonormal functions, emphasized in Lancaster (The Chi-Squared Distribution, 1969), and on the use of score statistics and generalizations of the smooth families suggested by Thomas and Pierce (1979) and Kopecky and Pierce (1979) in articles published in The Journal of the American Statistical Association.

_Dunedin, New Zealand_  
_J. C. W. R_  
_North Ryde, Australia_  
_D. J. B_
Preface to the Second Edition

We have embarked upon a second edition for various reasons. The first edition was well received, and since then goodness of fit testing has moved on, as has our own contribution to it. We believe that modelling, and assessing models, is of increasing importance. While we focus on constructing smooth tests for various classes of distributions, and particular distributions within those classes, the approach we advocate is of wider relevance. We hope readers will apply at least aspects of our approach to the development and assessment of the increasingly complex models that will form an essential part of understanding today’s world and the future beyond.

We are joined in producing this edition by Olivier Thas, whose youth and skills complement the intuition and experience of the two Johns. One of the most obvious benefits of the collaboration is that Olivier will maintain a web site with helpful software, information and new developments. This can be found at the following URL: www.wiley.com/go/rayner. Over time the URL may change, but modern search engines should have no trouble finding the site.

What is new in this edition? The book is permeated by a considerable updating of the literature and a number of important new topics. So we include a subsection outlining tests based on sample space partitions, work developed in and after Olivier’s PhD thesis. The development of data-driven tests is perhaps the most significant new development in smooth testing in the last 15 years, since it gives a solution to the problem of choosing the order of the smooth test. Rayner et al. (1995) made a contribution to the interpretation of smooth analysis. This was followed by a series of related papers from Henze and colleagues.

There is a new section in the ‘tools’ chapter, Chapter 3, on generalized score tests. Use of this tool enables us to develop a new class of smooth tests in Chapter 9, and hence to develop smooth tests for the logistic and other distributions.

Chapter 5 includes new material from the senior authors, published in the 1990s.

The old Chapter 6 has been divided into two Chapters: 6 and 8. The former includes new work on the bivariate Poisson that is of some interest because whether or not a user chooses to apply the smooth tests developed there, these tests enable existing tests to be better understood. The material on the Poisson, binomial and geometric distributions in the first edition needed to be updated, but that has been done using an idea from the tests developed by Chernoff and Lehmann (1954). Since this material is rightly developed in Chapter 7 on Neyman smooth tests for categorized composite null hypotheses, the Poisson, binomial and geometric distributions had to wait until after that chapter.

The smooth test statistics we derived in the first edition were all sums of squares of easily interpreted and powerful components that asymptotically are independent, and asymptotically have the standard normal distribution. However, after publication of the first edition we became aware of the fact that in all cases we were testing for distributions from exponential families of distributions, and matters were not nearly so convenient when testing for distributions not from exponential families. Moreover, the interpretation of our components is a little more involved than we first thought. These matters are resolved in this edition, and the tools for doing so are given in Chapters 9 and 10.

To obtain convenient components requires the use of generalized score tests to produce generalized smooth tests. This is done in Chapter 9. There we also consider questions of the efficiency of various possible components, and of the interpretation of the components.
In Chapter 10 we look at smooth testing in the light of model selection methods. The point is that if we test for a particular parametric model, we usually want to know more than that the model is rejected, or more than the model is rejected because certain moments of the data disagree with the model proposed. Usually we would like to know at least one model that is consistent with the data.

The new methods are pulled together in Chapter 11, where they are applied to produce generalized smooth tests for the logistic, Laplace, extreme value, negative binomial, zero-inflated Poisson and generalized Pareto distributions. The focus is on outlining the tests in a clear and informative way, and demonstrating their application to interesting data sets.

Our thanks to Paul Rippon, for making some of the text much clearer than it would have been otherwise and to Helen Best, who has helped with proof reading over many years. We also wish to thank Bert De Boeck for his contributions to the R package.

Our thanks to the University of Newcastle, including the Faculty of Science and Information Technology and the Centre for Complex Dynamic Systems and Control, for a New Staff Grant and for supporting leave through the Special Studies Program. The support, in both time and funding, enabled Olivier Thas to visit Newcastle, and John Rayner to visit Olivier in Antwerp. This enabled us to finish this second edition much sooner than otherwise would have been possible.

We are also grateful to Marie-Rose Van Cauwenberghe for her help with the conversion of the Word files of the first edition of the book to \LaTeX, which served as the basis for this edition.

J.C.W. Rayner  
*Newcastle, Australia*

O. Thas  
*Gent, Belgium*

D.J. Best  
*Carlingford, Australia*
1

Introduction

1.1 The Problem Defined

A number of statistical or probabilistic distribution models have found widespread use in business and commerce, law, science, medicine and engineering. This monograph is concerned with the assessment of the goodness of fit of such models, or seeing how well the data agree with the proposed distributional model.

One informal way of doing this is to draw a histogram of the frequency distribution of the data and judge the closeness of the histogram and the theoretical probability function ‘by eye’. Such graphical checks are subjective and cannot be recommended to assess goodness of fit on their own. The hypothesis tests we recommend may be complemented by graphical methods.

There are many examples of the use of the models we discuss. Many of the commonly used methods of statistical inference, such as t tests, determination of p-values and least significant differences, assume a normal or Gaussian distribution. The method of least squares is a common estimation technique but has optimum properties only when normality is assumed. More specific examples follow.

1. Quality control rules for the import and export of various foods often assume a normal distribution.
2. Safety limits for extreme rainfall used by hydrologists involved in flood control may assume a lognormal distribution.
3. Estimates of bacteria in sewerage may be based on an exponential distribution.

Failure of the distributional assumptions means failure of the model. The conclusions based on the model are then invalid. This possibility of failure can be objectively assessed by a goodness of fit test.

A number of previous books have recognized the importance of goodness of fit tests by devoting a section or chapter to the topic. For example, see Tiku et al. (1986, Chapter 6), Chambers et al. (1983, Chapter 6), Lawless (1982, Chapter 9), Shapiro and Gross (1981, Chapter 6), Gnanadesikan (1977, Chapter 5) and Stuart et al. (2004, Chapters 25 and 30). However, these treatments are somewhat limited in scope. The books of Read and Cressie (1988) and
Thode (2002) are devoted to generalizations of Pearson’s chi squared test and to testing for normality, respectively. D’Agostino and Stephens (1986) gave a much more comprehensive coverage. None offers the coherent approach of this monograph, which concentrates on one class of tests and develops both theory and applications. We also give more coverage to categorized models.

Without a statistical test, goodness of fit can only be assessed by visual subjective methods. Fisher (1925, p. 36) in his influential text, *Statistical Methods for Research Workers*, devoted large sections of his Chapters III and IV to goodness of fit and commented on the then common model assessment method:

> No eye observation of such diagrams, however experienced, is really capable of discriminating whether or not the observations differ from the expectation by more than we would expect from the circumstances of random sampling.

Kempthorne (1967) considered goodness of fit to be the ‘classical problem of statistical inference’. What then is a goodness of fit test? According to David (1966, p. 399)

> A goodness of fit procedure is a statistical test of a hypothesis that the sampled population is distributed in a specific way. . . for example, that the sampled population is normal.

This is the one-sample problem; the corresponding $S$-sample problem assesses whether or not $S$ independent random samples come from the same population. For a smooth treatment of this problem, see Rayner and Best (2001, Sections 9.4 and 9.5).

Subsequently we shall mainly be concerned with one-sample tests for goodness of fit. Formally, given a random sample $X_1, X_2, \ldots, X_n$, we test the null hypothesis that the sampled population has cumulative distribution function $F(x; \theta)$, $\theta \in \Theta$, against the alternative hypothesis that the cumulative distribution function is $G(x; \omega)$, $\omega \in \Omega$. All of $X$, $\Theta$ and $\Omega$ may be multidimensional. Frequently the alternative is simply ‘not the null hypothesis’.

What do we get from having applied a goodness of fit test? First, a compact description of the data. Saying that the data are binomial with parameters $n = 15$ and $p = 0.51$ is a valuable abbreviation of the available information. Second, powerful parametric procedures, such as the tests in the analysis of variance, are valid if the data are consistent with normality. And third, light may be shed on the mechanisms generating the data. For example, if the data cannot be viewed as a Poisson process, then we can expect that at least one of the axioms sufficient for a Poisson process has failed. For example, if lifetimes for cancer patients from the onset of ‘standard’ treatment have been exponentially distributed with mean 36 months in the past, and this distribution no longer holds under a new treatment, what has changed? It could be that either the mean only, or the distribution has changed in a more general way. In the latter case perhaps the new treatment is less effective than the standard treatment for some, and they die sooner than under the standard treatment; and the new treatment is apparently effective for others, who survive longer than previously.

What a goodness of fit test tells us is important, but so is what it does not tell us! Geary (1947) said that ‘Normality is a myth; there never was, and never will, be a normal distribution’. Strongly put perhaps, but given enough observations of virtually any generating mechanism, we could probably reject any specified hypothesis. As for normality, we do not observe
Introduction

arbitrarily large (or small) data; and as all data are rounded, we should ultimately be able to reject any continuous model. But although a distributional model may not hold precisely, it may hold sufficiently well for the three purposes outlined above. The important question is, are our data sufficiently well approximated by the distribution for which we test?

Some data sets are summarized by the sample mean and the sample standard deviation. This assumes a normal distribution, or at least a distribution that is completely determined by the mean and standard deviation. If the data were thought to be Poisson, then it would be sufficient just to quote the mean. But of course, in such cases the distribution should be assessed by a goodness of fit test.

Should the common tests of statistical inference, such as the $t$ test and the analysis of variance, be avoided by the use of more robust, distribution-free or non-parametric procedures? The latter minimize distributional assumptions and, at times, this minimization is a wise course to follow. However, in some cases, not using parametric tests can result in the use of inferior tests. We suggest that goodness of fit tests and other checks on the data should be employed before opting for robust or distribution-free techniques.

This leads to a difficulty. If a preliminary test of the assumptions for a parametric test is performed, does this affect the inferences made? We agree with Cox (1977) who proposed:

A combination of preliminary inspection of the data together with study at the end of the analysis of whether there are aspects of the data and assumptions reconsideration of which might change the qualitative conclusions.

We interpret this as meaning the parametric test is inapplicable if the distributional assumptions are not satisfied, so there is no need to incorporate the results of a goodness of fit test formally. The fact that a goodness of fit test is applied formally does not mean it is not, under some circumstances, part of the preliminary inspection of the data.

A distinction can be drawn between globally omnibus and directional tests. Globally omnibus tests are intended to have moderate power against all alternatives; directional tests are intended to detect specified alternatives well. Of course, against the specified alternatives, the directional tests are constructed to be more powerful than the globally omnibus tests, while against all other alternatives the globally omnibus tests should be superior. Consider the analogy of a search party. If all searchers are concentrated in a small target area, there is a much better chance of finding whatever is lost, provided it is in the target area; clearly there is no chance of detection in the balance of the search area. On the other hand, if the searchers are spread over the entire area, there is a reduced chance of detection in the target area, and an increased chance of detection in the balance of the search area. The smooth tests are constructed to be partially omnibus tests (between globally omnibus and directional), but their components provide powerful directional tests. See the more technical discussion in Section 9.5.

Finally, we mainly discuss formal statistical tests of significance. This is not to say that subjective methods are not valuable. Graphical methods may lead to insights that are not apparent otherwise, and methods such as quantile–quantile (Q–Q) plots or density estimates should be used alongside those we discuss here.

We now turn to a brief history of smooth goodness of fit tests.
1.2 A Brief History of Smooth Tests

Perhaps the most widely known test in statistical inference is Pearson’s \( X^2 \) goodness of fit test. An informal definition follows. Suppose observations may fall into \( m \) non-overlapping classes or cells. We hypothesize the cells should contain respectively \( E_1, \ldots, E_m \) observations, but the observed cell counts are \( O_1, \ldots, O_m \). Now define the Pearson test statistic by

\[
X^2_P = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j}.
\]

If this is larger than the 100\( \alpha \)% point of the \( \chi^2_{m-1} \) distribution then the hypothesized expectations can be rejected at the 100\( \alpha \)% level of significance. In particular, we use the following convention. If \( a_{m-1}(\alpha) \) is the 100\( \alpha \)% point of \( \chi^2_{m-1} \), then \( P \left( X^2_P > a_{m-1}(\alpha) \right) = \alpha \). The \( X^2_P \) test is more formally defined in Section 2.2. Pearson’s test is applicable for testing discrete data when there are no parameters that need to be estimated. The expansion of the methodology to cover more practical situations has occupied statisticians almost continuously since Karl Pearson introduced his \( X^2 \) test in 1900. In the next chapter we will devote some time to reviewing Pearson (1900) and the developments in \( X^2 \)-type tests. It is not widely known that Pearson’s test is a smooth test, but later in Chapter 5 we will demonstrate that this is the case.

According to Barton (1956) and Neyman (1937) himself, Neyman’s smooth test was developed to overcome presumed deficiencies in Pearson’s \( X^2 \) test. The test was called ‘smooth’ because it was constructed to have good power against alternatives whose probability density functions depart ‘smoothly’ from that specified by the null hypothesis. For example, the null hypothesis may specify the normal distribution with zero mean and unit variance, while the alternative may specify the normal distribution with small positive mean and unit variance. Smooth changes include slight shifts in mean, variance, skewness and kurtosis. See, for example, Figure 1.1. Data and analysis that underpin this figure are given in Example 1.4.4 later in this chapter.

Suppose we have a random sample from a continuous distribution with completely specified cumulative distribution function \( F(x) \). Applying the probability integral transformation, the null hypothesis \( H_0 \) specifies that \( Y = F(X) \) is uniformly distributed on \((0, 1)\). Neyman’s smooth alternative of order \( k \) to \( H_0 \), where \( k \) is integral, has probability density function

\[
g_k (y; \theta) = \exp \left\{ \sum_{i=1}^{k} \theta_i h_i (y) - K (\theta) \right\}, \quad 0 < y < 1, \tag{1.1}
\]

where \( \theta^T = (\theta_1, \ldots, \theta_k) \), \( K(\theta) \) is a normalizing constant, and the \( \{h_i(y)\} \) are orthonormal polynomials related to the Legendre polynomials. The first five such polynomials are:

\[
h_0(y) = 1,
\]
\[
h_1(y) = (2y - 1)\sqrt{3},
\]
\[
h_2(y) = (6y^2 - 6y + 1)\sqrt{5},
\]
Figure 1.1  Density of the fitted normal distribution (solid line) and the fitted improved density (dashed line) for the Mississippi river data (Example 1.4.4 in Section 1.4).

\[
\begin{align*}
  h_3(y) &= (20y^3 - 30y^2 + 12y - 1)\sqrt{7}, \\
  h_4(y) &= 3(70y^4 - 140y^3 + 90y^2 - 20y + 1).
\end{align*}
\]

Orthonormality is defined in Section 4.2 and in Appendix A.

The \( \{h_i(y)\} \) are constructed so that \( h_r(y) \) is of degree \( r \) and the \( \{h_i(y)\} \) is orthonormal on \((0, 1)\). See for example, Kendall and Stuart (1973, p. 444). To test the null hypothesis \( H_0 : \theta_1 = \ldots = \theta_k = 0 \), we use the Neyman statistic, given by

\[
\Psi_k^2 = \sum_{i=1}^{k} U_i^2 \quad \text{in which} \quad U_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_i(Y_j).
\]

The \( U_i \) are called components of \( \Psi_k^2 \).

Neyman’s conception for his smooth test was that it should be constructed to be locally most powerful, unbiased and of size \( \alpha \) for testing for uniformity against the order \( k \) alternative given by equation (1.1). Its power function was also constrained to be symmetric, depending on \( \theta \) only through \( \theta_1^2 + \ldots + \theta_k^2 \). Neyman (1937) noted that his solution is only approximate; only asymptotically is the test of size \( \alpha \), unbiased and most powerful. A detailed account of Neyman (1937) is given in Section 4.1.

Barton (1953, 1955, 1956) extended Neyman’s work. He used probability density functions asymptotically equivalent to \( g_{\varepsilon}(y; \theta) \). For example in Barton (1953) he used the probability
density functions

\[ g_k^*(y; \theta) = 1 + \sum_{i=1}^{k} \theta_i h_i(y), \quad 0 < y < 1. \]

His 1956 paper dealt with probability density functions involving nuisance parameters, but the statistic derived had an inconvenient distribution. As Kopecky and Pierce (1979) pointed out, the quadratic score statistic (see Chapter 3) has a more convenient distribution.

An interesting but little known result is that the Pearson $X^2$ test is a categorized form of the Neyman–Barton tests. Suppose a multinomial with $g$ classes is specified by the null hypothesis. Barton (1955) considered order $k$ alternatives of the form $g_k^*(y; \theta)$, but with the polynomials $h_r(y)$ replaced by an orthonormal system on the multinomial distribution. He then defined a statistic $B(g, k)$ that as $g \to \infty$, (i) approached $\Psi_k^2$, (ii) tended to be distributed as $\chi^2_k$, and (iii) was optimal in the limit. Moreover, $B(k + 1, k)$ was shown to be equivalent to the Pearson test statistic based on $k + 1$ classes. The importance of this result is that the Pearson $X^2$ test with $k + 1$ cells can be expected to have good power properties against order $k$ alternatives, especially for a moderate to large number of classes, when it will be very similar to the optimal $\Psi_k^2$. Kendall and Stuart (1973, p. 44) reviewed this material and showed that the $B(k + 1, k)$ may be obtained by partitioning the Pearson test statistic. This idea is taken up again in Chapter 4.

Watson (1959) extended a result of Barton (1956), and Hamdan (1962, 1963, 1964) considered smooth tests for various simple null hypotheses. He used the Hermite–Chebyshev polynomials to construct a test for the standard normal distribution, and an orthonormal set on the multinomial and the Walsh functions to construct tests for the uniform distribution.

These tests aroused little interest. They required computations that by hand would be considered heavy, and could not deal practically with the main interest in applications, composite null hypotheses. So it was not until the papers of Thomas and Pierce (1979) and Kopecky and Pierce (1979) that Neyman-type tests received much further attention.

Rather than work with orthogonal polynomials, Thomas and Pierce (1979) defined an order $k$ probability density function by

\[
\exp \left\{ \sum_{i=1}^{k} \theta_i y_i^j - K(\theta) \right\}
\]

or, in terms of the null probability density function,

\[
\exp \left\{ \sum_{i=1}^{k} \theta_i F_i(x) - K(\theta) \right\} f(x)
\]

where $f(x) = dF(x)/dx$. Their test statistic $W_k^*$ is a quadratic score statistic based on this model. The weak optimality of tests based on the quadratic score statistics is therefore conferred upon the $W_k^*$. 
If the probability density function $f(x)$ involves nuisance parameters, the model for an order $k$ alternative becomes

$$\exp\left\{\sum_{i=1}^{k} \theta_i F_i(x; \beta) - K(\theta; \beta)\right\} f(x; \beta).$$

The quadratic score statistic based on this model is $W_k$, given in detail in Thomas and Pierce (1979, p. 443). In particular, to test for normality with unspecified mean and variance, write $\beta = (\mu, \sigma)^T$, and write $F(x; \beta)$ for the cumulative distribution function. In testing for this distribution Thomas and Pierce suggested the statistics

$$W_1 = \frac{1}{n} \left\{ 16.3172 \sum_{j=1}^{n} \left( Y_j - \frac{1}{2} \right) \right\}^2,$$

$$W_2 = W_1 + \frac{27.3809^2}{n} \left\{ \sum_{j=1}^{n} \left( \frac{Y_j^2}{3} - \frac{1}{3} \right) - \sum_{j=1}^{n} \left( Y_j - \frac{1}{2} \right) \right\}^2,$$

where $Y_j = F(X_j; \hat{\beta})$. The statistics $W_1$ and $W_2$ are asymptotically distributed as $\chi_1^2$ and $\chi_2^2$, respectively. Thomas and Pierce (1979) showed that the small sample distributions are reasonably approximated by the limiting $\chi^2$ distributions.

Tests based on what might be called the Pierce approach include the test of Bargal and Thomas (1983) for the (censored) Weibull and the test of Bargal (1986) and Bargal and Thomas (1983) for the (censored) gamma. Unfortunately, a consequence of using powers instead of orthonormal functions is that tables of constants, such as 16.3172 and 27.3809 in $W_1$ and $W_2$ above, are needed to define the test statistics. This is somewhat offset by the need to know the orthonormal functions in the formulation we prefer. Those orthonormal functions may be obtained from recurrence relations. This is most convenient for computer implementation of the tests. See Appendix A.

Rayner and Best (1986), Koziol (1986, 1987) and Jarque and Bera (1987) all suggested smooth tests for the composite case when the parameters are of location–scale type. Their tests are based on orthonormal functions and are of a slightly simpler form than those of Thomas and Pierce (1979) in that

(i) they involve sums of squares and not quadratic forms,
(ii) numerical integration is not usually needed to specify constants in the test statistic,
(iii) the components are often identifiable with known moment-type statistics used in tests of fit, and
(iv) the components are asymptotically independent.

Given observations $X_1, \ldots, X_n$ from a location–scale distribution with probability density function $f_X(x)$, put $Z = (X - \mu)/\sigma$, where $\mu$ is the location parameter and $\sigma$ is the scale parameter. Suppose $f_Z(z)$ is the standardized probability density function.
Rayner and Best (1986) define the order $k$ alternative to be

$$C(\theta) \exp \left\{ \sum_{i=1}^{k} \theta_i h_i(z) \right\} f_Z(z),$$

(1.2)

where $\{h_i(z)\}$ are orthonormal on $f_Z(z)$, and $C(\theta)$ is a normalizing constant. If the densities $f_X(x)$ belong to an exponential family of distributions the appropriate test statistic is

$$\hat{S}_k = \sum_{i=3}^{k+2} \hat{V}_i^2$$

where $\hat{V}_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_i(\hat{Z}_j)$

in which $\hat{Z}_j = (X_j - \hat{\mu})/\hat{\sigma}$, $j = 1, \ldots, n$, and $\hat{\mu}$ and $\hat{\sigma}$ are maximum likelihood (ML) estimates of $\mu$ and $\sigma$, respectively. Here $\hat{V}_1 = \hat{V}_2 = 0$; the next two components assess, roughly, skewness and kurtosis.

In the $N(\mu, \sigma^2)$ case the orthonormal functions are the normalized Hermite polynomials, and those of order 3 to 6 are:

$$h_3(z) = \frac{1}{\sqrt{6}}(z^3 - 3z),$$
$$h_4(z) = \frac{1}{\sqrt{24}}(z^4 - 6z^2 + 3),$$
$$h_5(z) = \frac{1}{\sqrt{120}}(z^5 - 10z^3 + 15z),$$
$$h_6(z) = \frac{1}{\sqrt{720}}(z^6 - 15z^4 + 45z^2 - 15).$$

The test statistic and the squares of the components $\hat{V}_r$ all have asymptotic $\chi^2$ distributions, so the test is easy to implement. Moreover, the components may be individually informative. Examples are given in Best and Rayner (1985a). This approach can readily be extended beyond location–scale families, and to categorized data and multivariate distributions. This is done in later chapters.

To balance the picture, we should also provide the smooth model that is appropriate in the discrete case. In the absence of nuisance parameters, and if the null distribution specifies $\{p_j\}$, a smooth alternative, clearly analogous to (1.2) above, is

$$\pi_j = C(\theta) \exp \left\{ \sum_{i=1}^{k} \theta_i h_{ij} \right\} p_j, \quad j = 1, \ldots, m.$$  

Here $k$ is the order of the alternative, and the Pearson $X^2$ test results if in the score statistic we take $k = m - 1$ and choose the $h_{ij}$ appropriately. The appeal in this formulation is that if the null hypothesis is rejected, with significant components indicating particular $\theta_i$ non-zero, then $\{\pi_j\}$ above specifies an alternative model. Lack of a suggested alternative hypothesis has been a criticism of $X^2$-type tests. The formulation here, and its composite analogue, are investigated later in Chapters 5, 7 and 8.
An important issue in the application of the smooth tests is the choice of the order $k$. In finite samples, choosing $k$ too large results in a *dilution* effect. Such a test is seeking alternatives in a very general parameter space, and is often too *omnibus*. On the other hand, too small a choice of $k$ leads to a very *directional* test, with good power in a reduced parameter space, but poor power elsewhere.

As a solution to this problem, Ledwina (1994), focusing only on testing for uniformity, suggested incorporating the choice of order into the testing procedure. The resulting test is referred to as a *data-driven* smooth test. The order $k$ of the test is chosen from $\{1, 2, \ldots, d\}$, where $d$ is chosen before sighting the data. Given a random sample of size $n$ from (1.1) it would be natural to chose $k$ to maximize the likelihood from the admissible $\theta$. Since the likelihood cannot decrease by increasing the order, this approach would, however, always result in $k = d$. To avoid this problem Ledwina (1994) proposed maximizing the *Bayesian information criterion* (BIC) of Schwarz (1978), which is a trade-off between the quality of the fit (maximized likelihood) and the complexity of the model (the number of parameters, $k$ in this formulation). The BIC for the order $k$ model is

$$\text{BIC}_k = 2 \log L_k - k \log n,$$

where $\log L_k$ is the maximized logarithm of the likelihood for a random sample of size $n$ from the model (1.1). The optimal order is a random variable, $K$, the smallest order that maximizes the BIC. The data-driven smooth test statistic is now $U_1^2 + \ldots + U_K^2$, but its distribution is no longer asymptotically $\chi^2_K$. Simulation studies show this test is very competitive in terms of power. Kallenberg and Ledwina (1997b) conclude:

“In view of the consistency and the simulation results, we feel that the conclusion of Rayner and Best (1990, p. 9) – ‘don’t use those other methods – use a smooth test!’ – may be slightly sharpened to ‘use a data-driven smooth test.’

Kallenberg and Ledwina (1995) extended the system of orthonormal functions, while Kallenberg and Ledwina (1997b) showed that in BIC$_k$ the logarithm of the likelihood may be replaced by the score statistic. Inglof et al. (1997) and Kallenberg and Ledwina (1997b) extended this approach to composite null hypotheses. We shall give a more detailed discussion of these results in Chapters 4, 6 and 10.

1.3 Monograph Outline

The reader is now acquainted with what goodness of fit tests are, and why they are important. We have sketched the historical development of the smooth tests, and in future chapters we will return to that development in more detail.

In Chapter 2 we begin at the chronological beginning, with Pearson’s $X^2$ test. A review of Pearson (1900) is given, and also of the developments in $X^2$-type tests since then. This is not done from the viewpoint of smooth tests, since they were a later development. Certain $X^2$-type tests are smooth tests, as we have already mentioned. This will be demonstrated in Chapters 5, 7 and 8.

The main approach will be to define smooth models in various situations, and to derive tests that have good properties in large samples for these models. The machinery for doing this is
Smooth Tests of Goodness of Fit

given in Chapter 3, on asymptotically optimal tests. The likelihood ratio, score and Wald tests are introduced for models first without, and second with, nuisance parameters. These tests are asymptotically equivalent, and which is most convenient to apply will vary depending on the situation. In multivariate normal models it is usually the likelihood ratio test. For the smooth models we discuss, it is usually the score test.

Chapters 4 to 8 systematically work through derivations and properties of the score tests for categorized and uncategorized smooth models of the Neyman type, both when nuisance parameters are absent and when they are present. In this monograph uncategorized models will be either discrete or continuous; in categorized models the data are placed into a finite number of cells or classes. This involves some ambiguity, as the binomial, for example, could be treated as either categorized or uncategorized. Particular cases of the tests we derive include tests for the univariate and multivariate normal, exponential, geometric and Poisson distributions. Power studies are given to demonstrate the effectiveness of these tests in small samples. The tests are applied to real data sets.

The smooth tests given up to this point are sums of squares of easily interpreted and powerful components. However, for many distributions the current theory does not produce such convenient tests. It is necessary to move from score tests to generalized score tests. Theory for the resulting smooth tests, which we call the generalized smooth tests, is given in Chapter 9. If the hypothesized model is rejected, the order \( k \) family of alternatives is a natural alternative model for the data. However, modern model selection is deeper than this naive approach, and that is the topic of Chapter 10. The tools developed in Chapters 9 and 10 are implemented in Chapter 11 to develop generalized smooth tests for some non-exponential family distributions, specifically for the zero-inflated Poisson, logistic, extreme value, Laplace, negative binomial and generalized Pareto distributions.

Throughout the monograph we will need to calculate (approximate) \( p \)-values for certain data sets. Given a value of a test statistic for a data set, we usually need the probability of values of the test statistic at least as great as the observed under the null hypothesis. This is done by the parametric bootstrap. Details on the computations are given in Appendix B.

Throughout the monograph we will freely use graphical methods to augment our tests. For more information the reader is directed to Chambers et al. (1983) and to D'Agostino and Stephens (1986, Chapter 2).

1.4 Examples

In this section we give numerical examples demonstrating the use of some of the goodness of fit tests we have briefly discussed. Some will be considered again in more detail later.

Example 1.4.1 (Weldon’s dice data) This data set was discussed by Pearson in his classic 1900 paper. The data are reproduced in Table 1.1, and give the number of occurrences of a 5 or a 6 in 26306 throws of 12 dice. We return to these data again in Example 8.3.1.

If it is assumed that the dice are fair then the null hypothesis \( H_0 \) is that the probability of a 5 or 6 in one throw is 1/3, and so the probability of \( r \) occurrences of a 5 or 6 in 12 throws is given by the binomial probability, \( p_r = \binom{12}{r} \left( \frac{1}{3} \right)^r \left( \frac{2}{3} \right)^{12-r} \). From this probability the expected frequencies of \( r \) occurrences of a 5 or 6 in the 26306 throws of 12 dice can be calculated. To check whether the deviations between observed and expected frequencies are more than could be expected by chance, Pearson’s \( X^2 \) statistic \( X^2_P \) can be calculated. A visual comparison is given in Figure 1.2.
Table 1.1  Weldon’s data. For each number of dice in a cast of 12 that containe
Table 1.1  Weldon’s data. For each number of dice in a cast of 12 that contained a 5 or a 6, the observed
and expected frequencies are given, as well as the difference between observed and expected

<table>
<thead>
<tr>
<th>Number of 5s or 6s</th>
<th>Observed</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>185</td>
<td>203</td>
<td>−18</td>
</tr>
<tr>
<td>1</td>
<td>1149</td>
<td>1217</td>
<td>−68</td>
</tr>
<tr>
<td>2</td>
<td>3265</td>
<td>3345</td>
<td>−80</td>
</tr>
<tr>
<td>3</td>
<td>5475</td>
<td>5576</td>
<td>−101</td>
</tr>
<tr>
<td>4</td>
<td>6114</td>
<td>6273</td>
<td>−159</td>
</tr>
<tr>
<td>5</td>
<td>5194</td>
<td>5018</td>
<td>176</td>
</tr>
<tr>
<td>6</td>
<td>3067</td>
<td>2927</td>
<td>140</td>
</tr>
<tr>
<td>7</td>
<td>1331</td>
<td>1254</td>
<td>77</td>
</tr>
<tr>
<td>8</td>
<td>403</td>
<td>392</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>105</td>
<td>87</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1.2  The histogram of the observed frequencies with * indicating the expected frequencies (left),
and a plot of the difference between the observed and the expected frequencies (right).

Pearson (1900) obtained $X_P^2 = 43.9$ while our calculations give $X_P^2 = 41.3$. The difference
is almost certainly due to Pearson having used expected frequencies rounded to the nearest integer. In either case use of the $\chi^2_{12}$ distribution to obtain $p$-values indicates substantial deviations between observed and expected frequencies. With such a large number of throws of 12 dice this is hardly surprising. Closer inspection of Figure 1.2 indicates that there is a well-defined trend for 5s or 6s to occur more than they should; it appears the dice were slightly biased. Guttorp (1992) gives the following interpretation to these results.

However, there is a simple physical explanation of why dice are unbalanced, applying equally
to all the dice in Weldon’s experiment. The marking of the faces of dice is commonly done us-
ing little indentations, often with a thin coat of paint in them. The indentations change the center of
He further notes that a similar conclusion had already been given by Jeffreys (1939).

Like most, if not all, of the data sets to which Karl Pearson applied his $X^2$ test, this is a large data set. As Figure 1.2 shows, there is a definite trend or smooth alternative in these data. These may not have been picked up in a similar but smaller data set. Our next example will further highlight this. In later chapters we illustrate how components of $X^2_P$ complement the inspection of figures like Figure 1.2 and assist in specifying an alternative model.

**Example 1.4.2 (Birth-time data)** Unlike the previous example, data are often available ungrouped. In such cases application of Pearson’s $X^2$ test is dubious, because the data must be grouped and this loses information. Moreover, there is the problem of how to construct the groups. Suppose we consider a simple case in which no estimation is required and ask whether birth-times occur uniformly throughout the day. Hospital administrators would be interested in the answer to this question. Mood et al. (1974, p. 509) gave the following times for 37 consecutive births:

7.02 p.m. 11.08 p.m. 3.56 a.m. 8.12 a.m. 8.40 a.m. 12.25 p.m.
1.24 a.m. 8.25 a.m. 2.02 p.m. 11.46 p.m. 10.07 a.m. 1.53 p.m.
6.45 p.m. 9.06 a.m. 3.57 p.m. 7.40 a.m. 3.02 a.m. 10.45 a.m.
3.06 p.m. 6.26 a.m. 4.44 p.m. 12.26 a.m. 2.17 p.m. 11.45 p.m.
5.08 a.m. 5.49 a.m. 6.32 a.m. 12.40 p.m. 1.30 p.m. 12.55 p.m.
3.22 p.m. 4.09 p.m. 7.46 p.m. 2.28 a.m. 10.06 a.m. 11.19 a.m.
4.31 p.m.

Figure 1.3 gives histograms based on three and eight equal width classes for this small data set. Both histograms indicate a trend towards more births during the day. However, with such a small sample size, inspection ‘by eye’ can easily lead to false conclusions. In fact $X^2_P = 3.95$ for three classes and a smaller value for eight classes; neither of these is significant although the

![Figure 1.3](image_url) A histogram showing the birth time data grouped into three groups (left), and a histogram of the same data grouped into eight groups (right).